UNIVERSITÀ DEGLI STUDI DI GENOVA

SCUOLA POLITECNICA

DITEN

Department of Naval Engineering

MASTER'S THESIS

ENGINEERING TECHNOLOGY FOR STRATEGY AND SECURITY

2025

RESEARCH AND OPTIMIZATION OF DELIVERY PROCESSES BASED ON THE ANALYSIS OF LOGISTIC DATA

Student:

Alisher Juanyspay (5595745)

Supervisor:

Alessandro Sorce

2025

CONTENT

ABSTRACT **3**

INTRODUCTION **4**

MAIN PART **5**

1.1 Introduction to Logistics 5
1.2 Types of logistics 6
1.3 Data Science and Analytics 9
2.1 Collection and processing of historical data of Regal Export LLP 13
2.2. Data Analysis 18
2.2.1 Visualization *18*
2.2.2 Correlation Analysis *25*
2.3 Delivery modeling: time and cost forecast 33
2.3.1 Cost *34*
2.3.2 Delivery time *46*
2.4 Discount system 53
2.5 Implementation 57

CONCLUSION **61**

BIBLIOGRAPHY **62**

ABSTRACT

The thesis includes theoretical and practical parts. The theoretical part analyzes the roles of analytics and data science in logistics, which are becoming an important aspect of successful corporate management in modern conditions. The use of data, analytical methods and technologies to improve transport, warehousing and distribution processes, as well as to optimize supply chains is called logistics analytics. Logistics analytics allows companies to reduce costs, improve operational efficiency and make informed decisions.

Data science in logistics is used to analyze large amounts of data, discover patterns, and create models that can predict important metrics. In this study, data science is used to predict delivery times and transportation costs. Logistics companies can use machine learning and correlation analysis to uncover hidden relationships in data. This approach helps to increase forecasting accuracy and develop more effective flow management strategies. The main data science methods, such as statistical analysis, data visualization and machine learning, will be discussed in detail in the theoretical part. It will be demonstrated how these analytical methods help to solve problems related to demand forecasting, inventory management, cost reduction and route optimization in logistics. This chapter will also consider the evolution of analytics in logistics and its impact on modern procedures.

The analysis of logistics data is a key part of the applied aspect of this study. The main stages include the creation of predictive models for estimating delivery times and transport costs, studying correlations between factors affecting the delivery process, and collecting and pre-processing data. The result of this endeavor is a significant study that integrates theory and practice. The theoretical part will provide a deep understanding of analytical methods, while the practical part will demonstrate the application of these methods to solve real logistics problems. The outcome of the study is the development of a model that helps to improve the accuracy of forecasting key indicators of logistics operations and highlights the importance of analytics and data science in the modern economic context.

INTRODUCTION

Relevance of the work

In today’s world, where markets are constantly changing and becoming more competitive, business process efficiency is critical. As a vital link in the supply chain, logistics must be continuously optimized. There are new opportunities to increase productivity, reduce costs, and improve customer service when machine learning techniques are used in logistics operations.

The volume of data generated by logistics companies is growing exponentially, which creates a need for tools for analyzing big data and making informed decisions. Machine learning provides such opportunities, allowing not only to analyze current processes, but also to predict key parameters, such as delivery time and transportation costs.

The relevance of the work is due to the need to optimize logistics operations in a highly competitive environment and ever-increasing demands on quality and speed of delivery. The use of analytics and machine learning methods in logistics allows not only to increase operational efficiency, but also to strengthen the company's position in the market by improving customer experience.

The main goal of the project is to collect, process and study historical data of a logistics company to develop and evaluate machine learning models for predicting delivery time and transportation cost, as well as to study their impact on process optimization. This research contributes to the development of analytical approaches in logistics and demonstrates the practical value of using data science to solve real-world problems.

Objectives:

1. Explore and evaluate logistics data to find critical factors that impact delivery schedules and transportation costs, turning raw data into actionable insights that improve understanding of operational dynamics.
2. Create and apply advanced machine learning techniques to accurately forecast delivery schedules and transportation costs, turning complex data sets into useful tools for strategic planning and decision making.
3. Create a loyalty discount program with customized rewards to increase brand loyalty, retain repeat business, and gain a competitive advantage in the logistics industry.

MAIN PART

1.1 Introduction to Logistics

Logistics is a set of organizational, managerial, production and technological processes to effectively ensure the organization of the movement of material and other resources[1].

A broader definition of logistics interprets it as the study of planning, management and control of the movement of material, information and financial resources in various systems[2].

From the point of view of practical application, logistics is the choice of the most effective option for providing goods of the right quality, the right quantity, at the right time, in the right place with minimal costs[3].

The content of **logistics as a science** is the establishment of cause-and-effect relationships and patterns inherent in the process of commodity distribution, in order to determine and implement in practice effective organizational forms and methods of managing material and information flows.

The main **objects of research** in logistics are:

- logistics operations;
- supply chains;
- logistics systems;
- logistics functions;
- material flows;
- information flows;
- logistics costs.

Examples of **problems solved in logistics**:

- choosing the type of vehicle,
- determining routes,
- optimal packaging of goods in containers,
- determining the optimal placement in warehouse areas, marking, formation of group orders.

Depending on the specifics of the company's activities, various logistics systems are used. A logistics system is a set of actions of participants in the logistics chain (manufacturers, transport, trading organizations, stores, etc.), built in such a way that the main tasks of logistics are performed.

Some **management approaches and concepts** that include a logistics component or specific logistics strategies:

- MRP (materials requirements planning),
- DRP (distribution requirements planning),
- MRP II (manufacturing resource planning),
- ERP (enterprise resource planning);
- CSRP (customer synchronized resource planning),

- EOQ model,
- two level system,
- two-hopper scheme,
- model with constant order frequency,
- ABC analysis,
- non-stationary and stochastic models of stock management.
- Distribution (business)

Companies can develop their own logistics departments, or they can attract transport and logistics organizations to resolve supply, warehousing and procurement issues. Depending on the level of involvement of independent companies to solve business problems in logistics, different levels are distinguished: 1PL (from English first-party logistics) - an approach in which an organization turns to a specialist company in a separate logistics operation: warehouse (storage), post office (information exchange), taxi (transport); 3PL (third-party logistics) is an approach in which the full range of logistics services from delivery and address storage to order management and tracking the movement of goods is transferred to the side of the transport and logistics organization. The functions of such a ***3PL*** provider include organizing and managing transportation, accounting and inventory management, preparation of import-export and freight documentation, warehousing, cargo processing, and delivery to the end consumer.

The task of logistics management in practice comes down to managing several components that make up the so-called ***"logistics mix"***:

- warehouse buildings (separate warehouse buildings, distribution centers, warehouses combined with a store);
- reserves (volume of reserves for each item, location of the reserve);.
- transportation (types of transport, terms, types of packaging, availability of drivers, etc.);
- picking and packaging (simplicity and ease in terms of logistics services while maintaining an impact on purchasing activity);
- communication (the ability to obtain both final and intermediate information in the process of product distribution).

1.2 Types of logistics

Logistics is divided into types: procurement, distribution, transport, warehouse, production, information logistics[4].

Procurement logistics. The main goal of procurement logistics is to satisfy production with materials with maximum economic efficiency, quality and the shortest possible time. Purchasing logistics involves the search and selection of alternative manufacturing suppliers.

The main methods of procurement logistics are traditional and operational methods. The traditional method is carried out by supplying the required amount of goods at a time, and the operational method is carried out as the goods are needed.

Distribution logistics is an area of scientific research into the system integration of functions implemented in the process of distributing material and accompanying (information, financial and service) flows between various consumers, that is, in the process of selling goods, the main goal of which is to ensure the delivery of the right goods to the right place, at the right time with optimal costs. Thus, distribution logistics is a set of interrelated functions implemented in the process of distributing material flow between various wholesale buyers and sellers. Closely related to the concept of sales logistics is the concept of a distribution channel - a set of various organizations that deliver goods to the consumer.

Transport logistics is a system for organizing delivery, namely for moving any material objects, substances, etc. from one point to another along the optimal route. More detailed functions of this logistics are:

- personnel who carry out these tasks (loaders, drivers);
- classification of vehicles (by volume, speed, maneuverability);
- pricing policy (for labor, for fuels and lubricants, provision of transport services).

The transport and logistics system are understood as a set of consumers and producers of services, as well as the management systems, vehicles, communications routes, structures and other property used to provide them. Another definition states that a transport and logistics system is a set of objects and subjects of transport and logistics infrastructure, together with material, financial and information flows between them, performing the functions of transportation, storage, distribution of goods, as well as information and legal support of commodity flows[5].

Customs logistics is a set of activities aimed at moving cargo across the border while minimizing the costs of these procedures.

Customs logistics solves the following problems:

- transportation of imported and exported cargo;
- obtaining the necessary certificates for imported/exported goods;
- registration of customs documentation;
- assessment of the value, condition and compliance with customs requirements of the cargo;
- support for further movement of cargo that has passed the customs border.

Inventory logistics. Inventory management policy consists of decisions - what to purchase or produce, when and in what volumes. It also includes decisions about inventory allocation at manufacturing plants and distribution centers. The second element of inventory management policy concerns strategy.

You can manage the inventory of each distribution warehouse separately, or you can centrally (requires more coordination and information support).

Enterprise inventory management is an integrated process that ensures inventory transactions within the company and outside it - throughout the supply chain.

The main task of **warehouse logistics** is to optimize the business processes of acceptance, processing, storage and shipment of goods in warehouses. Warehouse logistics determines the rules for organizing warehousing, procedures for working with goods and the corresponding resource management processes (human, technical, information). In this case, the most common methods are used: FIFO, LIFO, FEFO, FPFO, BBD. For information and technical support of such processes, specialized WMS warehouse management systems can be used.

Complex logistics. A systematic approach to organizing the entire life cycle of a product and related activities in the period from the moment of production of its components to the moment of consumption. This is an effective system for managing material, information and financial flows associated with the product life cycle. An integrated approach to the logistics process makes it possible to reduce or neutralize the risks of uncertainty, which influence the functional life cycle of a product.

Environmental logistics ensures the movement of material during any production processes up to its transformation into a marketable product and, further, into waste, followed by waste management until disposal or safe storage in the environment[6]. Environmental logistics also ensures the collection and sorting of waste generated from the consumption of commercial products, their transportation, disposal or safe storage in the environment. It allows for radical cleanup of large areas contaminated with unauthorized waste.

Lean logistics. The synthesis of logistics and the concept of lean manufacturing made it possible to create a pull system that unites all organizations involved in the value stream, in which partial replenishment of stocks occurs in small batches. The principles of lean technology extend to the areas of logistics, warehouse, inventory and transport management within enterprises, and then to the management of flows external to factories[7]. Lean logistics uses the principle of total logistics cost (Total Logistics Cost, TLC), which makes it possible to reduce inventories throughout the chain, reduce transportation and storage costs, and establish logistics cooperation.

City logistics (city logistics, municipal logistics) is a complex of logistics solutions, actions, processes aimed at optimizing management decisions of the administration, flows of materials, vehicles, people, knowledge, energy, finance, information within the subsystems of the city and its infrastructure.

Logistics providers. When solving logistics problems, companies can rely on their own efforts or engage the services of logistics providers (logistics outsourcing). The following types of logistics providers are distinguished:

1PL is a small company operating locally or in its niche of logistics services.

2PL - organizes the transportation of goods from point to point, but at the same time remains only an intermediary (all contracts are concluded by the cargo owner).

3PL - a provider can organize several logistics services, that is, for example, become both a carrier and a warehouse operator (in some cases, the provider becomes a fulfillment operator).

4PL - 3PL + management logistics. Management logistics is based on optimization criteria (cost, safety, speed). The cargo owner can stipulate not only the transportation itself, but also additional criteria (for example, minimize the budget, or deliver as soon as possible, or ensure safety, or other).

1.3 Data Science and Analytics

In my thesis, using an analytical approach, working with data, I set myself the task of optimizing the work of the Regal Export company. Before proceeding directly to work, let's first understand what data science is, what it serves, what methods it uses and what tasks it solves. We will also compare it with data analytics, the processes of their work, and finally, reveal this fine line that separates analytics and data science.

Data science and analytics - both help extract valuable information from data, data analytics focuses more on analyzing historical data to make decisions in the present. In contrast, data science enables you to create data-driven algorithms to predict future outcomes.

Data Science is a set of specific disciplines from different areas responsible for analyzing data and finding optimal solutions based on it. Previously, only mathematical statistics dealt with this, then machine learning and artificial intelligence began to be used, which added optimization and computer science to mathematical statistics as methods of data analysis. The goal is to provide information that is not only descriptive (explaining what happened), but also predictive (predicting what might happen) and prescriptive (recommending action)[8].

Data science covers the entire lifecycle of data, from collection and cleaning to analysis and visualization. Data scientists use a variety of tools and techniques, such as machine learning, predictive modeling, and deep learning, to uncover hidden patterns and make predictions from data. Here are the essential components of data science:

Data collection: Accumulating data from various sources, such as databases, APIs, and web scraping.

Data cleaning and preprocessing: Ensuring data quality by managing missing values, removing duplicates, normalizing data, and preparing it for analysis.

Exploratory data analysis (EDA): Using statistical methods and visualization tools to understand the distribution and relationships of data[9].

Model building: Building and training machine learning models to predict outcomes and classify data.

Evaluation and optimization: Assessing model performance using accuracy, precision, and recall metrics, and refining models to improve accuracy.

Deployment: Deploy models in production for real-time forecasting and automated decision making.

While data analytics is a part of data science, it examines historical data to identify trends, patterns, and insights. It helps you systematically use statistical and quantitative methods to process data and make informed decisions.

4 Types of Data Analytics:

Descriptive Analytics - Make better predictions about future events based on historical data.

Diagnostic Analytics - Understand why something happened by analyzing data from previous events.

Predictive Analytics - Predict what will happen in the future by analyzing data from previous events.

Prescriptive Analytics - Predict & recommend how best to achieve those results by analyzing data from previous events.

The primary goal of data analytics is to analyze historical data to answer specific business questions, identify patterns, trends, and insights, and help companies make informed decisions.

For example, the goal of analytics might be to understand the factors that influence customer churn or optimize marketing campaigns to improve conversion rates.

Analysts use data analytics to create detailed reports and dashboards that help companies track key performance indicators (KPIs) and make data-driven decisions. Data analytics is typically simpler and less complex than data science because it does not require advanced machine learning algorithms or model building.

Both data science and analytics involve working with data and can be used to predict future outcomes. However, the crucial difference lies in the scope and depth of their approaches.

Data analytics is typically more focused and tends to answer specific questions based on past data. It is about analyzing data sets to gain actionable insights that will help businesses make informed decisions. While it may use predictive analytics to predict future trends, its primary goal is to understand what happened and why.

Data science, on the other hand, is a broader field that includes data analysis and other techniques such as machine learning, artificial intelligence (AI), and deep learning. Data scientists often work on more complex problems and use advanced algorithms and models to predict future events and automate decision making, resulting in new data-driven products and features.

In other words, while data analytics can provide insights and inform decisions, data science uses data to produce systems that can understand data and make decisions or predictions. It's like the difference between understanding data and creating new ways to interact with it. Both are valuable, but they serve different purposes and require different skills.

	Data science	Data Analysis
Scope and tasks	Broad and exploratory. The project aims to discover new ideas and create predictive models to forecast future trends.	Narrow and specific. The focus is on answering pre-defined questions and analyzing historical data to make decisions.
Methodologies	Uses advanced AI and machine learning algorithms and statistical models to analyze structured and unstructured data.	Uses statistical and data visualization methods, primarily working with structured data.
Results	Creates predictive models and algorithms that can automate decision-making processes and reveal hidden patterns.	Creates reports and dashboards that summarize past performance and provide actionable insights for business strategies.

The processes involved in data science and analytics also differ, reflecting their different goals and methodologies.

