Data-Driven Strategies for Identifying Underserved Markets and Optimizing Real Estate Investment in the U.S. Transportation Sector

by

Santiago Hernandez
BE, Chemical Engineering Food Science, Tecnológico de Monterrey

and

Shane Huisman
BA, Supply Chain Management, Michigan State University

SUBMITTED TO THE PROGRAM IN SUPPLY CHAIN MANAGEMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE IN SUPPLY CHAIN MANAGEMENT
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

© 2024 Santiago Hernandez and Shane Huisman. All rights reserved.
The authors hereby grant MIT permission to reproduce and to distribute publicly paper and electronic copies of this capstone document in whole or in part in any medium now known or hereafter created.

Signature of Author: _____

Santiago Hernandez
Department of Supply Chain Management
May 10, 2024

Signature of Author: _____

Shane Huisman
Department of Supply Chain Management
May 10, 2024

Certified by: _____

Dr. Ilya Jackson
Postdoctoral Associate
Capstone Advisor

Accepted by:

_____

Prof. Yossi Sheffi
Director, Center for Transportation and Logistics
Elisha Gray II Professor of Engineering Systems
Professor, Civil and Environmental Engineering

Data-Driven Strategies for Identifying Underserved Markets and Optimizing Real Estate Investment in the U.S. Transportation Sector

by

Santiago Hernandez

and

Shane Huisman

Submitted to the Program in Supply Chain Management
on May 10, 2024, in Partial Fulfillment of the
Requirements for the Degree of Master of Applied Science in Supply Chain Management

ABSTRACT

Investment in logistics-focused industrial real estate in the United States has traditionally been concentrated heavily in markets surrounding ports and large urban areas. Investors have continued to build up these markets over time, following growth in populations, e-commerce, and trade activity. Inherently, there is a relationship between the flow of goods and the location of logistics facilities that handle them. While the traditional markets continue to be worthy investment locations, trucks carrying goods travel through the entire United States, often in markets not captured by ports or large cities. In this capstone project, we explore the relationship between freight flows and high flow through industrial real estate facility locations to identify potentially underserved markets. Underserved markets are defined as those that could make use of additional logistics infrastructure to handle the goods moving through them. To achieve this, we collected public and private freight flow information, and industrial property location data. Using ordinary least squares regression, regularized regression, and XGBoost models, we evaluated market features capable of predicting the total asset area and loading positions needed to support the volumes of goods shipped through each market in the United States. Ultimately, market saturation statuses were assigned for high, medium, and low model sensitivity levels for each market. These results were displayed in an interactive visualization tool to use when exploring non-traditional industrial real estate locations. There is significant potential to leveraging the relationship between freight flow and logistics focused industrial real estate. This project serves as a starting point to understanding that relationship and how it can be applied to create value in markets that investors have historically overlooked.

Capstone Advisor: Dr. Ilya Jackson
Title: Post Doctoral Associate MIT Center for Transportation and Logistics

ACKNOWLEDGMENTS

We would like to extend our deepest gratitude to our advisor, Dr. Ilya Jackson, for his exceptional guidance, expertise, and support throughout our capstone project. His guidance and suggestions played a pivotal role in the success of this project. We are also thankful to our writing coach, Pamela Siska, for her commitment and assistance during the writing process. It was always motivating to get words of affirmation from Pamela.

We owe special thanks to the representatives of our corporate sponsor for their invaluable time, feedback, and expertise, which have significantly contributed to our project. Their focus and knowledge on industrial real estate has deeply inspired us, and we are thankful for the opportunity to engage in such an innovative and impactful project.

Our families deserve immense gratitude for their unconditional love and support during our academic journey at MIT.

Lastly, we wish to thank the entire MIT community, especially the Class of 2024, for providing an extraordinary educational experience and fostering lifelong friendships during our capstone project.

Table of Contents

1. INTRODUCTION

1.1 Motivation

The United States trucking industry serves as the backbone of supply chains in America. With $940.8 billion in gross freight revenue reported in 2022, trucking represented 80.7% of the entire transportation industry (Elgin, 2024). However, trucking relies on another industry to enable this vast movement of goods. Industrial real estate is the buildings and facilities connecting truck movements at the shipping origins and destinations. Some examples of industrial real estate are warehouses, cross-docks or transload facilities, manufacturing plants, cold storage, and industrial outdoor storage. It is crucial for the industrial real estate industry to place facilities in markets where freight flows are prevalent and growing, allowing for the most optimized and efficient supply chains. The United States freight and logistics market is expected to grow at a compound annual growth rate of 4.01% between 2024 and 2029 (Mordor Intelligence, 2023). With the trucking industry being responsible for most of this growth, it is evident that industrial real estate investors face a wealth of opportunities to expand their footprint. Specific areas of the US may already be seeing significant demand for transportation flows but lack the infrastructure to handle them efficiently. Macroeconomic factors can influence trucking and freight demand, such as the push in several industries to relocate production and supply chains from Asia to Mexico through a process known as nearshoring. Nearshoring results in transportation flow booms in border cities, increasing the need for additional trucking capacity, in the form of facilities that can handle additional traffic in those areas (Wolf, 2023).

Our sponsor company specializes in industrial real estate investment with a logistics focus. They acquire and develop assets that serve transportation networks across the world. Customers lease these assets and manage the operations of the supply chain activities. Currently, the company has capital allocated for future investments. Matching this budget to potential investments in underserved markets in the transportation sector presents an opportunity to provide valuable assets to customers in need of improving transportation efficiencies and a positive return on investment. A recent example of this is the sponsor company and its competitors investing in new assets in Laredo, Texas in response to the increased freight demand at the Mexico-United States border, resulting in significant positive returns. This capstone will focus on potential markets for new assets related to high flow through (HFT) logistics facilities. These are facilities that focus on efficient flow of goods, rather than storage and warehousing. Goods typically spend less than 24 hours in these facilities. HFT assets include the categories of industrial outdoor storage, final mile warehousing, and cross-docking facilities.

1.2 Problem Statement and Research Questions

The sponsor company's current strategy for identifying potential new asset locations could make use of more quantitative demand-driven approaches incorporating freight flow data. They consider qualitative factors such as macroeconomic trends and inputs from their customers to determine which markets to invest in. This investment strategy does not include the use of freight flow data and predictive models that can guide the company to future investments with quantitative justification.

The sponsor company hypothesizes the presence of underserved markets in the industry, that is, markets where the demand for HFT transportation such as final-mile and cross-dock facilities exceeds the supply. The company intends to transition to a more proactive strategy by leveraging private and public historic freight and other supplemental data to predict where capacity will be needed in the future.

With investor capital already allocated for future investments, the sponsor company now needs to identify the most suitable markets to build or purchase new assets. While the current approach would point the company to expand in the areas with the most historical freight volumes, such as major ports and urban areas on the coasts, there is a desire to explore the potential of overlooked markets. Macroeconomic factors have a significant impact on the attractiveness of these potential locations, such as manufacturers looking to shift away from importing from Asia and exploring more domestic manufacturing and nearshoring opportunities (Alvim and Averbuch, 2023). The quick shift in demand towards the continental United States means that the sponsor company can capitalize on this market shift where freight volumes are expected to increase. Existing infrastructure factors such as proximity to ports, intermodal rail terminals and highway interchanges also play a major role in the attractiveness of potential investment markets.

A challenge the company faces is obtaining external data to analyze flows of transportation in the United States. Currently, the company uses qualitative data to predict trucking freight market behavior. Such data sources include news articles, word of mouth, surveys, and customer insights.

Thus, the research questions to be answered through this capstone include:

1. Are freight volumes in the trucking industry a factor driving the demand for HFT real estate and how can the sponsor company leverage this relationship to capitalize on underserved markets?
2. How can current U.S. freight flow data be obtained from reliable sources?
3. What quantitative methodologies should the sponsor company follow to identify underserved markets in the continental United States?

1.3 Project Goal and Expected Outcome

The project goal is to provide the sponsor company with a quantitative data analysis including an output that determines which markets are underserved in the transportation real estate sector by HFT assets. The output would provide the company with key learnings applicable to capital investment decisions as a supplement to their current investment strategy. These learnings would position the sponsor company to identify similar opportunities as those presented by the increasing freight volumes at the Mexico-United States border in Laredo.

Through analyzing and connecting historic data from a digital freight brokerage company, the U.S. Department of Transportation (DOT), US Customs and Border Protection - Trade Statistics, the United States Census Bureau, and industrial real estate data from the sponsor company, the project's goal can be met, and the research questions answered. The data provided by the sponsor company captures HFT asset locations and features for both their own and competitors' properties. This real estate data served as a measure of supply, while freight flow data represented demand in markets in the United States. Understanding the relationship between the features in the supply and demand data allows the sponsoring company to identify potentially underserved markets. The project's expected outcome was a repeatable methodology that gives the sponsor company visibility to market opportunities. They would be able to apply this methodology with updated data sources in the future to continue identifying new opportunities in overlooked or underserved markets. As the transportation market shifts every 3 to 4 years (Fuller, 2022), keeping track of shifts in both freight flow trends and the HFT asset market is crucial for capturing the most value out of this project.

