Bike-Share Case Study

This report provides the results and step-by-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 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 on which the analysis was carried out is from January-December 2023 and was downloaded from https://divvy-tripdata.s3.amazonaws.com/index.html.

The library Pandas from Python was used to perform the analysis, and Matplotlib was used to plot the results. The code can be found in the Jupyter Notebook bike_share_analysis.ipynb.
Add hyperlinks between method names and cell in notebook

Cleaning:

- *read_data*:
  The original data was stored such that each month was in a separate .csv file. So in this method the data from each month is read and stored into a DataFrame (DF), and then the 12 DFs are concatenated into one multi-index DF. Using a multi-index DF has several advantages. First, the distinction between the different seasons/months can still be maintained (data from different months are not fused together into one large DF). Second, the multi-index DF facilitates finding and aggregating values across different months when needed. In Figure 1 we can see the first and last 5 entries of bike rides from the multi-index DF:

		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
Month	row_id													
January	0	F96D5A74A3E41399	electric_bike	2023-01-21 20:05:42	2023-01-21 20:16:33	Lincoln Ave & Fullerton Ave	TA1309000058	Hampden Ct & Diversey Ave	202480.0	41.924074	-87.646278	41.930000	-87.640000	member
	1	13CB7EB698CEDB88	classic_bike	2023-01-10 15:37:36	2023-01-10 15:46:05	Kimbark Ave & 53rd St	TA1309000037	Greenwood Ave & 47th St	TA1308000002	41.799568	-87.594747	41.809835	-87.599383	member
	2	BD88A2E670661CE5	electric_bike	2023-01-02 07:51:57	2023-01-02 08:05:11	Western Ave & Lunt Ave	RP-005	Valli Produce - Evanston Plaza	599	42.008571	-87.690483	42.039742	-87.699413	casual
	3	C90792D034FED968	classic_bike	2023-01-22 10:52:58	2023-01-22 11:01:44	Kimbark Ave & 53rd St	TA1309000037	Greenwood Ave & 47th St	TA1308000002	41.799568	-87.594747	41.809835	-87.599383	member
	4	3397017529188E8A	classic_bike	2023-01-12 13:58:01	2023-01-12 14:13:20	Kimbark Ave & 53rd St	TA1309000037	Greenwood Ave & 47th St	TA1308000002	41.799568	-87.594747	41.809835	-87.599383	member
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
December	224068	F74DF9549B504A6B	electric_bike	2023-12-07 13:15:24	2023-12-07 13:17:37	900 W Harrison St	13028	Racine Ave & Congress Pkwy	TA1306000025	41.874702	-87.649804	41.874640	-87.657030	casual
	224069	BCDA66E761CC1029	classic_bike	2023-12-08 18:42:21	2023-12-08 18:45:56	900 W Harrison St	13028	Racine Ave & Congress Pkwy	TA1306000025	41.874754	-87.649807	41.874640	-87.657030	casual
	224070	D2CF330F9C266683	classic_bike	2023-12-05 14:09:11	2023-12-05 14:13:01	900 W Harrison St	13028	Racine Ave & Congress Pkwy	TA1306000025	41.874754	-87.649807	41.874640	-87.657030	member
	224071	3829A0D1E00EE970	electric_bike	2023-12-02 21:36:07	2023-12-02 21:53:45	Damen Ave & Madison St	13134	Morgan St & Lake St*	chargingstx4	41.881396	-87.674984	41.885492	-87.652289	casual
	224072	A373F5B447AEA508	classic_bike	2023-12-11 13:07:46	2023-12-11 13:11:24	900 W Harrison St	13028	Racine Ave & Congress Pkwy	TA1306000025	41.874754	-87.649807	41.874640	-87.657030	member

5719877 rows × 13 columns

Figure 1: First and last 5 entries of bike rides from the original_data

In Figure 1 the multi-index of the DF is shown in the first two columns (month, row_id). Then looking at the entries themselves we can see that the data consists of 13 columns: 1) ride id, 2) type of bike, 3-4) date and time for the start and end of the ride, 5-12) the name, id, latitude and longitude of the start and end stations, and 13) whether the rider was a casual rider or a member.