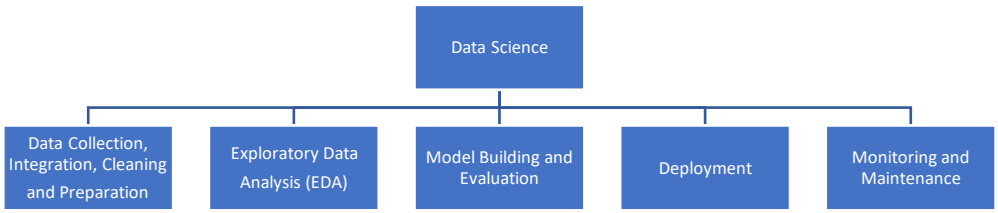
1. ***Setting Goals***: The first step in any analytics project is to establish clear and measurable goals with stakeholders. These goals should be aligned with overall business goals and should be specific, measurable, achievable, relevant and time-bound. Stakeholders can be anyone from executives and managers to end users who have a vested interest in the results of the analytics project.

2. ***Data Collection and Integration***: In this phase, we need to collect data from various sources such as databases, data warehouses, data lakes, online services, and user forms. Data warehouses and data lakes play a key role here. They store large amounts of structured and unstructured data respectively and provide a central repository for data that has been cleaned, integrated, and ready for analysis.

3. ***Data Cleaning***: Data cleaning ensures data quality by fixing errors, eliminating missing values, and standardizing formats. Tools like SQL for structured data and Hadoop or Spark for big data can be used in this process. It's all about making sure the data is reliable and ready for analysis.

4. ***Data Analysis and Visualization***: Now it's time to explore the data and discover patterns and trends. Using statistical methods, we then create a visual representation of the data to help understand the patterns and trends. Tools like Tableau, Power BI, or libraries like Matplotlib and Seaborn in Python, help create highly effective visualizations[10].

5. ***Data reporting***: Finally, we need to summarize our findings in reports and dashboards so that they are easy to understand and answer the business questions that started the process. Reporting tools like Tableau and PowerBI allow you to create interactive dashboards that decision makers can use to get the insights they need[11].



1. ***Data Collection and Integration***: In this phase, we need to collect large data sets from various areas such as unstructured sources, databases, APIs, and web scraping. Once the data is collected, it is then subjected to integration. Data integration combines data from multiple sources into a single view. This involves data transformation, cleaning, and loading to transform the raw data into the correct state. The integrated data is then stored in a Data Warehouse or Data Lake. These storage systems are important in the field of data analytics and data science, providing the necessary infrastructure to store and process large amounts of data.

Data Cleaning and Preparation: Data cleaning and preparation involves pre-processing the data to make it suitable for analysis. This includes handling missing values that can be filled in using various imputation methods and dealing with outliers that may skew the results. The data is also transformed into a suitable format for analysis, such as normalizing numeric data or coding categorical data.

2. ***Exploratory Data Analysis*** (EDA): The purpose of EDA is to uncover initial insights. It involves visualizing data using graphs and charts to identify patterns, trends, and relationships between variables. Summary statistics are also calculated to provide a quantitative description of the data.

3. ***Model building and evaluation***: This step uses machine learning algorithms to create predictive models. The choice of algorithm depends on the nature of the data and the problem being solved. Data science teams split the data into two sets: training and testing sets. They train the model on the training set. After building a model, teams evaluate its performance using metrics such as accuracy, precision, and recall. These metrics provide insight into how well the model correctly predicts outcomes.

4. ***Deployment***: Finally, you are ready to share your findings. Once the model is evaluated and tuned, it is deployed in a live environment to make automated decisions. You must plan the deployment, monitor and maintain the model, prepare a final report, and review the project.

5. ***Monitoring and Maintenance***: Teams continually monitor the model's performance after deployment to ensure it remains effective over time. If the model's performance degrades, it may need to be adjusted or retrained using new data. This step is vital to ensure the model remains relevant as new data arrives.

2.1 Collection and processing of historical data of Regal Export LLP

As we move on to the application of our research, we highlight the importance of data as one of the most important resources in contemporary business. increasingly needed for process analysis and optimization. The dataset utilized in this thesis offers a foundational framework for the implementation of machine learning techniques in the field of logistics.

This data set represents the actual operations of the company, including its efficacy, flaws, and strengths. It is more than just a collection of cold, hard figures

and facts. It has data that may be utilized to spot trends, forecast patterns, and—most importantly—optimize internal business operations. This data collection will be thoroughly examined in the parts that follow, including its structure, different data kinds, important metrics and indicators, and data pretreatment techniques required for additional modeling and analysis. We'll look at how this data may be used to gain insightful knowledge that can make a logistics company more competitive and adaptable in a world that is changing quickly.

The dataset used in this study is a rich set of two related files, each with different qualities and properties needed to evaluate and improve the performance of a logistics company. When combined, these files create a multidimensional data space that allows for detailed and in-depth analysis.

The data was collected based on the actual operations of the logistics company. All data was carefully collected and systematized in ***Excel*** format, which ensured flexibility and ease of use. The company's internal accounting systems were used to collect the data, providing information on delivery times, transportation costs, routes, cargo, and customer characteristics. Particular attention was paid to the harmonization of data between different sources to minimize errors and duplication of information.

Continuing to work with the dataset, the first step was the process of loading it into ***Python*** for further processing and analysis. Python was chosen as the main tool for data analysis due to its versatility and wide range of specialized libraries. The main goal at this stage is to access the data loaded in Excel format, then perform a preliminary inspection and ensure its correctness.

To load the data, I used the ***pandas*** library, which is a tool for working with tabular data. It allows you to easily load data from various formats, including Excel, and provides a convenient interface for their manipulation.

The first step is to import the pandas library into the workspace. The **pd.read_excel()** method is used to read data from the specified file. The file path is specified as a string, which allows Python to accurately find and open the file. The **.head()** method is used to check if the load was successful and to get an initial look at the data content. It displays the first five rows, which helps to understand the table structure, column names, and data types.

```
# Loading data and previewing top records

import pandas as pd

data = pd.read_excel('/Users/alisherbilyaluly/Desktop/Regal_Export/Regal_export_sh_d.xlsx')
data.head()
```

	SH_ID	Sent_date	Delivery_date	C_ID	SH_CONTENT	SH_DOMAIN	SER_TYPE	SH_WEIGHT	SH_CHARGES
0	1	2023-01-01	2023-01-11	4574	Electronics	International	Regular	27	77.22
1	2	2023-01-01	2023-01-08	6200	Hazardous Goods	Domestic	Express	123	498.15
2	3	2023-01-01	2023-01-09	8404	Industrial Equipments	Domestic	Express	2434	9200.52
3	4	2023-01-01	2023-01-10	5200	Construction	Domestic	Express	4935	15989.40
4	5	2023-01-01	2023-01-04	4571	Arts and crafts	Domestic	Express	1	2.16

source_latitude	source_longitude	destination_latitude	destination_longitude	distance_km	delivery_time
55.0084	82.9357	50.4241	80.2270	541.883826	10
50.2875	57.1798	51.1694	71.4491	1010.661766	7
43.3000	68.2500	49.9937	82.6127	1324.128087	8
49.9937	82.6127	49.8028	73.0877	684.208640	9
50.0377	72.9501	51.7226	75.3689	253.175071	3

The basis of the data structure is a unique shipment identifier (***SH_ID***), which allows each record to be uniquely identified and information about a specific shipment to be tracked.

Each shipment has a dispatch date (***Sent_date***) and delivery time (***delivery_time***), which allows for an assessment of service deadlines. This data is supplemented by an actual delivery date (***Delivery_date***).

An additional customer identifier (***C_ID***) links each shipment to a specific customer, which allows for an analysis of customer preferences and shipment volumes. The cargo category (***SH_CONTENT***) reflects the contents of the shipment, including types such as electronics, hazardous goods, industrial equipments, construction, art and etc. The data also contains information about the delivery domain (***SH_DOMAIN***), indicating whether the shipment is international or domestic, and the type of service provided (***SER_TYPE***), including regular and express delivery.

Key metrics for analysis are the weight of the shipment (***SH_WEIGHT***) and the cost of delivery (***SH_CHARGES***), which allows for an assessment of cost efficiency and the accuracy of calculations. The geographic coordinates of the origin (***source_latitude***, ***source_longitude***) and destination (***destination_latitude***, ***destination_longitude***) points allow for spatial analysis of routes, providing the basis for studying delivery routes and optimizing logistics processes.

	C_ID	C_FULL_NAME	C_CONT_N	C_EMAIL	DISCOUNT_%
0	4574	Amy Jackson	4712797612	xgay@example.com	5.0
1	6200	Joseph Myers	+1-755-999-5804x1179	kharris@example.org	1.5
2	8404	Sarah Nolan	001-750-597-5394x069	rodriguezdavid@example.net	7.0
3	5200	Heather Moore	508-260-3947x0376	banderson@example.net	0.0
4	4571	Tammy Rice	(244)218-8589x5821	noah59@example.org	7.0

The following data file contains information about the logistics company's customers, complementing the core delivery data set and enabling in-depth analysis of customer interactions and their impact on operational processes. Each record is identified by a unique customer identifier (***C_ID***), which serves as a key element for linking customer data to the corresponding deliveries.

Customer data includes the full name ***(C_FULL_NAME***), which enables personalized analysis of the customer base. Contact information is represented by phone numbers (***C_CONT_N***), which have a varied format reflecting the international nature of interaction, and email addresses (***C_EMAIL***), which provide opportunities for communication and marketing campaigns.

In addition, for each customer, the percentage of discount (***DISCOUNT_%)*** provided within the framework of the cooperation is indicated. This metric is an important element of analysis, allowing to assess the impact of the discount policy on the company's financial performance, as well as to identify customer segments that are of the greatest value to the business.

This data set is not only a means of identifying customers, but also a rich source for analyzing consumer behavior, creating customer profiles, and developing personalized strategies. It opens up great opportunities for integrating shipping and customer data, leading to more accurate forecasts, improved customer experience, and increased overall efficiency of a company's logistics operations.

An important stage of my work was calculating the delivery distance based on the geographical coordinates of the departure and destination points. Initially, the data set did not contain a column with distances, which limited the possibilities of analyzing logistics routes. Therefore, I developed a method for automatically calculating the distances between coordinates and adding this information to the data table.

```python
# Function for calculating distance
def calculate_distance(row):
    source_coords = (row['source_latitude'], row['source_longitude'])
    dest_coords = (row['destination_latitude'], row['destination_longitude'])

    # Check for coincidence of coordinates
    if source_coords == dest_coords:
        return 1  # minimum value 1 km
    else:
        # Calculate the distance between points
        return distance(source_coords, dest_coords).km

# Apply a function to each row of data
data['distance_km'] = data.apply(calculate_distance, axis=1)

# Saving the results back to Excel
data.to_excel('/Users/alisherbilyaluly/Desktop/Regal_Export/Regal_export_sh_d_updated.xlsx', index=False)

# Checking the result
data.head()
```

The ***calculate_distance*** function was used for calculations, which processed each row of the data set. The algorithm is based on the use of geographical coordinates - latitude and longitude of the starting and ending points of delivery. Coordinates were extracted from the corresponding table columns (source_latitude, source_longitude, destination_latitude, destination_longitude), after which they were passed to the distance function from the geopy library. This library allowed you to accurately calculate the distance between two points on the map in kilometers, taking into account the curvature of the Earth.

Particular attention was paid to handling cases where the coordinates of the departure point coincided with the coordinates of the destination. To avoid incorrect values, such as zero distance, it was decided to set the minimum distance value at 1 km. This logical assumption was based on the assumption that even in such cases, the minimum distance traveled still occurred. But, looking ahead, I will say that there were no such cases, and the minimum delivery distance was 27 km.

After the calculation, the new distance information was added to the dataset as a new column ***distance_km***. The updated dataset was then exported to a new Excel file to save the results and use them in subsequent analysis.

This step not only enriched the original dataset, but also opened up opportunities for further analysis, including studying the impact of distances on delivery costs, time metrics, and route efficiency.

2.2. Data Analysis

2.2.1 Visualization

After successfully collecting and pre-processing data containing detailed information on deliveries and transportation, the next stage of my work was devoted to data visualization. This stage plays a key role in the analysis, as it allows not only to see hidden patterns, but also to present the results in a clear and visual form. To implement this stage, I decided to use the BI tool Yandex DataLens, which provides powerful capabilities for creating interactive and informative visualizations.

Before starting to work with DataLens, I carefully prepared the data: checked it for gaps, eliminated outliers and made sure that all columns were in the correct format. This became an important basis for creating accurate and professional graphs.

Using Yandex DataLens, I developed various types of charts corresponding to the specifics of the data. For categorical variables, such as service type or cargo categories, bar and pie charts were created. Numerical data, such as weight, cost and distance, were visualized through histograms, scatter plots and line charts.

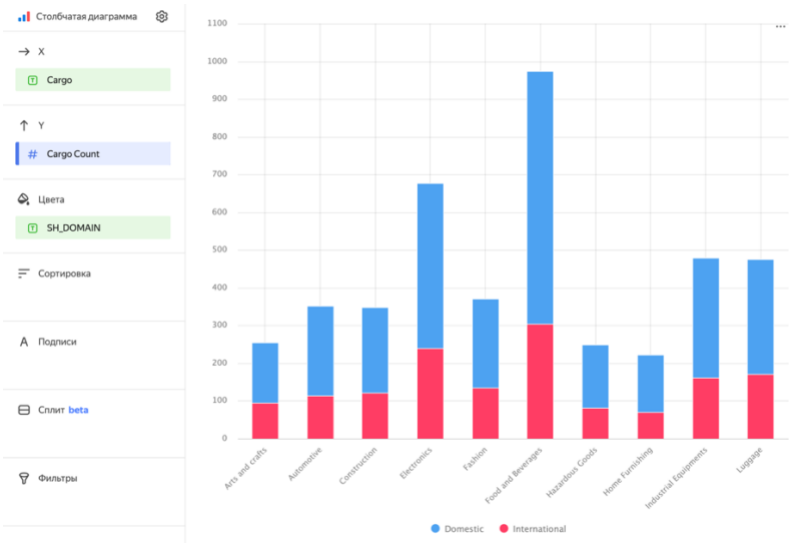
To connect Excel files to Yandex DataLens, I first exported the pre-processed data in Excel format and then loaded it into DataLens via the tool's interface. This process involved setting up the connection, uploading the file, selecting the desired tables and fields, and specifying the data types for each column. This step made the data available for further visualization.

The first visualization I created was an analysis of the distribution of cargo types. For this chart, I chose a bar graph since it is great for displaying categorical data. The creation process included the following steps:

Setting up the axes: For the X-axis parameter, I used the SH_CONTENT column, which I renamed "Cargo" for ease of perception. This allowed me to display the cargo types on the horizontal axis.

Setting the Y-axis parameter: For the Y-axis, I chose the same SH_CONTENT column, but in count format. This helped visualize the number of each cargo category.

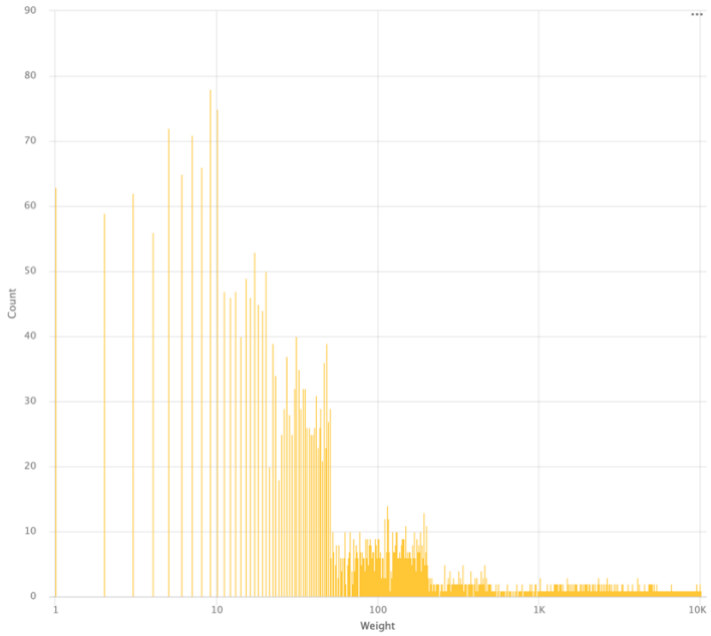
Color differentiation: For additional analytics, I added the SH_DOMAIN parameter to the color scale. This allowed me to separate the data into domestic and international shipments, highlighting the differences between them. The resulting chart clearly showed which cargo categories were most common, as well as their distribution between domestic and international shipments. This provides valuable insights for analyzing logistics trends, such as which cargo is most often delivered on international routes, and which - domestically. This approach allows us not only to estimate volumes, but also to identify the specifics of logistics processes for different types of cargo.



