**BCIS 5110**

**PROGRAMMING FOR BUSINESS ANALYTICS**

**Exploratory Data Analysis (EDA) and Regression Modeling for JD.com Customer and Sales Data**

**Project Group – 15**

*Harichandana Bhogaraju*

*Joel Madaram*

*Dattu Bhaskar Molakalapalli*

*Srikanth Pichu*

A green logo with a black background

Description automatically generated with low confidence

**Executive Summary**

The executive summary highlights the analysis of JD.com's extensive data, which has been used to develop predictive models for customer order delivery times. The comprehensive dataset includes multiple tables: Order, User, Delivery, Inventory, Network, SKU, and Click data, each offering unique insights. The report details the processes of data preparation, including cleaning and consolidation. An exploratory analysis has been performed to check the patterns and the trends of the data. The report also lays the groundwork for building predictive models, setting the stage for advanced analysis and practical applications in e-commerce logistics. The definition of the challenge is to determine whether, based on previous activity, the user level on the site influences price and expenditure going forward. In order to help with financial and marketing decisions, it entails posing perceptive queries that are addressed in the report.

**Project Motivation/ Background**

The project's motivation and background based on the objective: predicting customer order delivery times in the context of JD.com's e-commerce operations. This task is driven by a need to enhance customer satisfaction, optimize logistics, and improve operational efficiency. Several key aspects underline the significance of this objective:

**Customer Satisfaction:** Accurate delivery time predictions are important for providing a positive customer experience. When customers receive their orders within the given time frame, it leads to increased satisfaction and trust in the e-commerce platform.

**Operational Efficiency:** Timely deliveries also contribute to operational efficiency. By predicting delivery times accurately, JD.com can streamline its logistics operations, reduce waiting times, and allocate resources more effectively.

**Resource Allocation:** Predicting delivery times makes it easy for better resource allocation, such as delivery personnel and vehicles. It ensures that the right resources are available at the right time and location, reducing operational costs.

**Inventory Management:** It is important to understand customer behavior and product movement patterns aids in inventory management. This helps JD.com stock popular products adequately and manage inventory levels efficiently.

The dataset encompasses seven tables, capturing the interactions of 2.5 million users and 31,868 SKUs during March 2018. Tables include information on SKUs, users, clicks (browsing history), orders, delivery details, inventory, and the network structure of warehouses. This diverse set of data allows for a holistic examination of the customer journey from product selection to delivery.

The proposal aims to conduct an in-depth Exploratory Data Analysis (EDA) and subsequently build a regression model using the comprehensive dataset provided by JD.com. The dataset spans various aspects of user behavior, sales transactions, product attributes, and logistics, offering a rich landscape for exploration and predictive modeling. The primary objectives include understanding the underlying patterns in customer behavior, identifying key factors influencing sales, and ultimately constructing a regression model to predict sales based on various features.

In summary, the project's motivation and background revolve around the pivotal role that accurate delivery time predictions play in enhancing customer satisfaction, optimizing logistics, and driving operational efficiency in the e-commerce sector. Understanding customer behavior, product movement, and logistical efficiency is at the core of addressing this critical challenge.

**Data Description**

The data provided by JD.com is very unique and covers various facets of e-commerce operations. It is organized into several tables, each of which contains specific information about different aspects of the business. Below is a detailed description of the data tables and the types of information they contain:

**Orders Table:** This table contains information related to customer orders which includes data such as order IDs, timestamps, customer IDs, product IDs, quantities ordered, and order statuses. This table is important for understanding customer purchase behaviors and the timing of orders.

**Users Table:** The Users table provides an depth knowledge of customer demographics and attributes. It includes data on customer IDs, gender, age, location, and possibly other relevant demographic information. Analyzing this table can help identify customer segments and their preferences.

**Delivery Table:** This table focuses on delivery details which includes data on order IDs, delivery timestamps, delivery locations, and possibly delivery statuses. Understanding delivery logistics is very important to predict the delivery times.

**Inventory Table:** The Inventory table contains information about the availability of products in JD.com's inventory. It includes data on product IDs, stock quantities, and possibly information about product categories or attributes. Efficient inventory management is important for the order fulfillment at time.

**Network Table:** This table provides insights into JD.com's warehouse network. It may include data on warehouse locations, capacities, and other relevant network-related information. Understanding the network can help optimize product distribution.

**SKU (Stock Keeping Unit) Table:** The SKU table contains data related to product attributes. It includes information on product IDs, descriptions, prices, and possibly other product-specific attributes. This table provides great understanding of the product characteristics and the pricing strategies.

**Clicks Table:** The Clicks table captures data on customer interactions with products on the e-commerce platform. It may include information on customer IDs, product IDs, timestamps of clicks, and possibly other interaction-related data. This data can be valuable for studying customer browsing behaviors.

Overall, the data is rich and diverse, providing customer information, product details, order history, delivery logistics, inventory status, and network infrastructure. Such a comprehensive dataset provides a solid foundation for developing predictive models for customer order delivery. **Exploring the Data**

**Identifying Quality Issues**

An initial examination of the dataset reveals potential quality issues, with a focus on the delivery\_time variable. Issues such as missing data or extreme values could have a negative impact on predictive analyses.

**Proposed Solution for Quality Issue**

For addressing missing data, this analysis proposes employing imputation techniques, such as replacing missing values with the mean or median.

Extreme values can be handled through outlier detection methods, allowing us to either adjust or exclude them based on their impact on the dataset.

**Calculating Sales for Specific SKU on a Specific Day**

A critical operational metric involves calculating daily sales for a specific SKU. This would be accomplished by summing the quantities of that SKU sold each day.

**Determining Order Fulfillment by Distribution Center**

The distribution\_center variable in the orders table will be scrutinized to ascertain whether an order can be fulfilled by a specific distribution center. If the order exceeds the capacity of the distribution center, it may not be fulfilled, highlighting potential operational challenges.