- *count_entries*:
  The method *count_entries* is used to find the number of entries within each month, and the average per month. The result is shown in Figure 2 left. The dataset contains in total almost 5.7 Million entries, with an average of approximately 480,000 entries per month. From Decemeber to March the number of rides is relatively lower than the average, which is expected as these are cold months. This is confirmed by the peak highlighted in August. After retrieving this information for the original dataset, the duplicates are dropped, and the method is called again. The result of running the method after dropping the duplicates is shown in Figure 2 right. The number of entries before and after is identical, therefore the original dataset did not have any duplicates.

```
No of BikeRides Original:
Month          No of Entries
January        190301
February       190445
March          258678
April          426590
May            604827
June           719618
July           767650
August         771693
September      666371
October        537113
November       362518
December       224073
dtype: int64

Total in 2023:      5719877
Avg. per month:     476656
```

```
No of BikeRides without Duplicates:
Month          No of Entries
January        190301
February       190445
March          258678
April          426590
May            604827
June           719618
July           767650
August         771693
September      666371
October        537113
November       362518
December       224073
dtype: int64

Total in 2023:      5719877
Avg. per month:     476656
```

Figure 2: No of entries and average before and after removing duplicates

- *get_null_percentage*:
  This method calculates the percentage of null values for each column and month. The results are shown in Figure 3. As we can see the columns *start_station_name*, *start_station_id*, *end_station_name*, *end_station_id* in every month have 13-17% null values. The columns *end_lat* and *end_long* have less than 1% null values. In order to explore the dataset, as well as these null values a bit further, the number of unique (distinct) values for each column in original_data[“May”] is calculated and shown in Figure 4. The choice of the month of May is random and should not make a difference.
  - *ride_id_unique* = 604827: as expected, *ride_id* has as many unique values as the number of entries in the dataframe.
  - *rideable_type_unique* = 3: these 3 unique values are the types of bikes: [electric_bike, classic_bike, docked_bike].
  - *started(ended)_at_unique* = 503683, 505259: since these are datetimes (yy-mm-dd hh:mm:ss), one may have expected that they would have as many distinct values as the number of entries, since it seems unlikely for more than one rider to have rented a bike at the exact same time down to the second. However, the number of unique values in these columns is less than the number of entries by 17%. To ensure that these are not duplicate entries but with different

Month	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
January	0	0	0	0	14%	14%	14%	14%	0	0	< 1%	< 1%	0
February	0	0	0	0	13%	13%	14%	14%	0	0	< 1%	< 1%	0
March	0	0	0	0	13%	13%	14%	14%	0	0	< 1%	< 1%	0
April	0	0	0	0	14%	14%	16%	16%	0	0	< 1%	< 1%	0
May	0	0	0	0	14%	14%	15%	15%	0	0	< 1%	< 1%	0
June	0	0	0	0	16%	16%	17%	17%	0	0	< 1%	< 1%	0
July	0	0	0	0	16%	16%	16%	16%	0	0	< 1%	< 1%	0
August	0	0	0	0	15%	15%	16%	16%	0	0	< 1%	< 1%	0
September	0	0	0	0	15%	15%	16%	16%	0	0	< 1%	< 1%	0
October	0	0	0	0	15%	15%	16%	16%	0	0	< 1%	< 1%	0
November	0	0	0	0	15%	15%	15%	15%	0	0	< 1%	< 1%	0
December	0	0	0	0	15%	15%	16%	16%	0	0	< 1%	< 1%	0

Figure 3: Percentage of null values for each column across the months

```
Column Name          NUnique Values
ride_id               604827
rideable_type              3
started_at            503683
ended_at              505259
start_station_name      1287
start_station_id        1250
end_station_name        1254
end_station_id          1210
start_lat             188591
start_lng             185410
end_lat                 4759
end_lng                 4762
member_casual              2
```

Figure 4: No. of unique values for every column in original_data[“May”]

ride_id, one of these incidents has been retrieved and is shown in Figure 5. By looking at the values, it is clear that they are indeed different entries but with the exact same start time.