The deliverables to the company were as follows:

1. A repeatable methodology to identify underserved markets for HFT industrial real estate.
2. An analysis of the market features explaining the variability in HFT asset volumes in different areas of the United States.
3. A visualization tool summarizing the project findings, i.e. each market's saturation status and the features explaining this status.

1.4 Plan of Work

To achieve the project goal, we executed the plan of work depicted in Figure 1. First, a proper scope definition needed to be aligned between the sponsoring company and the research team. Once the scope was agreed upon, the data gathering stage began. Two main sources of data will be utilized for this project: endogenous (from within the project sponsor), and exogenous (data from external private and public sources). Next, the Analytical Assessment stage required deep analysis of the data and assessing whether

its quality is sufficient to be included in the methodology. Make use of the validated data, regression models were applied to output the desired outcome: market feature analysis and identifying underserved markets in the continental United States. Result validation in conjunction with the company was necessary to address the veracity of the model's predictions. This model assessment was needed to determine whether to iterate the process again starting from the data gathering stage. This process was designed with the expectation of iteration being necessary until the models yielded results sufficient to meet the project goals and answer the research questions. The plan's final step was to create a visualization tool to summarize the project's key learnings and conclusions.

2. STATE OF THE PRACTICE

Using flows of trucking freight in the United States to find attractive logistics infrastructure markets presents some challenges. Measuring and forecasting freight flows requires obtaining data sources representative of the total freight movement in the country. Similarly, understanding the infrastructure needs of the trucking industry requires adequate national real estate data. Thus, the key problem of this capstone project is to develop a methodology which takes inputs of freight flows as demand data and industrial real estate capacity as supply data and provides an output measure of the investment potential of different markets. Before building this methodology, we reviewed industry best practices for measuring freight flows, finding attractive asset locations, and other quantitative methods with relevant applications to freight and real estate data.

2.1 Measuring and Forecasting Freight Flows

2.1.1 Transportation Introduction

Transportation in the United States plays a crucial role in the continuity of the supply chains of thousands of companies (shippers and receivers). 75% of goods by weight in 2023 were moved from an origin to a destination via an inland transportation method (USDOT, 2023).

The common inland transportation methods in the United States are categorized as follows:

Over-the-road (OTR):

Dry-van: this method involves trucks pulling trailers in different sizes (ranging from 20 ft long to 53 ft long, 48 and 53 ft being the most common) (Rodrigue, 2020).

Container chassis: This method involves trucks pulling chassis that can fit either rail (53 ft long) or ocean containers (20 or 40 ft) (Rodrigue, 2020).

Rail: Transportation method that utilized the rail network in the United States to move freight. Trains pull railcars to move freight from origin to destination (Rodrigue, 2020).

Intermodal: Combined transportation method that utilizes both modes previously described. It is a combination of both OTR and rail. An example would be if a container travels OTR from the origin facility to a railyard to be loaded onto a railcar to the destination (Rodrigue, 2020).

Less-than-truckload (LTL): This transportation method is categorized as a consolidation method of goods for shippers that cannot ship a full-sized trailer fully. Multiple shippers' freight is combined by an LTL company in a single trailer to provide a lower shipping cost for each party. LTL companies have sorting facilities where freight from different shippers is consolidated, sorted, and shipped (Rodrigue, 2020).

2.1.2 Transportation Flows

Transportation flow is the direction in which freight moves in a particular lane. A lane is defined as all possible combinations between an origin and a destination (Lieu, 1999).

Based on this definition, a shipper with an Origin A and a destination of Destination B will generate Lane A-B. After a shipment has been made in this lane, the state of flow of the lane will change from x to x+1, adding a weight to the count of shipments made for this lane. The flow of goods can be measured with different metrics. Some examples of these metrics are count of shipments, unit of weight shipped, unit of volume shipped, cost of goods shipped, amongst others.

This project focuses on the count of shipments as the unit of measure to calculate flows from origins to destinations.

2.1.3 Current Public Transportation Flow Data Sources

The United States Department of Transportation's Federal Highway Administration (USDOT FHWA) and the Bureau of Transportation Statistics (BTS) partnered to create the Freight Analysis Framework (FAF). The FAF is the entity in charge of integrating data from various sources to analyze transportation flows in the continental United States. FAF's objective is to provide a comprehensive resource on national freight flow by combining multiple data sources from the United States Government database to understand and forecast freight flows from an origin to a destination. The FAF publishes results and forecasts on a publicly available dashboard on their website. The FAF results are updated every five years. At the time of this writing, the most recent FAF analysis was FAF5 published in 2022. However, FAF4 (published in 2017) provides a detailed explanation of the logic and procedure to build the model.

The Freight Analysis Framework has multiple advantages for understanding transportation flows in the United States, such as the fact it is publicly available, has multiple trustworthy Federal data sources,

has been implemented since 1999, considers both national and international trade, and excludes transportation methods outside of the scope of this project. However, it also has limitations for our research. These include the veracity of the data (the CFS is not audited; therefore, shippers may share inaccurate data), the consistency of the data (only published every 5 years), and the inability to access the raw data behind the analysis.

2.1.4 Current Private Transportation Flow Data Sources

DAT Freight and Analytics is the largest privately held trucking data company in North America. According to DAT, its database contains around 500 million transactions in over 68,000 lanes, and $150 billion in freight invoices analyzed each year, which is close to 20% of the total truckload transaction population in the United States (DAT, 2023).DAT iQ is one product from DAT offering different analytics tools that provide shippers with insights to rates, demand trends, freight forecasts, and compares network performance (shipper rates vs. comparable shipper rates). DAT iQ also offers customized analytics and reports for shippers, brokers, and carriers (DAT, 2023). The underlying data in DAT's database contains many attributes to satisfy all products it offers. These attributes include important time stamps, locations (origin, destination, stops), equipment sizes, weight, volume, and others. All these attributes are key to calculating transportation flows in different lanes across the United States, which are featured in other DAT products, such as TruckersEdge (DAT, 2023). DAT TruckersEdge offers truckers insights on the load density of trucks in and out of each US state. Load density is defined by DAT as the difference between truck inflow and outflow in a state. By observing load density, truckers can track the flow of goods from one state to another (DAT, 2019).

2.2 Industrial Real Estate Investment Practices

Much of the information reviewed on industrial real estate industry practices came from market reports generated by real estate investment management firms. The reports focus on industrial real estate, which can be defined as all land and buildings which accommodate industrial activities, including production, manufacturing, assembly, warehousing, research, storage, and distribution (Cauble, 2023). Our sponsor company's scope of HFT assets falls into the short-term storage and distribution categories, and is therefore represented by only part of the whole industrial real estate market. It is important to note that while the industry practices reviewed in this section are highly correlated to the practices in HFT asset investment, they may not be fully representative due to the inclusion of non-HFT asset categories in the reports. The industrial real estate reports were still used to understand industry practices, due to the limited market information on the individual subcategories of industrial real estate.

2.2.1 Major Industrial Real Estate Markets

Choosing where to invest in industrial real estate assets represents a significant part of this capstone's scope. To understand the current industry asset location practices, we reviewed the U.S. markets with the most significant industrial real estate footprints and ongoing investments. The key drivers for high investment performance in each market are large populations consuming many goods and proximity to distribution systems such as railways and other logistics crossroads. Mark Glagola, a senior managing director at Houston-based Transwestern Investment Group, stated in 2017 that industrial demand was linked to the nation's ports. From Seattle down to the Los Angeles Basin, then through Texas, Atlanta, and up to the East Coast to New York, a visualization of a U-shaped curve can be used to picture the markets with high demand for industrial real estate. The exception to this visualization is Chicago, which is also among these markets (Kirk, 2017). This U-shaped curve still holds true today, according to Cushman and Wakefield's Q3 2023 U.S. Industrial Marketbeat Report, which shows that the most square footage of under construction industrial real estate is in Dallas/Ft. Worth (TX), Phoenix (AZ), Inland Empire (CA), Savannah (GA), and Atlanta (GA) (Price, 2023). Dallas-based Real Estate Firm CBRE listed Inland Empire, Chicago, Atlanta, Savannah, and Dallas/Ft. Worth as the markets making up most of the 100 largest industrial lease transactions in 2022 (Berman, 2023). While the markets listed above represent the most common and highest performing areas for industrial real estate investment, this capstone's purpose is not limited to identifying these markets. Rather, this project focuses on finding potentially underserved markets. Therefore, the capstone methodology was developed to point the sponsor company to markets where there may also be high demand, but room for growth in supply of industrial real estate.