**Evaluating Delivery Priorities for PLUS Members**

An analysis of order data for PLUS members will be conducted, with a focus on variables such as delivery\_time and order\_priority. If PLUS members consistently experience shorter delivery times, it suggests the presence of delivery priorities for this customer segment.

Conclusion

This project encompasses a dual focus on both descriptive and predictive analytics, addressing fundamental questions about customer behavior and predicting future sales trends. The proposed research questions are designed to shed light on critical aspects of JD.com's business operations.

By acknowledging potential quality issues within the dataset and providing solutions, the study ensures the reliability and integrity of the subsequent analyses. The exploration of the data structure and variables is a crucial first step toward unraveling the modalities of the JD.com dataset, laying the groundwork for a more in-depth analysis and subsequent modeling.

**Data Preparation**

Data Preparation is a critical step in the data analysis process which plays a crucial role in ensuring that the data used for analysis is accurate, consistent, and ready for modeling and exploration. In the context of the JD.com data analysis, the Data Preparation process involves several key steps, each designed to make the data suitable for subsequent analysis. Here is a detailed description of the Data Preparation process:

**Importing Necessary Packages:** The first step in any data analysis project is to import the necessary libraries and packages in the chosen programming environment as these provide tools and functions required for the data manipulation, analysis, and visualization (e.g., here Python with libraries like Pandas, NumPy, and Matplotlib).

**Reading CSV Files:** The JD.com dataset is provided in CSV format, which is a common format for structured data. The Data Preparation process starts by reading these CSV files into data structures which are manipulated and analyzed. Each CSV file corresponds to one of the data tables mentioned earlier (e.g., Orders, Users, Delivery, Inventory, etc.).

**Handling Missing Values:** Real-world datasets often have missing or incomplete data, which can affect the quality of the analysis. During the Data Preparation phase, missing values are checked and analyzed. This can involve techniques such as imputation (replacing missing values with estimates) or removing rows or columns with a high proportion of missing data. The goal is to ensure that the dataset is as complete as possible.

**Data Cleaning:** Data cleaning involves identifying and correcting errors or inconsistencies in the dataset. In this case, the missing values have been identified and have been replaced or removed in the dataset.

**Data Merging:** Since the JD.com data is structured into multiple tables, it may be necessary to merge or join these tables to create a dataset that contains all the required information. In this case, the Order table and the Delivery table have been merged using the common key Order\_ID.

**Removing duplicate keys:** Once the data has been cleaned, the duplicate keys have to be checked and removed if any. Here, the item which is a single item and the item is a gift item has been removed. And the items with multiple packages on the same Order\_ID have been removed.

**Data Transformation:** Data transformation is used to create new variables or features that can be valuable for analysis. The time-related variables in this notebook are: order\_date, order\_time, ship\_out\_time, arr\_station\_time, and arr\_time. This can include aggregating data, creating calculated fields, and encoding categorical variables.

**Data Formatting:** Ensuring that the data is in the correct format for analysis is essential. This includes formatting dates, ensuring numeric data is in the correct units, and addressing any other formatting issues specific to the dataset.

**Data Validation:** Before proceeding with exploratory data analysis or modeling, it's important to validate the prepared dataset to ensure that it has the requirements of the analysis objectives.

**Data Splitting:** In machine learning and predictive modeling tasks, it's common to split the dataset into training and testing sets to assess model performance. This step typically occurs after Data Preparation and before Model Building.

By using the above steps in the Data preparation, the dataset has been changed into a clean structured, and well-prepared format. This ensures that subsequent analysis, including Exploratory Data Analysis and Model Building, can be conducted effectively and that the results are reliable and meaningful. Data Preparation is a critical foundation for any data-driven project, and it significantly influences the quality of insights and predictions derived from the data.

**Exploratory Data Analysis:**

Exploratory Data Analysis (EDA) is a crucial phase in any data analysis project. It involves examining and visualizing the data to gain insights, discover patterns, and identify trends. In the context of the JD.com dataset, EDA plays a pivotal role in understanding customer behaviors, product trends, and other significant patterns. Here's a detailed description of what the EDA process might entail:

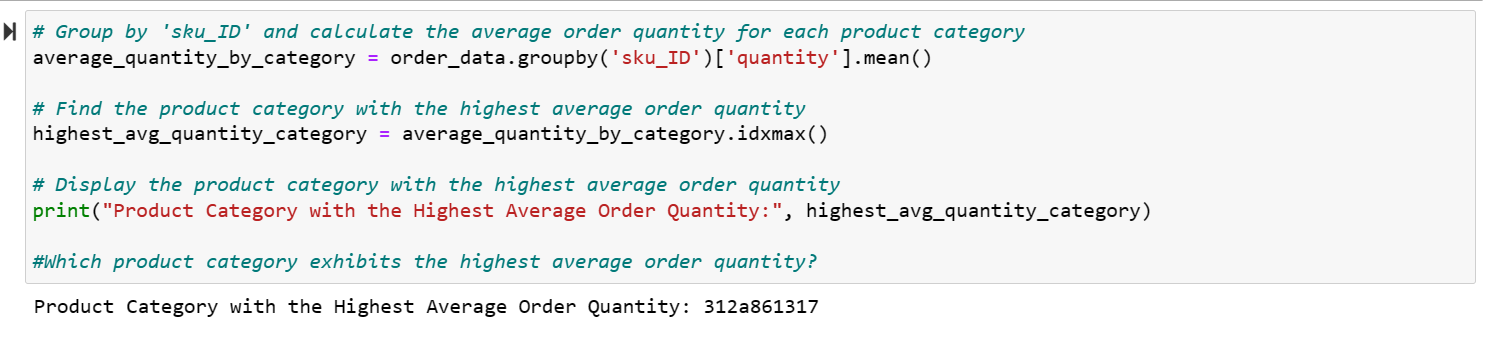
**Descriptive Statistics:** EDA often begins with the calculation of basic summary statistics for key variables in the dataset. This includes measures such as mean, median, standard deviation, minimum, maximum, and percentiles.

