Citi Bike Project Report

Diego Sanoja

2022-11-02

Objective of the Project

The objective was to analyze the customer data of Citi Bike, the nation's largest bike share program, with 10,000 bikes and 600 stations across Manhattan, Brooklyn, Queens, and Jersey City with the purpose of understanding how and which type of clients use the bikes of the company to provide recommendations of how to turn customers (users with a 24-hour pass or a 7-day pass) to subscribers (users with an annual membership).

Data Used

For this project the data used was the one stored in the Google Public data set called NYC Citi Bike Trips. The data was processed by Citi Bike to remove rents that are taken by staff to service and inspect the system, as well as any rents below 60 seconds in length, which are considered false starts.

Each row of the set consists of a moment where a bike was rented. Each observation contains information like the length of the trip in seconds, the time when it was rented from and returned to a station as well as the names of the stations, the type of rider, and gender of the person who used the bike.

The data has some limitations: it provides information starting from 2013 where the bike sharing program started, some of the clients don't reveal their gender or year of birth, there is no customer id, there is no information regarding the prices to rent the bikes for each type of customer nor season discounts from the last years, therefore, it is assumed that they were similar to the ones in the present.

Changelog

Version 1.0.0 (10-10-2022)

- Wrote queries in BigQuery to analyze the data.
- Divided the queries in 2 subgroups, queries to gather basic understanding of the data set and queries to be further analyzed.

The first group of queries provides the following information:

i. The number of observations in the data set.

ii. The date of the first observation.

iii. The date of the last observation.

iv. The years when the bikes were rented.

The second group of queries produced results which were saved for further analysis in Google Sheets are the one that provided:

i. The number of rents per day of the week, month, year, gender, and user type.

ii. The average lease time per rider type each year.

iii. The average lease time per rider gender each year.

iv. The maximum rent time per user type and gender each year.

v. The number of rents longer than 1 hour occurred each year and month by user type and gender.

vi. The average time of short leases (leases that were shorter than 1 hour) per user type each year.

vii. The average time of short leases (leases that were shorter than 1 hour) per rider gender each year.

viii. The average time of short leases per rider type and gender each year.

- Stored and saved the queries in a word document in the same order provided above.

Version 1.1.0 (10-10-2022)

The following changed were done in Google Sheets:

- For the table produced by the first query:

i. Changed the file's name file to Rents_per_rider_gender_and_type.

ii. Modified the Month column replacing the numbers with the names of the months using the Find and Replace option.

iii. Modified the Day column replacing the number 1 with Sunday, 2 with Monday, and so on until 7 with Saturday using the Find and Replace option.

- For the table produced by the second query:

i. Changed the file's name to Average_lease_time_per_user_type.

- For the table produced by the third query:

i. Changed the file's name to Average_lease_time_per_gender.

- For the table produced by the fourth query:

i. Changed the file's name to Longest_rent_time_per_user_type_and_gender.

- For the table produced by the fifth query:

i. Changed the file's name to Long_rents_per_user_type_and_gender.

ii. Modified the Month column replacing the numbers with the names of the months using the Find and Replace option.

- For the table produced by the sixth query:

i. Changed the file's name to Short_rents_average_time_per_user_type.

- For the table produced by the seventh query:

i. Changed the file's name to Short_rents_average_time_per_gender.

- For the table produced by the eighth query:

i. Changed the file's name to Short_rents_average_time_per_user_type_and_gender.

- Transformed each Google Sheets file to a csv document for further study using the Python programming language.

Version 1.2.0 (11-10-2022)

- Opened a jupyter notebook file with the name of Citi Bike data analysis
- Using the data of the Rents_per_rider_type_and_gender csv file:

i. Created a data frame which contains the data of the csv file.

ii. Added a column named Time Period which combined the Month and the Year columns.

iii. Created 5 visuals and 1 table shown in the following order:

1. A pie chart that shows the percentage of bikes rented made each year.
2. A grouped bar chart which shows the number of bikes used per year and user type of the client.
3. A grouped bar chart which displays the number of bikes rented per year and gender of the client.
4. A table that separates the number of rents per user type in columns and gender of the rider as rows.
5. A grouped bar chart which displays the number of bikes rented per day of the week and user type of the client.
6. A line chart that shows the number of bikes used during each month and year per user type.

- Using the data of the Average_lease_time_per_user_type csv file:

i. Created a data frame which contained the data of the csv file.

ii. Produced a grouped bar chart which displays the average time of the bike leases per user type each year using the data frame.

- Using the data of the Average_lease_time_per_gender csv file:

i. Created a data frame which stores the data of the csv file.

ii. Produced a grouped bar chart which shows the average time of the bike rents per gender each year using the data frame.

Version 1.2.1 (13-10-2022)

- Using the data of the Longest_rent_time_per_user_type_and_gender csv file:

i. Created a data frame which contains the data of the csv file.

ii. Produced 3 visuals in the following order:

1. A grouped bar chart which displays the duration of the longest time a bike was used each year by user type.
2. A dot plot which displays the time of the longest rent bike each year by gender of the customer user type.
3. A dot plot which displays the time of the longest rent bike each year by gender of the subscriber user type.

- Using the data of the Long_rents_per_user_type_and_gender csv file:

i. Created a data frame which stores the data of the csv file.

ii. Added a column to the data frame called group which contains the user type and the gender of the rider separated by a colon.

iii. Produced 2 visuals in the following order:

1. A table that separates the number of bike rents per rider group in the columns and year in the rows.

2. A line chart that shows the amount of bikes rented for long periods during each month and year per user type.

- Using the data of the Short_rents_average_time_per_user_type csv file:

i. Created a data frame which contains the data of the csv file.

ii. Produced a grouped bar chart which displays the average time a bike was rented each year by user type.

Version 1.2.2 (16-10-2022)

- Using the data of the Short_rents_average_time_per_gender csv file:

i. Created a data frame which stores the data of the csv file.

ii. Produced a grouped bar chart which shows the average time a bike was rented each year by gender.

- Using the data of the Short_rents_average_time_per_user_type_and_gender csv file:

i. Created a data frame which contains the data of the csv file.

ii. Produced 2 visuals in the following order:

1. A dot plot which displays the average time a subscriber rented each year by gender.
2. A dot plot which displays the average time a customer rented each year by gender.

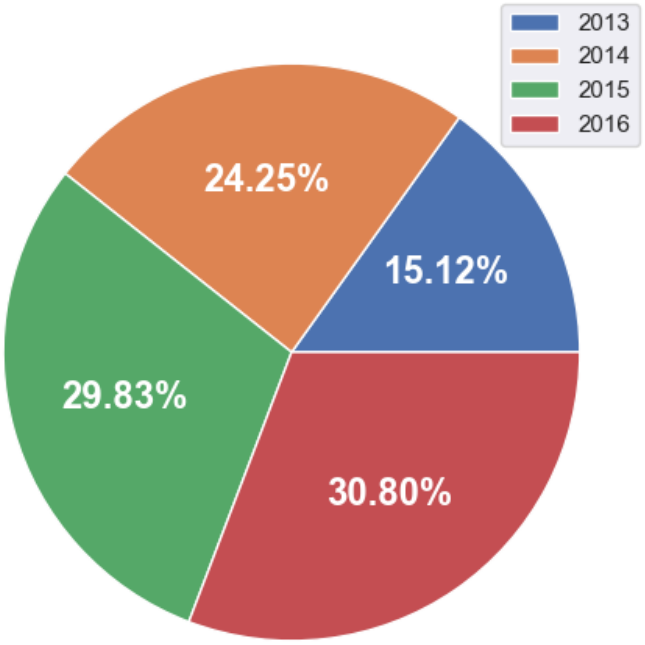
- Produced a table that separates the number of short bike rents per rider group in the columns and year in the rows using the data of the total number of rents and the number of long rents data frames.
- Saved each visual as a png file in the Citi Bike Project folder.

