“Weather” to Rent a Bike

Jiayi Gao
WISCONSIN SCHOOL OF BUSINESS
Dec 2021

Executive Summary

The purpose of the analysis in this case is to help PhillyCycle, a bike rental company, to explore some new financial strategies on bike rentals. The filtered data contains 2676 entries with 11 features including user groups, weather condition, seasons, and specific date time. To facilitate us on the later analysis, we created some new variables based on the existed variables. After these steps, we first had an overview of the distribution for the registered and casual users since all users were consisted with these two groups of users. We noticed that registered users account for 80% of all users. We then analyzed average hourly rentals in different situations. Few things need to be noticed in our analysis. First, the average hourly rental number in summer is the highest and almost twice as large as the number in winter. Spring and fall have about the same average hourly rental number as summer. The second thing need to be noticed is that registered users tend to have more bike rentals during the weekdays but casual users, on the other hand, tend to have more bike rental during the weekends. We also found that during the weekdays, the average hourly rental number for registered users reaches the highest at the 8am and 5pm, which implies the registered users rent the bike to go to or go from their workplaces. However, for weekend, both registered and casual users tend to rent a bike in the afternoon, which most of the rentals could be hang out with bikes. After we conducted a 95% confidence interval test to check whether there is evidence of different hourly rental volume on weekends compared to weekdays for users in each group, we found that we can say there is differences between the sample means for weekdays and weekends for both registered and casual users, however, there is no evidence to say the sample means are significantly different for all users. In our regression analysis, we first found that temperature and average hourly rental for all users have a positive linear relationship. And in the multiple regression analysis, we noticed that humidity, wind speed, mist, summer, fall, year 2012, and temperature all have a significant effect on average hourly rental number for all users. We also noticed an interesting fact that the coefficient for summer is negative, which we will have a detailed discussion in our multiple regression part.

Analysis

Bike Rental Patterns

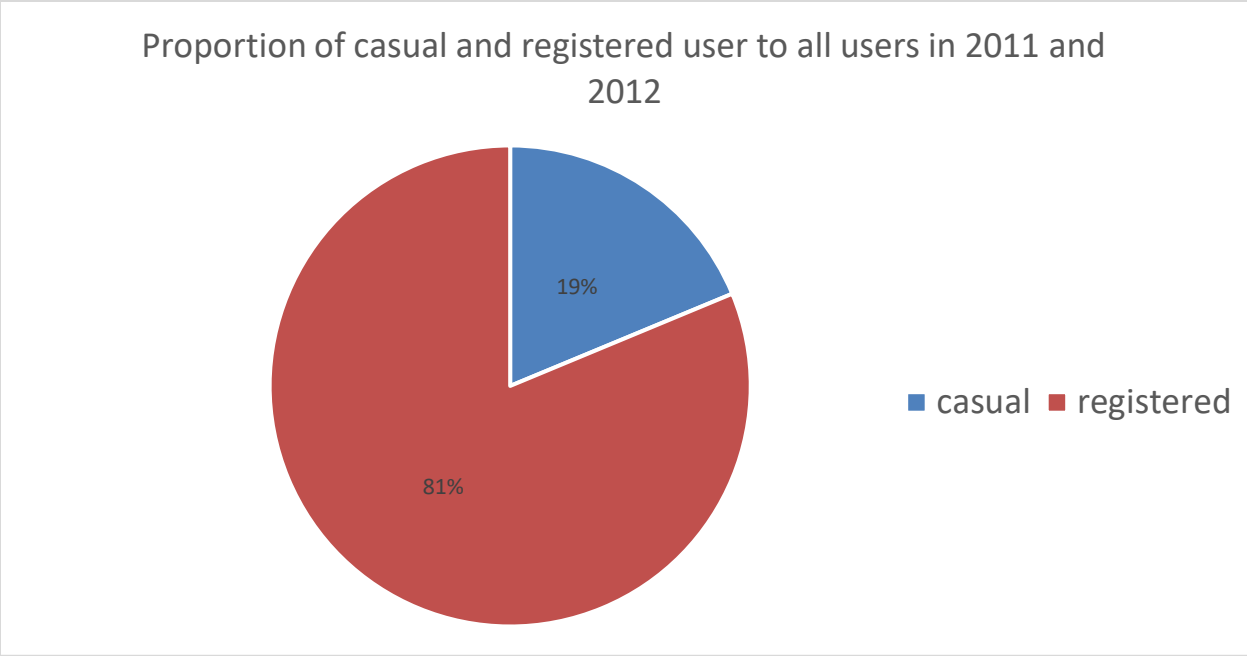
Table1

	casual user rentals	registered user rental
median	16	115
first quantile	4	35
third quantile	47	215.75

