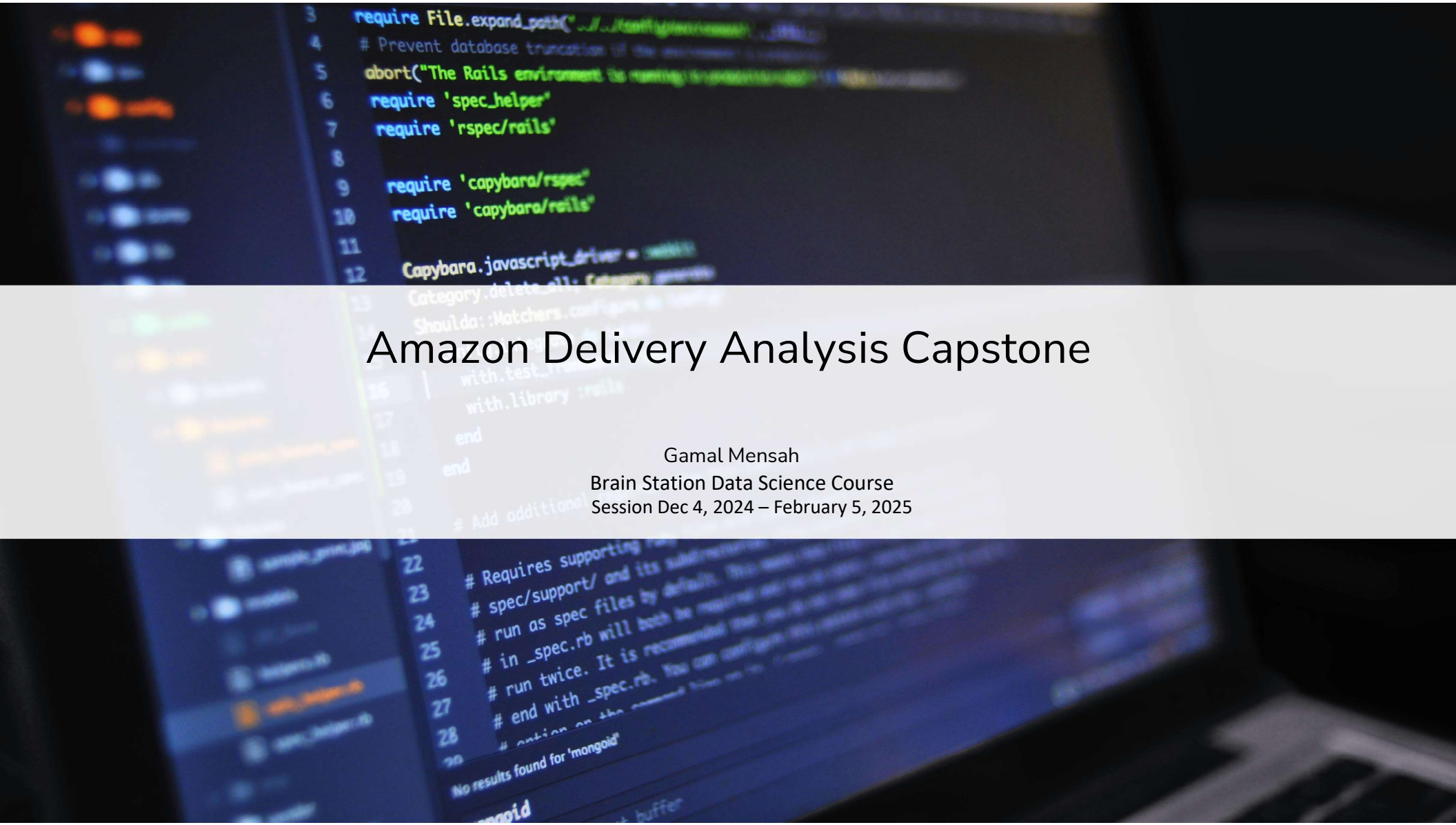


Amazon Delivery Analysis Capstone

Gamal Mensah

Brain Station Data Science Course

Session Dec 4, 2024 – February 5, 2025

Agenda

- Introduction : Overview of the problem and hypothesis
- Data Sourcing : Dataset description and relevant features
- Data Cleaning : Data preprocessing methods
- Exploratory Data Analysis: Visualizing trends and relationships, key visuals to explain insights
- Model Development : Chosen model and prediction approach
- Model Validation : Validation techniques and results
- Key Insights : Summary of findings and implications
- Conclusion and Next Steps : Recap and future recommendations

The Problem Statement

For the Data Science Course capstone project, I will investigate how weather conditions, area type (urban vs. rural), and traffic congestion impact Amazon delivery times.

By sourcing a deliver attribute relevant dataset, I will clean and explore the data, develop visualizations to identify key trends, and create a numerical model to predict delivery times.

I will apply model validation techniques and provide explanatory visuals to highlight insights, aiming to understand how these factors contribute to delivery delays.

Hypothesis

Hypothesis:

- Weather, area type (urban vs. rural), and traffic congestion affect delivery times.

Rationale:

- **Weather**: Adverse conditions like rain or snow can slow deliveries.
- **Area Type**: Urban areas face congestion, while rural areas have longer travel distances but fewer delays.
- **Traffic**: High traffic and road closures cause significant delays.

Hypothesis

Target variable for predictive modeling and ideal features:

- Understanding and modelling Delivery Time lead to recommendations to improve "operational efficiency", "cost reduction", and "route optimization" by enabling better resource allocation and scheduling.
- “Customer satisfaction" could be enhanced by providing the ability to accurate ETAs while helping detect "delays, or inefficiencies.
- The Business could gain a "competitive edge" by enhancements to logistics, reducing costs, and improving delivery reliability at scale.

Data and Tools Used