row_id	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
727	041763A703C94783	electric_bike	2023-05-28 14:59:58	2023-05-28 15:12:47	Kedzie Ave & Milwaukee Ave	13085	Kilpatrick Ave & Parker Ave	358	41.929673	-87.708045	41.930731	-87.744106	casual
2710	5AEC034DB275854E	electric_bike	2023-05-28 14:59:58	2023-05-28 15:26:56	Broadway & Belmont Ave	13277	NaN	NaN	41.940170	-87.645626	41.960000	-87.640000	casual
400665	A99D22D37DC92962	electric_bike	2023-05-28 14:59:58	2023-05-28 15:09:35	NaN	NaN	MTV Hubbard St	021320	41.880000	-87.660000	41.889779	-87.680341	member
557920	79A55702C0B6D246	docked_bike	2023-05-28 14:59:58	2023-05-28 16:24:12	Streeter Dr & Grand Ave	13022	Field Museum	13029	41.892278	-87.612043	41.865312	-87.617867	casual

Figure 5: Entries from original_data[“May”] that have the same *started_at*

– *start(end)_station_name(id)_unique* = 1287, 1250, 1254, 1210: since there is a limited number of stations, it is expected that these columns have a smaller number of unique values than the number of entries. However, one would have expected the number of unique station names and station ids to be the same, whereas the ids are less than the names by a small fraction. Which could either by accounted for by the null values or could mean that there are stations that have the different names but the same id.

– *start(end)_lat(long)_unique* = 188591, 185419, 4759, 4762: the start latitude and longitude numbers seem to be as expected, which is less than the total number of entries, but more than the number of stations (this is based on the assumption that the exact location where a bike is docked can vary within the station especially that the values are given to the 6^{th} decimal place). However, there is a large difference between the number of values in the start ($\sim$188000) and the numbers in the end ($\sim$4800) latitude and longitude. This difference cannot be accounted for by the null values in the end columns, since these were less than 1%. If these values are true, that would mean that users rode there bikes from many start points, but ended up in a much smaller set of points. Which cannot be the case since that would have been reflected in the number of end stations.

  – *member_casual_unique* = 2: the 2 unique types of riders are: [member, casual].

- *clean_data*:
  After exploring the dataset, we can see that the extractable information can be divided into information about the:

  1. rider (casual/member)
  2. bike (electric/classic/docked)
  3. ride (start-end: time, date, location)

  Since, the goal of the analysis is to find out whether to target converting casual riders into members or not, it seems that all the information is relevant to the analysis, with the exception of the ride location. The geographical location would have been important if for example the goal of the analysis was to find out whether more stations should be added and where to do so. Therefore, since the locations seem to be irrelevant, contain null values and discrepancies, these columns will be dropped. The column *ride_id* also does not provide any valuable information for the current analysis. Thus, the method *clean_data1* drops all columns related to geographical location and ride id, which leaves: *rideable_type*, *started_at*, *ended_at*, *member_casual*.

  Next, is data formatting. We have already looked at the columns *rideable_type* and *member_casual*, and ensured that they have the expected values. As for the columns *started_at* and *ended_at*, we need to make sure that the *ended_at* time always comes after *started_at* time. In order to compare the values, they are first converted in the method *clean_data1* from strings of characters to a numerical datetime format. Next, by comparing the values and filtering out the cases where the *ended_at* time is actually before the *started_at* time, we can see the results in Figure 6.

		rideable_type	started_at	ended_at	member_casual
Month	**row_id**				
February	189347	electric_bike	2023-02-04 13:08:08	2023-02-04 13:04:52	member
April	361967	electric_bike	2023-04-04 17:15:08	2023-04-04 17:15:05	member
	361983	classic_bike	2023-04-19 14:47:18	2023-04-19 14:47:14	member
	362063	electric_bike	2023-04-27 07:51:14	2023-04-27 07:51:09	casual
	363359	electric_bike	2023-04-06 23:09:31	2023-04-06 23:00:35	member
...	...	...	...	...	...
December	54495	electric_bike	2023-12-12 20:17:56	2023-12-12 20:17:55	casual
	64671	classic_bike	2023-12-11 19:31:28	2023-12-11 19:31:27	member
	117303	electric_bike	2023-12-07 16:43:01	2023-12-07 16:42:59	member
	133133	electric_bike	2023-12-05 18:04:30	2023-12-05 18:04:29	member
	220106	electric_bike	2023-12-06 16:07:40	2023-12-06 16:07:37	member

272 rows × 4 columns

Figure 6: Entries where the ended_at time is before the started_at time

  Looking at the entries in Figure 6, we can see that these are cases when the *ended_at* time is before the *started_at* time by just a few seconds. It can be that in these incidents the start and end time were switched due to some glitch, perhaps the bike rental time being only a few seconds (shorter than the server response time). In the entire dataset of approximately 5.5 Million entries, there is a total of 272 entries that have this issue. Since, the dataset is large, we can simply drop these entries.