Analysis Summary

For the analysis phase a total of 8 csv files generated using Big Query were used. The data set from which the files were generated contains 33319019 observations which happened between July 2013 and September of 2016.

The first file analyzed was the Rents_per_rider_type_and_gender csv file. It contains the amount of bikes rented per day, month and year the rents occurred as well as the gender and user type of the riders. From this file, 5 visuals and 1 table were created, each one provided different insights regarding the people who used bikes.

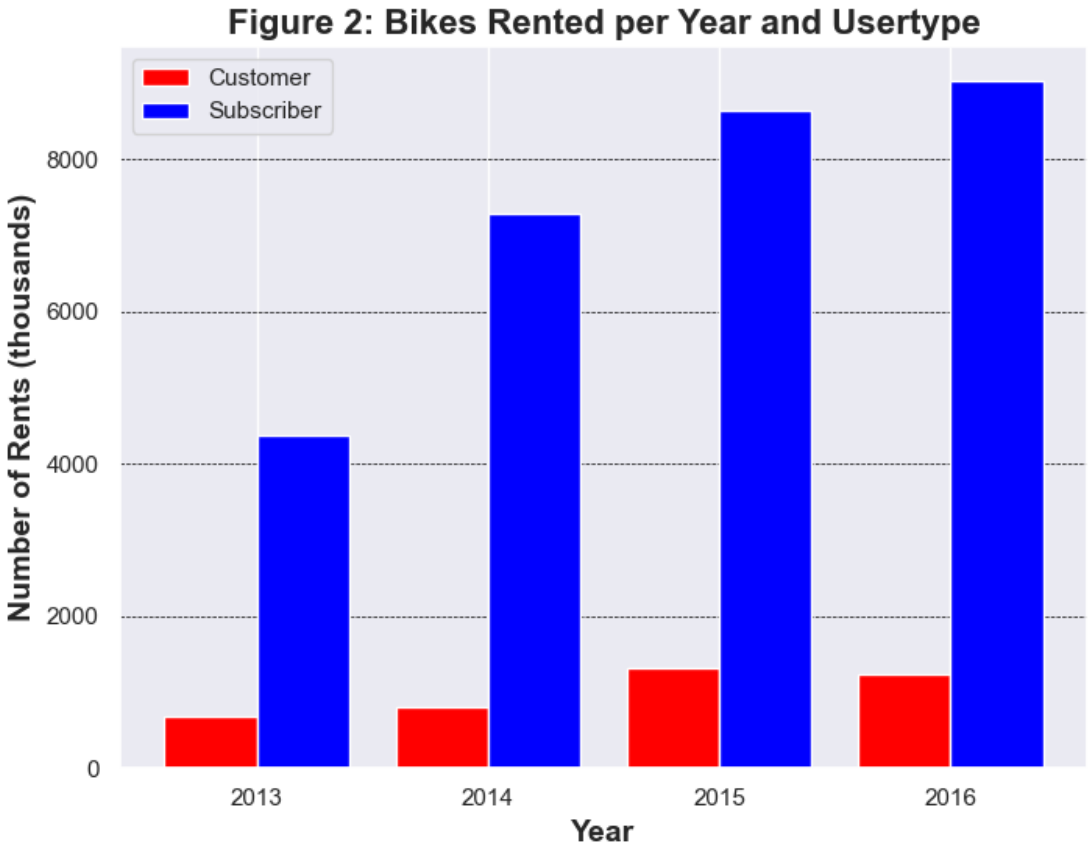
Figure 1: Percentage of Bike Rents per Year



The first of those visuals is Figure 1 displayed above which shows the percentage of rents made each year. The purpose of this pie chart was to understand how the leases had changed from the beginning of the program until the end of the period that the data set covers. This chart provided the following insights:

1. Every year, there was more people that used bikes from Citi bike.
2. The greatest increase was from 2013 to 2014. A possible reason for this could be that the program started during the later half of 2013.
3. The lowest increase was from 2015 to 2016. A possible reason for this could be that the data set contains information of the first 3 quarters of 2016.

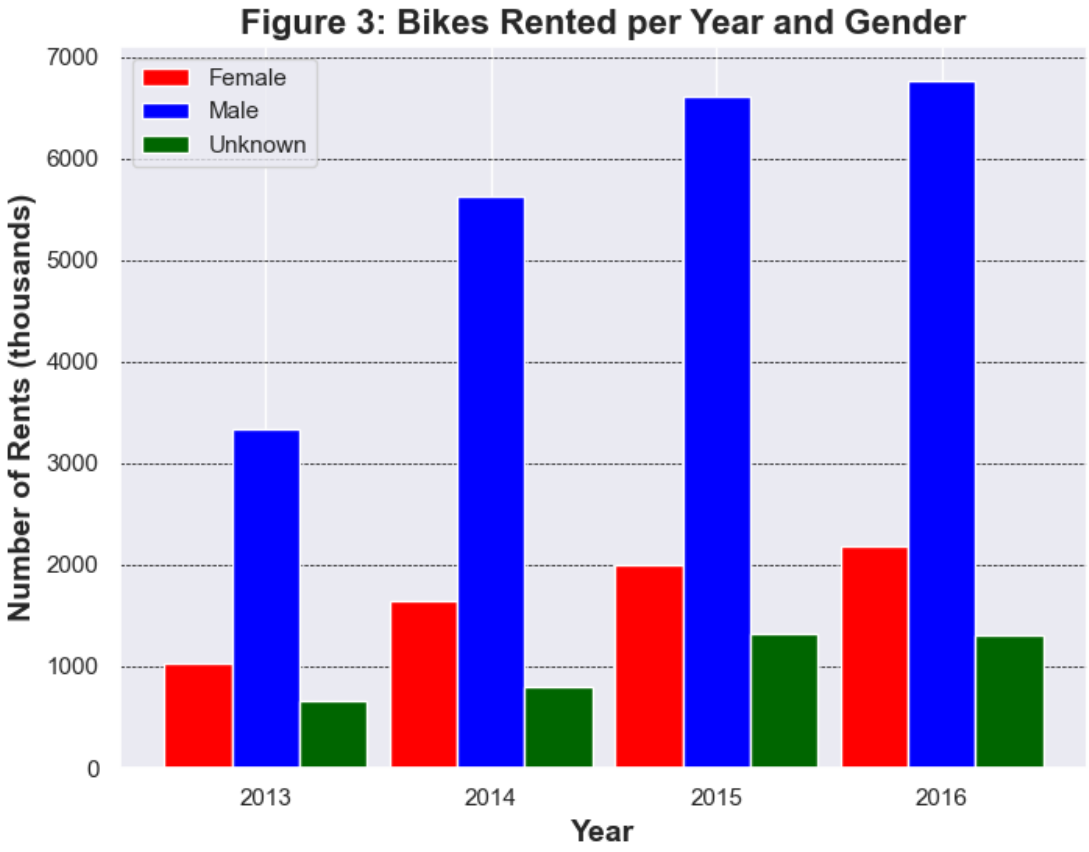
The next step analyzing this file was to know how many bikes were used by customers and subscribers through the years. To answer question, Figure 2 was built which shows the number of bikes rented each year for each group of riders.



By examining Figure 2 closely, the following conclusions were reached:

1. Subscribers used more bikes than customers each year with 2016 being the year that the difference was greater.
2. For subscribers, there was an increase in the amount of rents each year while for customers there was not a specified trend of increase nor decrease as time passed.
3. Approximately, 88% of the rents were done by subscribers.
4. Subscribers used more than 4 million bikes each year while customer rented between 0.5 and 1.5 million bikes each year.

