Exercise 6

Exercise 6.1

Data source

□ **Source** – Someone called Ryan Cummings uploaded the data set onto Kaggle.com. This data set was one of the options in our project brief and other than the name, there is no other information. I assume it can be trusted, because there would be no reason to give false information, because this is a non-profit website. But obviously we can never be sure.

□ **Data collection** – This can be classified as administrative data. The bike company is interested in their clients, the routes taken and time spent on the bikes.

□ **Data set** – 18 variables and 50,000 rows. The main variables are weekday, station start and end, trip duration and gender.

□ **Limitations** – There are some limitations. Birth_year has many NaN values. We are also reliant on people entering the right information upon signing up.

□ I've chosen this data set completely random. To avoid bias, I closed my eyes and moved my mouse around over the list and this is where it stopped. In future when working, I will not get to choose my data sets.

□ **Ethics** – This data is open source and published on Citibike's website. On their website they also laid out their Data licence agreement, there are no PII information on the data set, so I see no ethical issues with the data.

- Relevance – This data set meets the requirements of this task, simply because it was one of the sets I could choose in our project brief. It also contains all the information and variables I need to conduct my analysis.

Data Profile

Variable	**Description**	**Time Variant / Invariant**	**Structured / Unstructured**	**Quantitative / Qualitative**	**Nominal / Ordinal / Discrete / Continues**
trip_id	Unique identifier for trip	Invariant	Structured	Qualitative	Nominal
bike_id	Unique identifier for bike	Invariant	Structured	Qualitative	Nominal
weekday	Day of week bike was used	Invariant	Structured	Quantitative	Discrete
start_hour	Hour of day ride started	Invariant	Structured	Quantitative	Discrete
start_time	Time ride started	Invariant	Structured	Quantitative	Discrete
start_station_id	Unique identifier of start station	Invariant	Structured	Qualitative	Nominal
start_station_name	Name of station where ride started	Invariant	Structured	Qualitative	Nominal
start_station_latitude	Latitude of starting station	Invariant	Structured	Quantitative	Continues
start_station_longitude	Longitude of starting station	Invariant	Structured	Quantitative	Continues
end_time	Time ride ended	Invariant	Structured	Quantitative	Discrete
end_station_id	Station ride ended	Invariant	Structured	Qualitative	Nominal
end_station_name	Name of station where ride ended	Invariant	Structured	Qualitative	Nominal
end_station_latitude	Latitude of station where ride ended	Invariant	Structured	Quantitative	Continues
end_station_longitude	Longitude of station where ride ended	Invariant	Structured	Quantitative	Continues
trip_duration	Duration of trip in seconds	Invariant	Structured	Quantitative	Discrete
subscriber	Whether rider is subscriber or not	Variant	Structured	Qualitative	Binary
birth_year	Birth year of rider	Invariant	Structured	Qualitative	Ordinal
gender	Gender of rider	Invariant	Structured	Qualitative	Binary

Data wrangling

Columns dropped	Columns renamed	Columns data type changed	Comment/Reason
trip_id			Not needed
		gender	Changed from int to string. Also re-categorized
		start_time	Changed from string to datetime
		end_time	Changed from string to datetime
		start_station_id	Changed from int to string
		end_station_id	Changed from int to string
		birth_year	Changed from float to int

Consistency checks

Missing values	Action	Duplicates	Other
birth_year	had 6979 NaN values. Replaced them with column mean 1976		
		No duplicates found	
			23 birth years were inaccurate. I dropped those rows.

Questions to explore

1. What stations are busiest/least busy?
2. What day of the week is busiest/least busy?
3. What time of day is busiest/least busy?
4. How many riders are subscribers vs non-subscribers?
5. Usage of bikes, subscribers vs non-subscribers.
6. Users, male vs female
7. Gender vs subscriber/non subscriber

Answers to my questions:

1. Not yet
2. All days are equally busy, just Saturdays are slightly less busy.

3. 07:00 and 16:00 are the busiest start hours of the day. I assume that’s when people are going to work and going home from work. From 23:00 till around 05:00 – 06:00 it’s the least busy. That’s when people are sleeping.

4. Not yet

5. Not yet

6. Not yet

7. More subscribers are male than female.