Analysis:

- *prepare_data*:
  The method *prepare_data* adds two new columns: *ride_length*: the difference between the columns *ended_at* and *started_at* times, and *day_of_week*: extracted from the date in *started_at*. Then the columns *started_at* and *ended_at* are be dropped, since the new columns make them redundant.

- *mean_max_mode*:
  This method groups the data first by month and then calculates the mean and maximum ride length, as well as the mode of the day of the week. The method then aggregates the mean, mode and max values across the entire year. The results are shown in Figure 7. The mean ride length varies between 13 - 22 minutes across the months, whereas the maximum ride length is 68 days.

	ride_length		**day_of_week**
	mean	**max**	**mode**
Month			
January	00:13:00	23 day 08:03:44	Tuesday
February	00:13:31	13 day 02:25:46	Tuesday
March	00:13:04	11 day 16:08:04	Wednesday
April	00:17:12	12 day 18:35:29	Saturday
May	00:19:02	20 day 06:50:31	Tuesday
June	00:19:59	20 day 11:05:58	Friday
July	00:21:44	35 day 17:41:24	Saturday
August	00:22:25	68 day 09:29:04	Wednesday
September	00:17:52	01 day 01:07:46	Saturday
October	00:15:41	01 day 00:59:57	Tuesday
November	00:13:49	01 day 01:00:25	Thursday
December	00:13:24	01 day 00:59:57	Friday

Year 2023: 00:16:44 68 day 09:29:04 Tuesday

Figure 7: The mean and max ride length, and mode of day of the week

In order to understand how the data shown Figure 7 differs between members and casual riders, a pivot table is used to show the maximum and mean value per month by rider type. The table is shown in Figure 8. A quick look shows that the casual riders have longer mean and maximum ride lengths when compared to members. To see this more clearly the mean and maximum ride length are plotted in Figure 9.

	ride_length			
	mean		max	
member_casual	member	casual	member	casual
Month				
January	00:10:21	00:22:54	01 days 00:59:56	23 days 08:03:44
February	00:10:42	00:23:11	01 days 00:59:56	13 days 02:25:46
March	00:10:26	00:21:24	01 days 01:59:40	11 days 16:08:04
April	00:11:41	00:27:40	01 days 00:59:56	12 days 18:35:29
May	00:13:02	00:28:31	01 days 01:00:31	20 days 06:50:31
June	00:13:12	00:29:24	01 days 00:59:56	20 days 11:05:58
July	00:13:41	00:32:20	01 days 00:59:57	35 days 17:41:24
August	00:13:46	00:35:14	01 days 00:59:57	68 days 09:29:04
September	00:13:08	00:25:11	01 days 00:59:57	01 days 01:07:46
October	00:12:09	00:22:52	01 days 00:59:56	01 days 00:59:57
November	00:11:34	00:19:54	01 days 00:59:56	01 days 01:00:25
December	00:11:26	00:19:56	01 days 00:59:56	01 days 00:59:57

`Year 2023:  00:12:06  00:25:42  01 days 01:59:40 68 days 09:29:04`

Figure 8: Table of mean and max ride length divided by rider type

From the plot of the mean values in Figure 9 (left) we observe that casual riders consistently have longer rides, than members across the entire year. Furthermore, the mean ride length for casual riders varies across the year from approximately 20 minutes in the colder months to a peak of 35 minutes in August. In contrast, the mean ride length for members changes only from 10 to 13 minutes. When looking at the maximum ride length plot we can see a similar pattern; casual riders have longer rides, and their maximum ride lengths that changes from 1-68 days whereas members never have a ride that exceeds one day. However, it is important to note here that such longer rides (68 days), are outliers that occur quite rarely in the dataset. This can be seen when plotting the entire distribution of ride lengths in a box plot shown in Figure 10.