- Data was sources solely from **Kaggle** based on having some familiarity with it was a resource.
- The **Amazon Delivery Dataset** (amazon_delivery.csv) was the single data source consisting **43,739 records** and **16 columns**.
- Notably it has very few missing observations Agent_Rating (54 missing), Weather (91 missing).
- **The main features that you used :** Delivery_Time, Order_Time, Order_Time, Longitude and Latitude
- In my ideal scenario, I would have also found a source for “customer comments” related to delivers would have allowed for NLP analysis.

Data Processing

How Data Was Refined:

- Values were converted `Order_Date`, `Order_Time`, and `Pickup_Time` from object type to datetime format for proper time-based analysis.
- Haversine library was used to simplify distance calculations between two geographic points using latitude and longitude

Missing Values were addressed based on the data type:

  - Numerical data like `Agent_Rating`was replaced with the median to avoid skewing the distribution.
  - Categorical data `Weather`, mode was used to fill in the most common category, ensuring consistency.

Data Processing

- The dataset surprisingly had **no duplicates**, but if there were any, I would have handled them by identifying and removing exact duplicates.
- Boxplot was used to examine the data for **outliers**, however, the values appeared centered with no significant outliers.

Data Processing

Feature Engineering :

- Distance was calculated in kilometers using the **Haversine** library formula to ensure accurate measurements and removed any values with zero for both lat and long.
- **Order duration**, **extracted the day of the week**, and **categorized order time into bins**. Missing values were handled using `isna()` to identify and remove NaNs.
- The **day of the week** was derived from `Order_Date` and **categorized** `Order_Time` into morning, afternoon, and night **bins** for easier time-based analysis.

131.47

Exploratory Data Analysis

What are some main KPIs from your dataset?

- Average Delivery Time by Product Category
- Traffic impact of traffic on delivery time.

Were there any surprising insights or trends from your data exploration?