The results of the first visualization showed that the most frequently transported goods within the country are food products. This is understandable given the specifics of Kazakhstan: the country has a developed agriculture and a wide network of food producers and processors. Ensuring food security, maintaining supply chains and high domestic market needs create high demand for the transportation of such goods. This explains why food-related categories dominate in terms of transportation volumes. A wide range of customers, from large supermarkets and restaurants to small shops, actively use transportation services for food delivery.

At the other end of the spectrum are categories of goods such as hazardous goods and home furnishings, which are transported much less frequently. In the case of hazardous goods, this is due to their specificity: such goods require strict compliance with transportation rules, special equipment and certification, which increases the cost of logistics. In addition, the need for the transportation of hazardous goods is often limited to a narrow range of industries, such as the chemical industry or mining. This explains the low volume of shipments in this category.

As for home improvement products, their low share in the total volume of shipments also has its justification. The domestic market for furniture and interior items is relatively small, and most of these products are imported. In addition, due to their bulkiness and limited demand, shipments are often made on an individual basis rather than in large volumes, which also reduces the overall statistics for shipments in this category. These observations highlight the importance of visualization in analyzing and identifying the specifics of logistics processes for different types of cargo.

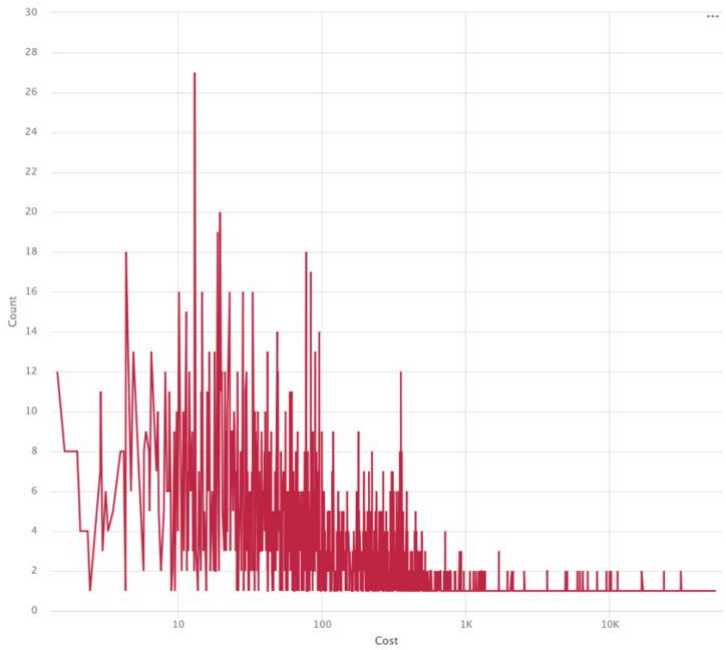


The following chart shows the weight distribution of the transported goods. To create this chart, I used the SH_WEIGHT parameter (renamed to "Weight" for convenience) as the X-axis value, and the Y-axis was the count of the number of shipments for each weight range.

The results showed that the largest volume of shipments falls on goods weighing up to 50 kg. This is not surprising, since small and light goods, such as documents, small consignments of goods or product samples, make up the bulk of daily logistics both domestically and in international transportation. Such goods are easy to pack, quickly deliver and are most often in demand among small and medium businesses.

In the weight range from 1 to 10 tons, there is a small number of shipments, which is associated with the delivery of large-sized goods, such as industrial equipment, construction materials.

Visualization of the distribution of cargo weights allowed us to identify key patterns and features of logistics processes depending on the mass of goods transported. This analysis helps to better understand the structure of demand for transportation and adjust the logistics strategy depending on the needs of customers.

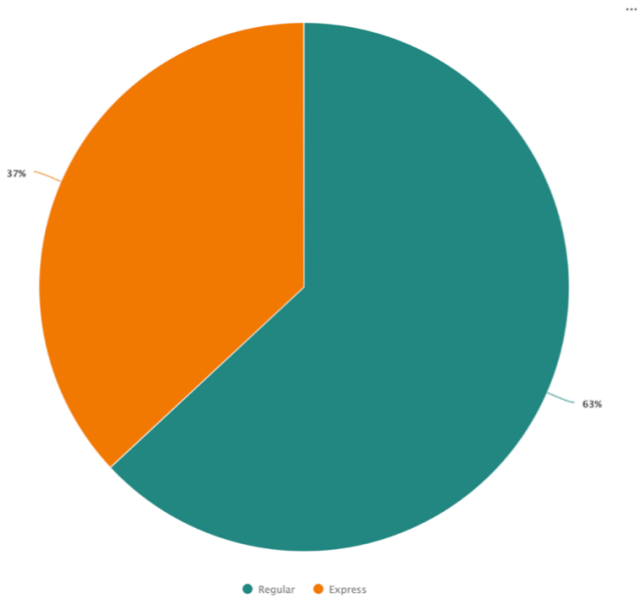


The next step in data visualization was to analyze the distribution of freight shipping costs. To plot this chart, I used the SH_CHARGES parameter (renamed to "Cost" for convenience) as the X-axis value, and the Y-axis value was the count of the number of deliveries for each cost range. Color parameters were not used in this case, as the main focus was on the overall distribution of shipping costs.

The results of this visualization showed that the bulk of deliveries fall in the segment with a cost of up to $100. This is explained by the fact that most of the shipments include small and lightweight goods, which are cheaper to ship. This segment also includes express domestic deliveries, which are popular among small and medium businesses for prompt receipt of goods.

The next step of data analysis was to visualize the distribution of the types of delivery services provided. To do this, I created a pie chart that shows the ratio of two types of services: Express and Regular. The data for the plot was taken from the SER_TYPE column. The percentage distribution of services was calculated, and the visualization immediately provided a clear picture of their popularity.

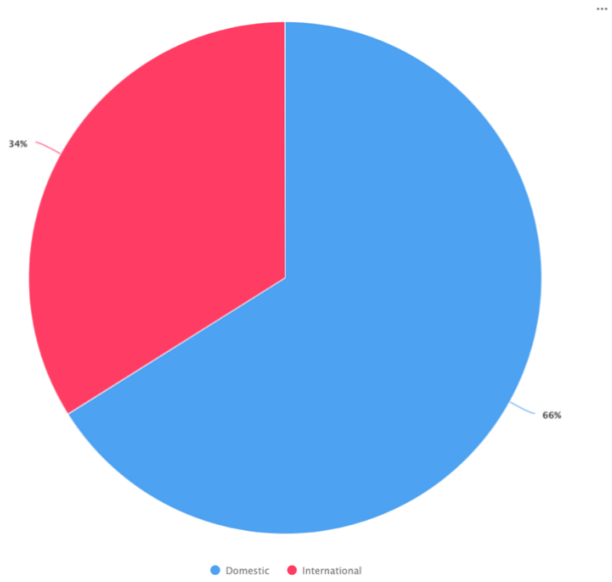
The results of the chart showed that 63% of all deliveries are Regular, and the remaining 37% are Express. This result allows us to draw several interesting conclusions.



A large share of Standard delivery is quite expected, since Regular services are usually used to transport less urgent goods. Such services are cheaper and popular among customers who do not have strict time frames for receiving their goods. For example, delivery of large or insignificant goods, such as building materials or artificial jewelry, is more often carried out in Regular mode to reduce costs.

On the other hand, Express services make up a significant share (37%), which indicates a high demand for fast and efficient delivery. This type of service is in demand among customers who value speed, such as perishable goods, medical supplies, or urgent commercial orders. Although express delivery is more expensive, customers are willing to pay for convenience and reliability, especially in cases where delays can lead to significant losses.

This visualization demonstrates the balance between cost-effectiveness and efficiency, which plays a key role in shaping demand for logistics services. Understanding this distribution helps make strategic decisions about the development of each type of service, from optimizing routes for regular deliveries to improving the speed and availability of express services.



The next visualization was dedicated to analyzing the distribution of deliveries by their geographic scope. For this, a pie chart was created showing the percentage of international (International) and domestic (Domestic) shipments. The data for the plot was taken from the SH_DOMAIN column, which was renamed to "Domain" for convenience. The results provided a clear understanding of the overall picture of logistics operations.

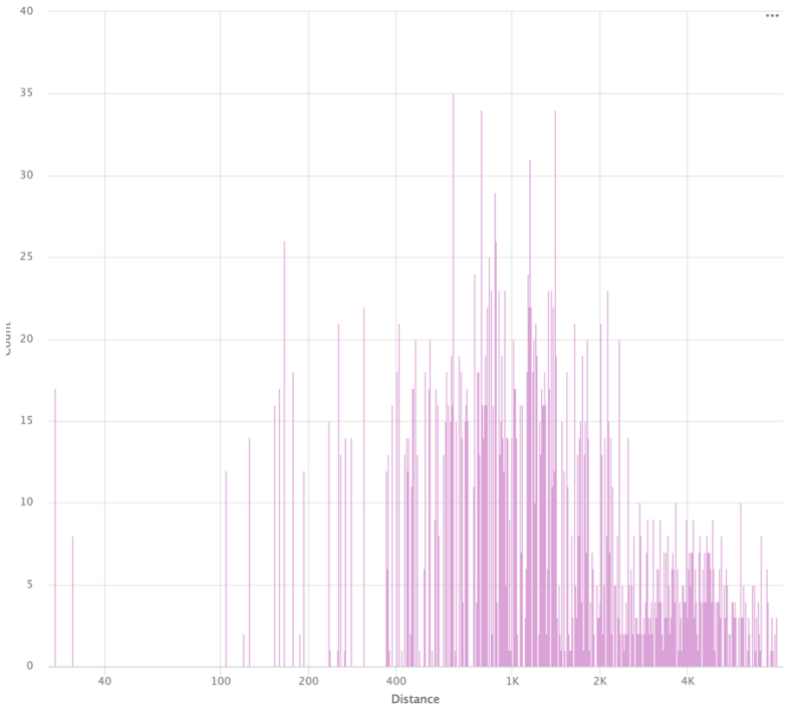
According to the visualization, 66% of all deliveries are domestic, and the remaining 34% are international. This data allows us to draw several interesting conclusions about the structure of transport operations and customer preferences.

The predominance of domestic shipments (66%) is not surprising, especially in a country like Kazakhstan, with its vast territory and an economic structure oriented towards the domestic market. This is explained by the high need for the delivery of goods between different regions of the country. For example, essential goods, construction materials and industrial equipment are often transported between large cities and remote regions. Such logistics support the functioning of the national economy and contribute to the even distribution of goods throughout the country.

International shipments account for 34%, which is also a significant figure. This indicates that the logistics company has a solid share of customers focused on export and import. International shipments may involve high-tech equipment, high-value-added goods, or rare materials that are not produced in the country. For example, exporting food products or importing specialized equipment for industry requires precise coordination and trust in the transport company.

This visualization demonstrates that the company successfully serves both segments: domestic and international. However, a higher percentage of domestic shipments indicates that the main focus is on national logistics. This knowledge can be useful for making decisions on developing international routes or introducing specialized services to meet the needs of import-export operations.

The following graph displays the frequency of deliveries in different distance ranges. The basis for the visualization was the distance_km column, which contains information about the distances calculated earlier. This analysis allowed us to identify which distance categories are most common and draw conclusions about the company's logistics priorities.



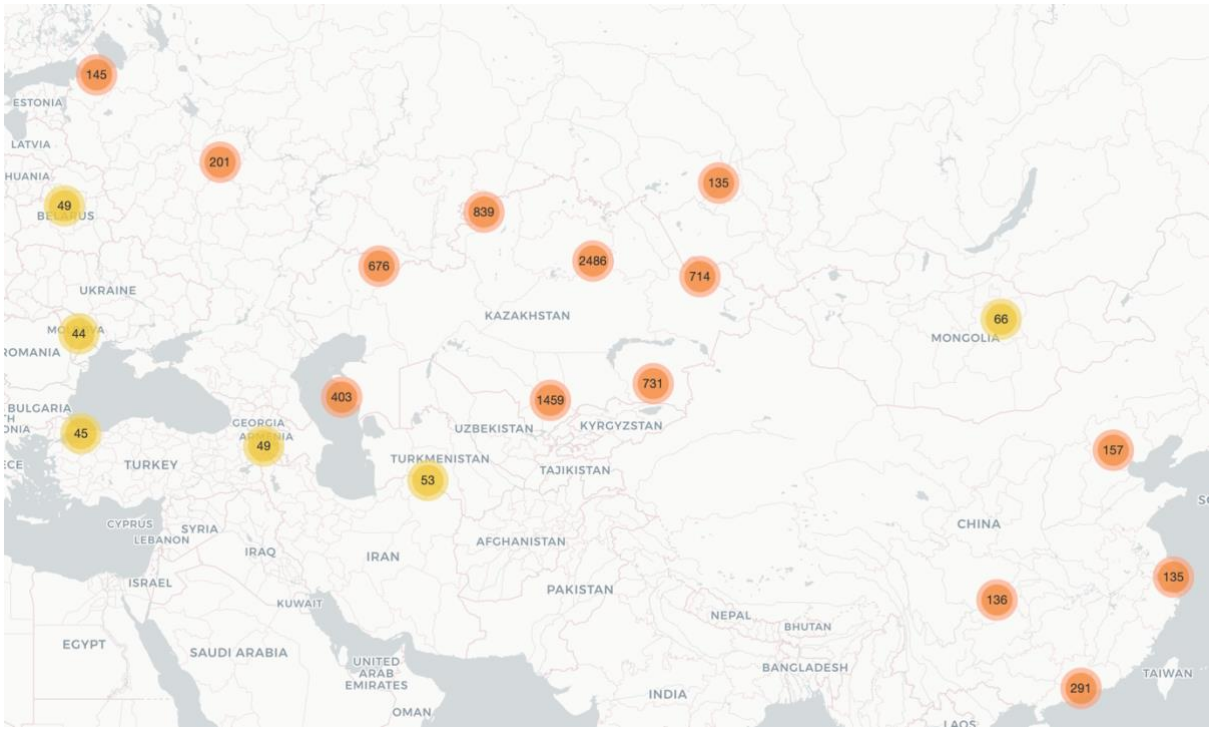
The results of the visualization showed that the largest number of deliveries falls on medium distances (from 400 to 2000 km). This is not surprising, given the geographical features of Kazakhstan. Medium distances are most often found within one region or between neighboring regions.

Medium distances (for example, from 500 to 1000 km) reflect logistics operations between the main economic centers of the country, such as Almaty, Astana, Shymkent and other cities. These routes are in demand due to the high concentration of industry, trade and population in these regions. Medium distances are also convenient for express deliveries that require efficiency, but are not associated with international routes.

Deliveries over long distances are less common, but still important. These routes may involve international shipments or complex logistics operations involving hard-to-reach regions of the country. For example, the delivery of industrial equipment or rare materials to remote locations requires careful planning and precise coordination. Visualizing the distribution of distances provided an opportunity to better understand the operational specifics of the company.

Lastly, a geospatial visualization was created using Folium with the "CartoDB Positron" map style to complement the dataset analysis. This

interactive map displays both the origin and destination points of each shipment, providing a clear overview of the company's logistics operations.



