REPORT – DELIVERY DATA ANALYSIS – FINAL PROJECT

INTRODUCTION:

(a) What is the name of your project?
Delivery Data Insights: Unlocking Efficiency and Customer Experience

(b) Please write it as a research question and provide a short synopsis/description.
Research Question: How do delivery times, ratings, and location-based patterns influence operational efficiency and customer satisfaction in delivery services?
Synopsis: This project explores delivery service data to uncover insights into customer ratings, delivery times, and location-based clusters. By analyzing these relationships, the project aims to identify factors impacting delivery efficiency and customer satisfaction. The study uses data scraping, cleaning, and advanced analytical techniques to provide actionable recommendations for improving delivery operations.

(c) What is the research question(s) that you are trying to answer?

1. How do delivery times affect customer ratings, and what trends can be identified?
2. What location-based clusters can be identified to optimize delivery routing?
3. What operational factors (e.g., traffic conditions, weather) correlate with delivery performance metrics?
4. How can actionable insights from the analysis be applied to enhance customer satisfaction and delivery efficiency?

DATA COLLECTION:

(a) Specify exactly where the data that you collected is coming from.
The data was collected from an online repository providing delivery-related information. It included three tables: Delivery Person Table, Order Details Table, and Location Details Table.

(b) Describe the approach that you used for data collection.
I performed web scraping to extract the data from the repository. The scraping process involved collecting, cleaning, and merging three tables into a single comprehensive dataset for analysis.

(c) How many different data sources did you use?
Three data sources were used, corresponding to the three tables: Delivery Person Table, Order Details Table, and Location Details Table.

(d) How much data did you collect in total? How many samples?
The dataset consisted of 45,594 rows and 20 columns, combining all three tables after merging.

(e) Describe what has been changed from your original plan as well as the challenges that you encountered and resolved.

- Nidhi Choudhary

REPORT – DELIVERY DATA ANALYSIS – FINAL PROJECT

Initially, I planned to collect data directly as a single table. However, the repository provided data in three separate tables, requiring additional effort to merge and clean them. Challenges included handling mismatched or missing IDs across tables and normalizing data formats, such as ensuring consistent coordinate systems in the Location Details Table. These were resolved through preprocessing steps like data type conversions, handling missing values, and verifying key relationships during table merging.

DATA CLEANING:

(a) Describe how and why you had to change the data before using it.
The data required cleaning to ensure consistency, accuracy, and usability for analysis.

- **Handling Null Values**: Missing values were replaced with appropriate defaults or derived from calculations to maintain data completeness. For instance, NaN values in critical fields like 'Delivery_person_Age' and 'Delivery_person_Ratings' were replaced with their respective averages. This step ensures that downstream processes are not interrupted by missing data.
- **Cleaning Coordinates**: Latitude and longitude values that were zero or near-zero were replaced with random values sampled within a realistic range for their respective fields. This correction removed potential inaccuracies in location-based analyses. The zeros values were replaced with random values in a range to reduce the effect of outliers.
- **Standardizing Data Types**: Columns with inconsistent or incorrect data types were converted to their appropriate types, like converting 'Time_Ordered' to a datetime format, converting latitude and longitude from string to numeric values etc. This was critical for time-based operations and analysis.

(b) What kind of format was the data in when you collected it?
The data was in tabular format, likely as a CSV file, with numeric, textual, and date fields. It contained missing values, non-numeric entries, and incorrectly formatted coordinates.

(c) What kind of format did you use for the processed data in the end?
The cleaned data was retained in tabular format but with consistent data types, accurate values for all records, and realistic ranges for key fields. It was suitable for direct use in analysis and modeling.

(d) Did this process affect the data that you collected in any way, shape, or form?
Yes, the cleaning process involved altering invalid or missing entries to maintain data quality. While this introduced some synthetic values (e.g., in latitude and longitude fields), the modifications were within logical and reasonable bounds to ensure analytical accuracy without compromising the dataset's integrity.

- Nidhi Choudhary

REPORT – DELIVERY DATA ANALYSIS – FINAL PROJECT

DATA VISUALIZATION

(a) Describe the type of data visualizations that you used.