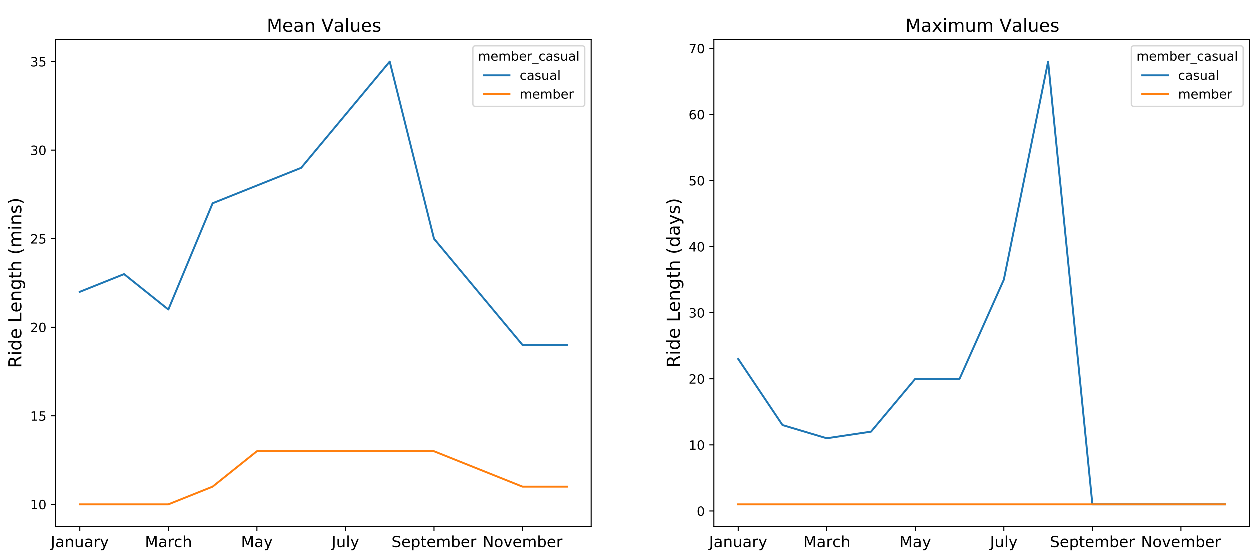

Figure 9: Plots of mean and max ride length divided by rider type grouped by month

A box plot typically shows the middle 50% of the distribution inside a box, then the lower and upper 25% are shown by the whiskers, and finally circles represent outliers. We can see that the entire distribution of ride lengths for casual rider is between 0-50 minutes, and the outliers are the ones that occupy the range 1-70 days. As for members, the distribution of ride length lies in the range 0-30 minutes, and its outliers 1- 25 hours. Therefore, if we ignore the outliers, the difference between the ride lengths for casual riders and members can be seen more clearly in Figure 11.