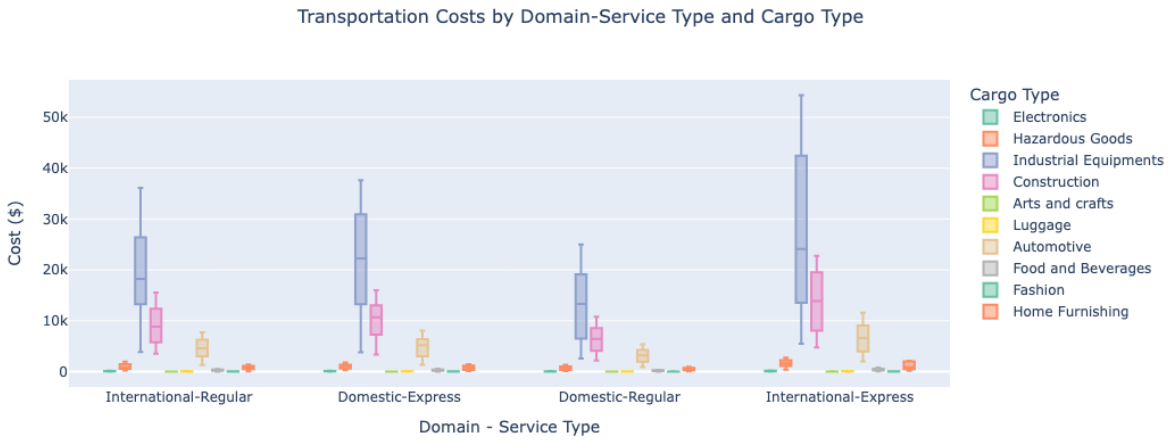
On the map, origins (the starting points of shipments) are marked with blue icons, while destinations (the delivery points) are marked with green icons. Most of these points are concentrated within Kazakhstan and its neighboring countries, with notable activity extending across Central Asia, Eastern Europe, China, and Mongolia. This visualization highlights the company's broad geographic reach and provides valuable insights into the spatial distribution of delivery routes. It helps to identify regions with high shipment density and frequently used corridors. Such insights are critical for optimizing delivery paths, reducing transportation costs, and improving overall operational efficiency. Integrating this geospatial data into the broader dataset allows for a multi-dimensional analysis and supports strategic decision-making and advanced logistics modeling.

2.2.2 Correlation Analysis

Correlation analysis is an important step in working with data that allows you to identify relationships between different variables in a data set. This method of analysis is especially useful for understanding which factors may influence each other and identifying key dependencies that can be used to optimize processes or make strategic decisions.

The goal of correlation analysis is to find pairs of variables that show a strong positive or negative relationship. For example, in logistics, this could be a correlation between cargo weight and shipping cost, distance and delivery time, or other parameters. Such dependencies help to gain a deeper understanding of operational processes and find hidden patterns.

As part of this analysis step, I created an interactive Box Plot to evaluate the distribution of shipping costs by domain (domestic or international) and service type (regular or express). I also added color coding by cargo type to visualize the differences between cargo categories.



To simplify the analysis and make it easier to visualize, I created a new category SH_CATEGORY that combines domain and service type information. For example, domestic shipping with regular service is displayed as "Domestic-Regular", while international shipping with express service is displayed as "International-Express".

To ensure the order of the categories, I set the display sequence to domestic shipments first. The chart displays:

X-axis: SH_CATEGORY categories (domain and service type).

Y-axis: SH_CHARGES shipping costs.

Color coding: SH_CONTENT cargo type.

To highlight outliers, the "outliers" display option was enabled. The chart title was centered and enlarged for ease of reading. The Set2 color palette from the ***Plotly*** library was used to clearly differentiate between cargo categories. Axis labels and legend were customized to improve readability.

The chart allows you to see how transportation costs vary depending on the combination of domain and service type. Evaluate the cost spread and identify outliers that may be abnormal values or reflect rare cases. Compare costs between different cargo types in the same SH_CATEGORY.

International shipments (especially express services) have a wider cost spread, which is explained by the complexity of logistics and high requirements

for delivery speed. Domestic shipments have more stable costs, especially for regular services.

The type of cargo also affects the cost: heavier or more specialized categories such as Industrial Equipments or Hazardous Goods tend to have higher transportation costs.

This chart highlights the importance of optimizing transportation costs based on cargo type, domain, and service. It also helps identify outliers that may indicate non-standard operations or calculation errors. This approach helps in making informed decisions to reduce costs.

The following analysis is an interactive scatter plot showing the dependence of shipping costs on the weight of the shipment, taking into account the service category and domain.

The graph showed the weight of the shipment (SH_WEIGHT) on the X axis, and the shipping costs (SH_CHARGES) on the Y axis. The points on the graph are colored depending on the service category, which allowed us to distinguish between domestic regular, domestic express, international regular and international express deliveries. The size of the points is proportional to the weight of the shipment, which added another dimension to the analysis. The color palette used ensured a clear separation of categories, and tooltips displayed additional information about the domain when hovering over the point.

The graph reveals a number of interesting patterns. First of all, there is a clear dependence of the cost of delivery on the weight of the shipment: the heavier the cargo, the higher its cost. At the same time, the influence of the service category remains significant. For example, international express deliveries consistently demonstrate the highest cost, which is due to logistical and administrative costs. On the other hand, domestic regular transportation remains the most economical option, especially for heavy cargo, where speed is not a priority.

This graph also helps to identify anomalies: points that deviate significantly from the main group may indicate rare cases or errors in the data. For example, shipments with a disproportionately high cost for their weight may require additional analysis.

Thus, this visualization illustrates the complex interaction between weight, service type and domain. It allows not only to better understand the existing rates and their logic, but also to identify possible areas for cost optimization, such as choosing more economical delivery options for certain types of cargo. The interactive format of the graph greatly facilitates studying the data, allowing you to quickly and conveniently analyze it from different angles.

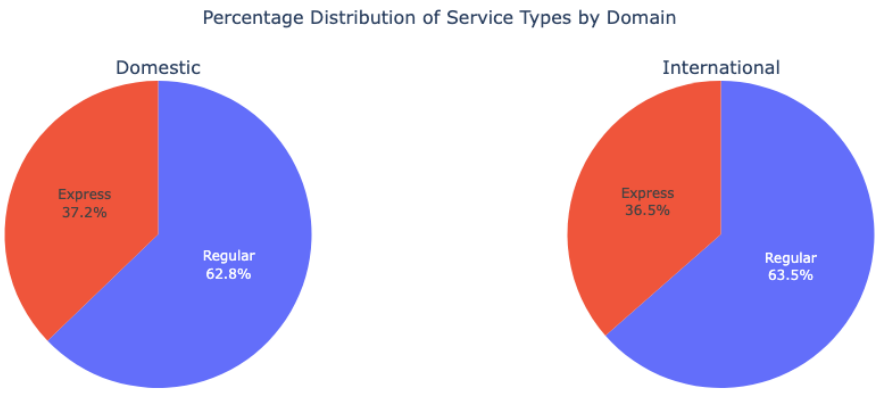
Next, I created a scatter plot showing the correlation between shipping distance and shipping costs by service category and domain.

The graph shows shipping distance in kilometers (distance_km) on the x-axis and shipping cost (SH_CHARGES) on the y-axis. Each dot is colored based on the service category and its size is proportional to the weight of the package (SH_WEIGHT). Tooltips when hovering over the dots provide additional information, including the shipping domain, service type, and type of cargo (SH_CONTENT).

The graph shows that there is a small increase in shipping costs with increasing shipping distance. However, the increase in cost is not uniform.

The visualization also highlights the impact of cargo weight on cost. Larger dots, which represent heavier packages, tend to be higher on the graph, confirming that costs increase with increasing weight. However, there is significant variation in cost among these dots, which may indicate the influence of additional factors, such as the specifics of the cargo or the chosen route. Another interesting feature of the graph is the identification of points that deviate significantly from the main dependence. These anomalies can be caused by atypical cases, for example, the delivery of expensive goods over short distances or, conversely, inexpensive transportation over long distances.

To analyze the percentage distribution of service types for domestic and international shipments, I created a double pie chart. First, the data was aggregated by delivery domain (domestic or international) and service type (regular or express). Then, a column with the percentage distribution calculated within each domain was added to visualize the relative shares of each service type.



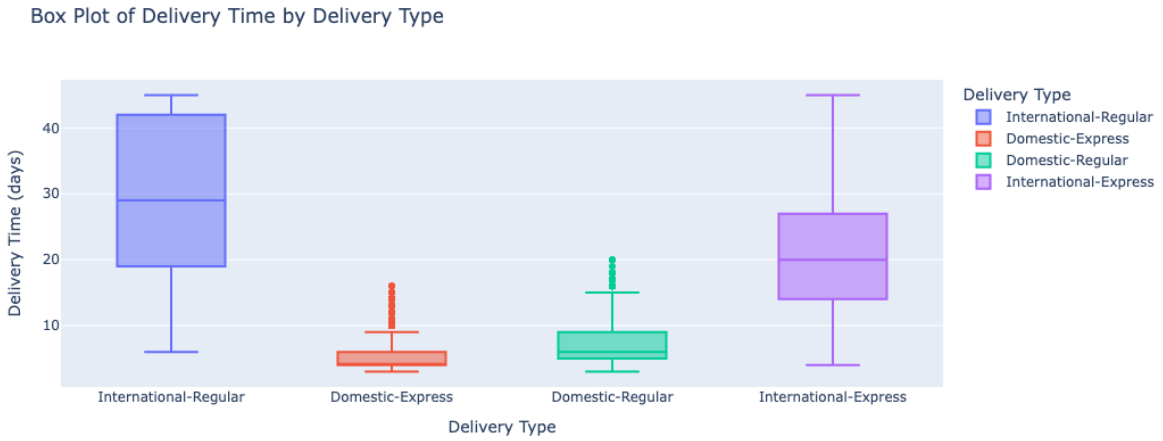
The charts feature two circles: the first represents domestic shipments, and the second represents international shipments. Each chart includes percentage values and labels for each sector, making the visualization as clear as possible.

The results of the chart showed that in both the domestic and international domains, regular shipments account for a larger share. This is because regular shipments are more cost-effective and often the preferred choice for most customers, especially for shipments that do not require urgent transportation. For example, in domestic shipments, these could be everyday goods or large consignments of industrial materials. In international shipping, regular deliveries also dominate, as they are suitable for less urgent shipments, such as large equipment or product shipments.

Express deliveries, in turn, occupy a smaller share in both domains. Their popularity increases only in situations requiring urgent delivery, such as documents, important product samples or perishable goods.

This visualization provides valuable information on the distribution of customer preferences depending on the type of delivery, which helps to better understand their needs and adapt the company's logistics strategy.

To analyze the delivery time by delivery type, I used a Box Plot chart, which allows you to visualize the distribution of delivery times for different categories. The X-axis contains delivery categories (e.g. domestic regular, domestic express, etc.), and the Y-axis contains the delivery time in days. The delivery categories were colored differently for easier reading.



The chart displays the minimum, maximum, and median delivery times, as well as the spread and outliers, if any. The results showed a clear difference between the delivery types. For example, express deliveries, as expected, show significantly lower delivery times, reflecting their priority and promptness. Domestic express deliveries turned out to be the fastest, as they usually occur over short distances.

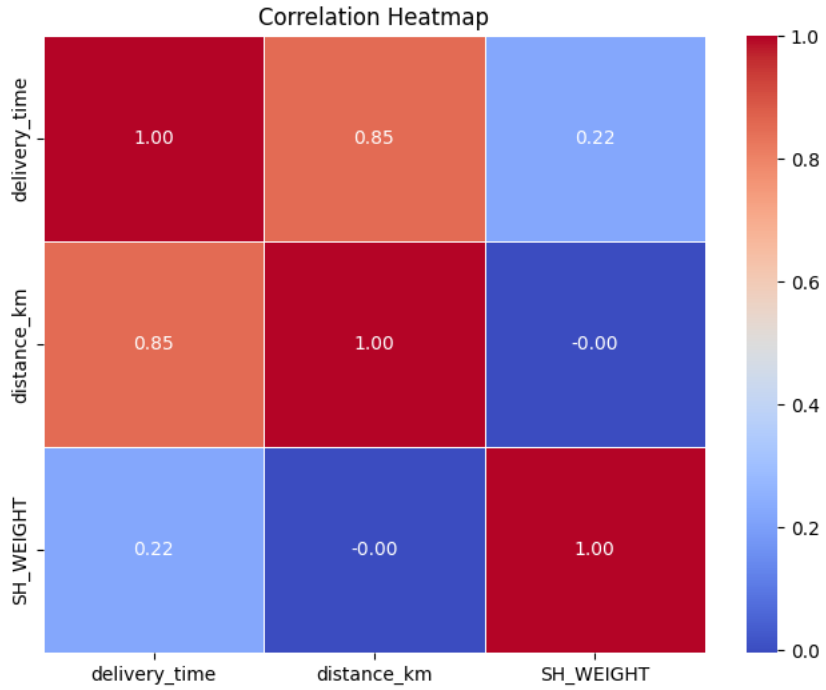
On the contrary, regular deliveries showed a wider spread of times, especially in the international segment, which is explained by the complexities associated with border crossings, customs procedures, and possible delays in transportation.

This visualization provides useful insights for improving the performance of the supply chain. For example, for categories with a large spread of delivery times, you can analyze the causes of delays and optimize processes. This also helps to more accurately predict delivery times for customers and adjust rates depending on urgency.

To analyze the relationship between the numeric variables in the data set, I created a correlation heat map. This tool helps determine whether there are strong linear relationships between variables that can be useful for further analysis or forecasting.

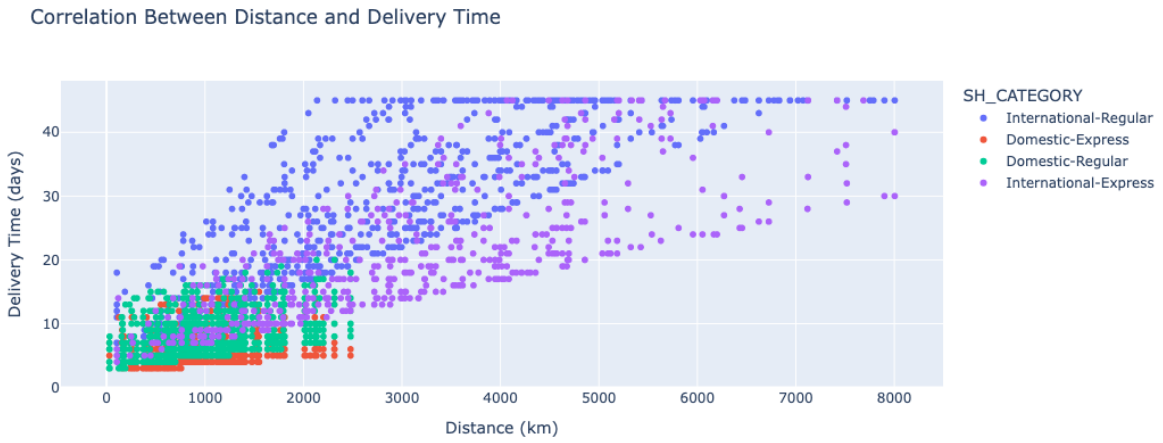
I chose three key numeric variables: delivery time, distance (in kilometers), and cargo weight. First, a correlation matrix was calculated showing the correlation coefficients between these variables. To visualize the matrix, I used the Seaborn library, which allows you to create a convenient and visual heat map.

On the heat map, the values of the correlation coefficients are represented by a color scale: warmer colors (red tones) indicate a positive correlation, and cold colors (blue tones) indicate a negative correlation. The map also indicates the exact numerical values of the correlation coefficients to make the analysis more accurate.



The results showed that there is a strong positive correlation between distance and delivery time, which is expected, since longer routes require more time for delivery. There was a weak correlation between cargo weight and delivery time, indicating that weight does not significantly affect transportation speed. Similarly, the correlation between cargo weight and distance was also weak.

To explore the relationship between distance and delivery time, I created a scatter plot. This type of visualization allows to visualize how two key parameters - distance in kilometers and delivery time in days - relate to each other depending on the delivery category.