**Data Visualization:** Visualization is a powerful tool for exploring data. Various types of charts and graphs can be created to provide a visual representation of the data. Common visualizations include histograms, bar charts, scatter plots, line charts, and pie charts.

**Correlation Heatmaps:** Explore relationships between variables by creating a correlation heatmap. This can reveal which variables are strongly correlated and which are not.

The values of -1 in the User table have been replaced with the value “New” and the values of 10 are been replaced with the “Bus”.

**Which product category exhibits the highest average order quantity?**



Product category 312a861317 exhibits the highest average order quantity for this we had used the variables sku\_ID, quantity which are related to the order data.

**What is the education level of the majority?**

Education level “3” has the majority over 200,000

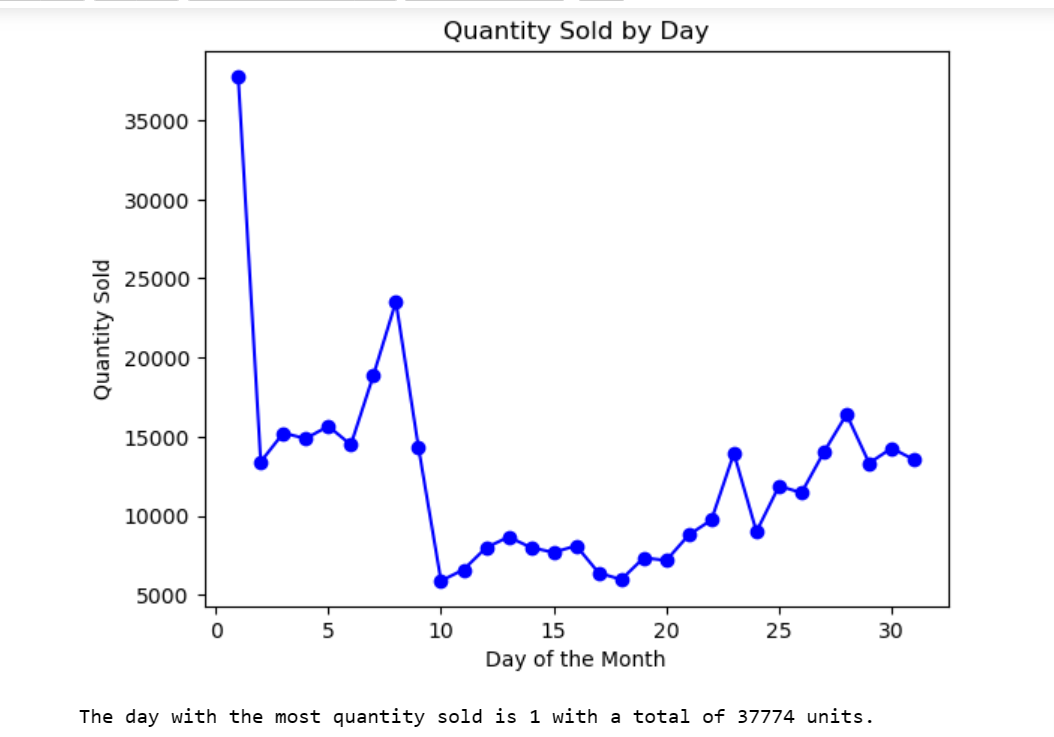
The bar graphs have been used to find the education level of the majority. We can see that in Jupyter notebook.

**Which age level has the most users?**

Age group between 26 to 35, has users over 175,000

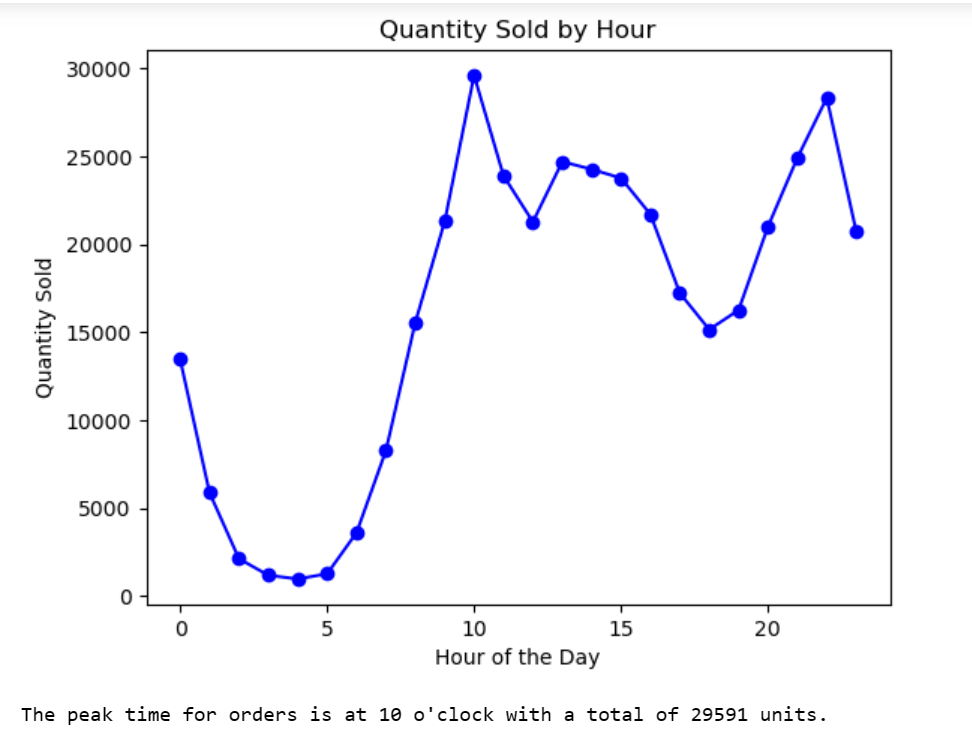
The bar graphs have been used to find the age level which has the most users. We can see that in Jupyter notebook.