1. Bar Plots: Illustrated delivery counts and average customer ratings across time intervals and parts of the day.
2. Line Plot: Showed trends in average delivery times across hours of the day.
3. Heatmap: Highlighted variations in delivery times by day of the week and part of the day.
4. Pie Chart: Depicted the contribution of vehicle types to multiple deliveries.
5. Histogram and Boxplot: Analyzed delivery time distribution and customer ratings by location.

(b) Explain why these visualizations make sense.

- Operational Efficiency: Bar plots and heatmaps pinpoint time intervals and parts of the day with higher delivery delays, helping optimize resource allocation.
- Customer Experience: The line plot and histograms identify patterns in delivery times and customer satisfaction, aiding in service improvement.
- Strategic Insights: The pie chart highlights vehicle efficiency, guiding decisions on fleet composition.
- Outlier Management: Boxplots reveal delivery anomalies, enabling targeted interventions to reduce delays.

These visualizations provide actionable insights to improve delivery performance, customer satisfaction, and operational strategy.

DATA ANALYSIS:

(a) Which analysis technique(s) did you use?

1. Linear and Polynomial Regression:
    - Analyzed the relationship between delivery time and customer ratings.
    - Found that longer delivery times correlate with slightly lower ratings, but the relationship is weak as shown by low R^2 values. Polynomial regression (degree 2) captured subtle variations but did not significantly improve the fit.
2. Clustering (K-Means):
    - Grouped delivery locations into clusters based on latitude and longitude.

- Nidhi Choudhary

REPORT – DELIVERY DATA ANALYSIS – FINAL PROJECT

- Used the Elbow Method to determine the optimal number of clusters (k=4).
- Identified delivery hotspots and patterns, which can help optimize delivery routing and resource allocation.

3. Elbow Curve:
    - Evaluated the optimal number of clusters by plotting inertia (within-cluster sum of squares) across different k values.
4. Visualization:
    - Scatter plots and cluster visualizations for delivery locations.
    - Regression plots for delivery time vs. ratings.

(b) What are your findings?

- Regression Analysis:
    - Delivery time impacts ratings minimally, with a slight downward trend in ratings for longer delivery times.
    - Delivery times between 10–30 minutes are optimal for maintaining higher ratings.
- Clustering Insights:
    - Clear clustering of delivery locations reveals natural groupings that can guide logistics planning.
    - Identified cluster centroids provide actionable data for streamlining delivery routes.

CONCLUSION:

(a) Observations and Conclusion:
Delivery times have a minimal but noticeable impact on customer ratings, with longer times slightly reducing satisfaction. Location clustering revealed delivery hotspots, highlighting areas for operational optimization. Key metrics like delivery times and ratings suggest opportunities to enhance efficiency and maintain customer satisfaction.

(b) Impact of Findings:
The analysis provides actionable insights to improve delivery routing, optimize resource allocation, and enhance customer experience. These findings can guide strategies to minimize delivery delays and improve service ratings, contributing to better overall operational performance.

- Nidhi Choudhary

REPORT – DELIVERY DATA ANALYSIS – FINAL PROJECT

FUTURE WORK

(a) Given more time, what would you do to further improve your project?
I would develop and deploy a real-time web application that provides business stakeholders with a dynamic dashboard to monitor delivery performance metrics such as average delivery times, customer ratings, and operational efficiency. This app would use live data feeds and predictive analytics to offer actionable insights, allowing businesses to respond proactively to delays or inefficiencies. Additionally, I would integrate better machine learning models to forecast peak demand times and recommend optimized delivery routes. Exploring the impact of factors like weather and traffic on delivery efficiency could also provide deeper insights.

(b) Would you use the same data sources next time? Why yes or why not?
Yes, I would use the same data sources as they provide a comprehensive view of delivery operations. However, I would enhance them by incorporating external data, such as real-time traffic conditions and weather updates, to enrich the analysis. Combining internal and external data would allow for more precise decision-making and improved predictions, ultimately driving better business outcomes. To make it more engaging, I'd consider adding gamification elements, like performance leaderboards for delivery personnel, to motivate higher efficiency and satisfaction. I would like to include a dataset with more number of columns which will help us conclude more co-related features for determining the models.

- Nidhi Choudhary