Jordan Henghold
Task 6.1
Data Immersion

Sourcing Open Data

Data Source

Data Source Summary:

I found this data source on the Kaggle website, which was a reliable and recommended source through Career Foundry Lessons. Kaggle's retrieval of the data set was actually from the Citi Bike Website, which is a direct link. On the website there are a consistent chronological data values spanning across demographics of information regarding every single archive they have in reference to Citi Bike's monthly trip data. The data is suitable to be alleged as reliable for a source, as it is from Kaggle—recommended by this school. I do not see a ton of variation in the information from the first look but managed to change several components of the data to make it more useful. Additionally, the data is relatively easy to digest because it is laid out in such a format that we are getting individual usage data alongside administrative access to data. This means I can see how long rides took, timing of rides, locations, and many bits of information regarding clients—per a direct link to their application information. Misuse of the application (inaccurate age or even a different person) could potentially lead to information inaccuracies that would be quite hard to test against.

Why I Chose this Dataset:

I chose this dataset because my fiancé has recently started driving for Uber—which is a taxiing service to drive patrons around on short trips. He has told me that most of the drives he completes are short, around-town drop offs, which is a similar clientele to that of Citi Bike. I wanted to complete a fully analytical investigation on Citi Bike, and after this Achievement is completed, I will go on to compare if the two have similar patterns. In order to fully investigate Citi Bike, I have a desire to compare the monthly and even weekly patterns so that after my fiancé has been working for Uber for a month, I can begin running data checks. If I am able to figure a way out for my fiancé to maximize on making money at his third job to bring into my personal household, I am more than happy to do so by conducting a throughout initial analysis of Citi Bike.

Data Profile

Clean & Understand Data (6,7):

*Consistency Checks are in Jupyter Notebook for checking accuracy ☺

Columns Dropped	**Columns renamed**	**Columns Type Changed**	**Reason**
Unnamed:0, Trip_id, Bike_id			Unneccessary
Start/end time		Changed to start and end of ride & data type (datetime64)	Data Type Change
	weekday -> day_of_week		Understandability
		Bike_id -> string	String/object is used bc values unnecessary for descriptive analysis

Variable	Description	Time Variant/ Invariant	Structured/ Unstructured	Quantitative/ Qualitative	Nom/ Ord/ Disc/ Cont.
Trip_id	Unique identifier	Invariant	Structured	Qualitative	Nominal
Bike_id	Unique identifier	Invariant	Structured	Qualitative	Nom
Weekday	Ride DOW	Invariant	Structured	Quantitative	Discrete
Start_hour	Hour of ride	Invariant	Structured	Quantitative	Disc
Start_time	Time of ride	Invariant	Structured	Quantitative	Disc
Start_station_id	Station id bike left from	Invariant	Structured	Qualitative	Nom
Start_station_name	Station name: begins ride	Invariant	Structured	Qualitative	Nom
Start_station_latitude	Latitude of start station	Invariant	Structured	Quantitative	Continuous
Start_station_longitude	Longitude od start station	Invariant	Structured	Quantitative	Cont
End_time	Time ride ends	Invariant	Structured	Quantitative	Disc
End_station_id	Station id bike ends at	Invariant	Structured	Qualitative	Nom

End_station_name	Station name: ends ride	Invariant	Structured	Qualitative	Nom
End_station_latitude	Lat of end station	Invariant	Structured	Quantitative	Cont
End_station_longitude	Long. Of end station	Invariant	Structured	Quantitative	Cont
Trip_duration	Trip length (seconds)	Invariant	Structured	Quantitative	Disc
Subscriber	Rider has subscription?	Variant	Structured	Qualitative	Ordinal
Birth_year	Rider year born	Invariant	Structured	Quantitative	Ord
Gender	Rider gender	Invariant	Structured	Quantitative	Disc

Limitations & Ethics (8):

As far as limitations are considered, this data set possesses information regarding rides—and the surrounding demographics—which took place from the original launching of May 27, 2013- and the set carries on until the start of October of 2013. This New York City launch gives us information regarding riders' information and bikes according to their unique identifiers, alongside information regarding locations and times of rides to and from specific stations—per their unique identifiers, as well. This set of data is inclusive of many components, however, there is still potential for limitations of accurate customer ids ad ages, as mentioned earlier. Additionally, there were what seems to be a set number of integers downloaded to the set because it is exactly fifty- thousand rides. The data set would tremendously benefit from holding a column for identifying the rider more uniquely to characterize data in a more accurate manner. If we had access to this, we would be able to further assess for rider weekly and monthly riding frequencies and we could personalize the riding experience and offer incentives. Since we cannot see the subscription status in every case this ill be a limitation, as well. For all we currently know, there are no rider identifiers, so we are missing out on perfectly valuable information and just about anyone could use another person's application.

Exploratory Questions

- Subscriber trends and they compare to non-subscribers?
- Most frequent trip days?
- Most frequent trip hours?
- Most popular stations to start trips?
- Ending trips most popular stations?
- Frequent age groups using the application?
- Do the popularities/ frequencies of these qualifying questions fluctuate?
  -In which directions, and at which times?

  Resources:
  https://www.kaggle.com
  https://citibikenyc.com