Project Management Summary: Analyzing New York Citi Bike Dataset – Raul Brand

Project Overview:

The project aimed to analyze the New York Citi Bike dataset sourced from Kaggle, consisting of approximately 50,000 rows and 18 columns. The dataset provides detailed information about Citi Bike trips, including trip ID, start/end station details, trip duration, user demographics, and more. The analysis sought to gain insights into Citi Bike usage patterns, user behaviors, and potential areas for optimization.

Data Cleaning and Validation:

The data underwent thorough cleaning and validation procedures in Jupyter Notebook. Missing values in the birth_year column were handled by replacing them with NaN for further investigation. Unrealistic birth years were identified and replaced with NaN. Missing birth years were imputed using mean and median values. Final data quality checks ensured successful cleaning and prepared the dataset for analysis.

Limitations and Ethics Considerations:

Various limitations and ethical considerations were addressed, including sampling bias, privacy concerns, data accuracy, temporal bias, ethical use of data, and data ownership. It was essential to ensure fair and ethical treatment of the data, respect user privacy, and acknowledge data ownership and attribution.

Questions for Further Analysis:

A set of questions for further analysis was outlined, covering various aspects such as temporal patterns, demographic insights, station usage, and trip durations. These questions aimed to delve deeper into the dataset, uncover trends, and inform future decisions related to bike-sharing infrastructure and services.

Exploratory Data Analysis:

Exploratory data analysis involved understanding variable definitions, data structure, descriptive statistics, and visualizations. Insights were derived regarding trip duration distributions, start/end station patterns, user demographics, and temporal trends, laying the foundation for further analysis.

Geospatial Analysis:

Geospatial analysis provided insights into the distribution of Citi Bike start stations across New York City. The analysis highlighted station clusters, trip duration variations, and potential factors influencing usage patterns, contributing to a deeper understanding of spatial dynamics.

Clustering Analysis:

Clustering analysis segmented Citi Bike users into distinct groups based on behavior patterns. These clusters were instrumental in targeting marketing strategies, optimizing services, and guiding product development efforts tailored to different user segments.

Linear Regression Analysis:

Linear regression analysis examined the relationship between trip durations and distances between start and end stations. Insights from the analysis informed predictions and reflections on data biases, emphasizing the importance of data quality and representation.

Tableau Analysis:

Tableau analysis provided visualizations comparing subscriber and non-subscriber behaviors, identifying usage patterns, peak hours, and trip duration differences. Insights gleaned from the analysis guided recommendations for service allocation, marketing strategies, promotions, and user surveys.

Conclusion:

The project yielded valuable insights into Citi Bike usage dynamics, user behaviors, and potential areas for improvement. By leveraging data-driven approaches and addressing ethical considerations, the analysis contributed to informed decision-making and optimization of bike-sharing services in New York City.