**Which day has the most quantity sold?**

****

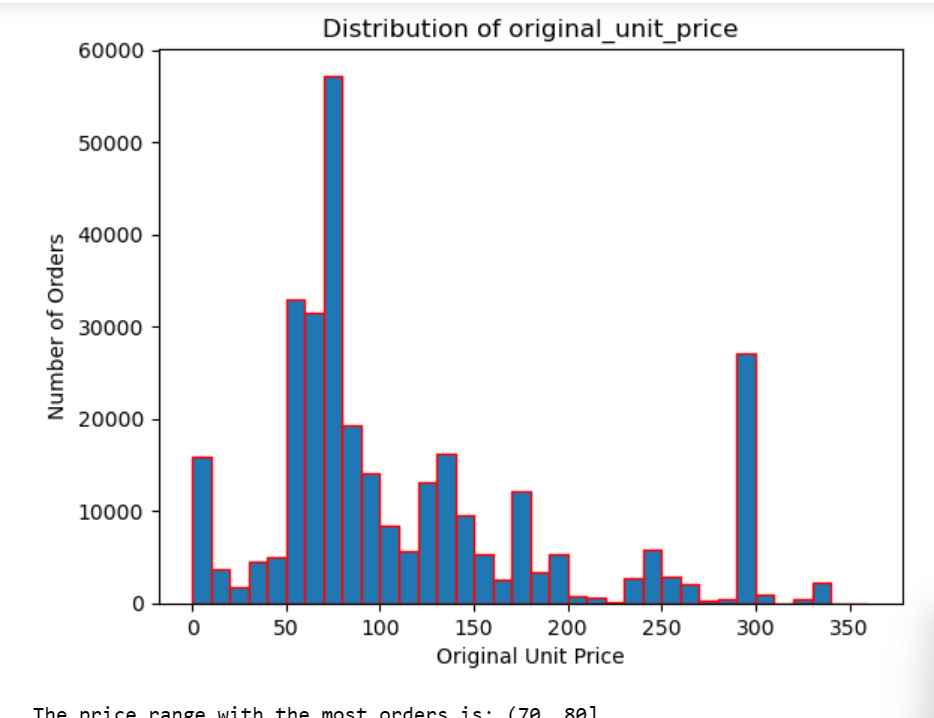
We had used the variables order\_day and quantity to provided the day on which the most quantity has been sold. We can see that Day 1 has the most quantity sod with 37774 units.

**When is the peak time for orders during a day?**

****

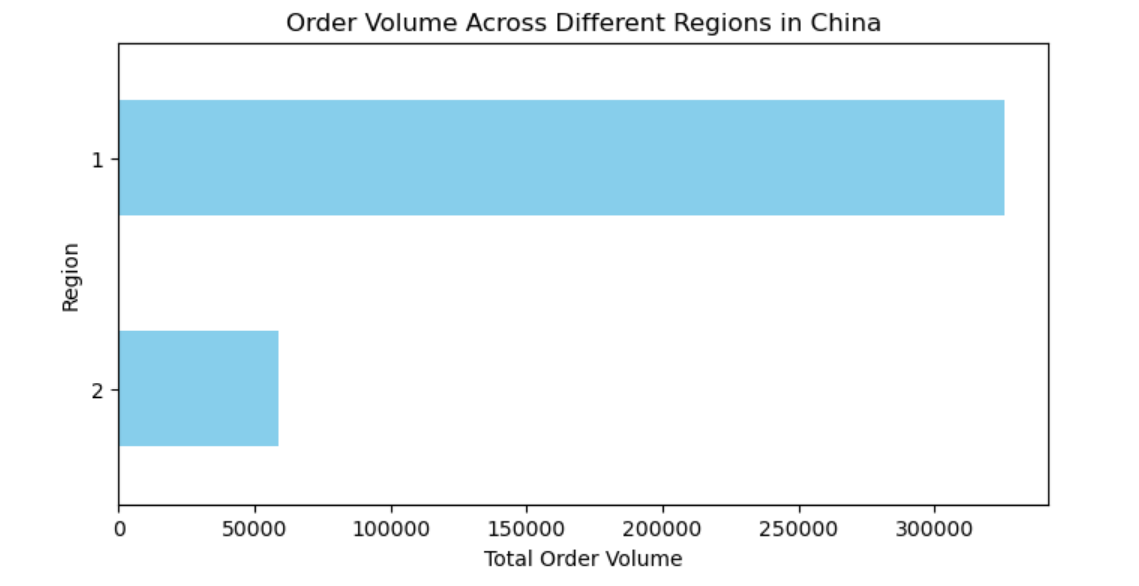
We had used the variables order\_hours and quantity to provides the peak time for orders during a day. We can see that at 10 o’Clock we had more number of order compare to the other timings.

**Which price range has the most orders?**

****

We had used the variable Original Unit price to get the details for what price range has more orders from the bar graph we get to know that price range between 70-80 we have more orders.

**How does order volume vary across different regions in China?**

****

We have used the type and quantity variable to get the order volume vary across different regions in China. The values type and quantity have been multiplied to display the varying order value across different regions in China.

The EDA phase is essential for shaping the direction of the analysis and helping stakeholders understand the data's characteristics. It lays the foundation for subsequent modeling and analysis, enabling data analysts and data scientists to make informed decisions and derive actionable insights from the dataset.