To continue with the process it was important to know the gender of the riders and how much each one used a bike. To achieve the objective, Figure 3 was designed to show the number of bikes rented each year by the gender of the riders.



After studying Figure 3 closely, the following conclusions were reached:

1. Males were the ones who rented bikes the most.
2. During each year, the combination of rents of both female riders and the ones who kept their gender private composed less than 50% of the riders through each year.
3. For each group, the number of rents increased every year (except for the unknowns group in 2016).

Furthermore it was important to understand which gender used more bikes in each group. To discover that, Table 1 shown below was created and it separates the number of rents both by the gender and user type of the client.

Table 1: Distrbution of Rents per Group of Clients

	Customer	Subscriber
Female	13148	6865636
Male	20060	22329254
Unknown	3974823	116098

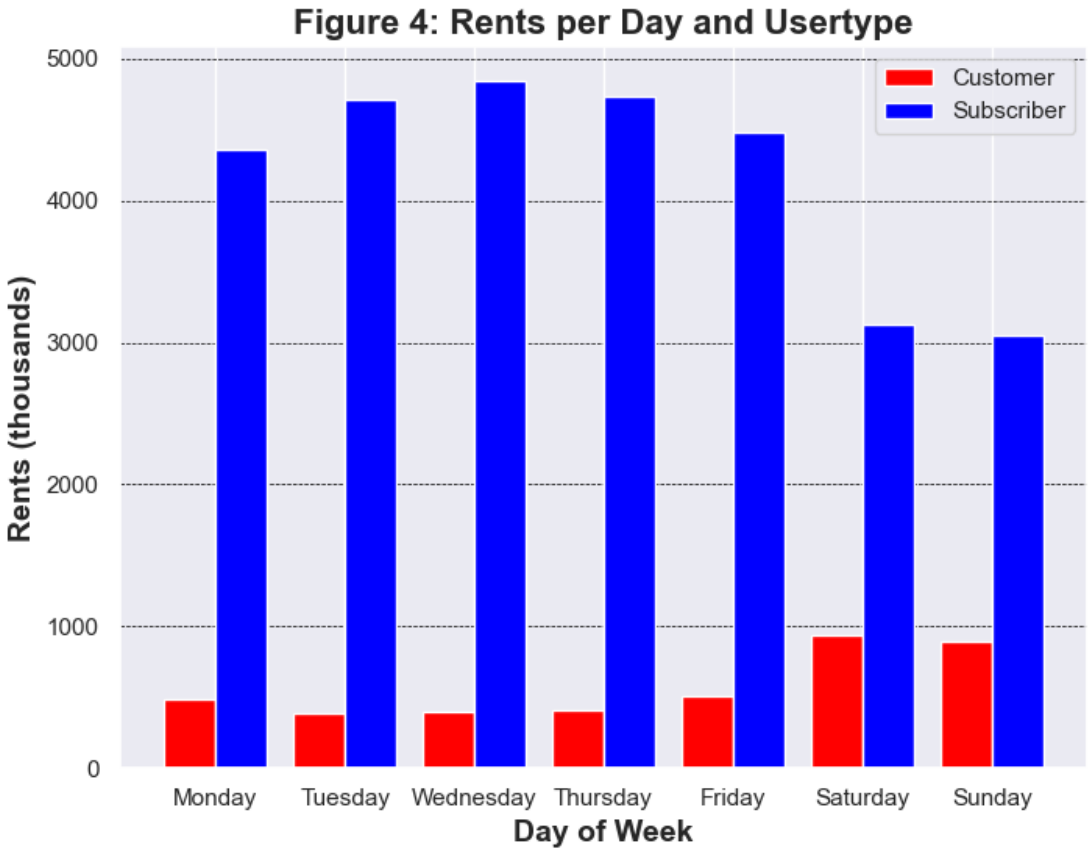
Table 1 provided the following insights regarding the riders:

1. Most of the members of the customer group didn't provide their gender when they payed to use the

bikes.

2. Most subscribers were males.
3. For both groups (mainly for the subscriber group), men used more bikes than women.

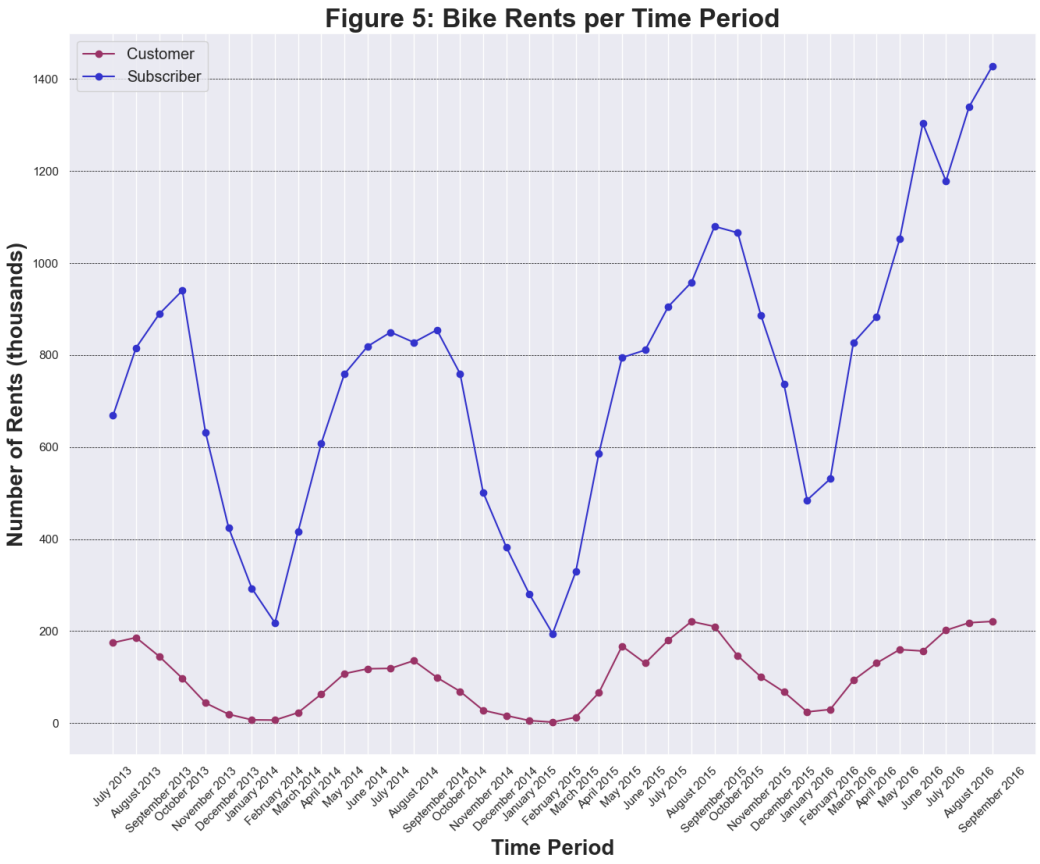
In addition to knowing how many riders fell in each group of gender and user type, it was of great interest discovering which day of the week a rider rented a bike. To discover that, Figure 4 was built to display the number of rents per group for each day of the week starting from Monday.



By examining Figure 4 closely, the following conclusions were reached:

1. Subscribers rented bikes mainly from Monday to Friday while the customers rented them mostly during weekends.
2. The day where bikes were rented mainly by subscribers and customers were Wednesdays and Saturdays respectively.

