Bike-Share Case Study

This report provides the results and step-step explanation of the data analysis performed for a bike sharing case-study. The data belongs to a bike-sharing company that has two kinds of users: annual members and casual riders. The goal of the case-study was to identify how annual members and casual riders use the bikes differently in order to help the stake-holders decide whether or not to target converting casual riders into annual members in the next marketing campaign. The data about the bike-rides used in this case-study was between January-November 2023, each month was stored in a csv file, and was downloaded from https://divvy-tripdata.s3.amazonaws.com/index.html.

Data cleaning:

The code that was used to perform the data cleaning can be found in the Jupyter Notebook cleaning.ipynb. Here are the main functions and what they do.

- *read_data*:
  Here the csv files are read and stored into a dictionary called "data". Each element in the dictionary has a key (the name of the month) and a value (the panada dataframe that holds the csv entries). This way the data for each corresponding month can be easily accessed by using the month as the key (e.g. data["February"] retrieves the dataframe that holds the entries from February). Below we can see the first 5 entries of bike rides from February:

	ride_id	rideable_type	started_at	ended_at	start_station_name	start_station_id	end_station_name	end_station_id	start_lat	start_lng	end_lat	end_lng	member_casual
0	CBCD0D7777F0E45F	classic_bike	2023-02-14 11:59:42	2023-02-14 12:13:38	Southport Ave & Clybourn Ave	TA1309000030	Clark St & Schiller St	TA1309000024	41.920771	-87.663712	41.907993	-87.631501	casual
1	F3EC5FCE5FF39DE9	electric_bike	2023-02-15 13:53:48	2023-02-15 13:59:08	Clarendon Ave & Gordon Ter	13379	Sheridan Rd & Lawrence Ave	TA1309000041	41.957879	-87.649584	41.969517	-87.654691	casual
2	E54C1F27FA9354FF	classic_bike	2023-02-19 11:10:57	2023-02-19 11:35:01	Southport Ave & Clybourn Ave	TA1309000030	Aberdeen St & Monroe St	13156	41.920771	-87.663712	41.880419	-87.655519	member
3	3D561E04F739CC45	electric_bike	2023-02-26 16:12:05	2023-02-26 16:39:55	Southport Ave & Clybourn Ave	TA1309000030	Franklin St & Adams St (Temp)	TA1309000008	41.920873	-87.663733	41.879434	-87.635504	member
4	0CB4B4D53B2DBE05	electric_bike	2023-02-20 11:55:23	2023-02-20 12:05:48	Prairie Ave & Garfield Blvd	TA1307000160	Cottage Grove Ave & 63rd St	KA1503000054	41.794827	-87.618795	41.780531	-87.605970	member

- *check_entries*:
  This method performs a few preliminary calculations. It finds the number of entries

per file as well as the number of columns. From these it calculates the total number of bike rides in the data set. There is also an option within the method to remove duplicates. Therefore, the method is first called with the remove duplicates option deactivated, in order to get a preliminary feel of the dataset, how big it is, how the entires varies across the months. And then the method is called again with the remove duplicates option activated. The results are then written to output files and shown below.

Original_BikeRides

Month	No Of Entries	No Of Cols
January	190301	13
February	190445	13
March	258678	13
April	426590	13
May	604827	13
June	719618	13
July	767650	13
August	771693	13
September	666371	13
October	537113	13
November	362518	13
Total:	5495804	
Average:	499618	

BikeRides_without_Duplicates

Month	No Of Entries	No Of Cols
January	190301	13
February	190445	13
March	258678	13
April	426590	13
May	604827	13
June	719618	13
July	767650	13
August	771693	13
September	666371	13
October	537113	13
November	362518	13
Total:	5495804	
Average:	499618	

We can see that all the files have the same number of columns. So that is a good preliminary check on the consistency of the data across the months. In total the dataset contains 5.5 Million entries, with an average of 500,000 entries per month. The number of entries before and after removing duplicates are identical, so the original dataset did not have any duplicates.