**Models and Analysis**

In the "Models and Analysis" section of the report, the primary objective is to develop and evaluate predictive models for customer order delivery times. This section outlines the methodologies employed, the process of model training, and the evaluation of model accuracy.

**Methodologies Selection:** The Methodologies used for the models have been mentioned in this section. The methodologies for predicting delivery times include:

**Regression Analysis:** Decision tree regression, technique have been used to model the relationship between various input features (such as order details, customer information, and product attributes) and delivery times.

**Model Training:** In the model training, the training data has been used to provide the selected algorithms or models.

**Model Evaluation:** Different metrics can be used to evaluate the performance of the predictive models in general. But here, only Root Mean Squared Error has been used.

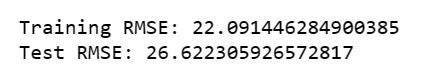
***Root Mean Squared Error (RMSE):*** RMSE is the square root of MSE, providing an interpretable metric in the same units as the target variable.

**Model Performance**: The model evaluation results have been provided.

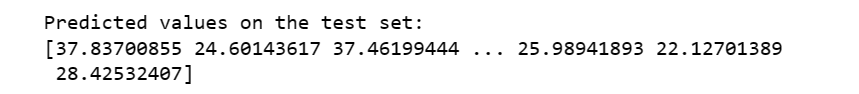
The "Models and Analysis" section provides a comprehensive view of the modeling process, ensuring that stakeholders understand how predictive models were developed and their performance in predicting customer order delivery times.

To build a predictive model we had used the Decision tree regression model for that at firstly we had taken the variables original unit price and quantity to calculate the original value and final unit price and quantity to calculate the final value of orders. We had used the data related to order and user table for the decision tree regression model, for this we can observe the variables which had across the same order, we use the 'first' method to keep the value in the groupby result. We aggregate the variables and we used the filtered data later we merge the order table with using user table.

We need to transform the delivery time to hours to get the out put in hours format, we had calculated the dis\_rate and busy\_hour too, which is used as a variable for the model. Split the data in to train and test and we can see the output for the train and test data. After that we had used the train and test data to build a decision tree regression model we got the out put is



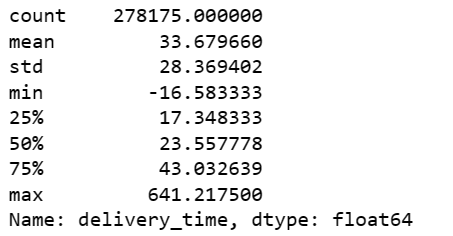
We made the predictions on testing data the output is



Evaluated the model performance by using RMSE (Root Mean Squared Error)



Delivery\_time for the decision tree regression model is different from the actual mean of the original data



As per the results predicted mean value of delivery\_time and actual delivery\_time are varies.

**Findings and Managerial Implications:**

In the "Findings and Managerial Implications" section of the Jupyter Notebook analysis of JD.com data, the focus is on summarizing the key insights obtained from the analysis and discussing their impact on business strategies and decision-making. Here's a detailed description of what this section might entail:

**Key Insights Summary:** The most important findings and insights derived from the analysis like patterns in customer behavior, peak order times or preferred product categories, correlations or trends discovered in the data, start by providing a concise summary of the most significant findings and insights derived from the analysis. These insights should relate directly to the project's objectives, which include predicting customer order delivery times. Some example key insights might include:

**Impact on Business Strategies:** The key insights have a great impact on business strategies like operational efficiency, customer experience

**Operational Efficiency:** If the analysis identifies inefficiencies in delivery times, JD.com can use this information to streamline its supply chain operations and improve delivery speed.

**Customer Experience:** Insights into customer behavior and preferences can guide JD.com in tailoring its services to meet customer expectations, leading to improved customer satisfaction and loyalty.

**Inventory Management:** Findings related to inventory levels and product movement can inform inventory management strategies, reducing carrying costs and ensuring product availability.

**Marketing and Promotions:** Understanding customer behavior can help JD.com design more effective marketing campaigns and promotions, targeting specific customer segments with relevant products.

**Data-Driven Decision-Making:** Emphasize the importance of data-driven decision-making within the organization. Highlight how the insights obtained from the analysis demonstrate the value of leveraging data for strategic planning.

**Conclusion:**

This report provides an insight on how the JD.com has provided its data and deep analysis on the data has been done. The questions have been answered successfully. Different data analysis methods have been used, different cleaning methods have been used to clean the data and then the data was merged using the common keys. The duplicate keys have been eliminated and the graphs have been created using the parameters in the tables. The Root Mean square error have been identified in the data. The success of internet firms has been greatly attributed to business analytics. Any company that wants to grow and expand must use business analytics to guide its decision-making process. Compared to other analytics tools, Python has greatly aided in managing large amounts of data and processing it smoothly with minimal resources. There were hundreds of thousands of records in this procedure, making the data frame large. The programming language allows one to explore, edit, and create various data structures. Thus, it is feasible to acquire meaningful attribute data with minimal effort and time investment.