2.2.2 Build-to-Suit vs Speculative Buildings

When analyzing the industrial real estate market and particularly new construction, it is important to consider two different types of assets. 'Speculative' or 'Spec' buildings are built by developers with the goal of attracting tenants during or shortly after construction, while 'Build-to-suit' (BTS) buildings are purpose-designed, built and typically owned by a specific organization to accomplish specific goals (Jacobs, 2021). Spec buildings represented over 80% of assets under construction in Q3 2023 (Price, 2023). These assets are typically favored by companies like our sponsor, because they are flexible to suit many different tenants, which have a variety of needs. BTS buildings have a risk of meeting the needs of only a specific tenant, and thus not being attractive to future tenants when the original lease ends. The growth or decline in spec construction provides commercial real estate firms insight to changes in supply and demand for industrial properties, to be discussed more in section 2.2.3.

2.2.3 Supply and Demand Metrics for Industrial Real Estate

Although this project's goal of using freight flows as a demand metric for logistics infrastructure is not currently a common practice in the industrial real estate industry, several metrics are used as proxies. Tracking trends in vacancy, absorption, rent, and construction allow real estate investment firms to quantify the shifting of supply and demand over time.

2.2.3.1 Vacancy

The vacancy rate represents the total amount of unoccupied space, as a percentage of the total inventory of buildings in the market. According to Cushman and Wakefield's Q3 2023 U.S. Industrial Marketbeat Report, new supply of buildings in the United States coupled with a decrease in demand caused the vacancy rate to increase to 4.7%. The report notes that while vacancies are rising, it still sits below the 15-year average of 6.8% (Price, 2023). Real estate firms therefore use vacancies to explain the impact of new builds, especially spec builds on the supply and demand health of different markets, comparing against different points in time as a reference. Vacancy rate growth indicates a slow-down in demand (or an increase in supply) for industrial space, while low vacancies suggest high demand and potentially undersupply. For HFT properties specifically, vacancy rate has minimal impact on the properties trucking companies use.

2.2.3.2 Construction

As mentioned in 2.2.2, BTS and Spec construction make up the total square footage under construction in a market. Real estate investors can use this square footage to understand growth in each market from a supply perspective. This metric can be used to anticipate changes in other metrics, such as vacancy, rent and net absorption. The breakdown of construction by BTS and Spec gives investors an idea of how other metrics will change in more detail. If most of the new construction is Spec, vacancy rates would be expected to rise more than if BTS made up the majority, since having tenants signed before construction would lower vacancy rates once the building is completed.

2.2.3.3 Leasing Activity

Leasing activity is the sum of all leases over a period. This includes pre-leasing activity as well as expansion. It does not include renewals (Price, 2023). This simple metric helps investors understand another input into vacancy rate changes and net absorption rates. Strong positive trends in leasing activity suggest that demand for industrial real estate is growing.

2.2.3.4 Net Absorption

Net absorption is the sum of square feet that became physically occupied, minus the sum of square feet that became physically vacant during a specific period (Georgules, 2017). This demand metric is used to understand the balance between vacancy, construction, and leasing activity, and serves as an overall

measure of a market's health. Growth in this metric indicates growing demand and the potential need for additional supply in the future, while negative trends suggest shrinking demand.

2.2.3.5 Asking Rent

Rent trends are important for investors to understand the attractiveness of investment opportunities and can also be predicted by changes in the previously mentioned metrics. As demand for industrial space (as measured by leasing activity and net absorption) increases, asking rent tends to rise. Conversely, as demand decreases or supply increases (as measured by construction), the asking rent of a market tends to decrease. Therefore, understanding all the supply and demand metrics and how they relate to one another aids real estate investors in predicting rent changes, which ultimately determines the return on the investment opportunity in a market.

2.2.4 SF/Door Ratio

When analyzing the individual investment opportunity, our sponsor company uses additional metrics to determine the attractiveness of the asset. One of these metrics is the square foot per door ratio. This metric considers the total square feet of the building and the number of dock doors accessible to trucks using it. From the perspective of HFT asset investment, this is an important figure to consider because a building being used to handle a lot of freight movement will need sufficient dock doors to avoid idle trucks and long wait times for picking up and dropping off loads. Customers seeking to occupy and operate in HFT assets will be more attracted to buildings with a square foot to door ratio appropriate for their expected freight flows. For the sponsor company, this represents one of the most significant factors when assessing the value of potential property investment.

2.3 Other Quantitative Methods

Additional quantitative methods were explored and deemed relevant to this project: Network Science and Natural Language Processing (NLP).Network science has many applications in the logistics industry and can be used to generate useful market features, which may help identify underserved markets and optimal zip codes for asset location within those markets. NLP "combines computational linguistics—rule-based modeling of human language—with statistical and machine learning models to enable computers and digital devices to recognize, understand and generate text and speech" (IBM, 2024) NLP is utilized in this project as text summarization to categorize commodities specified in the transportation flow data acquired.

2.3.1 Network Science

Network Science was used to understand the interconnectedness of freight flows. A network is a way to describe a set of points, called nodes, and the connections between them, called links (Menczer et al., 2020). There are many different types of networks, but a logistics network is one of the most common and

powerful. In a freight flow network, the nodes are the origin and destination points, and the links are the movement of freight between the nodes. More specifically, the logistics networks relevant to this project are both directional and weighted, meaning the links between nodes can be uni- or bi-directional, and have an associated weight, such as the count of shipments, time, cost, or tonnage. In network science, several useful metrics can help explain the behavior of a network and the importance of individual nodes. The network density is a value between 0 and 1 that represents the fraction of all possible links that exist (Menczer et al., 2020). Network Density shows to what extent the nodes in a network are connected to all other nodes. At the node level in a directional network, an in degree and out degree can be calculated. These are the number of links coming into a node from predecessors, and the number of links going out of a node as successors, respectively. Incorporating the weight of each link to the degree calculation yields the in strength and out strength of the node (Menczer et al., 2020). The relative importance of a node in a logistics network can be quantified and compared to other nodes with centrality measures. Some common centrality measures are:

- **Closeness Centrality:** based on the network distance between a node and each other node, using the list of distances between a node and all other nodes in the network as an input (Bloch, 2023).
- **Katz-Bonacich Centrality:** based on the number of walks emanating from a node. A discount factor is applied to the sum of all walks coming from the node. Shorter walks are weighted much heavier than longer walks. This measure counts the total number of walks from a node to all other nodes, and discounts this sum based on the walk lengths (Bloch, 2023). A "walk" is a way of traversing the network, starting at one node and moving to other nodes on the links.
- **Eigenvector Centrality:** The importance of a node is related to the importance of its neighbors. The node's centrality is proportional to the sum of its neighbors' centralities. (Bloch, 2023). The computation is performed iteratively until the scores do not significantly change with more iterations.
- **Betweenness Centrality:** calculates the shortest path from each node to every other node in the network, and nodes that are crossed most frequently have higher betweenness (Menczer et al., 2020).

These measures help identify which nodes are 'central,' meaning that they provide most of the connections in the network. These nodes are known as hubs. They have many neighbors, while other nodes have just a few. One way to test whether a network has many hubs is to compute the heterogeneity parameter, which compares the variability of the degree across nodes to the average degree. (Menczer et al., 2020). In the context of freight flows, hubs are extremely useful to identify, as these areas are the key points where a lot of the flow travels through, and likely represent areas with high demand for industrial

real estate. Network science can be applied to freight flow data using a python package called networkx, along with Geopandas and matplotlib for plotting the network on a visual map of the United States.

2.3.2 Natural Language Processing for Commodity Categories

Natural Language Processing is a branch of Artificial Intelligence (AI) that focuses on processing language by an algorithm the way a human brain processes language and words. Common use of NLPs includes text translation, text generators, respond to typed commands, classify blocks of text, among others (Gruetzemacher, 2022).

NLP was utilized to categorize commodities shipped in the transportation flows analyzed in this project. Categories of commodities in transportation flows were generated to understand the movement of goods in the United States. Understanding which types of commodities moved through a particular lane helps determine the impact of certain types of commodities to a HFT asset. The use case that this project focuses on is the classification of blocks of text. Ambiguity, homonyms, incorrect grammar, idioms, and typographical errors can be reasons why inconsistency in user entry attributes can lead to misinterpretation of data or not make the attribute useful at all (IBM, 2024). Classification of blocks, also known as text classification, uses NLP techniques to break down a sequence of words or characters into tokens. This process is called tokenization. A token is “an instance of a sequence of characters in a document that are grouped together as a useful semantic unit for processing” (Manning, 2008). Tokens are the building blocks for NLP that will help train the NLP model for text classification.

There are three primary approaches to text classification (Manning, 2008):

1. **Rule-based System:** This approach utilizes handmade linguistic rules to organize text into groups. It involves defining a list of words associated with specific categories, like categorizing "pizza" and "fries" under food, or "toothpaste" and "paper towels" under consumer goods.
2. **Machine-based System:** This method learns to classify text based on prior data observations. It involves using prelabeled data as training and test sets, allowing the system to develop a classification strategy from previous inputs and to continually improve its accuracy.
3. **Hybrid System:** The hybrid approach combines rule-based and machine-based systems. It starts with the rule-based system to create tags, then employs machine learning to refine and expand the rule set. If discrepancies arise between the machine-generated and rule-based tags, human intervention is used to manually improve the list. This approach is considered the most effective for implementing text classification.