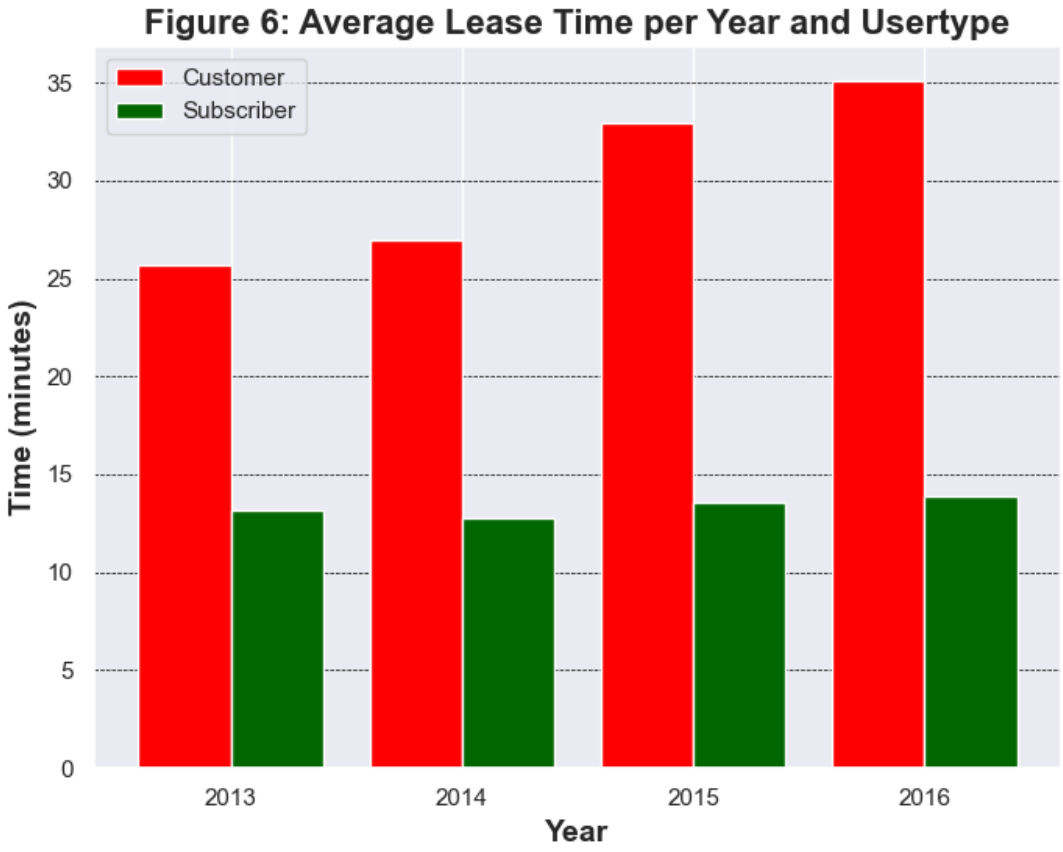
To conclude the analysis on this particular file, the month and year where the bikes were rented were combined to create time periods with the objective of determining if there were patterns or trends in the data set. To achieve this goal, Figure 5 was created which shows the number of bikes rented each month of each year per user type.



After studying Figure 5 closely, the following insights were discovered:

1. Every month of every year, subscribers used more bikes than customers.
2. For each group of clients there was a seasonality trend (mainly for the subscribers group) since the number of rents was at its maximum during the months of summer and at the lowest during the months of winter each year.
3. For almost every time period, subscribers and customers rented more than 200 thousand bikes and less than 200 thousand bikes respectively.

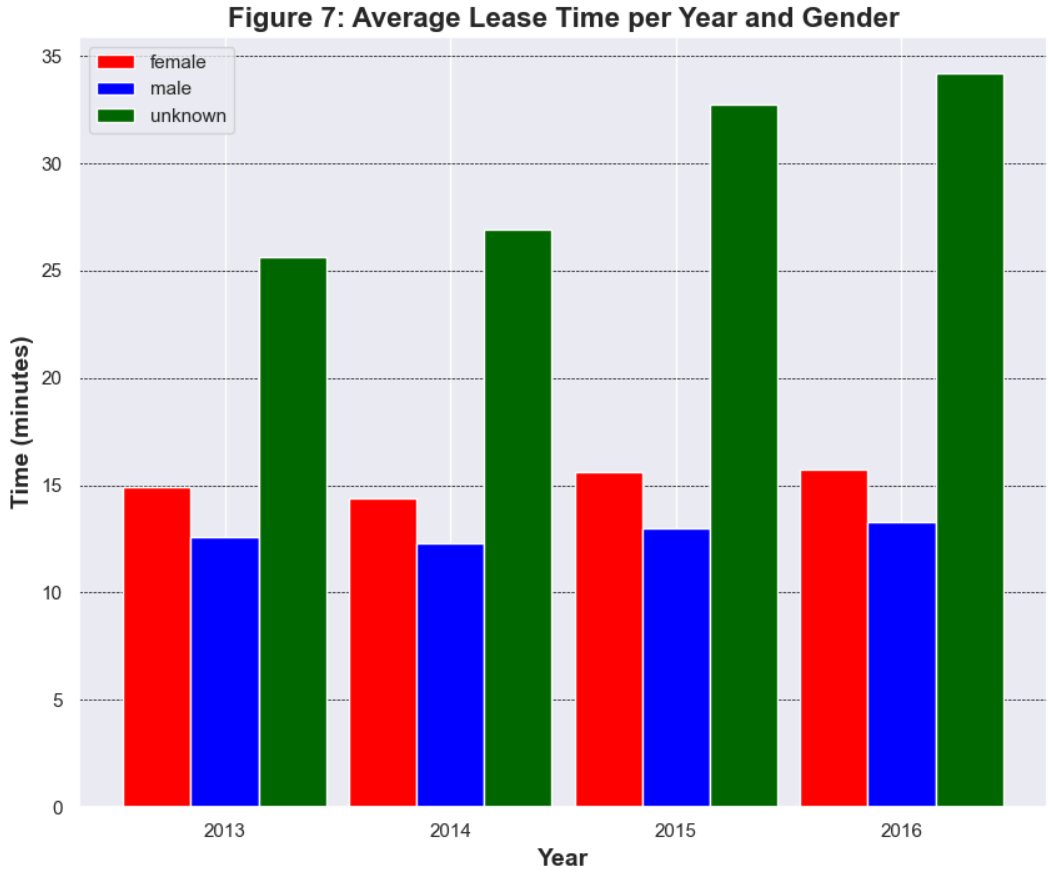
The Average_lease_time_per_user_type file was the second document analyzed, it contains the average time that a bike was used each year per user type. Using this data, Figure 6 was designed which displays the average lease time of each rider type per year.



By examining Figure 6 carefully, the following conclusions were obtained:

1. The average lease time was higher for customers than for subscribers.
2. Each year, the average lease time increased for customers with the greatest increase being from 2014 to 2015.
3. The average lease time for subscribers had no defined trend of increase nor decrease and was always between 10 and 15 minutes.