Appendix 1

Orders Table Data structure

|  |  |  |  |
| --- | --- | --- | --- |
| # | Column | Non-Null Count | Dtype |
| 0 | order\_ID | 549989 non-null | object |
| 1 | user\_ID | 549989 non-null | object |
| 2 | sku\_ID | 549989 non-null | object |
| 3 | order\_date | 549989 non-null | object |
| 4 | order\_time | 549989 non-null | object |
| 5 | quantity | 549989 non-null | int64 |
| 6 | type | 549989 non-null | int64 |
| 7 | promise | 549989 non-null | object |
| 8 | original\_unit\_price | 549989 non-null | float64 |
| 9 | final\_unit\_price | 549989 non-null | float64 |
| 10 | direct\_discount\_per\_unit | 549989 non-null | float64 |
| 11 | quantity\_discount\_per\_unit | 549989 non-null | float64 |
| 12 | bundle discount per unit | 549989 non-null | float64 |
| 13 | coupon\_discount\_per\_unit | 549989 non-null | float64 |
| 14 | gift\_item | 549989 non-null | int64 |
| 15 | dc\_ori | 549989 | int64 |
| 16 | do des | 549989 | int64 |

SKUs Table Data Structure

|  |  |  |  |
| --- | --- | --- | --- |
|  | | |  |
| # | Data columns | Non-Null Count | Dtype |
| 0 | sku\_ID | 31868 non-null | object |
| 1 | type | 31868 non-null | int64 |
| 2 | brand\_ID | 31868 non-null | object |
| 3 | attribute) | 31868 non-null | object |
| 4 | attribute2 | 31868 non-null | object |
| 5 | activate\_date | 3058 non-null | object |
| 6 | deactivate date | 1141 non-null | object |

Appendix 2 Users Table Data Structure

|  |  |  |  |
| --- | --- | --- | --- |
| # | Column | Non-Null Count | Dtype |
| 0 | user\_ID | 457298 non-null | object |
| 1 | user\_level | 457298 non-null | int64 |
| 2 | first\_order\_month | 457298 non-null | object |
| 3 | plus | 457298 non-null | int64 |
| 4 | gender | 457298 non-null | object |
| 5 | age | 457298 non-null | object |
| 6 | marital\_status | 457298 non-null | object |
| 7 | education | 457298 non-null | int64 |
| 8 | city\_level | 457298 non-null | int64 |
| 9 | purchase\_power | 457298 non-null | int64 |

**Deliery Table Data Structure**

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 package\_ID 293229 non-null object

1 order\_ID 293229 non-null object

2 type 293229 non-null int64

3 ship\_out\_time 293229 non-null object

4 arr\_station\_time 293229 non-null object

5 arr\_time 293229 non-null object

**Inventory Table Dat Structure**

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 dc\_ID 136079 non-null int64