For this project, a hybrid approach was used with Tokenization and Preprocessor libraries in Python. Two of the most common libraries used for tokenization and text classification are Natural

Language Toolkit (NLTK) and Hugging Face. NLTK was used in the methodology of this project for categorization of commodities shipped.

3. METHODOLOGY

Based on our review of the literature, we developed a methodology to answer the research questions. Our methodology takes freight demand and industrial real estate supply data as inputs, applies techniques from network science and performs a regression analysis to determine market feature importance. After iteratively improving the regression models, the generated predictions were used to identify potential underserved markets.

3.1 Data Collection and Feature Engineering

To achieve the goal of identifying potentially underserved markets, a comprehensive data frame containing key market features needed to be created. These features included the volumes of freight moving through the market, network science centrality scores, HFT real estate metrics, and demographic data. Collected from various sources, both raw data and engineered features were added to the market data frame for the analysis.

3.1.1 Freight Flows

To answer the research questions, one of the most important features utilized in the project was freight flows in the United States. A variety of freight flows were considered in the analysis to incorporate different modes of transportation significant for the sponsor company's interest.

The freight flows considered for the project were:

- **Domestic freight flows:** inland shipments that had an origin and a destination in the continental United States hauled by a truck trailer.
- **Ocean Port Freight Flow:**
    - Total port flows: containers flowing into or out of the United States that were imported or exported by one of the ocean ports managed by the United States Customs and Border Protection (USCBP).
- **Inland Border Crossing Freight Flow:**
    - Imports: trailers and intermodal containers flowing into the United States via an inland border crossing managed by the USCBP.
    - Exports: due to the nature of inland exports in the United States (most exports do not need sorting at the port of exit), export quantification at port of exit was omitted for the purpose of this project. The assumption that the same import industrial real estate facilities were

used for exports was also made. Hence, only one metric was needed due to the high correlation between import and export inland flows.

Details of each freight flow and data sources are explained below.

3.1.1.1 Domestic Freight Flows

A dataset of a privately held digital transportation broker, was obtained from the Massachusetts Institute of Technology Center for Transportation and Logistics. This dataset contained over 250,000 loads spread over eight years (from 2016 to 2023) managed by the broker. Each row in the dataset represented an executed load by the broker that contained the relevant information for the analysis. Some columns, omitted from Table 3.1, were excluded from the analysis, due to irrelevance to the research questions. The relevant columns utilized in the project are outlined in Table 3.1:

Table 3.1

Relevant Columns from Domestic Freight Flow Dataset

Column Name	**Description**
equipment_type	Type of trailer used by truck to move a load.
origin_kma	Key Market Area (KMA)
dest_kma	KMA of the current load's destination.
direct_mileage	Mileage from origin to destination.
commodity	Commodity being carried by the truck.
origin_latitude	Latitude of the pickup location.
dest_latituted	Latitude of the destination location.
origin_longitude	Longitude of the pickup location.
dest_longitude	Longitude of the destination location.
origin_city	City of origin of the shipment.
origin_state	State of origin of the shipment.
origin_country	Country of origin of the shipment.
origin_zipcode	Zipcode of origin of the shipment.
dest_city	City of destination of the shipment.
dest_state	State of destination of the shipment.
dest_country	Country of destination of the shipment.
dest_zipcode	Zipcode of destination of the shipment.

pickup_date	Date of pickup of the load.
direct_mileage	Mileage from origin to destination of the quote
quote_date	Date of the quote
weight	Weight of the load
origin_drop_trailer	If the trailer needs to be dropped at the origin
dest_drop_trailer	If the trailer needs to be dropped at the destination
team_driver	If the shipment requires more than one driver.

An exploratory data analysis was performed on the dataset. Due to the confidentiality of the data, the findings cannot be fully disclosed here. However, the decision was made to use four years' worth of data from 2019 to 2022. The data for this timeframe was more representative of total U.S. freight flows than earlier years in the dataset. The data was then aggregated by 3-digit zip code based on the load's origin and destination to be further analyzed. After aggregating data at a 3-digit zip code level, more generalized analyses could be performed. For example, the data could then be rolled up at Key Market Area levels, providing increased layers of visibility. Once the data was aggregated metrics for each 3-digit zip code were established. These metrics included average annual inflow (number of loads coming into the zip code), average annual outflow (number of loads going out from the zip code), average annual internal flow (loads shipping within the same zip code).

As outlined in Table 3.1, commodity was an attribute present in the domestic flow data set. During the exploratory data and analysis for this attribute, we concluded that the 'commodity' attribute from the dataset was a free-form text entry from the Broker's user. The unique value count for 'commodity' was > 30,000. Also, when sorting the data, it was observed that the commodities were often the same but with different unique values. For example, a unique commodity type could be 'cases of beer' and another unique commodity could be 'pallets of beer'. Both unique values clearly describe the same product: beer.

Natural Language Processing (NLP) was applied for text classification to the commodities present in the dataset. The library utilized was Natural Language Toolkit (NLTK) in Python. The Hybrid System was trained with all the 'commodity' data. The model was instructed to create categories based on preferred category names. Table 3.2 shows the resulting categories and their respective percentage representation of the data.

Table 3.2

Resulting Commodity Category Distribution from NLP Application

Category	Percent
Food and Beverages	37.30%
Unknown	28.48%
Consumer Goods	10.36%
Automotive and Machinery Parts	8.90%
Construction and Building Materials	7.24%
Industrial Goods	6.95%
Agricultural Products	0.60%
Healthcare and Pharmaceuticals	0.18%
Total	**100%**

It is important to note the following observations:

1. Unknown category: representing almost 29% of the dataset, the unknown category outlines loads where the 'commodity' column contained null values or key words such as none, N/A, not provided, various, FAK (Freight All Kinds), and miscellaneous.

2. The dataset from the digital freight broker may be biased towards a certain industry that might cause the commodity categories to be inflated.

Once the commodities were categorized, the average annual total flows of each category were calculated for each 3-digit zip code 3.1.1.2 Ocean Port Freight Flow

To capture the import and export ocean freight flows outlined in the section above, a dataset from the United States Customs and Border Patrol (USCBP) was obtained. This dataset provided import and export information from each port in the United States. The report was later filtered to only include data from 2019 to 2022 (the same time window for domestic freight flows). However, the dataset from the USCBP did not contain information about individual loads or number of containers imported or exported. The only two metrics provided by the USBCP regarding quantifiable assets were kilograms and value of goods.

Since the value of goods can be hard to interpret and quantify for a flow analysis, it was decided to analyze the weight of goods moved through the port.

Once the import and export weight by ports by years were identified, the flow metrics were assigned to all zip codes in the specific Key Market Area where the port was located. For example, the Port of Long Beach in California is in zip code 90802. However, the imports and exports that flow through the Port of Long Beach not only affect industrial real estate facilities in zip code 90802: they also affect facilities located in the neighboring zip codes. Therefore, the decision to include the average yearly import and export weights for the whole Key Market Area for the port location was made.

3.1.1.3 Inland Border Crossing Freight Flow

Inland imports and exports were captured in a similar way to ocean port flows. A dataset from the USCBP containing a list of inland ports of entry to the United States was obtained. The dataset was filtered to analyze the same timeframe as the previous datasets (2019-2022) to keep consistency. However, the dataset for inland imports and exports contained different flow metrics than the ocean ports. The inland border crossing dataset contained monthly data by port of entry for different transportation asset types. The dataset included loaded trucks, empty trailers, buses, cars, pedestrians, freight trains, containers, and empty containers. For this project's purpose, buses, cars, pedestrians, and freight trains were removed from the dataset. Freight trains were decided to not be included in the dataset since boxcars in freight trains are not handled by the facilities our sponsor company is interested in. In contrast, trucks, trailers, and containers will all flow through industrial real estate facilities for loading, sortation, or unloading.

The same assumption that was used for ocean port flows was implemented for inland port flows. The flow metrics were assigned to the Key Market Area where the port of entry into the United States was located.

3.1.2 Network Science – Centrality

Using the networkx package in Python, the freight flow data obtained from the digital freight broker was converted into a network. Because each row of this data contained a shipment origin and destination, it was possible to establish nodes. The origins and destinations were available as 5-digit zip codes, 3-digit zip codes, and key market areas (KMA). KMA's are a collection of zip codes that form an economic market, as established by DAT (Pitz, 2017). To ensure a higher number of nodes in the network and enable detailed market analysis, 3-digit zip codes were chosen as the nodes in the network. Once the network was created, networkx was again utilized to calculate centrality scores for betweenesss, closeness, eigenvector, and katz. These scores were then included as features in the subsequent regression analyses. However, only one centrality score was ultimately included in each model to avoid multicollinearity. Details on the centrality score selected are shared in section 4.

3.1.3 Market Demographics

U.S. Census Bureau data was used to add demographic market features to the dataset. The two key features explored were total households and median family incomes. These two features were selected based on discussions with the sponsor company, leveraging their expert knowledge on drivers of industrial real estate demand. The Census data was initially grouped by 5-digi zip code; therefore, these features were aggregated at the 3-digit zip code level. For total households, a simple sum of all the 5-digit zip codes with a common 3-digit zip code sufficed, while a weighted average of each node was calculated for the median family incomes. These two features were selected based on discussions with the sponsor company, leveraging their expert knowledge on drivers of industrial real estate demand.