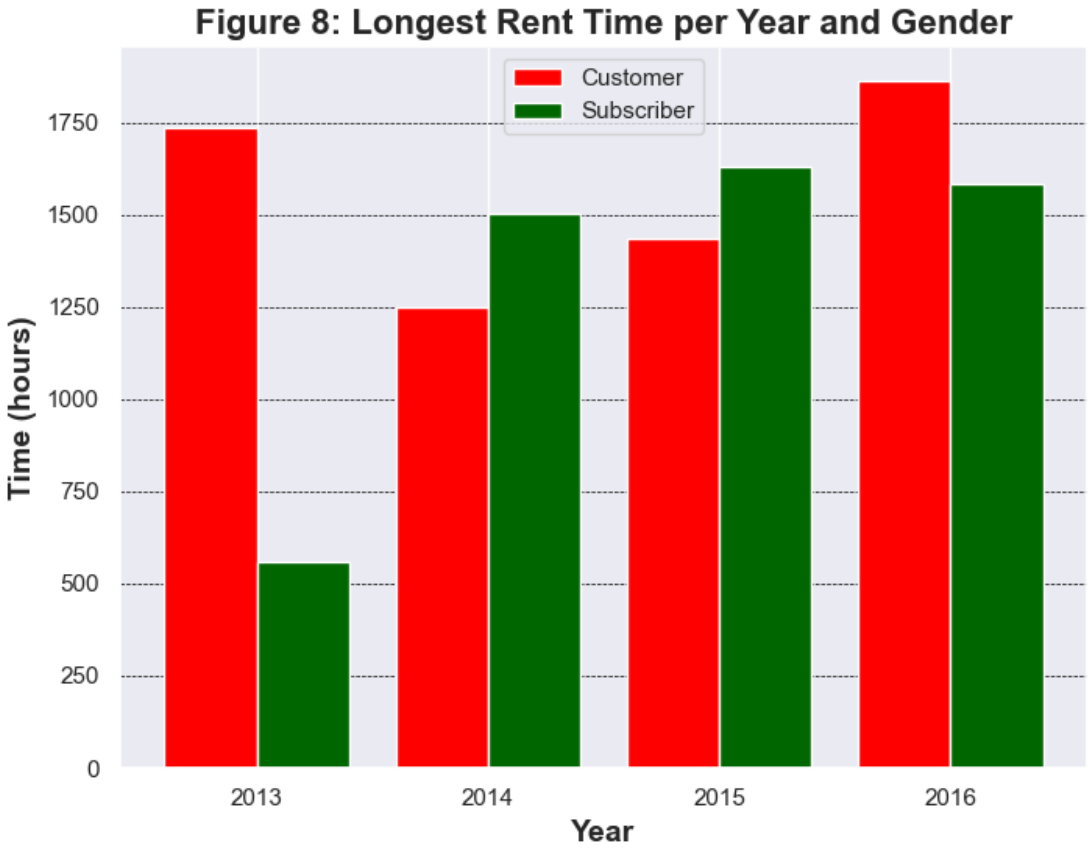
The third csv file analyzed was the Average_lease_time_per_gender document which contains the average time that a bike was rented each year per gender of the client. Using this data, Figure 7 was designed which shows the average lease time of each rider by gender per year.



The examination of Figure 7 provided the following insights:

1. Every year, riders of unknown gender made leases that lasted longer, while the ones that did shorter leases were males.
2. The average lease time of the clients who made longer leases increased every year. The average for this group is between 25 and 35 minutes through the time period of interest.
3. The average lease times for males and females had no defined trend of increase or decrease and were always between 12 and 14 minutes for males and between 14 and 16 minutes for females.
4. The difference between the average for males and females was within a range of 2 to 3 in minutes.

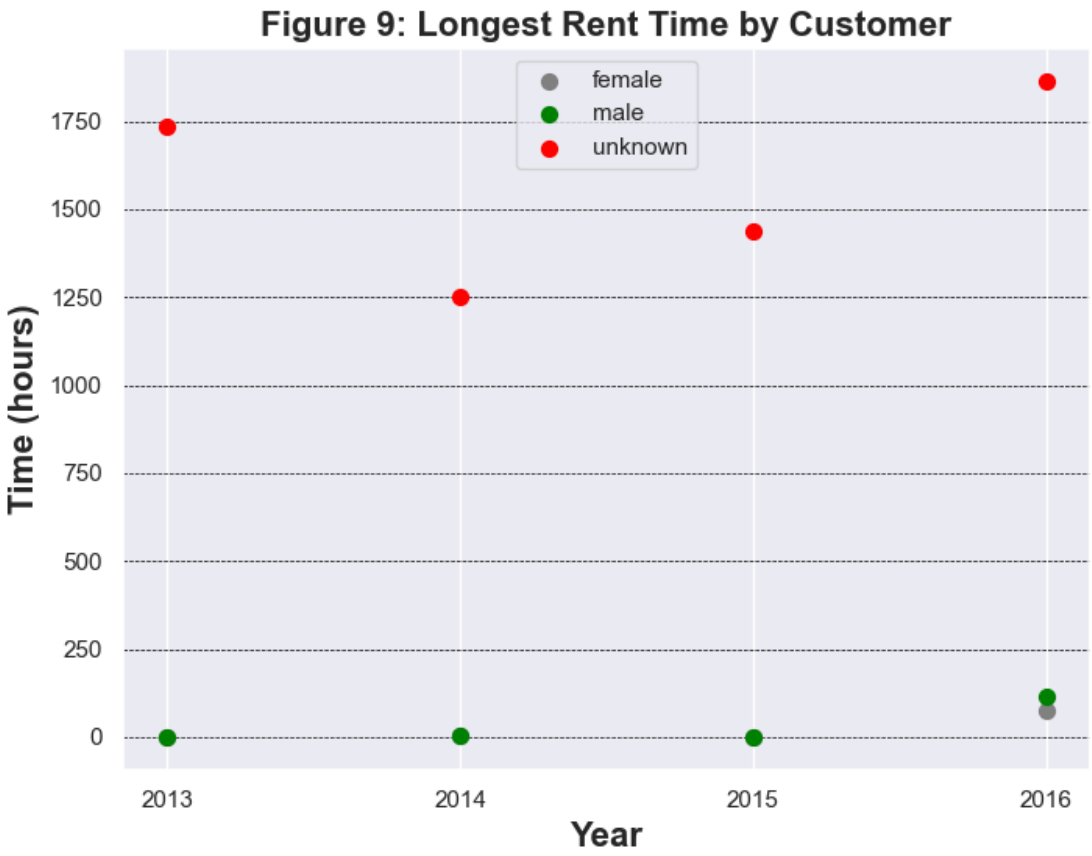
To discover the time that a bike was rented the most each year both by user type and gender of the client, the Longest_rent_time_per_user_type_and_gender was created and studied. Using the data of this file, 3 visuals were created. The first one was Figure 8 which displays the time of the longest rent per user type each year.



After studying Figure 8 closely, the following conclusions were reached:

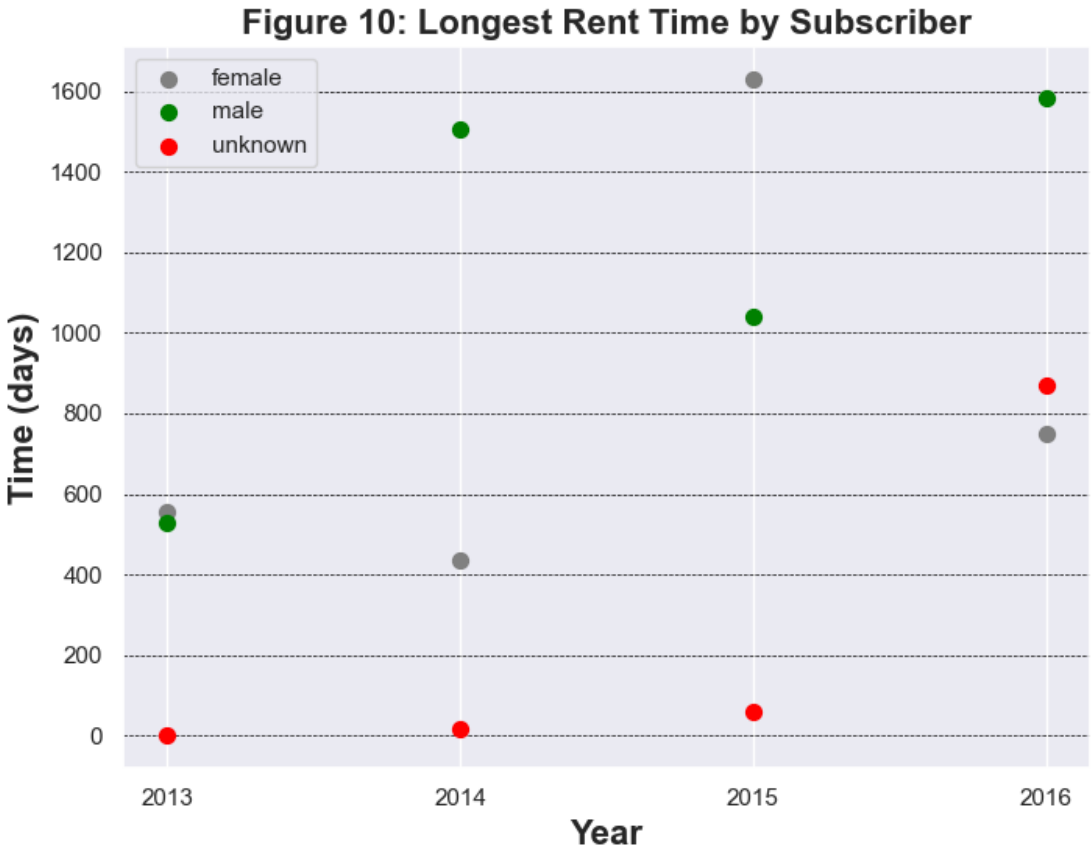
1. In 2013 and 2016, a customer was the rider who rented the bike for the longest time while in 2014 and 2015 the opposite happened.
2. There wasn't a trend of increase nor decrease for the time of the longest rent each year for each group.
3. The year with the greatest difference in the longest rents between groups was 2013.