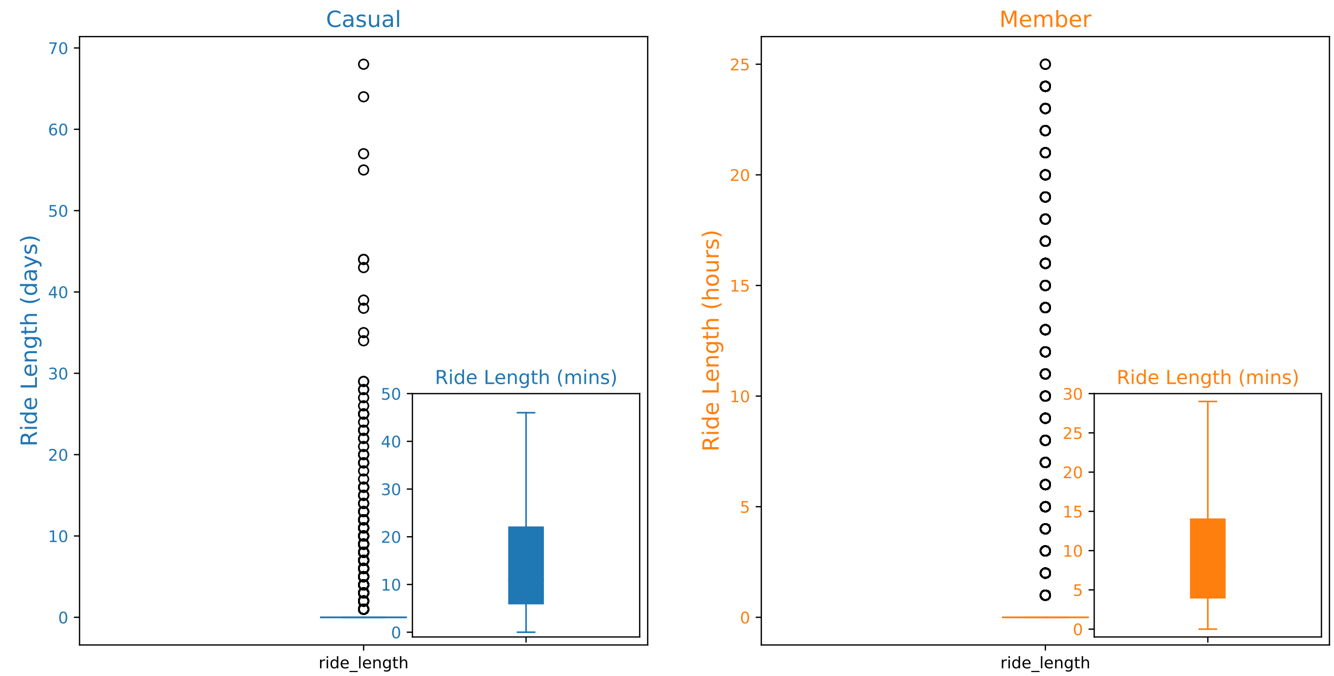

Figure 10: Box plots of the ride length distributions. Left: casual riders, Right: members

Figure 11 shows that, after ignoring outliers, the main distribution of ride lengths for casual riders is indeed In order to determine how significant this difference is we need to look at the number rides for the different rider types.

So far we have been looking at the data by first grouping it using the month. In order to get a different perspective, we now look at the mean value of the ride length, however when the data is grouped using the day of the week. The result can be seen in Figure 12. Here we see the same higher average for the casual riders when compared to the members.

Conclusion:

In general, casual riders always have a longer rides. And the behaviour across seasons or days varies with a peak in August and the weekends. We can see that the entire distribution of ride lengths for casual rider is between 0-50 minutes, with outliers in the range from 1-70 days. As for members, the ride lengths vary between 0-30 minutes, with outliers in the range 1- 25 hours. Therefore, if we ignore the outliers, the casual riders have a higher ride length by 66%.

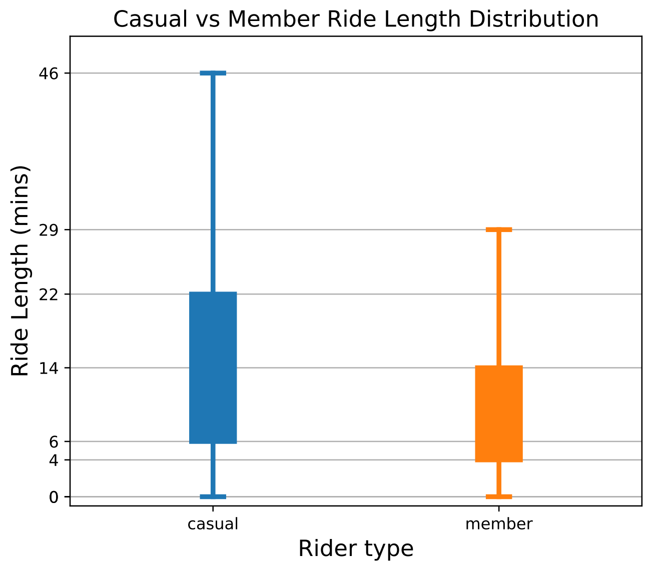

Figure 11: Box plot comparison between the ride length distributions of casual and member riders

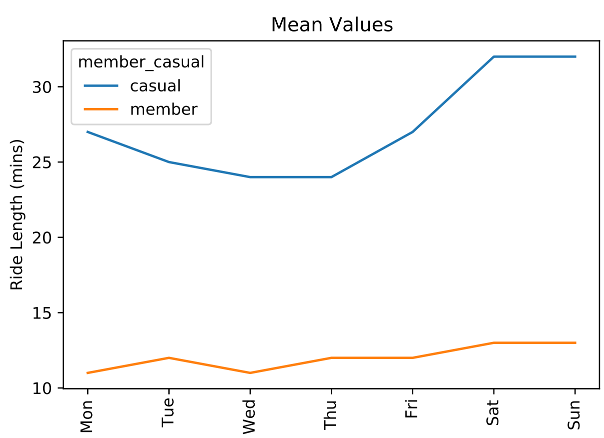

Figure 12: Plot of mean ride length divided by rider type grouped by day of the week