3.1.4 Real Estate Features

As discussed in section 1.3, the sponsor company provided the research team with a dataset containing their HFT real estate properties as well as those of their competitors. This data required pre-processing and cleaning before being used to add features to the market dataset. A key step was determining the property types. Each row of data represented a single property, which was categorized as either a Cross-Dock, Final-Mile Warehouse, or Industrial Outdoor Storage. The geopandas python package was used to convert property addresses to coordinates and zip codes. Key features extracted from this dataset include Total Area (square feet), Total Loading Positions, and Total Property Count. These features were aggregated at the 3-digit zip code level and added to the market dataset to be used in the regression models.

3.2 Regression Analysis

For each property type, two sets of regression models were run to predict two different target variables: Total Area and Total Loading Positions. For each target variable, several different models were created to find the best performing model. For example, to predict the total cross-dock area a market, an Ordinary Least Squares Regression (OLS), LassoCV regularized regression, and a non-linear XGBoost model were created and compared to one another. The best performing models in terms of R^2 and error metrics were chosen to generate the predictions needed to complete the methodology. Splitting the data into train and test sets, as well as cross validation were performed to ensure robust models and results. This process was then repeated with another set of models using the volumes of commodity flows described in section 3.1.1.1, rather than the total flow into and out of a market. This provided useful insight into the commodity categories and their relationship to the target variables.

3.3 Evaluating Feature Importance

Extensive model tuning was performed by including and excluding different sets of features to find the most significant and important market characteristics to predict the target variables. This took multiple

iterations of feature combinations. It was ensured that each model contained features pertaining to freight flows, network science, market demographics, and real estate (information regarding the other two types of HFT real estate not being targeted by the model). The significance or importance of each feature was judged by p-values in OLS models, coefficients in LassoCV models, and SHAP (Lundberg et al., 2020) values for XGBoost (Chen et al, 2016) models.

3.4 Target Variable Predictions

After selecting the model with the best performance in predicting the target variable, the features shown to be insignificant in predicting the target variability were removed. The models were then run again on the full dataset of all market nodes with the features deemed to be significant. The predictions generated by these final models were captured in the market dataset.

3.5 Tagging Underserved Markets

After capturing the model predictions, the errors were calculated and included in the market dataset. The errors were used to determine the status of each market as either underserved, overserved, or neutral. The mean and standard deviation of the errors were calculated and used to tag each market with these statuses. If the prediction error of a market was above the mean error + 1 standard deviation, it was labeled as underserved. Conversely, if the prediction error was less than the mean error – 1 standard deviation, it was labeled as overserved. Otherwise, the market was considered neutral and likely contains an accurate amount of HFT real estate in relation to demand. This process was repeated using 1.5 and 2 standard deviations in the market status calculation, to provide the sponsor company with different sensitivity levels for their analysis.

4. RESULTS

As mentioned in section 3.2, three different types of regression models were used to evaluate the market feature importance in predicting the total area and loading positions for Cross-Docks, Final-Mile Warehouses, and Industrial Outdoor Storage. The OLS and LassoCV linear regression models were compared to the XGBoost non-linear model to determine which type of fit was best suited in predicting the target variables. The OLS and LassoCV models consistently outperformed the XGBoost models in terms of R^2 and error metrics. Therefore, linear models were chosen to evaluate feature importance. Ultimately, OLS was selected for generating the final target variable predictions due to ease of interpretability to the sponsor company over LassoCV. The feature importance evaluation was consistently matching between OLS and LassoCV. Variance Inflation Factors (VIF) were calculated for the features input into each model in combination with the iterative approach described in section 3.3, to check for multicollinearity. As a

result, Katz centrality was used as the sole centrality measure in all models, and average annual total flow (named Total Flow in the models) as the freight flow measure.

4.1 Example Model Walkthrough

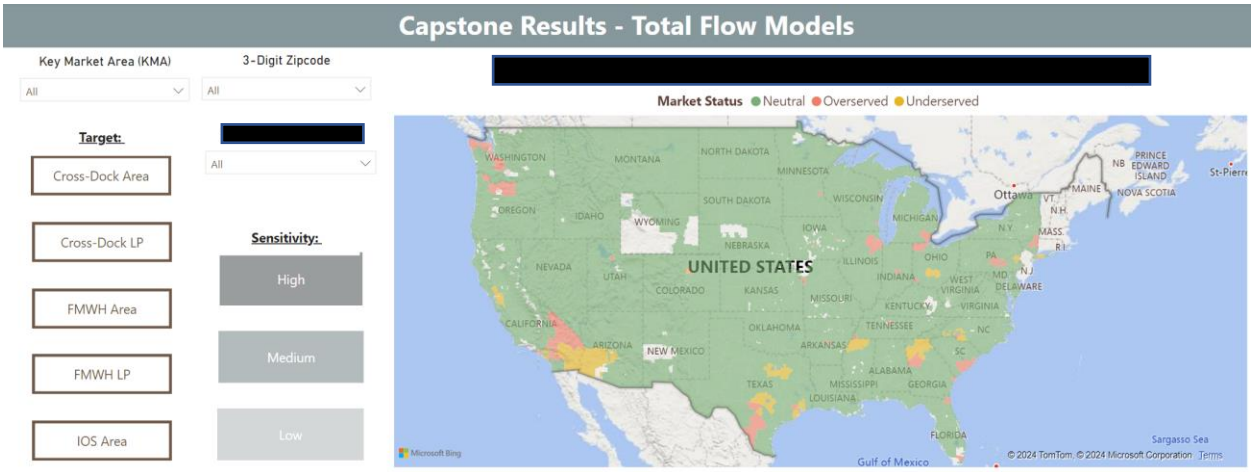
To demonstrate how feature importance was assessed, Appendix A1 serves as an example. In this model, the target variable was Cross-Dock Area, and the feature inputs are listed in Appendix A1. This model's strength is demonstrated by a training R^2 of 0.67 and a test R^2 of 0.51, meaning 67% and 51% of the variance in the target variable is explained by the significant features for the train and test datasets, respectively. To evaluate feature importance, both the feature coefficients and their p-values were analyzed. Features with a p-value's greater than 0.05 were deemed insignificant and excluded from the subsequent final prediction models. Features with p-values less than 0.05 were both included in the final prediction model as well as ranked by the absolute value of their coefficients to understand which market characteristics contributed most to the target variable prediction. In this example, Total Industrial Outdoor Storage Area, Total Final-Mile Warehouse Area, Total Households, Katz Centrality, and Total Flow were significant in predicting the Total Cross-Dock Area of each market, in order of most to least importance. These features were then input into the final prediction model, which led to the market status evaluations described in section 3.5.

4.2 Visualization Tool – PowerBI Dashboard

Visualizing the results of this project was a crucial step to ensuring the sponsor company's ability to implement the results into their future investment strategies. A dashboard was created using PowerBI to achieve this. The dashboard features freight flow mapping, market feature analysis, and most importantly, the results of this capstone. These results were displayed as an interactive map of the United States, highlighting the markets, or nodes, which were tagged underserved or overserved. The company can now select a target, such as Cross-Dock Area or Loading Positions, and the desired sensitivity levels. Filters for KMA and 3-digit zip codes allow them to focus on a specific market of interest. Toggling between sensitivity levels gives the company an idea of how robust a market's status is. A sample of this dashboard is provided in Figure 4.1.

Figure 4.1:

Sample of PowerBI Dashboard



5. DISCUSSION

5.1 Variability in Models' Strengths

Analyzing the results in section 4 generated discussion of the strengths and usefulness of the models to the sponsor company. Comparing outputs and metrics for the different target variables and models, some models performed better than others. For example, models with cross-dock area as the target variables performed best out of all the models, while models with IOS area performed worst. However, the outputs of all predictive models for each HFT type were still provided to the sponsor company as a demonstration of the methodology's potential, with the understanding that there are actionable next steps for improvement. The variability of results for the different property types should be considered by the sponsor company when comparing different property types in the same market. A property type with less explainable models than others should be further analyzed and researched by the sponsor company before using this project's output to supplement an investment decision. The model and dashboard should serve only as a starting point to generate discussions and exploratory analysis for the sponsor company, not a final source of truth.

5.2 Sponsor Company Application

The model predictions, output, and visualization in the dashboard should be used by the sponsor company as a starting point to generate conversations and exploratory analysis for potential investment into underserved markets. It should not be utilized as a single source of truth to formulate an investment strategy. In the visualization, a sensitivity analysis was included with different levels of sensitivity: high, medium, and low. The sensitivity analysis allows the company to set stricter parameters on categorizing markets. High sensitivity yields more underserved markets, while low sensitivity yields fewer underserved markets.

