Data Source Document – Raul Brand

Project Summary: The New York Citi Bike dataset was downloaded from Kaggle and contains information about Citi Bike trips. It consists of around 50,000 rows and 18 columns, including details such as trip ID, bike ID, weekday, start hour, start time, start station ID, start station name, start station latitude, start station longitude, end time, end station ID, end station name, end station latitude, end station longitude, trip duration, subscriber status, birth year, and gender. This dataset provides us with a wealth of specific information, offering a detailed look into how people interact with the Citi Bike system in New York City. By examining this data, we can gain insights into various aspects of Citi Bike usage, such as when and where trips occur, how long they last, and who is using the bikes. This level of detail allows us to understand patterns in user behavior, identify popular biking routes and times, and even explore factors like age and gender that may influence how people use Citi Bike.

With such comprehensive coverage, this dataset serves as a valuable tool for analyzing Citi Bike's impact on transportation in New York City. By uncovering these insights, we can potentially identify areas for improvement and optimization within the bike-sharing system, ultimately enhancing the overall experience for users and contributing to the success of Citi Bike in the city.

Explanation why this data was chosen: I've chosen this dataset because I have a personal interest in Citi Bike, as I am a user of the service and live in New York. Citi Bike is a prominent bike-sharing system in New York City, providing convenient and eco-friendly transportation options for residents and visitors alike. By delving into this dataset, I would gain deeper insights into various aspects of Citi Bike usage, such as popular biking routes, peak hours, subscriber demographics, and trip durations. Analyzing this data can not only satisfy my curiosity about the service but also potentially uncover valuable patterns and trends that could inform future decisions and optimizations for Citi Bike operations and services. Overall, my personal connection to Citi Bike and my desire to explore its data further serve as strong motivations for selecting this dataset for analysis and next project.

Data Cleaning Summary: (Technical procedures in JUPITER NOTEBOOK)

Below the steps (cleaning) I performed on the New York Citi Bike dataset:

- **Handling Missing Values:** We identified missing values in the birth_year column and decided to replace them with NaN for further investigation.
- **Data Validation:** Upon review, we found unrealistic birth years (e.g., 1899) that likely represented incorrect or missing data. We replaced these unrealistic values with NaN to be addressed later.
- **Imputation of Missing Values**: We imputed missing birth years with the mean, followed by the median, to ensure that all missing or incorrect birth years were adequately handled.
- **Final Data Quality Check:** After cleaning and imputing missing values, we verified the summary statistics of the birth_year column to ensure that the cleaning process was successful and that the data was ready for further analysis.

Consider limitations and ethics: When working with the Citi Bike data, it's essential to consider its limitations and ethical considerations, particularly concerning its source and collection methods. Below are some potential limitations and ethical considerations for the dataset of New York Citi Bike:

- **Sampling Bias:** The dataset may suffer from sampling bias as it only represents individuals who have used the Citi Bike service. This may not be representative of the entire population of New York City, leading to biased insights and conclusions.
- **Privacy Concerns:** The dataset likely contains sensitive information such as users' birth years and gender. Care must be taken to anonymize or aggregate this data to protect users' privacy.
- **Data Accuracy:** There may be inaccuracies or errors in the data, including missing values, incorrect entries, or inconsistencies. It's important to thoroughly clean and validate the data to ensure the reliability of any conclusions drawn from it.
- **Temporal Bias:** The dataset covers a specific time period (2013), which may not be representative of other time periods. Trends or patterns observed in the data may be influenced by external factors such as seasonality, weather conditions, or events happening in New York City during that time.
- **Ethical Use of Data:** Any analysis or interpretation of the data should be conducted ethically, considering the potential impact on individuals, communities, or society as a whole.
- **Data Ownership and Attribution**: (Kaggel) It's important to respect the ownership of the data and provide proper attribution to the source when using or sharing the dataset. Researchers should adhere to any terms of use or licensing agreements associated with the data.
- **Informed Consent:** If the dataset includes personally identifiable information or data collected from human subjects, researchers must ensure that proper informed consent was obtained for its collection and use. This includes informing individuals about how their data will be used and obtaining their explicit consent.

Questions for further analysis:

Adjoining Questions:

- Are there specific days of the week or times of the day that show increased bike usage during September 2013?

Elevating Questions

- What insights can we derive from the September 2013 bike trip data to inform future bike-sharing infrastructure planning?
- How does bike-sharing in September 2013 contribute to the city's goals of promoting sustainable transportation and reducing carbon emissions?
- Can we identify any notable events or trends during September 2013 that influenced bike-sharing usage?

Clarifying Questions:

- What were the primary factors influencing the duration of bike trips in September 2013?
- How did the usage of bike-sharing services vary across different neighborhoods and boroughs of New York City during September 2013?
- What is the distribution of user ages in the September 2013 bike trip data?
- Are there any age groups that show a higher propensity for using bike-sharing services during September 2013?
- What is the distribution of user ages in the September 2013 bike trip data?
- Are there any age groups that show a higher propensity for using bike-sharing services during September 2013?

Funneling Questions:

- What were the most common start and end points for bike trips during September 2013?
- How did subscriber status impact the frequency and duration of bike trips in September 2013?

Privacy and Ethics Questions:

- How can we ensure the anonymity and privacy of individuals represented in the September 2013 dataset?
- What measures were in place to protect user data during the collection and storage of the September 2013 bike trip data?
- Are there any potential biases in the September 2013 dataset that need to be addressed to ensure fair and equitable analysis?