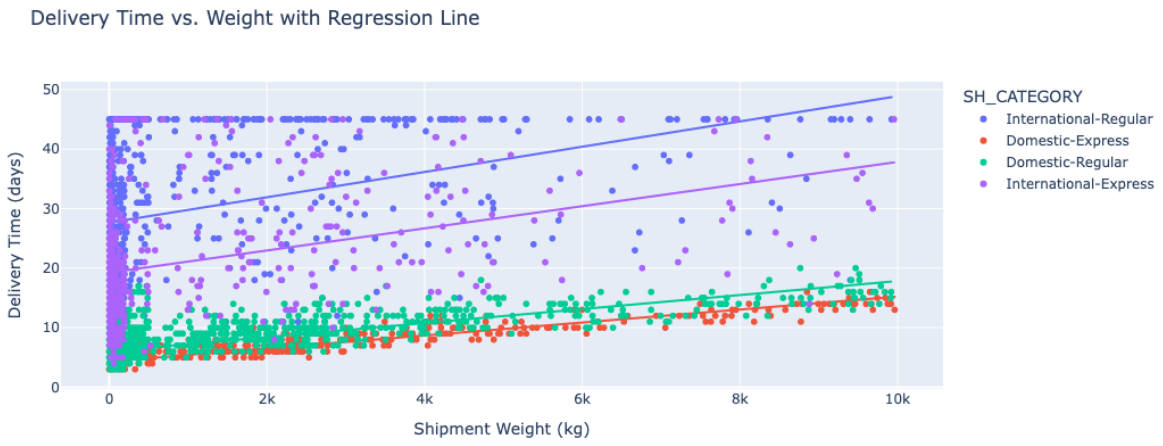


In the graph, the dots represent individual deliveries, where their position on the X-axis shows the distance traveled, and on the Y-axis - the time it took to

deliver. Color differentiation by delivery category (combining "Domain" and "Service Type") helps to distinguish between delivery types, such as domestic or international shipments, and express or regular. Additionally, when hovering over the dots, the "Domain" and "Service Type" parameters are displayed, which makes the graph more informative.

The results of the visualization confirmed the obvious relationship: the greater the distance, the longer the delivery time. However, there are differences in the slopes of the trends for different categories. For example, international deliveries on average take longer for the same distance compared to domestic ones. Express deliveries for both types usually have a smaller spread in time, indicating clear delivery times.

Next, I looked at the relationship between cargo weight and delivery time. I created a scatter plot with a linear regression line added. This approach allows not only to visualize the data, but also to identify possible trends and patterns between the parameters under consideration.



In the graph, each point represents a separate delivery, where the X-axis shows the weight of the shipment in kilograms, and the Y-axis shows the delivery time in days. A color palette is used to indicate the delivery category (a combination of "Domain" and "Service Type"), which makes it easy to distinguish between domestic and international shipments, as well as express and regular services. The added regression line helps to assess the overall trend, demonstrating the average effect of weight on delivery time.

The results of the analysis show that an increase in cargo weight can have an insignificant effect on delivery time, especially in the regular services category. At the same time, express deliveries, regardless of weight, have a relatively narrow spread in time, which confirms their priority status.

I created a scatter plot to look at delivery costs and delivery times. This visualization helps us understand how delivery times vary depending on shipping costs, and explore possible differences between different service categories.

In the graph, the X-axis represents shipping costs in dollars, and the Y-axis represents delivery times in days. Each dot represents a single shipment, and the color differentiation indicates the delivery category, which is determined by a combination of the service type (e.g. Regular or Express) and the region (Domestic or International). For additional analysis, tooltips display information about the distance in kilometers and the weight of the shipment.

Analysis of the graph shows that, in general, express deliveries have higher costs and take less time compared to regular services. However, for international shipments, even with a high cost, delivery times can be significantly longer due to factors such as customs checks or logistics complexity.

2.3 Delivery modeling: time and cost forecast

In the final part of the thesis, I delve into a complex and critical forecasting problem: building models to predict delivery times and costs using historical data from a logistics company. Using the knowledge gained from previous correlation studies, I identified key parameters such as cargo type, weight, distance, service type, and delivery category that significantly affect delivery times and costs. These parameters became the basis for developing predictive models.

To ensure a robust and comprehensive analysis, I chose three different machine learning algorithms: linear regression, random forest, and neural networks. These models were chosen for their complementary strengths. Linear regression provides simplicity and interpretability, random forest is excellent at capturing nonlinear relationships and interactions, and neural networks provide the flexibility to model complex patterns in the data. Together, they provide a range of approaches to the forecasting problem, allowing for a thorough comparison of their performance across different scenarios and data configurations.

Throughout the process, I evaluated the models using key metrics such as mean absolute error (MAE), root mean square error (RMSE), and R-squared. These metrics provided a comprehensive view of each model's accuracy, precision, and ability to generalize to unseen data. Additionally, I analyzed the feature importance in the random forest model and examined the weight distribution in the neural network to gain deeper insights into the factors that influence shipping costs and times.

The ultimate goal was to not only create models that could make accurate predictions, but also provide actionable insights that could improve logistics operations. For example, understanding how distance and weight of a shipment impact costs can help optimize pricing strategies, while accurate shipping time predictions can improve customer satisfaction and operational planning.

In the following sections, I'll detail each stage of this work, from data preprocessing to training, evaluation, and interpretation of results, ultimately highlighting how these models can make measurable improvements to logistics decision making.

2.3.1 Cost

Before building the model, the categorical data was encoded, which is an important step in data pre-processing. Such cargo types as SH_DOMAIN, SER_TYPE, and SH_CONTENT were presented in a categorical format, which needed to be converted to a numeric format for the machine learning algorithms to work correctly. For this task, I used the ***OneHotEncoding*** technique, which is a standard method for converting categorical data to numeric.

```
from sklearn.preprocessing import OneHotEncoder

# List of categorical columns to encode
categorical_columns = ['SH_DOMAIN', 'SER_TYPE', 'SH_CONTENT']

# Create OneHotEncoder
encoder = OneHotEncoder(drop='first')

# Transform the categorical columns using OneHotEncoder
encoded_data = pd.DataFrame(encoder.fit_transform(data[categorical_columns]).toarray())

# Get the names of the new columns
encoded_columns = encoder.get_feature_names_out(categorical_columns)

# Add the encoded columns to the original DataFrame
encoded_data.columns = encoded_columns

# Combine the original DataFrame with the encoded columns
data = data.drop(columns=categorical_columns)  # Drop the original categorical columns
data = pd.concat([data, encoded_data], axis=1)

# Check the result
data.head()
```

The essence of the method is that for each unique value in the categorical column, a separate binary column is created. The value “1” in such a column indicates that the given observation belongs to the corresponding category, and “0” indicates that it does not belong. For example, if the SH_CONTENT column has three unique values: Automotive, Construction, and Electronics, then after applying OneHotEncoding, they will be converted into three new columns: SH_CONTENT_Automotive, SH_CONTENT_Construction, and SH_CONTENT_Electronics. If a particular row in the original data belongs to the Automotive category, the new column SH_CONTENT_Automotive will contain "1" and the rest will contain "0".

SH_CONTENT_Automotive	SH_CONTENT_Construction	SH_CONTENT_Electronics
0.0	0.0	1.0
0.0	0.0	0.0
0.0	0.0	0.0
0.0	1.0	0.0
0.0	0.0	0.0

To improve the efficiency of the models and prevent data redundancy, the drop='first' option was used. This means that for each categorical variable, one column is skipped, since its information is already contained in other columns. For example, if the variable SH_DOMAIN has the categories Domestic and International, then after encoding, only one column SH_DOMAIN_International will be created. If the value is "0", this means that the row belongs to the Domestic category.

After encoding was complete, the new binary columns were added to the dataset and the original categorical columns were removed. This approach allows the model to correctly analyze categorical information without the risk of distortion or interpretation of order that may have been erroneously introduced using other encoding methods such as Label Encoding.

To build a model that predicts shipping costs, I started by splitting the data into features and a target variable. The features were the parameters that previous studies have shown to have the greatest impact on shipping costs: cargo weight (SH_WEIGHT), distance (distance_km), and coded categorical variables such as shipping type (SH_DOMAIN_International), service type (SER_TYPE_Regular), and cargo content (e.g. SH_CONTENT_Automotive, SH_CONTENT_Food and Beverages, etc.). The target variable in this problem is shipping cost (SH_CHARGES).

```
# Split the data into features (X) and target variable (y)
X = data[['SH_WEIGHT', 'distance_km', 'SH_DOMAIN_International',
          'SER_TYPE_Regular',
          'SH_CONTENT_Automotive', 'SH_CONTENT_Construction',
          'SH_CONTENT_Electronics', 'SH_CONTENT_Fashion', 'SH_CONTENT_Food and Beverages',
          'SH_CONTENT_Hazardous Goods', 'SH_CONTENT_Home Furnishing', 'SH_CONTENT_Industrial Equipments',
          'SH_CONTENT_Luggage']]

y = data['SH_CHARGES']  # Delivery cost (target variable)

# Split the data into training and testing sets (80% for training, 20% for testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Next, I split the data into training and test sets in an 80:20 ratio to ensure fairness in evaluating the model. This step is standard practice in machine learning and is necessary to ensure that the model learns on one part of the data and tests its predictions on another, previously "unseen" part. This split helps to avoid a situation where the model simply memorizes the training data instead of identifying general patterns that apply to new input data.

For the split, I used the train_test_split function from the sklearn library, setting the test_size parameter to 0.2, which means that 20% of the data is allocated for the test set. The remaining 80% is left for training the model. In addition, I fixed the random_state parameter, setting its value to 42. This made the data splitting process reproducible: each time the code is run, the data will be split in the same way, which is especially important for analyzing and comparing results at different stages of the work.

This approach achieves two important goals. First, the training set is large enough for the model to learn patterns in the data and tune its parameters. Second, the test set remains representative enough to objectively test the quality of the model, since it contains data that it has not seen before. This allows us to assess how well the model will cope with real-world problems, where new data may differ significantly from the data it was trained on.

As the first model for forecasting, I chose **linear regression**. This algorithm is simple, interpretable, and is one of the most popular methods for problems related to forecasting numerical data. Linear regression allows you to determine how a change in each input feature (for example, cargo weight or delivery distance) affects the target variable - in this case, delivery cost. In addition, it provides the ability to identify and analyze the relationships between parameters, which makes it especially useful for logistics analytics.

The main task of the linear regression algorithm is to find the optimal weights (coefficients) for each of the features. These weights determine the degree of influence of each parameter on the final cost of delivery. To find these weights, the model uses the least squares method, which minimizes the sum of the squares of the errors between the actual values of the cost of delivery and the values predicted by the model.

```python
# Build a linear regression model
model = LinearRegression()

# Train the model on the training data
model.fit(X_train, y_train)

# Make predictions on the test data
y_pred = model.predict(X_test)

# Evaluate the model using metrics
mse = mean_squared_error(y_test, y_pred)  # Mean Squared Error
r2 = r2_score(y_test, y_pred)  # R-squared coefficient
rmse = np.sqrt(mse)  # Root Mean Squared Error
mae = mean_absolute_error(y_test, y_pred)  # Mean Absolute Error
evs = explained_variance_score(y_test, y_pred)  # Explained Variance Score

n = len(y_test)  # Number of examples
p = X_test.shape[1]  # Number of features

print(f"Mean Squared Error: {mse}")
print(f"Root Mean Squared Error (RMSE): {rmse}")
print(f"Mean Absolute Error (MAE): {mae}")
print(f"R-squared: {r2}")
print(f"Explained Variance Score: {evs}")
```

The learning process can be described as an iterative search for the best line or hyperplane in a multidimensional space that most accurately describes the dependence of the target variable on the features. Each feature, be it weight, distance, or cargo category, receives its own coefficient, which reflects its contribution to the formation of the final cost.

After training the model, I tested it on the test set to assess its ability to predict shipping costs on new, previously unseen data. I used several key metrics to test the quality of the model, each providing a unique perspective on the accuracy and performance of the model:

1. ***Mean Squared Error*** (MSE) measures the average of the squared differences between the actual and predicted values. The lower the MSE, the more accurate the model. In this case, the MSE was 3668497, which indicates some error in the predictions, but the value of this metric alone is difficult to interpret due to its dependence on the scale of the target variable.

2. ***Root Mean Squared Error*** (RMSE) is the square root of the MSE, interpreted in the same units as the target variable. In the context of shipping costs, RMSE is a measure of the average prediction error in absolute terms. In this case, the RMSE is 1915.33, indicating that the model is off by that dollar amount on average when predicting shipping costs.

3. Mean Absolute Error (MAE) calculates the average of the absolute differences between the actual and predicted values. MAE shows how much the model deviates from the actual values on average, regardless of the direction of

deviation. In this case, the MAE was 1061, indicating that the model is off by about $1,061 on average when predicting shipping costs.

4. R-squared (R^2): shows how much of the variance in the target variable is explained by the model. An R^2 value close to 1 indicates good explanatory power for the model. For this model, the R^2 was 0.93, indicating that 93% of the variation in shipping costs can be explained by the model. This is a high score, indicating that the model does a good job.

5. Explained Variance Score (EVS) evaluates the degree to which the model explains the variability in the data. An EVS value close to 1 indicates that the model effectively interprets the data. In this case, the EVS was 0.93, confirming the high accuracy of the model.

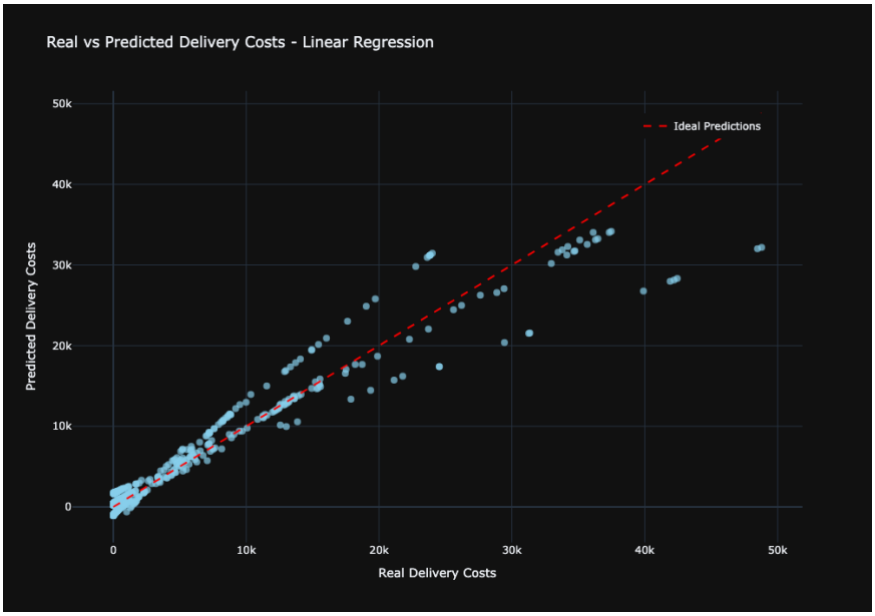
Taken together, these results show that linear regression was able to effectively use the input data to predict shipping costs. The high R^2 and EVS values demonstrate that the model explains most of the variability in the data, and the error metrics (MSE, RMSE, and MAE) provide a quantitative representation of the average deviation of the predictions from the actual values. However, errors such as the RMSE of $1.915 indicate that there may be significant deviations for some cases.

To further analyze the model's performance and assess its quality, I created an interactive scatter plot visualization using the Plotly library. The purpose of this visualization is to visualize how the model's predicted shipping costs compare to the actual values from the test dataset, as well as to identify potential outliers and areas for improvement.

In the plot, each dot represents one record from the test dataset. The x-axis represents the actual shipping costs, and the y-axis represents the model's predicted values. This way, the position of each dot demonstrates the accuracy of the prediction for a particular example. To make the visualization more visually clear, the dots were stylized in skyblue, with a size of 8 and a transparency of 0.7 to avoid clutter in dense areas of the graph.

To assess the perfect match of the predictions to the actual values, a red dotted line was added to the plot, which shows the perfect predictions (the $y = x$ line). This line is drawn between the minimum and maximum values of the actual data, setting a benchmark for the dots to gravitate towards. If the dots are close to this line, this indicates high accuracy of predictions. Significant deviation of the dots from the line indicates cases where the model was wrong.

The graph has been adapted for ease of analysis: its dimensions have been increased (width 1000 pixels, height 700 pixels), and the legend with the description of the graph elements is placed in the upper right corner to minimize overlap with the data.



The main purpose of this visualization is to help interpret the model's performance. It allows you to see how well the model's predictions match the actual values in different ranges of shipping costs. Ideally, the dots should form a cluster along the red line. If there is a strong deviation, this may indicate the need for additional model tuning or revision of the features used. This approach provides a deeper understanding of where and why the model makes mistakes, which is important for its further improvement.

To improve the accuracy of predictions and better account for the complex relationships between features, I used the **Random Forest** algorithm. This method is an ensemble approach based on the use of many decision trees. The essence of its work is that each tree is trained on a random subsample of the data, and then the predictions of all the trees are combined to obtain the final result. In regression problems, as in this case, the final prediction is formed by averaging the values obtained from all the trees.

Random Forest uses two key mechanisms that make it effective: first, the bagging method, which consists of creating random subsamples of data from the training set to train each tree. This reduces the likelihood of overfitting, since each tree only sees a part of the data. Second, when constructing each tree, Random Forest selects a random subset of features for splits at each node. This approach reduces the correlation between trees, which makes the ensemble more robust and accurate.

To implement the model, I used the RandomForestRegressor class from the sklearn library. I set the n_estimators parameter to 100, which means that 100 trees will be created in the forest. This value is standard and provides a balance between the quality of the model and the computation time. I also fixed the random_state parameter to 42 so that the results of the experiment could be reproduced. After initializing the model, I trained it on the training dataset using

the fit method, which allows the algorithm to determine the optimal parameters for constructing each tree.