When a market is categorized as underserved in all three sensitivity levels, the sponsor company can be more confident that there is truly a need for more HFT real estate. However, it should still be researched and analyzed further using the company's traditional investment practices. Further scenario analysis could be simulated by the sponsor company if desired. By altering the data frames behind the dashboard, different simulations in target markets could be run. For example, the total flow in a market could be increased or decreased by an arbitrary percentage to help simulate a potential growth or contraction of a market. Similarly, the number of households could be increased by population growth trends in a market. With these alterations, the sponsor company could analyze the effect certain scenarios have on the categorization of a market as underserved or overserved.

5.3 Recommendations and Improvements

Throughout the implementation of this project, multiple areas of opportunity were identified to enhance the project's reliability and replicability. One of the limitations of this project was the size and scope of the datasets obtained. The two main data frames with transportation flow and real estate data were obtained from single sources. Recall that the transportation flow data was obtained from a digital broker representing only a small subset of the total transportation flow in the market. It is recommended that the sponsor company invests in data sets that include more sources to capture more of the total flow in the market. As noted in the State of Practice, section 2.1.4, DAT could be a better source of transportation flow data for analyzing this project. Using a dataset with more balanced commodity categories would improve the strength of the models using specific commodity flows to predict the target variables. The real estate data set was obtained directly from the sponsor company. The total area from the properties in the data set represent close to a billion square feet. According to Jones Lang LaSalle (JLL), the total industrial real estate market in the United States is 15 billion square feet (Jones Lang LaSalle, 2024). Thus, the data set from the sponsor company represents close to 6% of the industrial real estate area. Even though the data set contains over 13,000 properties, it only represents a small subset of the total industrial real estate market that may be biased towards markets where the sponsor company already invests or wants to invest. Additional public information from real estate broker sites could be integrated into the model to make it more robust and understand the supply of real estate in markets where currently the sponsor company is not investing. Thus, the model could now not only categorize markets as underserved or overserved but also understand if the market is underserved or overserved by only the company (using the company's data set) or by the whole industry (combining both the data set from the company and the public real estate data set).

An area of opportunity observed during the analysis was to ensure the sponsor company updates the data sets used for the analysis frequently. As mentioned in the Methodology section, data up to 2022 was used. It is recommended that the sponsor company updates the data sets with the most up-to-date data.

Since this project started in 2023, full-year 2022 was utilized. However, if the sponsor company decides to utilize this project's results in the future, it should be considered that the results might be outdated. Therefore, if the sponsor company replicates the analysis, they should ensure they update the data sets with the most up-to-date full-year data sets available. It is recommended that the company not limit the analysis to the current data sets obtained privately and publicly. Additional market features could help increase the model's robustness and provide better predictions. Some features worth looking at include:

- Parcel and small package flows in airports
- Rail flows, both intermodal and train, in major rail hubs
- E-commerce spends per capita by zip code
- Industrial property tax rates per zip code
- Industrial real estate vacancy rates per zip code

These larger, up-to-date data sets and additional features could provide the model with stronger training data to predict more trustworthy results to be used in future investment decisions.

6. CONCLUSION

This project was successful in answering the three research questions outlined at its inception. A repeatable methodology was created to enable the sponsor company to refresh and improve the outcomes in the future. An exploration of market feature relationships with HFT real estate volumes was conducted, providing key learnings to create new avenues for further research. Finally, a visualization tool was given to the sponsor company so that they can view and understand the project's results while considering future investment locations. While there are suggestions for further improvement and exploration, this project served as a starting point to explore the potential use of freight flow data to find underserved industrial real estate markets in the United States.

REFERENCES

Alvim, L. and Averbuch, M. (2023). US Nearshoring Wave Grows as Mexico Exports Jump Close to Record. Bloomberg. https://www.bloomberg.com/news/newsletters/2023-06-28/supply-chain-latest-us-nearshoring-proof-grows-as-mexico-exports-jump

Berman, J. (2023). Megawarehouse demand hits record high in 2022, reports CBRE. Logistics Management (2002), 62(3), 12–13.

Bloch, F., Jackson, M. O., & Tebaldi, P. (2023). Centrality measures in networks. Social Choice and Welfare, 61, 413-453. https://doi.org/10.1007/s00355-023-01456-4

Caplice, C. & Ponce, E. (2023). MITx MicroMasters Program in SCM Key Concepts, Facility Location Problems p. 128.

Cauble, T. (2023). The Ultimate Guide to Industrial Real Estate. The Cauble Group. https://www.tylercauble.com/blog/industrial-real-estate

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

DAT (2023). DAT iQ. https://www.dat.com/iq

DAT (2019). How to use a load board: DAT TruckersEdge [Video]. https://truckersedge.support.dat.com/video-load-board-tutorial-37310711

Elgin, S. (2022). Trucking Industry Trends, Statistics, & Forecast – 2023 Edition. https://www.truckinfo.net/research/trucking-statistics

Fuller, C. (2022). What causes the booms and busts of the trucking cycle? FreightWaves. https://www.freightwaves.com/news/what-causes-the-booms-and-busts-of-the-trucking-cycle

Georgules, J. (2017). How is net absorption calculated? https://www.jll.ca/en/trends-and-insights/cities/how-is-net-absorption-calculated

Gruetzemacher, R. (2022). The power of natural language processing. Harvard Business Review. https://hbr.org/2022/04/the-power-of-natural-language-processing

IBM. (2024). What is natural language processing (NLP)? https://www.ibm.com/topics/natural-language-processing

Jacobs, K. J. (2021). Build-to-Suit vs. Spec: Which Building is Right For a Specific Company? | NAIOP | Commercial Real Estate Development Association. https://www.naiop.org/research-and-publications/magazine/2021/summer-2021/development-ownership/build-to-suit-vs.-spec-which-building-is-right-for-a-specific-company

Jones Lang LaSalle IP, Inc. (2024). United States industrial outlook | Q1 2024. JLL Research. https://www.us.jll.com/en/trends-and-insights/research/industrial-market-statistics-trends

Kirk, P. (2017). Top Challenge for Industrial Investors is Finding Assets to Buy. National Real Estate Investor. https://www.proquest.com/docview/1895937370/abstract/54ECE4CB9B4F4A33PQ/1

Lieu, H. (1999). Traffic-Flow Theory. Public Roads Vol. 62 No. 4

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature machine intelligence, 2(1), 56-67.

Menczer, F., Fortunato, S., & David, C. A. (2020). A First Course in Network Science. Cambridge University Press.

Mordor Intelligence. (2023). United States Freight & Logistics Market - Growth, Trends, COVID-19 Impact, and Forecasts (2021 - 2026). https://www.mordorintelligence.com/industry-reports/united-states-freight-logistics-market

Nix, B., Osborne, R., Radharamanan, R. (2012). Supply Chain Management: Center of Gravity Method for the Placement of Distribution Centers p. 87. IEMS Conference, Mercer University

Pitz, P. (2017). Free zip zone and KMA wall map available. DAT Freight & Analytics. https://www.dat.com/blog/free-zip-zone-and-kma-wall-map-available

Price, J. (2023). Q3 2023 U.S. INDUSTRIAL MARKETBEAT. https://cw-gbl-gws-prod.azureedge.net/-/media/cw/marketbeat-pdfs/2023/q3/us-reports/industrial/us_industrial_marketbeat_q3_2023.pdf?rev=428e0b91ec574639986d4e9eb29aaae5

Rodrigue, J. (2020). The Geography of Transport Systems, 5th Edition. Chapter 5.1: Transportation Modes, Modal Competition and Modal Shift

USDOT (2023). Weight of shipments by transportation mode. U.S. Department of Transportation, Bureau of Transportation Statistics and Federal Highway Administration, Freight Analysis Framework, version 5.5. https://data.bts.gov/stories/s/Moving-Goods-in-the-United-States/bcyt-rqmu/

Wolf, C. (2023). Nearshoring Boom Faces Growing Pains. Transport Topics. https://www.ttnews.com/articles/nearshoring-growing-pains

APPENDICES

Appendix A – Cross-Dock Models

A1 – Cross-Dock Area Prediction Using Total Flow

OLS Regression Results

Dep. Variable:	y	R-squared:	0.672
Model:	OLS	Adj. R-squared:	0.667
Method:	Least Squares	F-statistic:	125.9
Date:	Wed, 01 May 2024	Prob (F-statistic):	8.56e-114
Time:	01:14:02	Log-Likelihood:	-430.54
No. Observations:	500	AIC:	879.1
Df Residuals:	491	BIC:	917.0
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.776e-17	0.026	1.07e-15	1.000	-0.051	0.051
Total Flow	0.1422	0.054	2.623	0.009	0.036	0.249
Katz Centrality	0.1433	0.056	2.564	0.011	0.034	0.253
Total_Households	0.2152	0.036	5.967	0.000	0.144	0.286
Weighted_Avg_Family_Income	-0.0717	0.028	-2.524	0.012	-0.128	-0.016
Total_FMWH_Area	0.2485	0.039	6.411	0.000	0.172	0.325
Total_IOS_Area	0.3084	0.038	8.046	0.000	0.233	0.384
border_crossing_import	0.0493	0.026	1.882	0.060	-0.002	0.101
Total_Port_Flows	0.0250	0.030	0.830	0.407	-0.034	0.084