To further analyze the data in the file, Figures 9 and 10 were designed which display the longest ride time but by gender each one by a group of riders in particular.



The data used to create Figure 9 provided the following insights:

1. Customers of unknown gender made the longest rent each year.
2. For customers that were either female or males, their longest bike rent each year was close.
3. The year where a customer rented a bike the longest was in 2016.



The data used to create Figure 10 provided the following insights:

1. Through most of the time period of interest, (except in 2016), subscribers of unknown gender had the shortest rent in this group of clients. However, the time of the longest rent made by this group increased every year.
2. Male subscribers rented a bike the longest in 2014 and 2016 while females did this in 2013 and 2015.
3. The longest rent made by a subscriber was in 2015.

Since most of the longest rents shown in Figures 8 to 10 are above 200 hundred hours each year, the insights and conclusions from Figures 6 and 7 could be skewed by those values and others close to them. To discover if the results were skewed, the leases were divided in 2 categories, short and long. Shorts leases are the ones were the a bike was used by a client for an hour at most while long leases are the ones that lasted more than 1 hour.

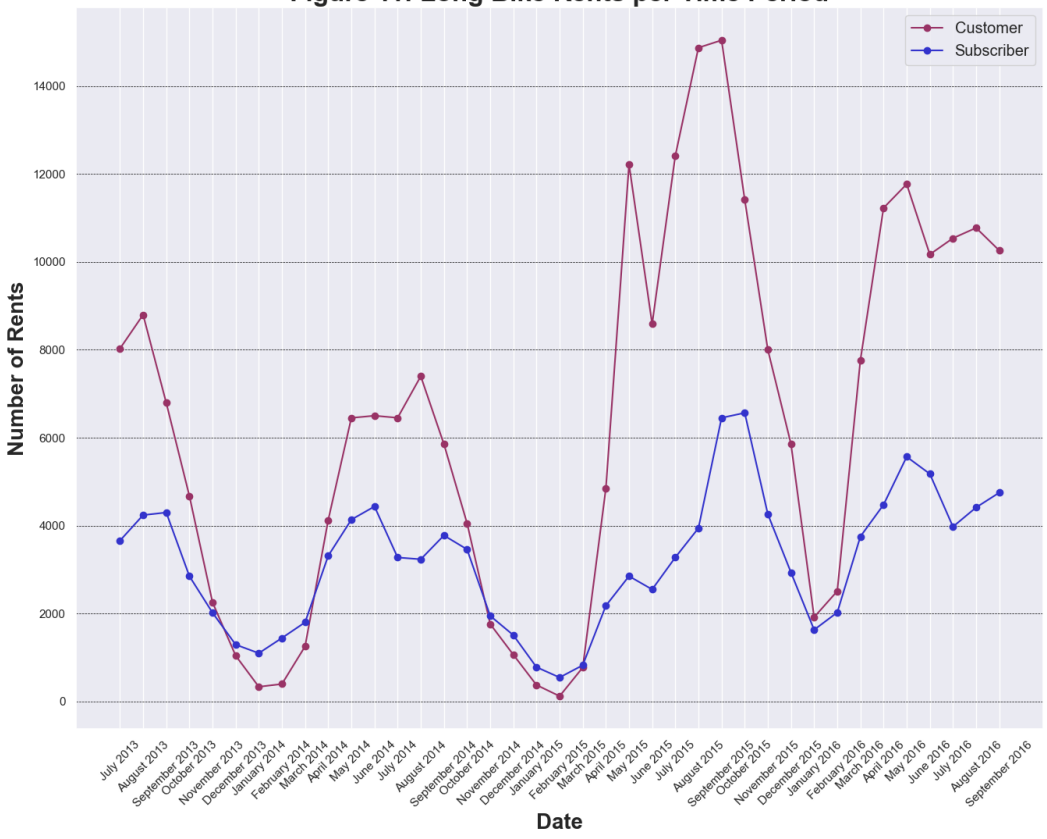
To confirm if the results were skewed or not, 4 csv files were generated, the first one of them that was analyzed was the Long_rents_per_user_type_and_gender file. It contains the amount of bikes rented per day, month and year the rents occurred as well as the gender and user type of the riders. Using the data of this file, the table and line chart shown below were designed.

Table 2 separates the number of leases both by the gender and user type of the client in the columns and the year by row. Figure 11 is a line chart which displays the number of long rents each month of each year per user type and gender.

Table 2: Distribution of Long Rents per Group of Clients

	Customer: Female	Customer: Male	Customer: Unknown	Subscriber: Female	Subscriber: Male	Subscriber: Unknown
2013	0	1	31587	5447	12907	1
2014	3	3	45637	9511	23839	72
2015	0	0	94520	11577	25513	61
2016	511	614	75805	11078	23062	1619

Figure 11: Long Bike Rents per Time Period



After studying both Table 2 and Figure 11 closely, the following conclusions were obtained:

1. The greatest number of long rents was done by customers.
2. The year that had the greatest number of long leases was 2015.
3. Customers of unknown gender were the ones who made more long rents through each year while for the group of subscribers, males were the ones that did long rents mostly each year.
4. For each year, the customer group had a greater number of long leases than the subscriber group during the months of summer, especially in 2015.
5. There was a seasonal trend of increase for both groups with the greatest and lowest amounts of long leases during the months summer and winter respectively.

6. During the months of winter in 2013 and 2014, subscribers made more long leases than customers. The opposite happened in 2015.

Furthermore, after adding the numbers of each category in the table, it was discovered that 373368 rents were longer than an hour which is equivalent to 1.12% of the observations in the data set. Therefore, 98.88% of the leases were short. The next step of the analysis was to discover how this affected the rent averages.

By using the data in the Trips_per_customer_gender_and_type file and Table 2, Table 3 shown below was designed. It has the same structure as Table 2 but it displays the number of short leases instead.