```
from sklearn.ensemble import RandomForestRegressor

# Initialize and train the Random Forest model
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# Make predictions on the test set
y_pred_rf = rf_model.predict(X_test)
```

A feature of the random forest is its ability to take into account nonlinear dependencies between features without the user explicitly specifying such dependencies. For example, the model can automatically detect that the effect of distance on shipping costs depends on the type of cargo contents or other factors. In addition, the algorithm is robust to outliers, since the errors of individual trees are compensated by the overall ensemble.

After training, I used the predict method to perform predictions on the test dataset and saved the results in the y_pred_rf variable. Comparing these predictions to actual values will help assess the quality of the model and its ability to generalize patterns in the data. It is important to note that random forest, despite its flexibility, can be more computationally intensive than simpler models such as linear regression. However, its ability to efficiently handle large amounts of data and account for complex dependencies makes it a suitable choice for this task.

After training the random forest model and making predictions on the test dataset, I analyzed its performance using key regression metrics. The results were impressive, significantly outperforming linear regression. The ***Root Mean Squared Error*** (RMSE) was **222.09**, indicating that the model's predictions on average deviate from the actual values by approximately 222 arbitrary units. This metric, expressed in the same units as the target variable, is an intuitive indicator of the model's accuracy. The ***Mean Absolute Error*** (MAE) was **63.83**, demonstrating the model's average absolute error and highlighting its high accuracy in predicting shipping costs.

The most telling result is the ***R-squared*** (R^2), which was **0.9991**. This means that the model explains 99.91% of the variance in the data, which is almost perfect and demonstrates the ability of random forest to effectively capture complex relationships. The ***Explained Variance Score*** (EVS) metric was also calculated, the value of which is **0.9991**. This indicator confirms the high explanatory power of the model and emphasizes its reliability.

Such results are explained by the specifics of the random forest. The model builds many decision trees, each of which is trained on a random data sample and a random subset of features. The final forecast is formed by averaging the

predictions of all trees, which reduces the probability of errors and minimizes the risk of overfitting. Due to its ensemble nature, the random forest effectively copes with problems containing nonlinear dependencies and provides high prediction accuracy for both low and high values of the target variable.

To evaluate the performance of the random forest model, I again built a scatter plot visualization that compares the actual shipping cost values with those predicted by the model. In the plot, each point represents an individual example from the test dataset, where the position of the point relative to the diagonal line of ideal predictions (y = x) clearly demonstrates the accuracy of the model.

A distinctive feature of this plot for the random forest model is that most of the points are located almost close to the line of ideal predictions, which emphasizes the high accuracy of the model. It is also visually noticeable that, unlike linear regression, the random forest model copes with predictions more accurately in the range of high shipping cost values, where the deviations are minimal.

This visualization makes it easy to identify individual examples with the largest prediction errors. For example, several points located far from the line may indicate rare cases where the model failed due to outliers or missing important features. These examples can be used for further analysis and data improvement. The graph also confirmed that random forest provides stable prediction quality across the entire range of shipping cost values, making it a more robust tool than linear regression, especially for complex data dependencies.

The TensorFlow and Keras libraries were then used to build and train a neural network. Neural networks are a useful tool for task prediction because they can identify intricate non-linear dependencies in data. Standardization, which is essential to the neural network's correct operation, was done during the data

preparation phase using StandardScaler. This makes it possible to reduce all functions to a single scale, which enhances model convergence and expedites learning.

```
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout
from tensorflow.keras.optimizers import Adam
from sklearn.preprocessing import StandardScaler
```

```
# Step 1: Scaling the data
scaler = StandardScaler()

X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Step 2: Creating the neural network model
model = Sequential([
    Dense(64, input_dim=X_train_scaled.shape[1], activation='relu'),  # Input layer + 1st hidden layer
    Dropout(0.2),  # Dropout to prevent overfitting
    Dense(32, activation='relu'),  # 2nd hidden layer
    Dropout(0.2),
    Dense(1, activation='linear')  # Output layer (delivery cost prediction)
])

# Step 3: Compiling the model
model.compile(optimizer=Adam(learning_rate=0.001), loss='mse', metrics=['mae'])

# Step 4: Training the model
history = model.fit(
    X_train_scaled, y_train,
    validation_data=(X_test_scaled, y_test),
    epochs=50,  # Number of epochs
    batch_size=32,  # Batch size
    verbose=1  # Display the training process
)
```

The architecture of the model is three-level. 64 neurons in the first hidden layer have the ReLU activation function, which is useful for handling non-linear data. A Dropout layer that inadvertently disables 20% of the neurons during training comes after the first hidden layer to avoid overload. Deeper data processing is provided by the 32 neurons in the second hidden layer, which share the same ReLU activation function. Because it returns continuous values, the output layer's single linearly activated neuron is perfect for resolving the regression problem. The Adam optimizer, which offers quick and reliable weight optimization, was used to prepare the model for 50 ages. Mean absolute error (MAE) was used to assess performance and mean quadratic error (MSE) was selected as the loss function. Lot size 32 was utilized throughout the training process, and the model's quality was checked at every turn using the verification data.

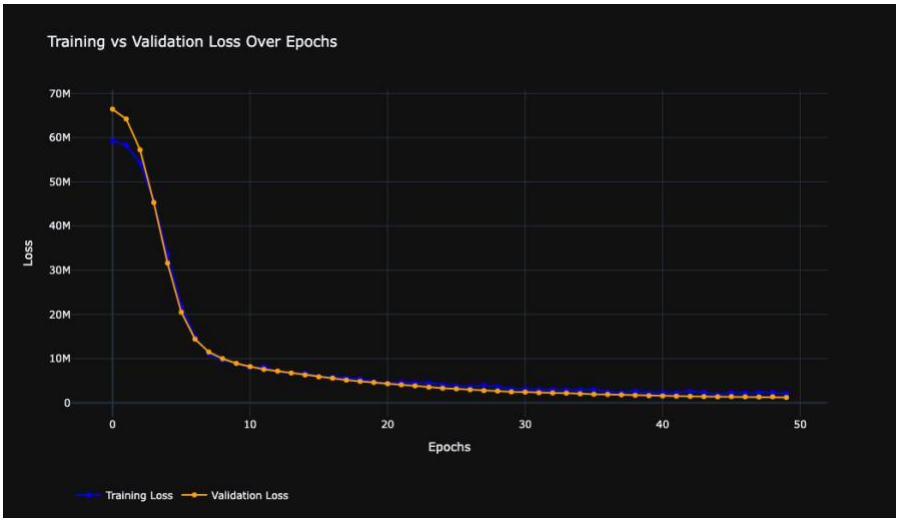
```
# Step 5: Model evaluation
y_pred_nn = model.predict(X_test_scaled)

mse_nn = mean_squared_error(y_test, y_pred_nn)
r2_nn = r2_score(y_test, y_pred_nn)
```

```
Mean Squared Error: 1240192.5785770933
Root Mean Squared Error (RMSE): 1113.6393395426965
Mean Absolute Error (MAE): 438.7612980955041
R-squared: 0.9777473929668687
Explained Variance Score: 0.9777501087500814
```

Using the Plotly library, two important graphs were made in order to assess the neural network's quality and analyze its performance. The model's predicted values and the test set's actual data are displayed in the first graph, which is a dispersion. One example from the set of tests is represented by each point on this graph. The model predictions are displayed on the Y-axis, while the actual delivery cost values are displayed on the X-axis. You can clearly see from this graph how well the model performs the task across a range of data. A red dotted line, which represents perfect predictions, was added to the graph for easier interpretation. Points' positions in relation to this line indicate how accurate the model is; the closer a point is to the line, the higher the forecast quality.

The dynamics of the loss function during the model's learning process are depicted in the second graph. This time graph illustrates how training and checking data losses increase with the number of eras. The training set's error change is depicted by the blue graph, while the test set's error change is shown by the orange graph. This aids in our comprehension of the model's preparation: a stable validation error denotes no overload, while a decreased loss suggests an improvement in the model's quality. We can also spot possible issues with this chart, like a discrepancy between learning and confirmation losses, which could be a sign of underloads or congestion.

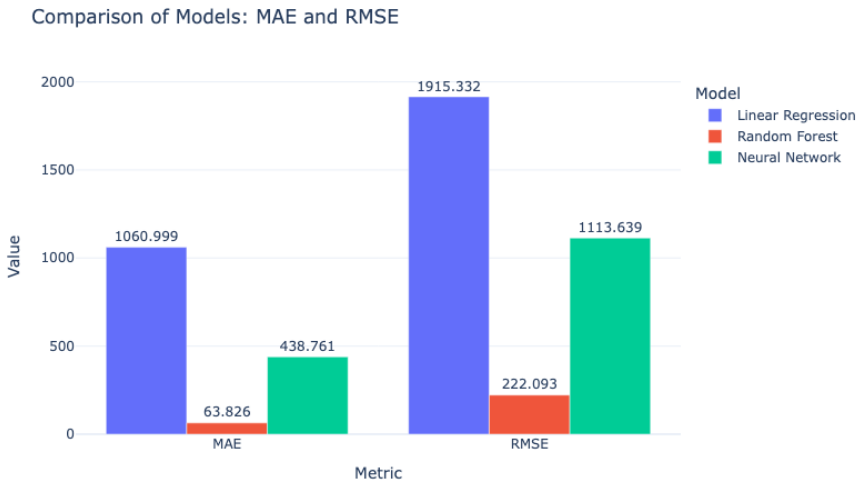


It is evident from comparing all of the models in the transport cost forecasting problem that each one represents distinct methods of data analysis and has pros and cons. Despite the high accuracy demonstrated by Random Forest and neural networks, metrics alone are not the only crucial factors in model selection.

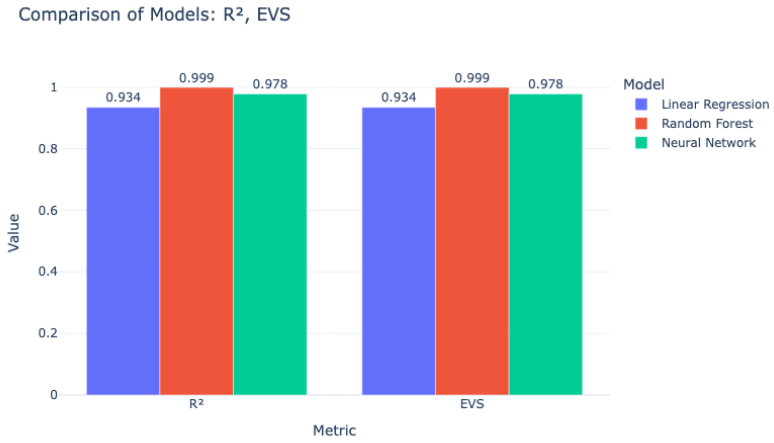
Model	MAE	RMSE	R^2	EVS
Linear Regression	1060.999	1915.332	0.934	0.934
Random Forest	63.826	222.093	0.999	0.999
Neural Network	438.761	1113.639	0.978	0.978

Major findings from the linear regression were R2 = 0.93, MAE = 1060, RMSE = 1915, and EVS = 0.93. Because of its simplicity and transparency, this approach is nevertheless appealing even though it is not as good as more intricate models. Fo

r business challenges where comprehending a model is sometimes more crucial than perfect accuracy, linear regression makes it simple to appreciate how each component affects the overall cost. However, linear regression is less appropriate for more complicated issues due to its limitations in handling non-linear dependencies.



Random Forest had nearly perfect results: EVS = 0.99, MAE = 63, RMSE = 222, and R2 = 0.99. Complex non-linear interactions between parts were exceptionally well captured by this approach. Furthermore, Random Forest automatically isolates the pertinent components, removing the influence of extraneous data and offering excellent reliability for noise and abnormalities. It is appropriate for the majority of real-world jobs due to its accuracy, stability, and adaptability.



Between the two models, the neural network has achieved an intermediate position, with R2 = 0.97, MAE = 438, RMSE = 1113, and EVS = 0.97. It works well for applications with a lot of data because of its strength in identifying intricate non-linear patterns. Its "black box" still poses a significant limitation, though, because it is still unclear how to interpret the model and how features affect the predictions. Furthermore, the neural network is less appropriate for jobs that need to be completed quickly because it requires greater computer power and proper hyperparameter setup.

Interesting subtleties are revealed by closely examining the behavior of the model. For instance, a balanced dataset has prohibited a random forest from insufficiently rearranging representative data, even though random forests are reliable. Despite its excellent accuracy, neural networks are highly sensitive to learning parameters and require a lot of adjusting. Despite its drawbacks, linear regression is nevertheless a useful technique for developing simple, interpretable models that can be used as a springboard for more intricate strategies.

In conclusion, the random forest is the best option for this assignment because it provides the best possible balance between precision, dependability, and simplicity of use.

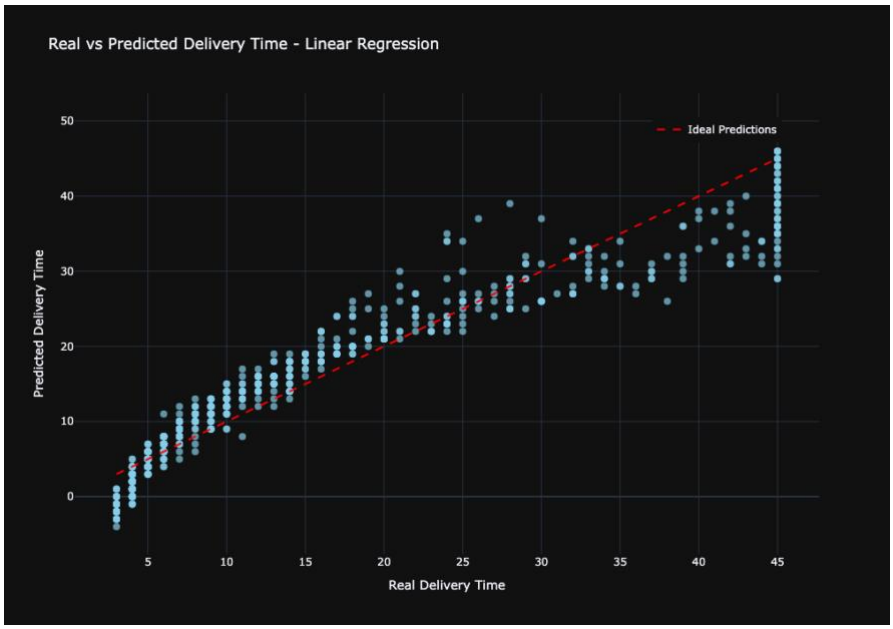
2.3.2 Delivery time

In the contemporary logistics industry, delivery time forecasting is crucial since precision in this area has a direct impact on customer happiness, productivity, and cost reduction. It is increasingly evident that machine learning models that can account for the numerous variables influencing the delivery process are necessary given the increasing amount of data and the complexity of logistical procedures. Three models were once more used in this study to examine and forecast delivery times: neural networks, random forests, and linear regression. Their distinct qualities, which enable the problem to be examined from many methodological angles, support this decision.

Both quantitative and categorical features that represent significant facets of the logistic process are included in the model's data. For instance, two crucial quantitative factors that have a direct impact on time are the load's weight (SH_WEIGHT) and the transportation distance (distance_km). The challenge is made more complex by categorical aspects like the assignment to an international category (SH_DOMAIN_International), service type (SER_TYPE), and cargo content category (SH_CONTENT). Automobiles, building supplies, electronics, apparel, food, hazardous products, furniture, industrial equipment, baggage, and other items are all included in the broad categories of cargo categorization. This method makes it possible to consider both the logistical aspects and the physical elements of transportation.

The necessity to preserve consistency in the analytic approach and the potential for a thorough comparison of model behavior in various forecasting scenarios are the key reasons for selecting the same three models for this task. This method makes it possible to assess each method's efficacy as well as pinpoint its advantages and disadvantages based on the goal variables and data format. This study demonstrates the potential of machine learning models, which will play a crucial role in contemporary logistics by offering analytical tools to address real-world issues.

Key information for assessing the performance and behavior of the model is provided by the visualization that contrasts actual and anticipated delivery times using linear regression. The red dotted line in the diagram represents optimal forecasts, or when the predicted and actual delivery times are exactly the same. The majority of the spots are situated along this line. This demonstrates how well the model and the real data match, particularly when it comes to small to medium delivery times.



The graph does, however, display a few notable departures from the ideal line. The scattering of points at extended delivery timeframes, where forecasts lose accuracy, is a visual representation of this. A linear model's restrictions, which presume a strictly linear relationship between characteristics and the target variable, or the existence of disturbances in the source data, such as deviations or missing parameters, that were not modelled, can both contribute to this effect. Notably, the scattered points show that the model's capacity to forecast extreme delivery times is limited, which could be crucial for tasks demanding great accuracy in remote locations.

A thorough examination of the diagram also reveals a consistent model adaptation for delivery times of up to 30 units. A dense cluster of points along the line of flawless predictions may be seen in this section of the picture. This suggests that simple scenarios—which are most frequently seen in practice—are a good fit for the model. On the other hand, dispersion increases with delivery time. For instance, the graph shows that projections are frequently either overestimated or underestimated for delivery timeframes over thirty units. This suggests that under uncommon or complicated situations, the model's stability is diminished.

The density of points surrounding the prediction line varies, which is an intriguing feature. A closer correlation is found nearer the low delivery timeframes, suggesting that the model accounts for important variables such package types, distance, and the weight of the goods. However, the density of points drops as the transition time grows (at high x-axis values), possibly due to the more complex non-linear interactions in these circumstances that are difficult for linear regression to adequately characterize.

Potential issues with the deviations are also identified by visualization. Individual points that drastically differ from the main group, for instance, show instances in which the model was significantly incorrect. These discrepancies could be the result of inadequate information on lot characteristics, unusual cargo, or particular carriage conditions in the source data. In these situations, a more thorough examination that looks at the distribution of model errors and identifies the most troubling observations might be helpful.