Omnibus:	286.323	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4565.677
Skew:	2.140	Prob(JB):	0.00
Kurtosis:	17.172	Cond. No.	5.16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Test R2 Score: 0.5055110873910726
Test MSE: 236590020048.47208
Test RMSE: 486405.20150227845
Test MAE: 262616.6197880041
Test MAD: 355859.0479435882

A2 – Cross Dock Area Prediction Using Commodity Flows

OLS Regression Results

Dep. Variable:	y	R-squared:	0.687
Model:	OLS	Adj. R-squared:	0.678
Method:	Least Squares	F-statistic:	76.14
Date:	Wed, 01 May 2024	Prob (F-statistic):	1.23e-112
Time:	01:38:44	Log-Likelihood:	-418.86
No. Observations:	500	AIC:	867.7
Df Residuals:	485	BIC:	930.9
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.776e-17	0.025	1.09e-15	1.000	-0.050	0.050
Agricultural Products	-0.0354	0.026	-1.370	0.171	-0.086	0.015
Automotive and Machinery Parts	0.0413	0.029	1.416	0.157	-0.016	0.099
Construction and Building Materials	-0.0237	0.029	-0.823	0.411	-0.080	0.033
Consumer Goods	0.1268	0.033	3.819	0.000	0.062	0.192
Food and Beverages	0.0130	0.031	0.427	0.670	-0.047	0.073
Healthcare and Pharmaceuticals	-0.0934	0.028	-3.314	0.001	-0.149	-0.038
Industrial Goods	-0.0124	0.028	-0.435	0.664	-0.068	0.044
Katz Centrality	0.2142	0.042	5.072	0.000	0.131	0.297
Total_Households	0.1977	0.036	5.537	0.000	0.128	0.268
Weighted_Avg_Family_Income	-0.0674	0.028	-2.401	0.017	-0.123	-0.012
Total_FMWH_Area	0.2908	0.042	6.896	0.000	0.208	0.374
Total_IOS_Area	0.2833	0.040	7.107	0.000	0.205	0.362
border_crossing_import	0.0436	0.026	1.669	0.096	-0.008	0.095
Total_Port_Flows	0.0218	0.030	0.726	0.468	-0.037	0.081

Omnibus:	269.940	Durbin-Watson:	2.034
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3938.892
Skew:	2.002	Prob(JB):	0.00
Kurtosis:	16.154	Cond. No.	4.30

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Test R2 Score: 0.39759924591261064
Test MSE: 288220833374.813
Test RMSE: 536862.0245228871
Test MAE: 277544.33464283904
Test MAD: 355859.0479435882

A3 – Cross Dock Loading Positions Prediction Using Total Flow

OLS Regression Results

Dep. Variable:	y	R-squared:	0.578
Model:	OLS	Adj. R-squared:	0.571
Method:	Least Squares	F-statistic:	85.67
Date:	Wed, 01 May 2024	Prob (F-statistic):	9.95e-89
Time:	01:15:06	Log-Likelihood:	-502.56
No. Observations:	509	AIC:	1023.
Df Residuals:	500	BIC:	1061.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.606e-17	0.029	1.59e-15	1.000	-0.057	0.057
Total Flow	0.1389	0.057	2.430	0.015	0.027	0.251
Katz Centrality	0.1868	0.060	3.102	0.002	0.068	0.305
Total_Households	0.2956	0.040	7.402	0.000	0.217	0.374
Weighted_Avg_Family_Income	-0.0686	0.032	-2.150	0.032	-0.131	-0.006
Total_FMWH_Loading_Positions	0.0428	0.061	0.704	0.482	-0.077	0.162
Total_IOS_Area	0.3573	0.058	6.178	0.000	0.244	0.471
border_crossing_import	0.0363	0.029	1.236	0.217	-0.021	0.094
Total_Port_Flows	-0.0682	0.034	-2.036	0.042	-0.134	-0.002

Omnibus:	239.236	Durbin-Watson:	1.981
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2113.886
Skew:	1.838	Prob(JB):	0.00
Kurtosis:	12.282	Cond. No.	5.13

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Test R2 Score: 0.5724042982651893
Test MSE: 172468.21268238107
Test RMSE: 415.2929239493265
Test MAE: 244.54605721762294
Test MAD: 422.938720703125

A4 – Cross Dock Loading Positions Prediction Using Commodity Flows

OLS Regression Results

Dep. Variable:	y	R-squared:	0.577
Model:	OLS	Adj. R-squared:	0.566
Method:	Least Squares	F-statistic:	52.00
Date:	Wed, 01 May 2024	Prob (F-statistic):	7.45e-84
Time:	01:39:21	Log-Likelihood:	-503.09
No. Observations:	509	AIC:	1034.
Df Residuals:	495	BIC:	1093.
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.606e-17	0.029	1.58e-15	1.000	-0.057	0.057
Agricultural Products	-0.0453	0.030	-1.526	0.128	-0.104	0.013
Automotive and Machinery Parts	0.0558	0.034	1.647	0.100	-0.011	0.122
Construction and Building Materials	-0.0318	0.032	-0.983	0.326	-0.095	0.032
Consumer Goods	0.0461	0.034	1.360	0.174	-0.020	0.113
Food and Beverages	0.0509	0.035	1.437	0.151	-0.019	0.121
Healthcare and Pharmaceuticals	-0.0185	0.031	-0.591	0.555	-0.080	0.043
Industrial Goods	-0.0130	0.032	-0.407	0.684	-0.076	0.050
Katz Centrality	0.2672	0.047	5.724	0.000	0.175	0.359
Total_Households	0.2757	0.039	7.061	0.000	0.199	0.352
Total_FMWH_Loading_Positions	0.0644	0.064	1.014	0.311	-0.060	0.189
Total_IOS_Area	0.3297	0.061	5.383	0.000	0.209	0.450
border_crossing_import	0.0334	0.030	1.115	0.266	-0.025	0.092
Total_Port_Flows	-0.0761	0.034	-2.263	0.024	-0.142	-0.010

Omnibus:	243.204	Durbin-Watson:	2.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2145.384
Skew:	1.878	Prob(JB):	0.00
Kurtosis:	12.330	Cond. No.	5.55

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Test R2 Score: 0.5578853534535699
Test MSE: 178324.34372283553
Test RMSE: 422.28467142774144
Test MAE: 246.89211315318357
Test MAD: 422.938720703125

Appendix B – Final Mile Warehouse Models

B1 – Final Mile Warehouse Area Prediction Using Total Flow

OLS Regression Results

Dep. Variable:	y	R-squared:	0.632
Model:	OLS	Adj. R-squared:	0.620
Method:	Least Squares	F-statistic:	51.14
Date:	Wed, 01 May 2024	Prob (F-statistic):	1.52e-47
Time:	01:15:25	Log-Likelihood:	-226.94
No. Observations:	247	AIC:	471.9
Df Residuals:	238	BIC:	503.5
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-5.464e-17	0.039	-1.39e-15	1.000	-0.077	0.077
Total Flow	-0.0589	0.081	-0.725	0.469	-0.219	0.101
Katz Centrality	0.0229	0.090	0.254	0.799	-0.155	0.200
Total_Households	0.1680	0.052	3.233	0.001	0.066	0.270
Weighted_Avg_Family_Income	0.0324	0.043	0.759	0.448	-0.052	0.116
Total_Cross_Dock_Area	0.4019	0.065	6.221	0.000	0.275	0.529
Total_IOS_Area	0.3958	0.056	7.047	0.000	0.285	0.506
border_crossing_import	-0.0774	0.040	-1.916	0.057	-0.157	0.002
Total_Port_Flows	-0.0422	0.044	-0.968	0.334	-0.128	0.044

Omnibus:	171.908	Durbin-Watson:	1.694
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2485.685
Skew:	2.538	Prob(JB):	0.00
Kurtosis:	17.689	Cond. No.	5.16

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Test R2 Score: 0.5730265778291377
Test MSE: 24204122323677.48
Test RMSE: 4919768.523383746
Test MAE: 2001131.5653510948
Test MAD: 4027802.051630138