- The delivery time was consistently average across all products (131.47), except for groceries (26.48), which had a significantly lower delivery time.
- A positive correlation between delivery time and driver age, older drivers tend to have longer delivery times.
- A strong negative correlation between agent rating and delivery time, lower delivery times correspond to higher ratings.
- Younger drivers generally receive higher ratings.
- Deliveries appear faster when it is cloudy or foggy.

Do you have plots showing these insights?

- **See Appendix**

Modeling

Model:

- I compared Random Forest and Linear Regression models for performance evaluation.

Optimizing Model Performance

- To optimize performance, I tuned parameters like the number of trees and depth for Random Forest

Which one performed the best?

- Random Forest was more accurate than Linear Regression with a lower Mean Absolute Error and Mean Squared Error , and a higher R-squared score (0.812 vs. 0.719).
- RF was best due to its ability to capture non-linear relationships and reduce overfitting.

How can you improve the accuracy of your best model?

- I could further fine-tune parameters, do try feature engineering.

Summary Conclusions

My hypothesis was validated, with a few surprising findings.

The analysis showed that weather impacts delivery time, with cloudy or foggy days leading to more deliveries.

Additionally, ordering groceries speeds up delivery time.

Interestingly, suburban areas had longer delivery times than both urban and rural areas.

Key Learnings

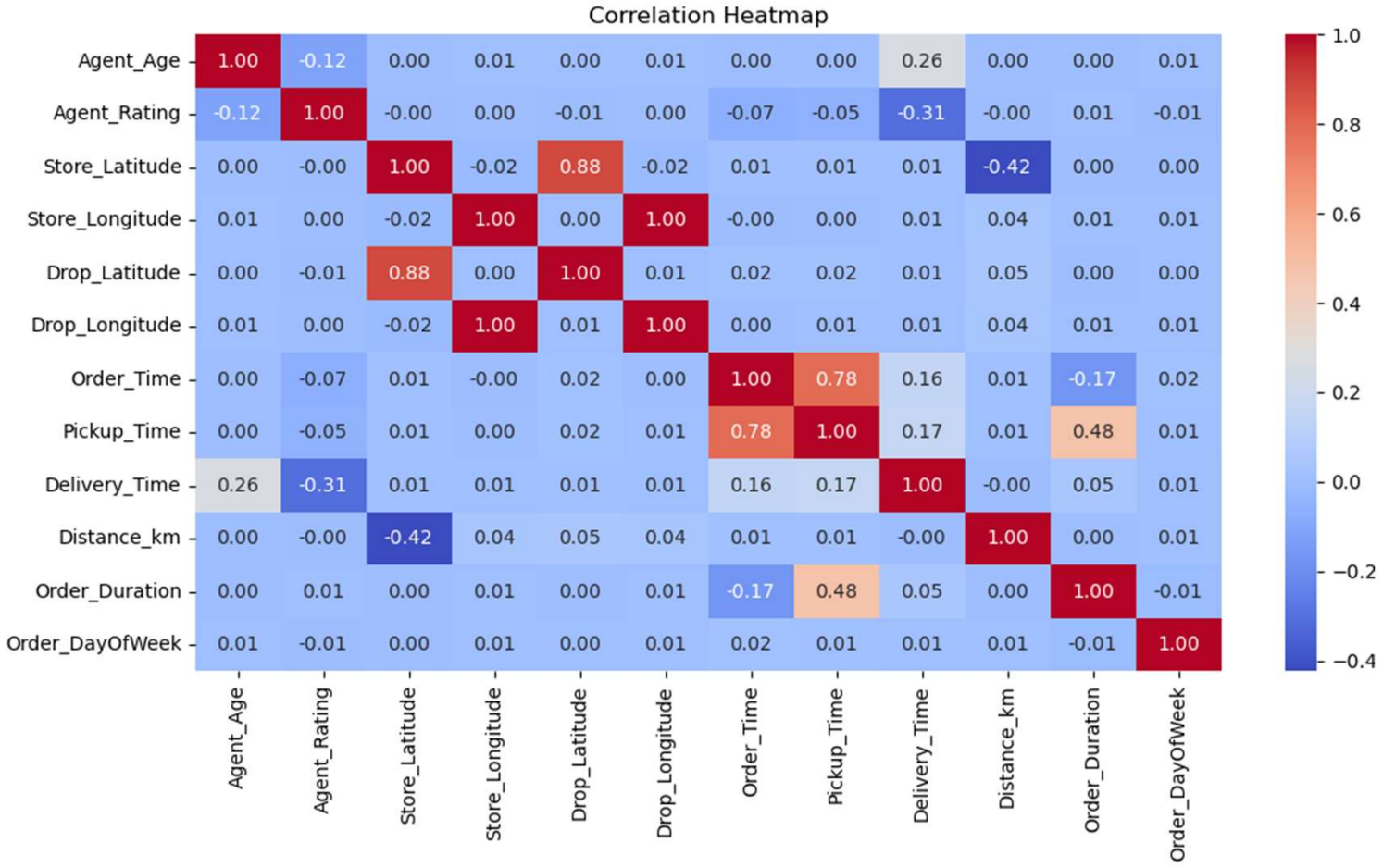
What would you change in this project process if you could do this all over again?