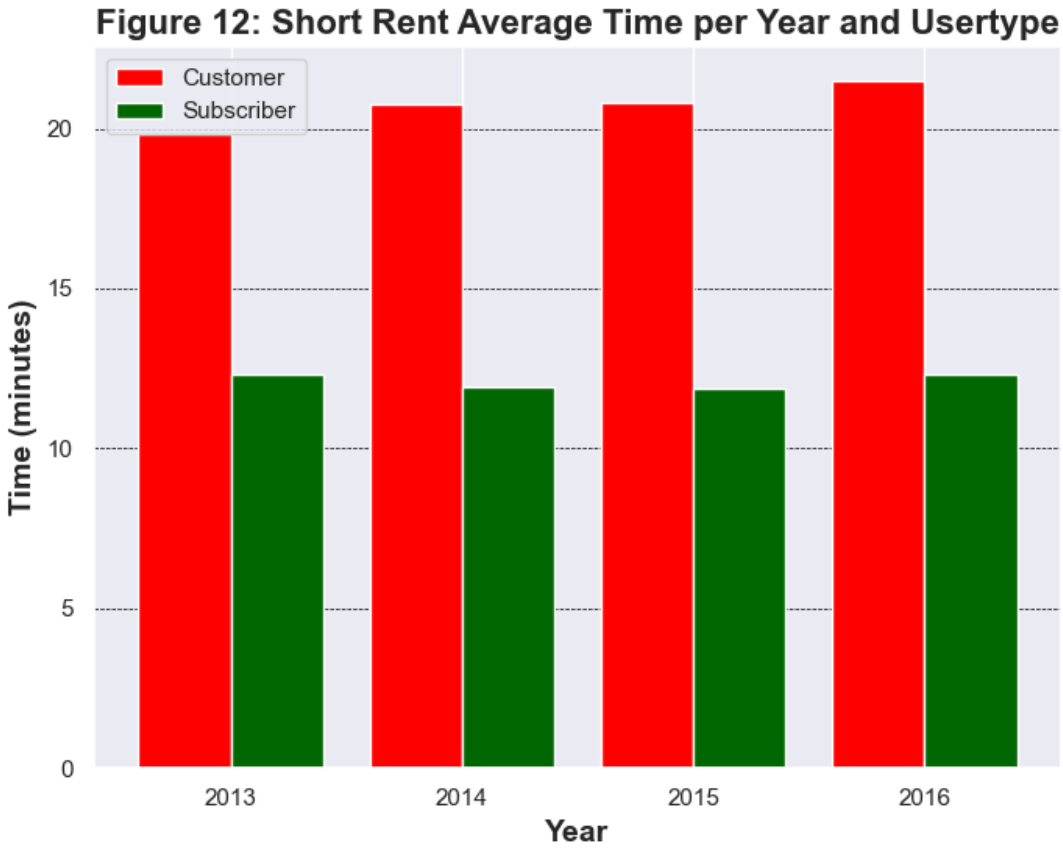
Table 3: Distribution of Short Rents per Group of Clients

	Customer: Female	Customer: Male	Customer: Unknown	Subscriber: Female	Subscriber: Male	Subscriber: Unknown
2013	1	41	635310	1029183	3322442	265
2014	19	47	747786	1640741	5612073	1485
2015	0	0	1216811	1995385	6583416	10686
2016	12614	19354	1127367	2162714	6726002	101909

Table 3 provided the following insights regarding riders who made short rents:

1. Each year for the customer group, most of the rents were done by riders who refused to reveal their gender.
2. With the exception of 2015, the number of female and male customers increased every year.
3. Only customers who refused to provide their gender used bikes in 2015.
4. Each year for the subscriber group, the number of leases increased, especially for males.

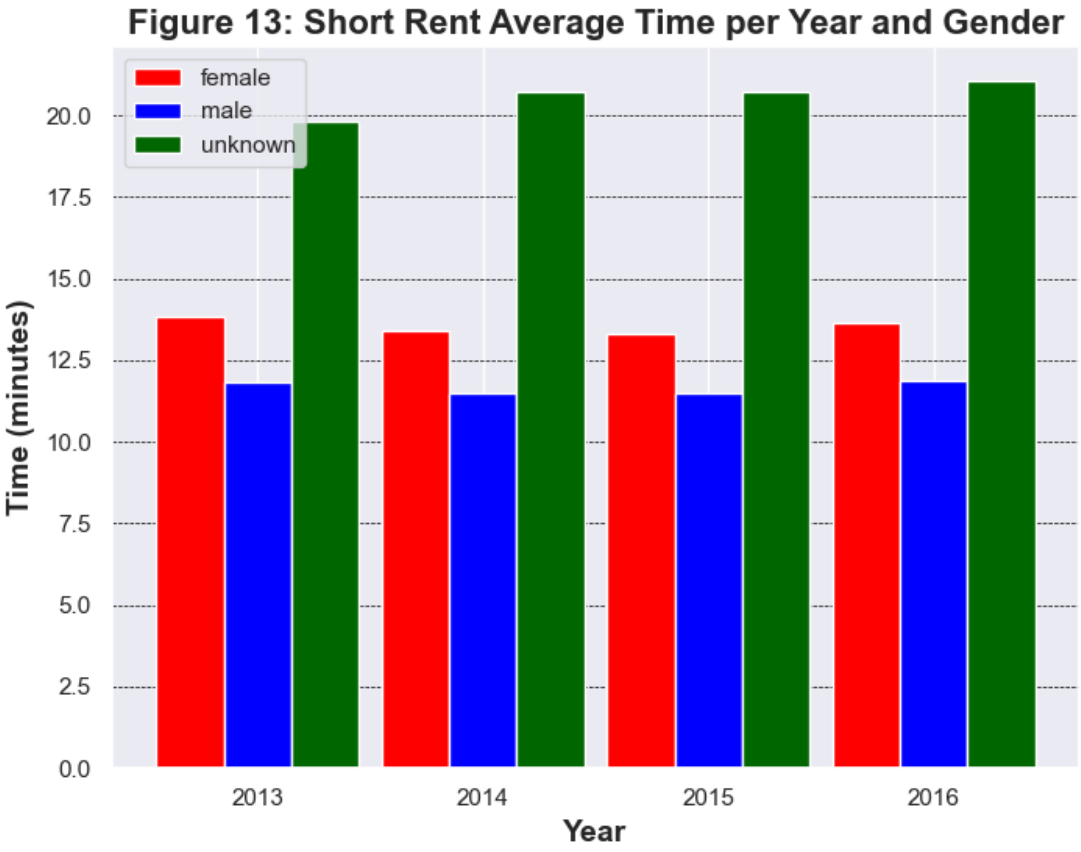
The next file analyzed was the Short_rents_average_time_per_user_type csv, which contains the average of the short rents per year and user type of the clients. Figure 12 shown below displays the data in the file.



By examining Figure 12 and comparing it to Figure 6, the following insights were obtained:

1. The average of each group was reduced each year, especially for the customer group.
2. The average rent time still was lower for subscribers than for customers but the difference was lower than before.
3. The average lease time of subscribers was between 11 and 13 minutes each year while for customers was between 19 and 22 minutes.