B2 – Final Mile Warehouse Area Prediction Using Commodity Flow

OLS Regression Results

Dep. Variable:	y	R-squared:	0.707
Model:	OLS	Adj. R-squared:	0.691
Method:	Least Squares	F-statistic:	43.24
Date:	Wed, 01 May 2024	Prob (F-statistic):	1.12e-54
Time:	01:39:32	Log-Likelihood:	-198.89
No. Observations:	247	AIC:	425.8
Df Residuals:	233	BIC:	474.9
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-5.464e-17	0.035	-1.54e-15	1.000	-0.070	0.070
Agricultural Products	0.0222	0.036	0.610	0.542	-0.049	0.094
Automotive and Machinery Parts	-0.1082	0.042	-2.555	0.011	-0.192	-0.025
Construction and Building Materials	0.2075	0.041	5.117	0.000	0.128	0.287
Consumer Goods	-0.0128	0.044	-0.288	0.774	-0.100	0.075
Food and Beverages	-0.0670	0.044	-1.516	0.131	-0.154	0.020
Healthcare and Pharmaceuticals	0.2011	0.037	5.433	0.000	0.128	0.274
Industrial Goods	0.0174	0.041	0.429	0.668	-0.063	0.098
Katz Centrality	-0.0834	0.061	-1.378	0.170	-0.203	0.036
Total_Households	0.1588	0.046	3.477	0.001	0.069	0.249
Total_Cross_Dock_Area	0.3599	0.061	5.947	0.000	0.241	0.479
Total_IOS_Area	0.4787	0.052	9.212	0.000	0.376	0.581
border_crossing_import	-0.0513	0.038	-1.354	0.177	-0.126	0.023
Total_Port_Flows	-0.0642	0.040	-1.599	0.111	-0.143	0.015

Omnibus:	88.812	Durbin-Watson:	1.828
Prob(Omnibus):	0.000	Jarque-Bera (JB):	655.075
Skew:	1.209	Prob(JB):	5.65e-143
Kurtosis:	10.603	Cond. No.	3.95

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Test R2 Score: 0.6794516247308062
Test MSE: 18171135913389.19
Test RMSE: 4262761.536068982
Test MAE: 1898405.5467966208
Test MAD: 4027802.051630138

B3 – Final Mile Warehouse Loading Positions Prediction Using Total Flow

OLS Regression Results

Dep. Variable:	y	R-squared:	0.445
Model:	OLS	Adj. R-squared:	0.428
Method:	Least Squares	F-statistic:	26.11
Date:	Wed, 01 May 2024	Prob (F-statistic):	4.18e-26
Time:	01:15:32	Log-Likelihood:	-265.40
No. Observations:	236	AIC:	546.8
Df Residuals:	228	BIC:	574.5
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.735e-17	0.049	3.52e-16	1.000	-0.097	0.097
Total Flow	0.0058	0.120	0.049	0.961	-0.231	0.243
Katz Centrality	0.0567	0.126	0.450	0.653	-0.191	0.305
Total_Households	0.2228	0.062	3.573	0.000	0.100	0.346
Total_Cross_Dock_Loading_Positions	0.2184	0.075	2.906	0.004	0.070	0.367
Total_IOS_Area	0.3623	0.059	6.152	0.000	0.246	0.478
border_crossing_import	0.0083	0.050	0.165	0.869	-0.091	0.108
Total_Port_Flows	-0.0139	0.054	-0.257	0.797	-0.120	0.092

Omnibus:	253.242	Durbin-Watson:	2.124
Prob(Omnibus):	0.000	Jarque-Bera (JB):	10353.546
Skew:	4.301	Prob(JB):	0.00
Kurtosis:	34.287	Cond. No.	6.01

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Test R2 Score: 0.8216434903211736
Test MSE: 189605.27757680372
Test RMSE: 435.4368812776471
Test MAE: 290.18489786733466
Test MAD: 484.71999999999997

B4 – Final Mile Warehouse Loading Positions Prediction Using Commodity Flow

OLS Regression Results

Dep. Variable:	y	R-squared:	0.579
Model:	OLS	Adj. R-squared:	0.554
Method:	Least Squares	F-statistic:	23.45
Date:	Wed, 01 May 2024	Prob (F-statistic):	7.94e-35
Time:	01:39:39	Log-Likelihood:	-232.88
No. Observations:	236	AIC:	493.8
Df Residuals:	222	BIC:	542.2
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.735e-17	0.044	3.98e-16	1.000	-0.086	0.086
Agricultural Products	0.0218	0.046	0.475	0.636	-0.069	0.112
Automotive and Machinery Parts	-0.2430	0.051	-4.743	0.000	-0.344	-0.142
Construction and Building Materials	0.3044	0.051	5.954	0.000	0.204	0.405
Consumer Goods	0.0546	0.059	0.925	0.356	-0.062	0.171
Food and Beverages	-0.0966	0.054	-1.779	0.077	-0.204	0.010
Healthcare and Pharmaceuticals	0.1393	0.046	3.000	0.003	0.048	0.231
Industrial Goods	0.0266	0.049	0.546	0.585	-0.069	0.123
Katz Centrality	-0.0571	0.081	-0.706	0.481	-0.217	0.102
Total_Households	0.1640	0.055	2.961	0.003	0.055	0.273
Total_Cross_Dock_Loading_Positions	0.2395	0.067	3.586	0.000	0.108	0.371
Total_IOS_Area	0.4872	0.058	8.419	0.000	0.373	0.601
border_crossing_import	0.0290	0.045	0.646	0.519	-0.059	0.117
Total_Port_Flows	-0.0562	0.048	-1.161	0.247	-0.152	0.039

Omnibus:	200.783	Durbin-Watson:	1.954
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4887.293
Skew:	3.135	Prob(JB):	0.00
Kurtosis:	24.394	Cond. No.	4.17

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Test R2 Score: 0.709083829601859
Test MSE: 309263.96428842243
Test RMSE: 556.1150638927365
Test MAE: 338.53465382451384
Test MAD: 484.71999999999997

Appendix C – Industrial Outdoor Storage Models

C1 – Industrial Outdoor Storage Area Prediction Using Total Flow

OLS Regression Results

Dep. Variable:	y	R-squared:	0.618
Model:	OLS	Adj. R-squared:	0.598
Method:	Least Squares	F-statistic:	31.35
Date:	Wed, 01 May 2024	Prob (F-statistic):	8.20e-29
Time:	01:15:40	Log-Likelihood:	-153.79
No. Observations:	164	AIC:	325.6
Df Residuals:	155	BIC:	353.5
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-6.939e-18	0.050	-1.4e-16	1.000	-0.098	0.098
Total Flow	0.0054	0.112	0.048	0.962	-0.216	0.227
Katz Centrality	0.0787	0.125	0.629	0.530	-0.168	0.326
Total_Households	-0.1284	0.065	-1.962	0.052	-0.258	0.001
Weighted_Avg_Family_Income	0.0515	0.053	0.970	0.334	-0.053	0.156
Total_FMWH_Area	0.4796	0.073	6.574	0.000	0.336	0.624
Total_Cross_Dock_Area	0.3018	0.083	3.635	0.000	0.138	0.466
border_crossing_import	0.0195	0.052	0.373	0.710	-0.084	0.123
Total_Port_Flows	0.2336	0.055	4.249	0.000	0.125	0.342

Omnibus:	53.460	Durbin-Watson:	1.793
Prob(Omnibus):	0.000	Jarque-Bera (JB):	310.755
Skew:	1.020	Prob(JB):	3.31e-68
Kurtosis:	9.428	Cond. No.	5.96

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Test R2 Score: 0.1318213072186566
Test MSE: 12771720298397.564
Test RMSE: 3573754.3701823666
Test MAE: 1423572.518211438
Test MAD: 1722561.6365195254

C2 – Industrial Outdoor Storage Area Prediction Using Commodity Flow

OLS Regression Results

Dep. Variable:	y	R-squared:	0.674
Model:	OLS	Adj. R-squared:	0.648
Method:	Least Squares	F-statistic:	25.97
Date:	Wed, 01 May 2024	Prob (F-statistic):	6.87e-31
Time:	01:39:47	Log-Likelihood:	-140.89
No. Observations:	164	AIC:	307.8
Df Residuals:	151	BIC:	348.1
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-6.939e-18	0.046	-1.49e-16	1.000	-0.092	0.092
Automotive and Machinery Parts	0.1680	0.062	2.727	0.007	0.046	0.290
Construction and Building Materials	-0.1508	0.055	-2.757	0.007	-0.259	-0.043
Consumer Goods	-0.0497	0.067	-0.737	0.462	-0.183	0.083
Food and Beverages	0.0036	0.058	0.062	0.951	-0.112	0.119
Healthcare and Pharmaceuticals	-0.1309	0.051	-2.569	0.011	-0.232	-0.030
Industrial Goods	0.0664	0.055	1.199	0.233	-0.043	0.176
Katz Centrality	0.0898	0.091	0.984	0.327	-0.091	0.270
Total_Households	-0.0926	0.061	-1.507	0.134	-0.214	0.029
Total_FMWH_Area	0.6102	0.072	8.467	0.000	0.468	0.753
Total_Cross_Dock_Area	0.1557	0.083	1.870	0.063	-0.009	0.320
border_crossing_import	-0.0087	0.051	-0.170	0.865	-0.110	0.093
Total_Port_Flows	0.2501	0.052	4.800	0.000	0.147	0.353

Omnibus:	60.199	Durbin-Watson:	1.650
Prob(Omnibus):	0.000	Jarque-Bera (JB):	245.739
Skew:	1.329	Prob(JB):	4.35e-54
Kurtosis:	8.376	Cond. No.	4.76

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Test R2 Score: 0.6069269003687947
Test MSE: 5782472810097.273
Test RMSE: 2404677.2777437875
Test MAE: 1241061.1635053742
Test MAD: 1722561.6365195254