- Data cleaning definitely seems to follow the 80/20 rule, where 80% of the time is spent on cleaning and preparing the data, and the remaining 20% on modeling and analysis.
- Proper data cleaning is crucial to ensure the quality and accuracy of the insights drawn from the dataset.
- The process was very iterative, and I found that commenting and using pseudocode were essential, especially during data cleaning.
- I often had to move debugging steps earlier in the process during data exploration to catch issues sooner and ensure the integrity of the dataset.
- Do more statistical analysis

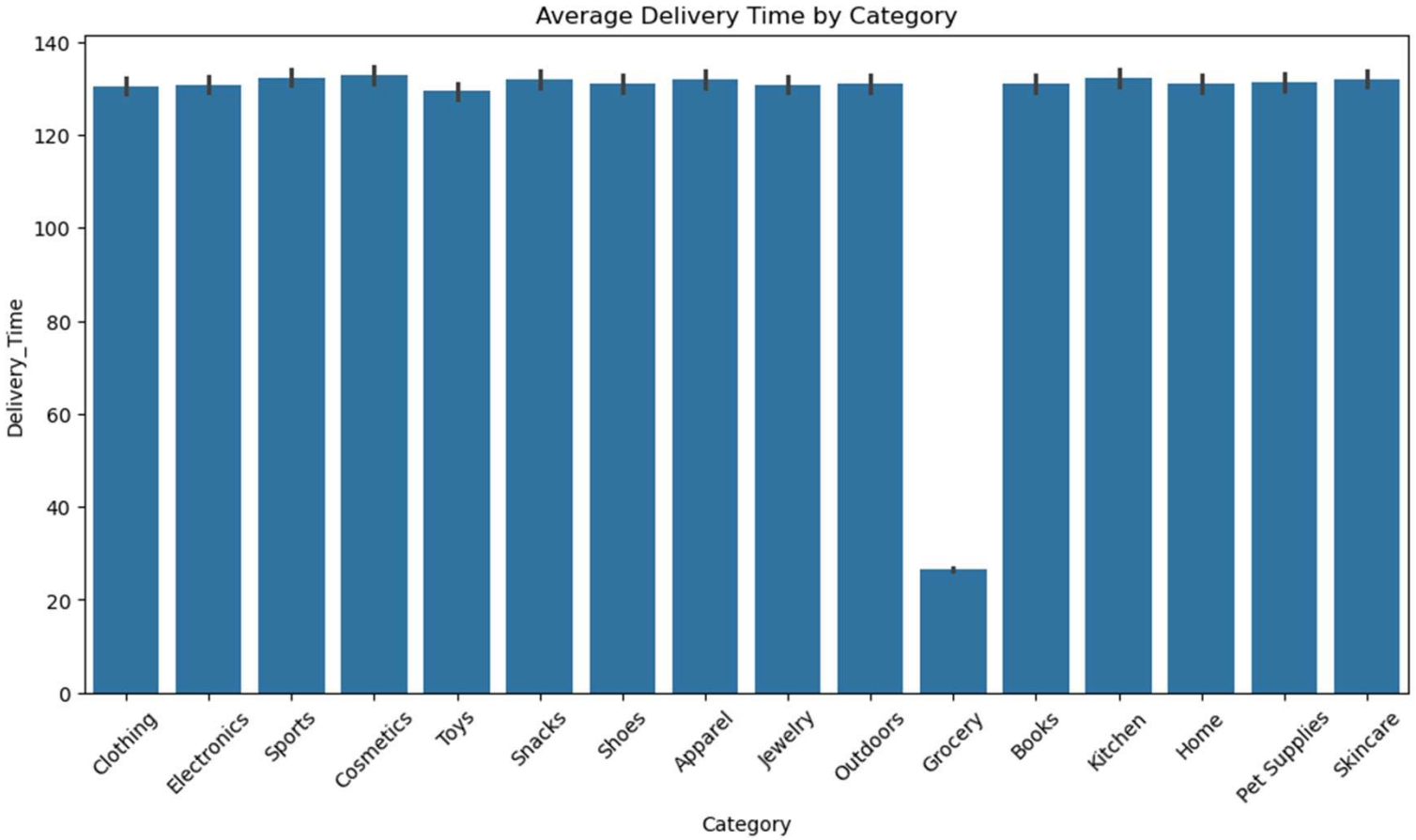
Thank You

Appendix

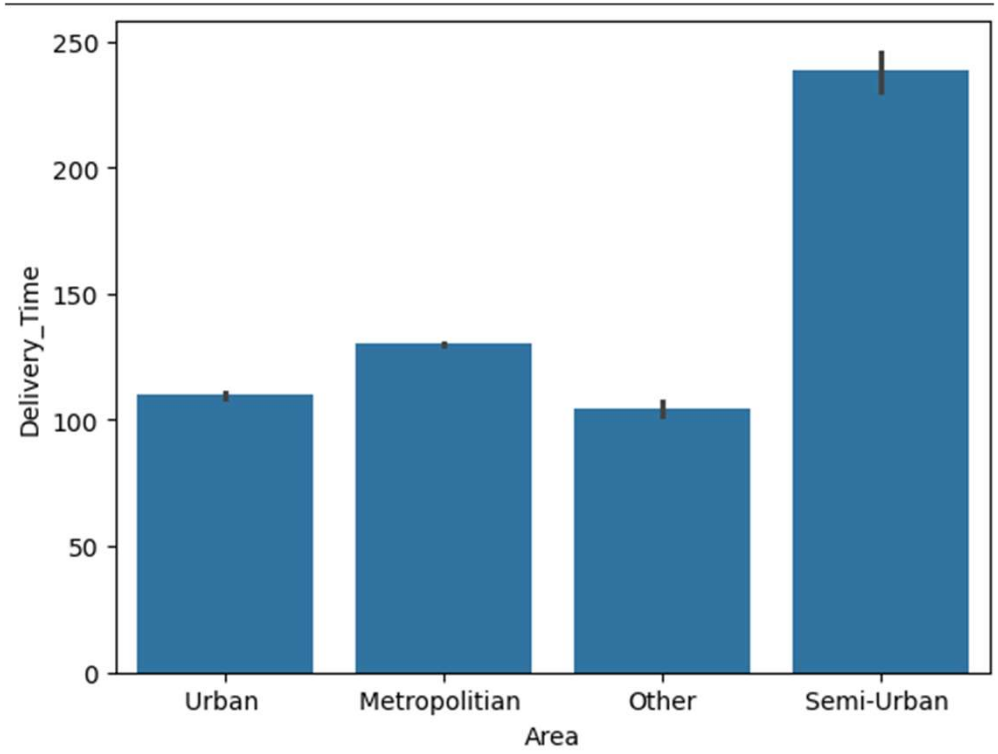
Exploratory Data Analysis – Comparative



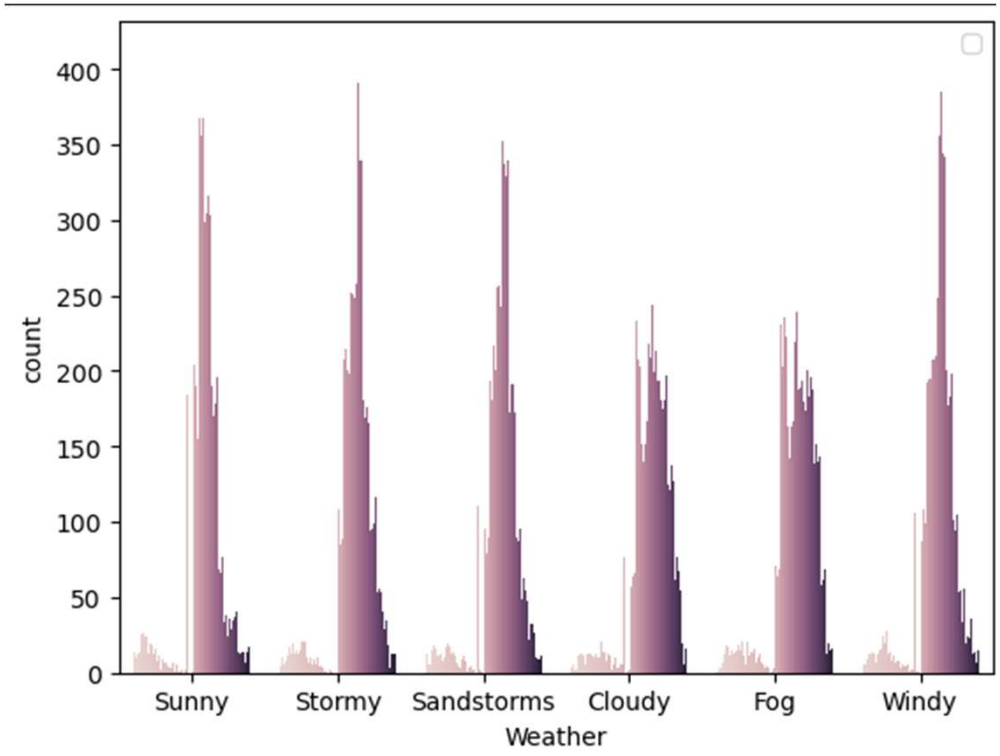
Exploratory Data Analysis – Product Category