1 sku\_ID 136079 non-null object

2 date 136079 non-null object

**Clicl Table Data Structure**

# Column Dtype

--- ------ -----

0 sku\_ID object

1 user\_ID object

2 request\_time object

3 channel object

**Network Data Table Structure**

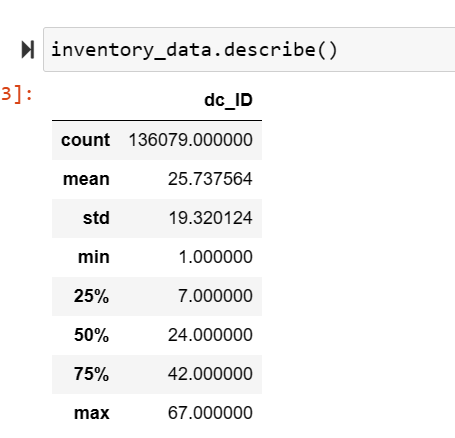
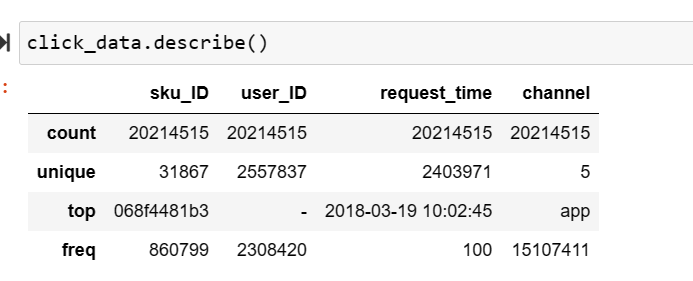
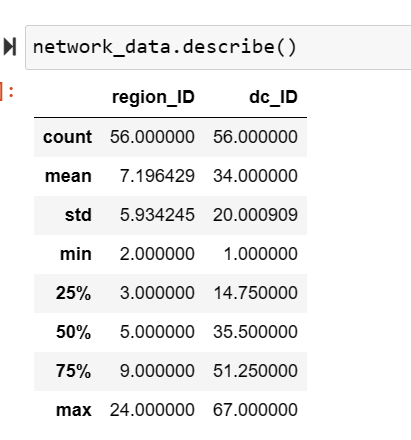
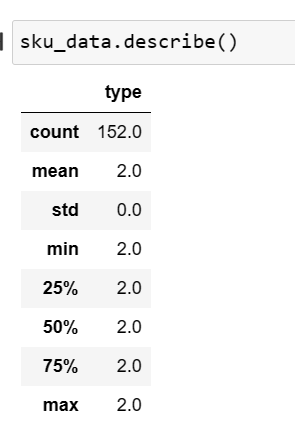
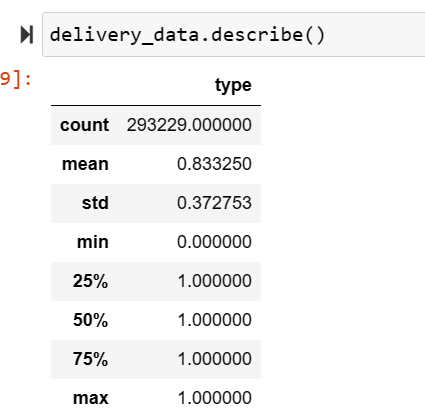
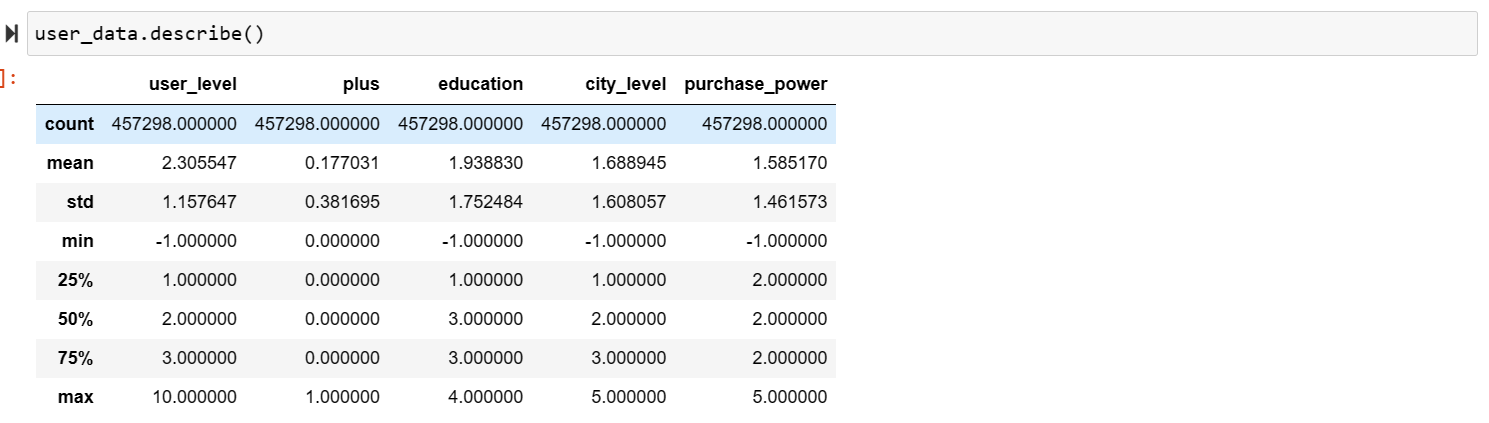
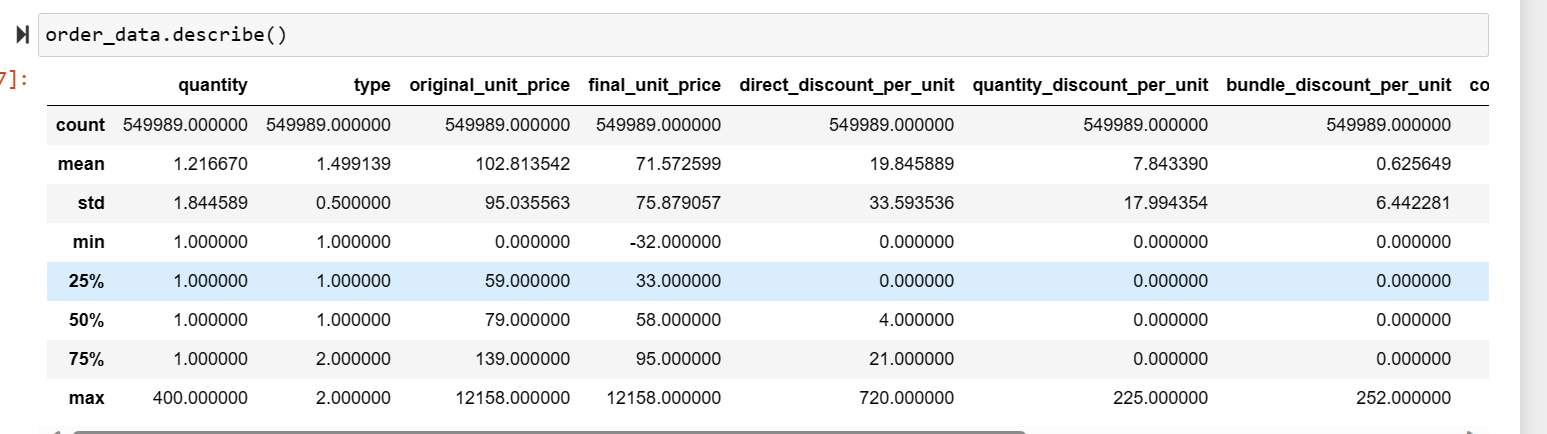
# Column Non-Null Count Dtype

--- ------ -------------- -----

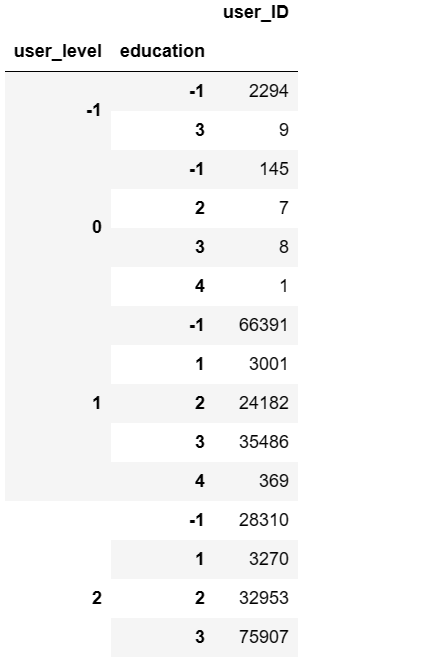
0 region\_ID 56 non-null int64

1 dc\_ID 56 non-null int64

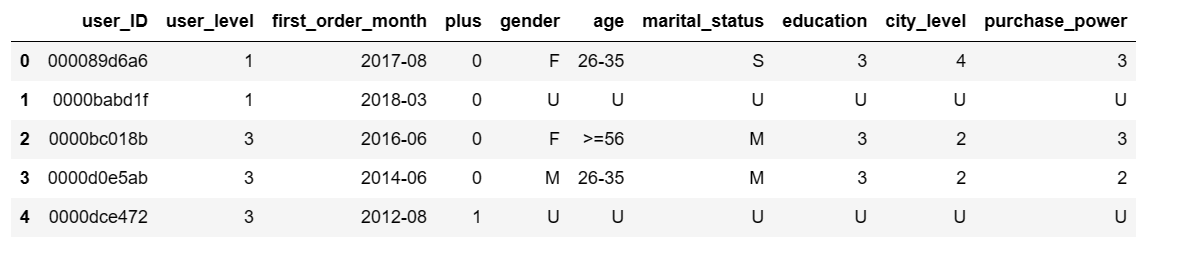
Appendix 3 Descriptive statistics of all tables