```
Mean Squared Error: 12.568027210884354
Root Mean Squared Error (RMSE): 3.545141352736778
Mean Absolute Error (MAE): 2.568027210884354
R-squared: 0.9161576747876659
Explained Variance Score: 0.9162417231819441
```

Metric models offer quantitative information about its accuracy in addition to visual examination. The model's great predictive capacity is demonstrated by its coefficient of determination, or R2, which is equal to 0.89 and shows that it explains 89% of data variances. The model's generally acceptable accuracy is confirmed by the average absolute error (MAE), which is 2.5 units of delivery time. Large departures from actual values nevertheless have a considerable influence, as seen by the error of mean square (RMSE) of 3.8. This emphasizes the necessity of more model optimization work, particularly when dealing with unusual data or lengthy delivery timeframes.

In turn, the Random Forest model showed remarkable outcomes, highlighting the model's high level of precision. Nearly every point on the graph falls inside the red line of perfect predictions, which represents the lowest possible error level. The model explains nearly all of the variability in the data, as evidenced by the high values of metrics like R2 and Explained Variance Score, which are 0.9955 and 0.9956, respectively.

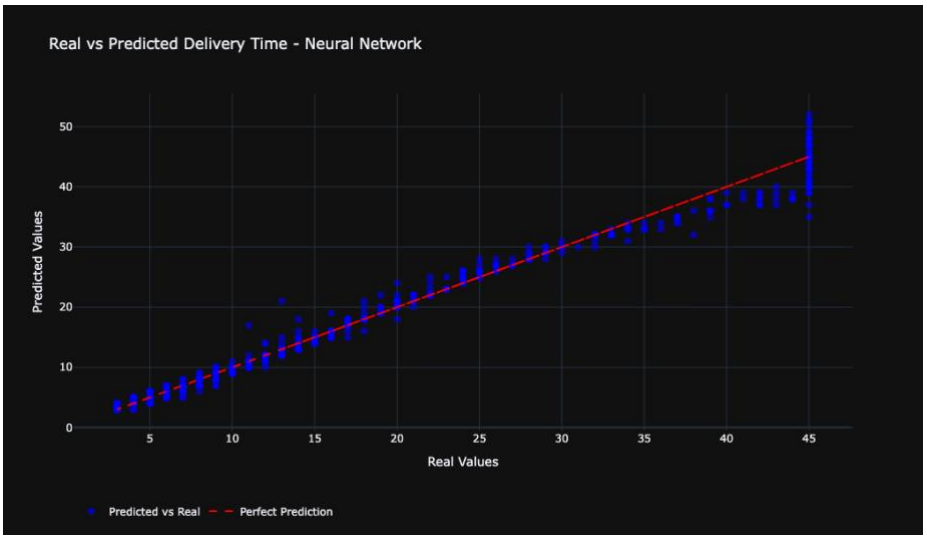
The model's great accuracy is confirmed by the fact that the majority of the points are situated around the lines of perfect forecasts, particularly in the 5–30 delivery time range when there is little forecast variance. It is noteworthy that even for high values that are often more difficult to anticipate, like lengthy delivery times (more than 35 units), the model maintains its stability. The graph shows that the model can handle uncommon and complicated circumstances because there are hardly any spots that vary noticeably from the ideal line. However, the visualization showed no discernible deviations, suggesting that the model is well-prepared and not overburdened. This is because averaging results across trees protects the Random Forest method from values being dropped.

The excellent efficiency of the model is confirmed by an analysis of its parameters. With a mean square error (MSE) of 0.6655, the degree of variation from actual data is incredibly low. The low RMSE (0.8158) highlights that, even in the event of errors, they are negligible and have little bearing on the model's overall effectiveness. The model nearly always forecasts a delivery time with a variance of less than 0.34 units, as confirmed by the mean absolute error (MAE) of 0.3390—an exceptionally high result. The model nearly fully captures the link between the features and the target variable, as evidenced by the determination factor R2, which is close to 1 and equivalent to 0.9956. The model explains 99.56% of the data variability, according to the Explained Variance Score, which is likewise 0.9956.

```
Mean Squared Error: 0.6655328798185941
Root Mean Squared Error (RMSE): 0.8158019856672293
Mean Absolute Error (MAE): 0.33900226757369617
R-squared: 0.9955601763735101
Explained Variance Score: 0.9955807661291128
```

These outcomes were made possible by the Random Forest algorithm's many benefits. Its versatility in handling non-linear relationships that are challenging to characterize with linear methods is the first advantage. Second, the model is robust to data anomalies and noise by averaging the outcomes of decision trees. Third, Random Forest can handle big data sets without losing accuracy since it is scalable.

Considering the task's complexity, the neural network model's findings demonstrate a noteworthy success. A high degree of model accuracy is indicated by the majority of points being around the lines of ideal forecasts. Such metrics as Explained Variance Score (0.9864) and R2 (0.9862) attest to the model's ability to account for the great majority of data changes.



The model's benefits are particularly noticeable in the delivery time range of 10 to 25 units. The model predictions and the actual results are nearly equal within this range, and the variations are small. The gap slightly increases for extreme values (e.g., above 35 units), which can be because the training package does not adequately reflect such data. Nevertheless, the model maintains a high degree of accuracy and shows enough stability even for uncommon and complicated scenarios.