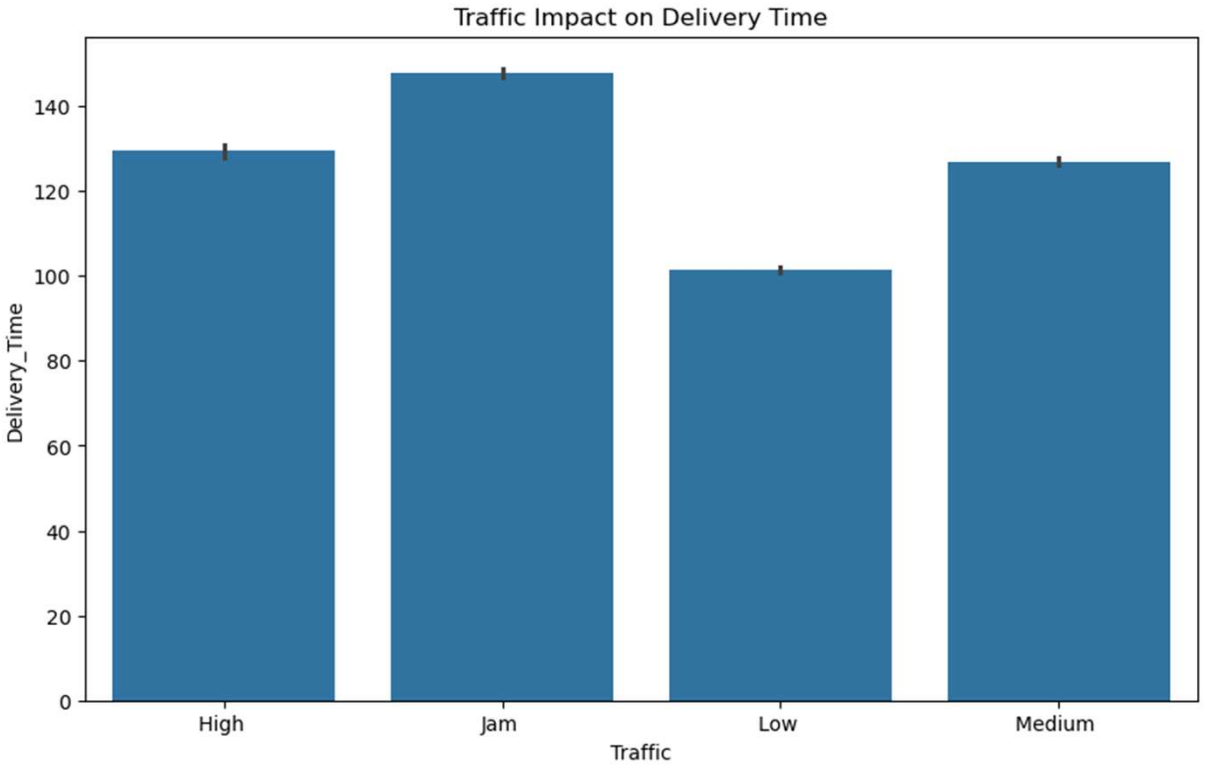
Exploratory Data Analysis - Area

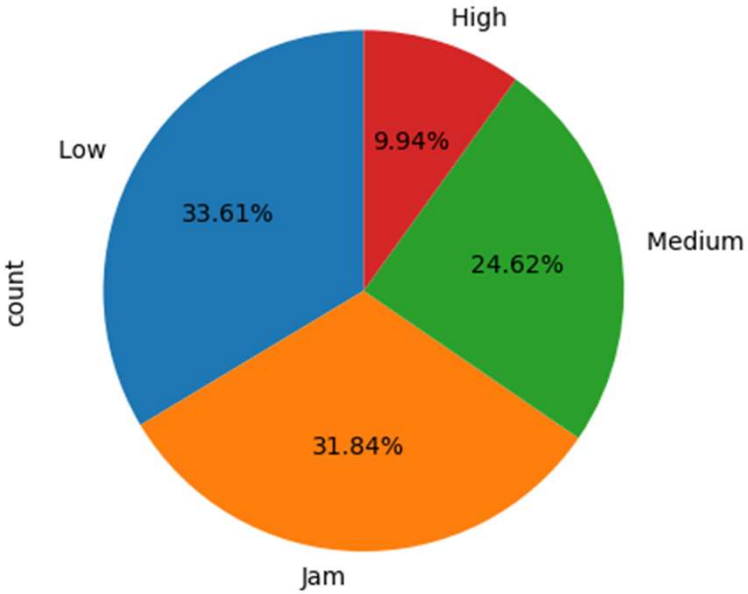


Exploratory Data Analysis - Weather



Exploratory Data Analysis – Traffic Impact





References

- https://www.kaggle.com/code/kyuhwankim/amazon-delivery-dataset
- https://pypi.org/project/haversine/
- https://datasciencedojo.com/blog/heatmaps/
- https://pieriantraining.com/seaborn-pie-chart-a-tutorial-for-data-visualization/
- https://www.datacamp.com