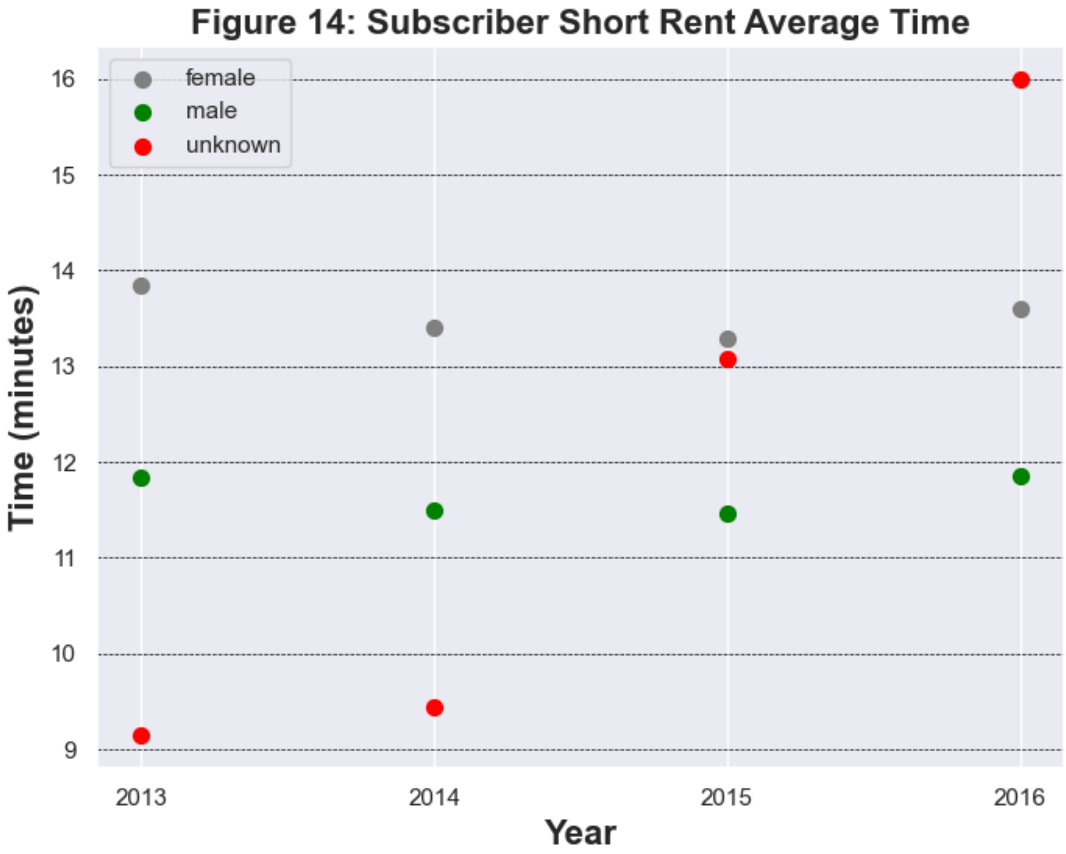
The Short_rents_average_time_per_gender was the second to last csv analyzed, which contains the average of the short rents per year and gender of the clients. Figure 13 shown below displays the data in the file.



By studying Figure 13 and comparing it to Figure 7, the following conclusions were obtained:

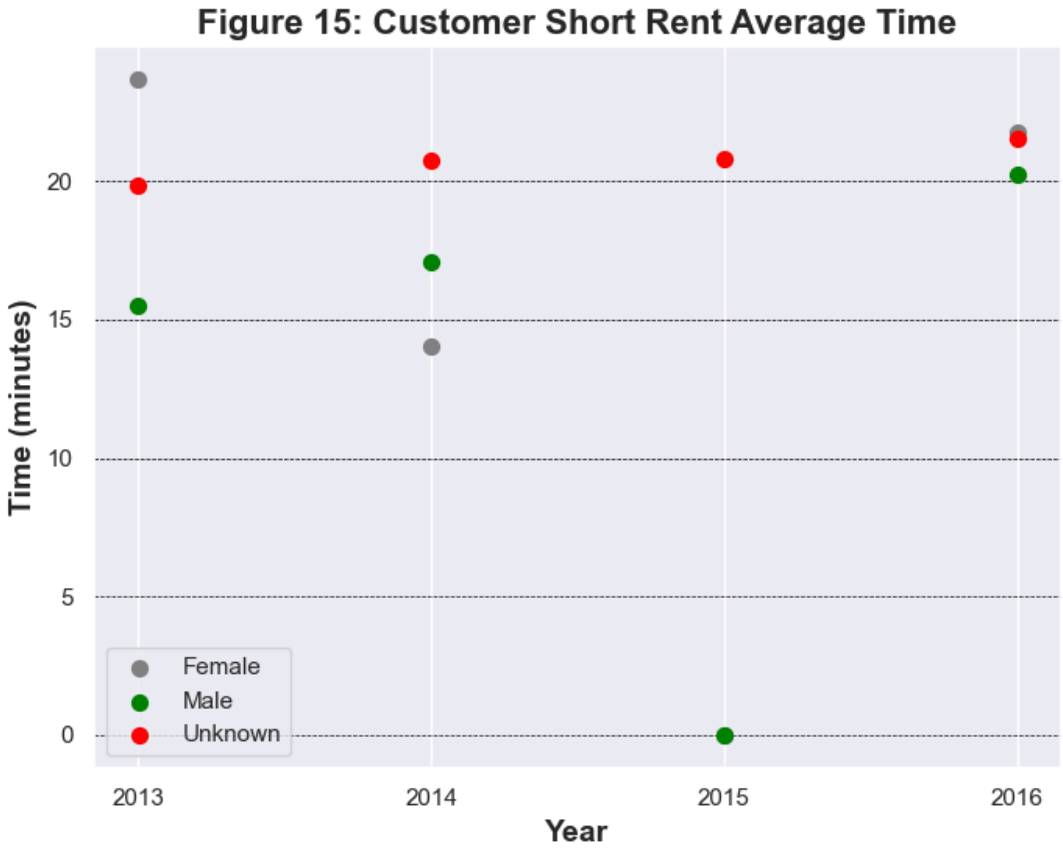
1. The average of each group was reduced each year, especially for riders who didn't provide their gender when renting a bike. Also, the increase in the average each year was lower.
2. The average lease time still was lower for males than for females and there was no trend in the average as the time passed. Furthermore, the average was between 11 and 12 minutes for males and between 13 and 14 minutes for females.
3. The average rent time of riders of unknown gender now was between 19 and 22 minutes. Furthermore, the trend of increase in the average time each year still remained.

The Short_rents_average_time_per_user_type_and_gender was the last csv file analyzed, it contains the average time of the bike leases per year and both the gender and user type of the client. By using the data of this document, Figures 14 and 15 were designed to show the longest lease time by gender, each one by a group of riders in particular.



After studying Table 3 and Figures 12 to 14 closely, the following conclusions were obtained:

1. The average of subscribers of unknown gender increased every year while the averages of females and males had no trend.
2. Through most of the years except in 2016, female subscribers had a higher average than the other 2 groups.
3. The average of female subscribers was always higher than the one of male subscribers.
4. The average of each year for the customer group was mainly influenced by the lease time of male riders.



After analyzing Table 3, Figures 12, 13 and 15 carefully, the following conclusions were obtained:

1. The average of customers of unknown sex was between 19 and 22 minutes and it increased every year.
2. In 2013 and 2016, female customers had a higher average than male ones. The opposite happened in 2014.
3. The average of 2015 was made only by customers of unknown gender.
4. The average of each year for the customer group was mainly influenced by the lease time of riders of unknown gender.

Conclusions

- Each year, the amount of bikes used by each group increased, especially for male subscribers.
- About 98.88% of the clients rented bikes for periods shorter than or equal to 1 hour.
- The months where bikes were used the most are the ones of the summer season (from June to September), where subscribers used more bikes than customers each month of every year.
- Most customers didn't provide their gender when they rented bikes and they used them mostly on weekends. Furthermore, customers were the clients who used bikes for longer periods of time in average.
- Subscribers rented bikes mainly through the week. Also, male subscribers rented bikes for lower periods of time than female subscribers in average.

Recommendations to turn customers into subscribers.

a. Propose a discount or a special offer during the months of summer to turn customers into subscribers.

b. Reduce the price of the annual membership for first-year subscribers.

c. Create a user name with a piece of information like the phone number or email for each client when they use a bike. Then use this information to send the users messages and updates that show them the benefits of becoming subscribers.

d. Collect further information about the clients like the user name or phone number to discover how many times each rider uses a bike and what type of bike they rent. Then, repeat this analysis after 1 year with the additional information to make further examinations, reach stronger conclusions and develop new and better ideas to increase the number of subscribers.