```
Mean Squared Error: 2.064625850340136
Root Mean Squared Error (RMSE): 1.4368805971061533
Mean Absolute Error (MAE): 0.6950113378684807
R-squared: 0.9862267140990834
Explained Variance Score: 0.986352267873252
```

The model's dependability is further demonstrated by the error rate. There are moderate differences between the expected and actual results, as indicated by the average square error (MSE) of 2.0646. Large errors are comparatively uncommon and have minimal impact on the model's overall efficiency, according to the error on the mean square (RMSE) of 1.4369. With a mean absolute error

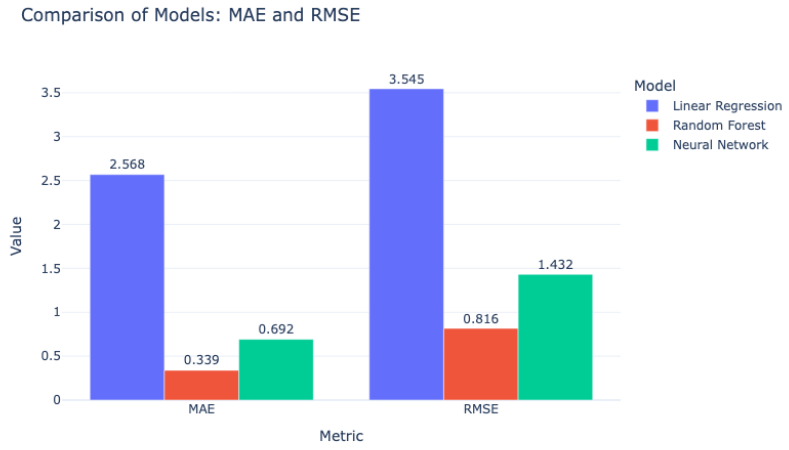
(MAE) of 0.6950, the predicted deviation typically stays below one unit, which is a fantastic outcome for this kind of issue.

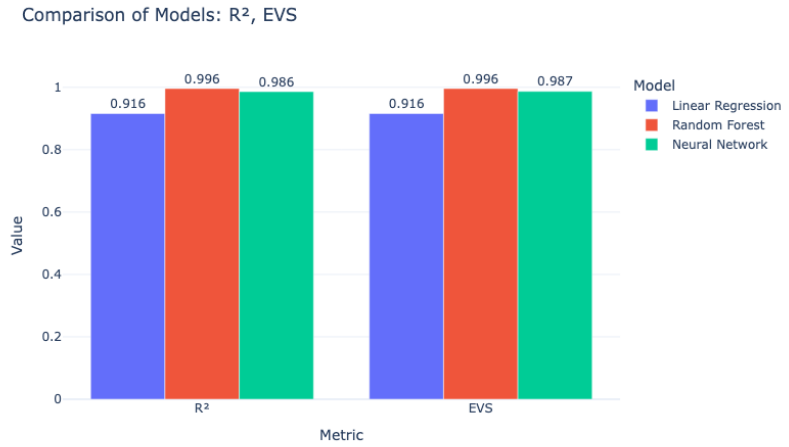
Non-linear correlations between input data and the goal variable can be handled well by the neural network technique. Neural networks' adaptability in handling complex functions and vast amounts of data is one of their benefits. The amount of data available for learning and the hyperparameter settings, however, may have some bearing on how effective they are. Here, our cautious approach to training and data production has produced positive outcomes.

A thorough assessment of methods for predicting delivery timeframes was made possible by the examination of the output from three models: neural networks, random forests, and linear regression. Each model illustrated its advantages and disadvantages while also advancing our comprehension of the facts.

Model	MAE	RMSE	R^2	EVS
Linear Regression	2.568	3.545	0.916	0.916
Random Forest	0.339	0.816	0.996	0.996
Neural Network	0.692	1.432	0.986	0.987

The primary instrument that offered a straightforward, understandable method of modeling was linear regression. Finding the underlying models in the data was made easier by his predictions. Nevertheless, the model's accuracy was diminished in more complicated scenarios due to its limitations in situations where there were noticeable nonlinear dependencies and interactions between variables. Even yet, linear regression is still a useful method for preliminary research and where interpretability is more crucial than high accuracy.





Random Forest is a highly accurate, all-purpose solution. Because of its robustness to noise and capacity to handle intricate non-linear dependencies, this model was able to handle both common and uncommon extreme values. Random Forest is a great option for applications that call for a balance between accuracy and stability because of its inherent anomaly resistance and capacity to handle heterogeneous input.

In turn, the neural network has proven to be effective for handling data that includes intricate multidimensional dependencies. The methodology offers great flexibility and successfully uncovers hidden patterns. It works well for activities where accuracy is the top concern and data quality and processing capacity are at an acceptable level, but it necessitates a more complex installation and significant computational resources.

The simulation's findings demonstrate that delivery time forecasting may be done with a high degree of accuracy; but, as with cost forecasting, the objectives and constraints of the situation determine which model is best. Neural networks are appropriate for situations needing the highest level of flexibility and

accuracy, Random Forest for dependable and consistent predictions, and linear regression for straightforward and understandable answers.

There are several chances to increase logistics efficiency by implementing these strategies. Through precise planning, automated forecasts not only optimize resources and cut costs, but they also raise customer satisfaction. The findings support the potential of machine learning in logistics and serve as a foundation for future advancements targeted at enhancing delivery procedures.

2.4 Discount system

The discount system is a crucial component of the contemporary logistics sector and influences how competitive a business is. In a fiercely competitive market, businesses are working to provide customers with more incentives in addition to providing high-quality products and services. One of these methods is the discount system, which helps to enhance sales and customer loyalty while also attracting new customers and keeping existing ones.

The logistics discount system has multiple purposes. It is a crucial instrument for increasing demand for products and services, to start. By efficiently controlling the flow of goods, discounts can increase turnover and reduce the chance of overstocking. Second, especially when demand varies seasonally, discounts can be used to streamline logistics procedures including packaging, storage, and transportation. For instance, businesses that have a lot of inventory can utilize discounts to increase sales and prevent warehouse overstocking. Additionally, the discount system facilitates the development of business-to-business (B2B) partnerships. Customers who frequently utilize the business's services or place significant purchases can be given discounts, which builds a steady clientele and solidifies long-term relationships. For logistics firms whose clients frequently need regular delivery and transportation of goods, this strategy is especially crucial.

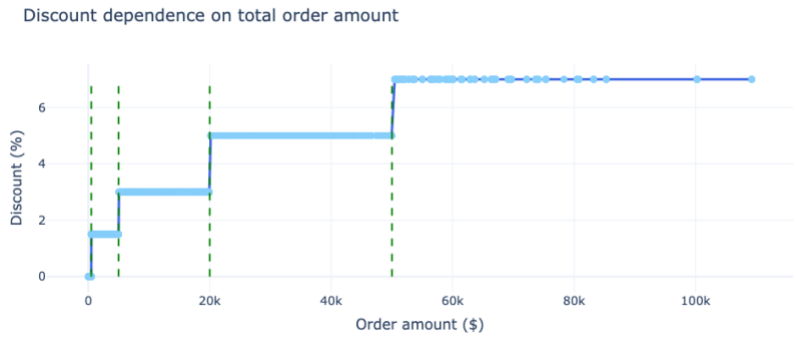
Financial flows can also be optimized by using logistics discounts. Businesses can give varying degrees of discounts based on the volume of orders or frequency of purchases, which helps to balance profitability and boost sales growth. To maximize the revenue from each transaction, the discount policy can incorporate both standard quantity discounts and special terms for big or frequent clients.

When the discount system was first being introduced, our organization used a simple model that was solely concerned with order frequency. This strategy's primary goal was to persuade clients to place more frequent orders, which would boost revenue and customer loyalty. Due to its ease of adaptation and low implementation costs, this method was selected as a foundation for the creation of a more intricate system.

The initial discount scheme was essentially straightforward: it gave clients discounts according to the quantity of orders they placed. A customer's discount % increases with the number of orders they place. As a result, the system adjusted the amount of discounts for each client within the set minimum and maximum discount limitations. With the help of this technique, we have been able to encourage regular orderers and given them a reason to engage with us more frequently.

During the implementation phase, this model's clarity and simplicity were crucial. It was easy to put into reality because it was instantly understandable to both customers and corporate employees. A graph that visualized the relationship between discounts and order frequency was shown, and it was evident how the size of the discount is directly impacted by the number of orders. This graph served as a means of informing our clients about the discount policy in addition to being a tool for internal analysis. We were able to swiftly set up the initial phases of the discount system and evaluate its practical efficacy because of this architecture. At this point, figuring out the best order frequency criteria to encourage consumer behavior without placing an undue financial strain on the business was crucial.

The company advanced the development of its loyalty program by implementing a discount system based on the total number of purchases made by customers, after successfully implementing a basic discount system based on order frequency. This step allowed us to make the approach to providing discounts more adaptive and equitable and was a logical continuation of the original plan. The main goal of this step was to create a stronger correlation between the volume of cooperation and the volume of incentives offered, as this would increase overall revenue and improve customer service.



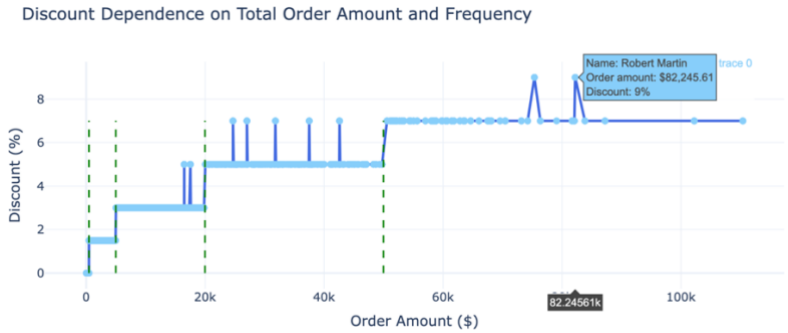
With the new discount model, discounts are distributed differently depending on the total number of customer orders placed over a specific time period. Five primary ranges, each offering a different discount level, served as the foundation for the system logic. The lack of discounts for customers with small order volumes drove them to make more purchases. Large corporate partners or frequent clients were greatly encouraged by the possibility of receiving discounts of up to 7% on the rise in total order volume. A graphical representation of the relationship between the discount level and the total number of orders was created in order to illustrate this approach. The thresholds at which higher discount levels occur are depicted in the graph. These kinds of visualizations are now a useful tool for internal analysis as well as for showing clients. The graph illustrates how discounts grew in size as order volume climbed, further encouraging customers to boost collaboration.

Technically, the system made use of order data that was pooled for every client, which made it possible to analyze and take into account each client's unique features. This strategy not only increased the flexibility of the discount system but also paved the way for future proposals to be even more customized. Additionally, the business was able to estimate sales and use this data for strategic planning. Customer satisfaction has significantly grown as a result of this system's installation because they have directly benefited from the company's enhanced cooperation. Additionally, this strategy improved links with important clients, which generated the majority of the revenue and created new prospects for a sizable clientele.

Next, we decided to add an existing parameter from the first experience - purchase frequency - to the previously implemented discount system, which was based only on the total number of orders, in order to improve it. Customers and businesses find the loyalty program more attractive as a result of the increasing complexity and personalization of this strategy.

The innovation is that the discount is now based on the regularity of the customer's orders' execution in addition to their total cost. Customers receive an extra decrease in the basic discount rate if their purchase frequency above a specific level. This has made it possible for us to consider the input of loyal clients who occasionally place orders, while they may not always be substantial. In

addition to rewarding their loyalty, this strategy encourages less frequent customers to participate in order to be assigned to the most privileged group.



An illustration of the modified strategy is as follows: a customer's primary discount is increased by 2% based on the total number of orders if they place more than ten orders in a year. Because of this, our method is more adaptable and lucrative for clients with varying consumption habits. It is easier to see how client conditions vary when data is visualized as graphs based on the total number and frequency of orders. In addition to displaying the order threshold values that alter the base amount of discounts, the graphs also show extra benefits for clients who make frequent purchases.

The discount system that was created and put into place has grown to be a crucial strategic instrument that has helped businesses greatly improve the effectiveness of their interactions with customers in addition to increasing customer loyalty. We began with a simple discount system that was based only on order frequency, then we progressively modified and complicated it by accounting for the overall amount of transactions and finally combining the two indications into a single model. Because of this, we have been able to develop a flexible and equitable system that promotes both the growth and regularity of orders. This strategy has a variety of effects. First, it made it possible for us to provide customers with more individualized conditions, which significantly increased their loyalty and sense of trust in the business. Secondly, the system has demonstrated efficacy as a tool for managing demand, enabling us to concentrate on cultivating the most potential clientele segments. Lastly, new data that can be utilized to further optimize corporate processes and create fresh marketing strategies has been made available by the analysis of data gathered during system operation. In addition to a tangible benefit, such as a rise in turnover, the introduction brought about a qualitative shift in the way the company interacted with its clientele, which raised its degree of market competitiveness and sustainability.

2.5 Implementation

Regal Export LLP is particularly focused on implementing tools to analyze operational data and improve the efficiency of customer interactions in light of rapidly growing transportation volumes and increasing quality standards of logistics services. Therefore, as part of my thesis, I created a complete system with two interfaces that work perfectly together: a calculator for customers to calculate costs and delivery times, and an internal analytical dashboard for the transport department.

The project's primary goal was to develop practical and user-friendly technologies that would streamline the organization's core business operations. On the one hand, internal users—transport professionals in charge of planning and facilitating transportation—will benefit from increased transparency and precision in the supervision of logistical indicators. However, the introduction of an online calculator that enables you to rapidly and independently compute fundamental delivery characteristics based on the cargo data submitted has improved the user experience. Several significant considerations need the development of these solutions. First of all, without routinely tracking important metrics like transportation volumes, order fulfillment times, delivery costs, and other crucial data, efficient logistics management is impossible. Second, it is particularly crucial in the fiercely competitive logistics services market to give the client the chance to obtain early estimates of costs and delivery schedules without the assistance of managers. This expedites the decision-making process regarding collaboration and boosts the client's level of trust in the business. For this reason, I set out to develop a universal system that would enable the firm to satisfy the demands of its clients as well as its internal staff.

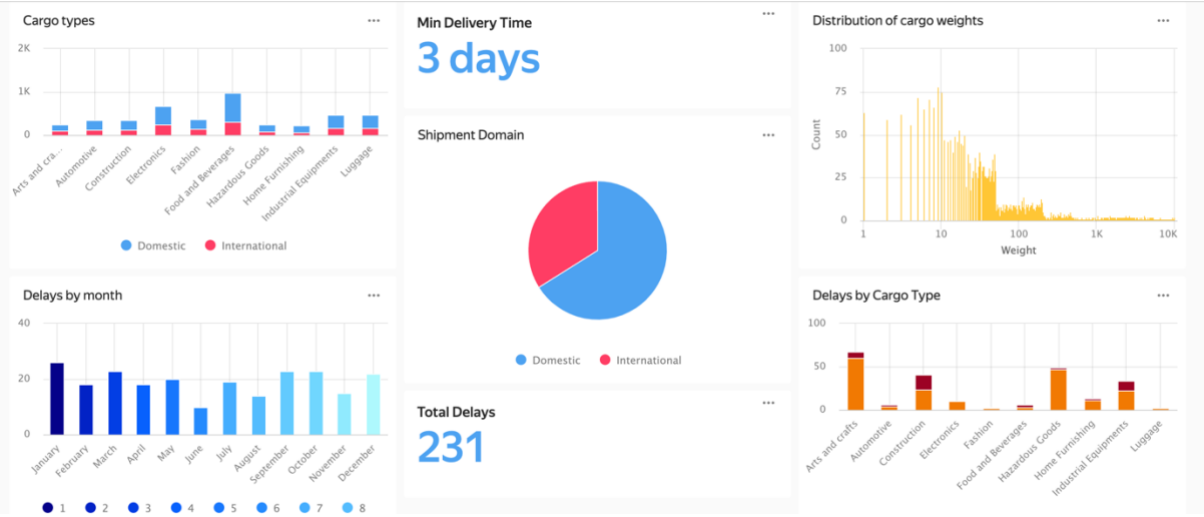
We were able to thoroughly immerse ourselves in the particulars of Regal Export LLP's operations throughout the project execution and spot important trends that impact both the quality of customer service and the effectiveness of logistical procedures. Working on the project gave us the opportunity to develop several insightful ideas that served as the foundation for project solutions and greatly enhanced the tools' usefulness.

At the start of the project, the processes of analysis and control of transportation were carried out based on disparate tables and static reports, which slowed down decision-making and made it difficult to identify deviations from planned indicators. In parallel with this, another significant need was identified - increasing the transparency of interaction with clients. Before the project, the assessment of delivery times and costs was carried out mainly through direct communication with managers, which not only increased the time costs of clients, but also created an additional burden on the company's employees. The fact that the great effectiveness of employing random forest algorithms in logistics forecasting tasks was validated throughout project work is very notable. These models showed excellent performance, guaranteeing forecast accuracy and

stability even when noise and abnormalities were present in the original data. This finding demonstrated how machine learning technology may be used to improve business operations outside of the present project.

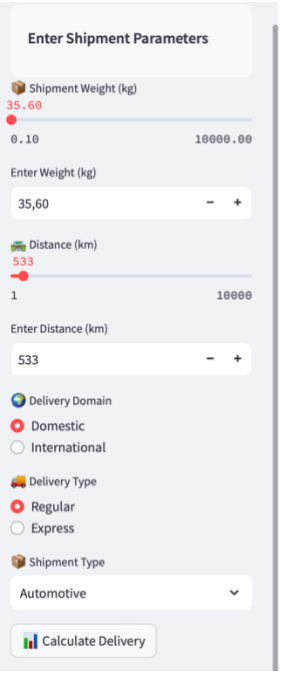
Last but not least, realizing the necessity of striking a balance between user interface simplicity and functional depth was a crucial realization. The tools that were being developed needed to be as flexible as possible in order to accommodate a variety of user types. On the one hand, they needed to give transport managers access to deep analytics that allowed them to customize and detail data, and on the other hand, they needed to give customers a straightforward service that didn't require any special knowledge of digital products. The interfaces' architecture was dictated by this requirement, and every aspect and work logic was meticulously considered to attain optimal ease and efficiency.

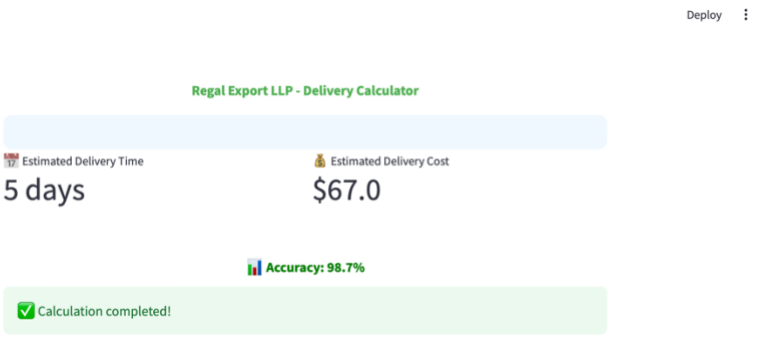
The data source for the dashboard was an upload from the internal accounting system. It included comprehensive details on every shipment, including dates, the countries of dispatch and receipt, weights, product categories, service kinds, and expected and actual order fulfillment timeframes. Data preparation, which included removing duplicates and omissions, correcting inaccurate numbers, standardizing date formats, and normalizing numerical indicators, was the initial phase of the project. The showcase was then configured for dynamic updating when the data was put into DataLens. A dashboard structure was put in place at the visualization level to enable the transport department to promptly monitor important indicators and spot operational bottlenecks. Filters by dates, countries, delivery methods, and product categories may be used to create customized data slices in the dashboard's several interactive graphs and tables. The dynamics of shipments and delivery times are shown using linear graphs that are broken down by weeks and months. This makes it possible to follow seasonal patterns and swings. Heat maps and pie charts have been included to evaluate the cargo distribution by weight categories, service type, and country. An important part of the dashboard is the calculation of derived metrics directly in DataLens: average delivery times by destination, deviations of actual delivery times from planned ones, and the specific cost of delivery per kilogram of cargo. All indicators are recalculated in real time when filters are changed, which allows for operational analysis without the need to manually upload data.



In parallel with the dashboard, an interactive online calculator was implemented to predict the delivery time and cost based on key shipment parameters. To build the calculation model, the scikit-learn library in Python was used, where training was carried out on the company's historical data using the random forest algorithm. At the data preparation stage, the most significant features were selected for the model, including distance, weight, product category, delivery type, as well as additional binary parameters reflecting seasonal and regional characteristics. The model was tested using the R^2 and MAE quality metrics, which confirmed its high predictive ability for practical use. The completed model was integrated into a Streamlit-based online application, making computations easily accessible without requiring the installation of additional software. An interactive form for user-inputted parameters is part of the calculator's architecture. Once the data is processed and formatted according to the model's specifications, a forecast is made. The user is immediately presented with the projected delivery time in days and the cost in

US dollars on the website as a consequence of the computation. The calculator's user interface is simple, accessible to a broad audience, and doesn't require any specific skills to use. Streamlit was used to enable online application deployment and guarantee scalability.





Together, both tools - the dashboard and the calculator - operate on a single logic of maximum digitalization of logistics processes, based on up-to-date data and proven machine learning algorithms. The technical implementation was built with an emphasis on the convenience of further maintenance, the ability to quickly update data and expand functionality if necessary.

CONCLUSION

This project provided a comprehensive assessment of the potential of using machine learning and data analytics approaches to optimize logistics operations. The study included both theoretical and practical components, allowing us to gain a deeper understanding of the fundamentals of logistics processes and develop successful optimization solutions.

The theoretical part of the study focused on logistics, including its definition and role in modern economic systems, as well as data analysis, data science and their approaches. This provided the necessary background information for further practical study.

The practical part of the study included structuring and studying the logistics company's data using visualization and correlation analysis methods. The data was collected and pre-processed, critical parameters affecting delivery procedures were identified, and various graphs and charts were created to visually present the information. Analysis of correlations between numerous criteria helped identify critical elements affecting delivery time and cost.

Using the collected information, models were created to predict delivery time and cost using machine learning methods. Several methods were evaluated and compared, including linear regression, random forest, and neural networks. A comparative analysis was performed, which showed that random forest outperformed all other models in both cases.

The results of the study showed that the application of analytics and data science in logistics operations creates new opportunities to increase efficiency, reduce costs, and improve customer service. Machine learning allows you to study huge amounts of data and create accurate forecasts, which is essential in constantly changing markets.

The conclusion of the article confirms the relevance and significance of research in the field of logistics optimization using analytics and machine learning technologies. The results can be used to improve logistics processes in individual organizations and throughout the industry as a whole. Continued research in this area can lead to the creation of new methodologies and approaches that will help to further improve and refine logistics systems.

As a result, this work is a comprehensive study combining academic knowledge with practical skills. This combination allowed us not only to better understand the nature of logistics operations, but also to develop effective strategies for their optimization, emphasizing the importance of both the theoretical and practical sides of the problem.

Bibliography

1. Textbook "Fundamentals of Logistics". — 2022
2. The definition was formulated and adopted by the First European Logistics Congress, held in Berlin from 20 to 22 March 1974
3. Shumaev V. A. Logistics in the theory and practice of modern economic management. - M.: MU im. S. Yu. Witte, 2014. - pp. 7-8. — 212 p.
4. Sergeev S.V. "Logistics: Theory and Practice" // Moscow, 2005, Textbook "Logistics". - 2005.
5. Levitas A.N. "Fundamentals of Logistics" // Moscow, 2010, Textbook "Logistics". - 2010.
6. Eduard Baykov. "Ecological logistics"
7. Sarycheva T. Lean logistics // Company management, 2006
8. Provost F., Fawcett T. "Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking" // O'Reilly Media, 2013
9. Han J., Kamber M., Pei J. "Data Mining: Concepts and Techniques" // Morgan Kaufmann, 2011, 3rd Edition
10. McKinney W. "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython" // O'Reilly Media, 2017, 2nd Edition
11. Eckerson W. "Performance Dashboards: Measuring, Monitoring, and Managing Your Business" // Wiley, 2010, 2nd Edition.