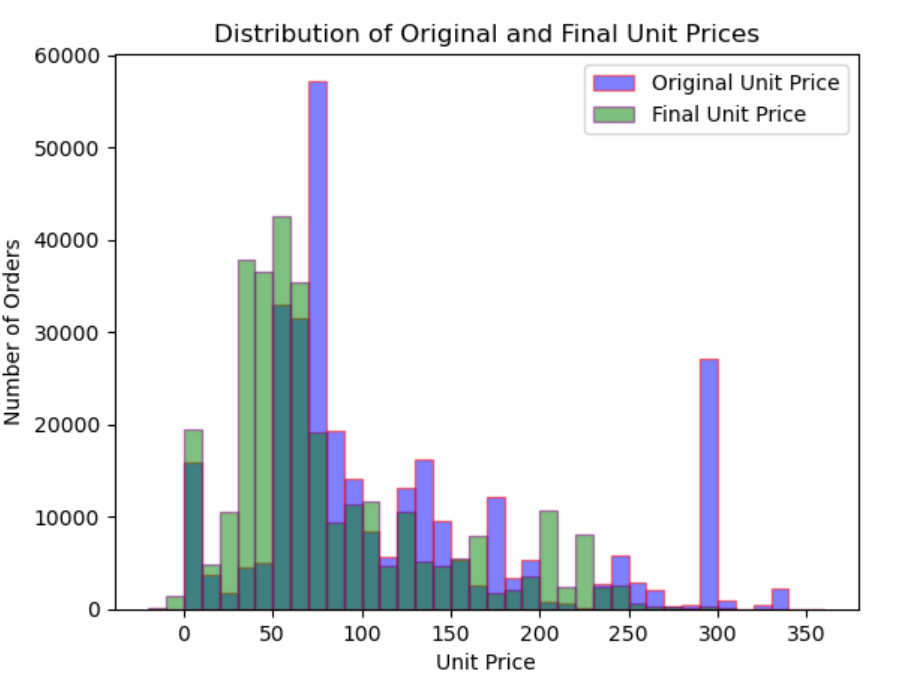
Pivot table that counts the customers on their user level and Education.



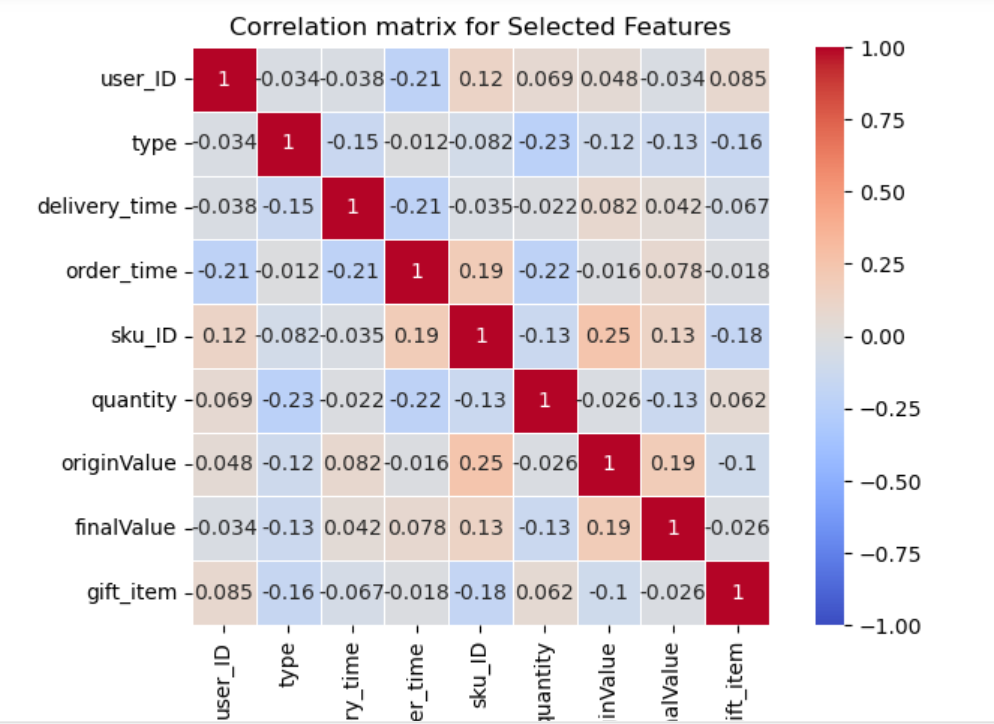
The meaning of -1 in education, city\_level and purchase\_power is missing values. We will replace it with 'U', as missing value indicator of other variables like 'age', 'gender', etc.



Comparing to original unit prices, how they are different



**Appendix 4 correlation for the model data frame**



**References**

Tamboli, N. (2022, November 30). Tackling missing value in dataset. Retrieved from https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/

VanderPlas, J. (n.d.). Combining datasets: Merge and join. Retrieved from https://jakevdp.github.io/PythonDataScienceHandbook/03.07-merge-and-join.html

Weinstein, A. (2013). *Handbook of market segmentation: Strategic targeting for business and technology firms* (3rd ed.). Routledge.

Shatby, S. E. (2022, May 18). How to build a predictive model in Python? Retrieved from <https://365datascience.com/tutorials/python-tutorials/predictive-model-python/>