Comparing to mean value, the median is a more fit central tendency measurement in this case since the outliers in the dataset might cause variations to the mean value. The reason for these variations is that mean value is calculated by the average of all samples in the dataset. In this case, some extreme values may cause the central of the data skew to one side. Thus, median is the central tendency measurement that better fit this case. In addition, the first and third quantile tell us the range of the 50% of the data at the same time let us have a general idea of the variation in the dataset.

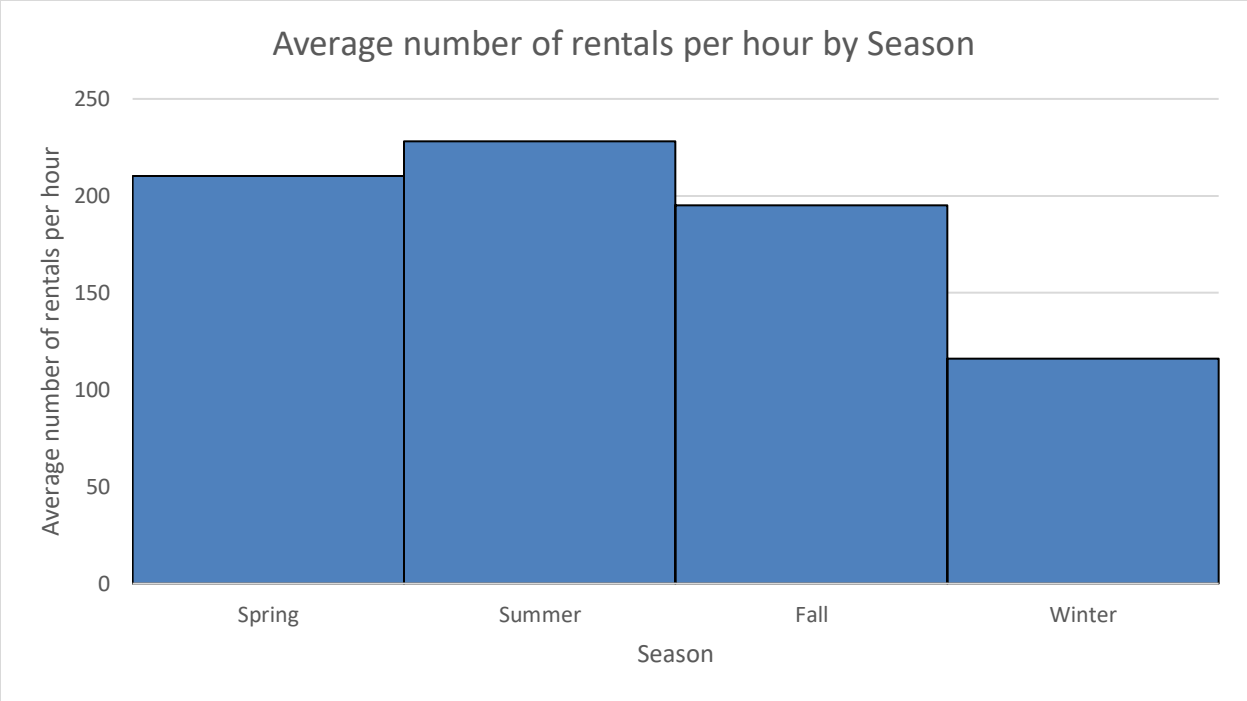
According to the table 1 above, one can conclude that the bike rental number of registered users are way higher than the bike rental number of casual users. Furthermore, there are 50% of the time that the number of casual user rentals are between 4 and 47 per hour. The other 50% of the time are either below 4 rentals per hour or above 47 rentals per hour. For the registered user, there are 50% of the time the rental numbers are between 35 to 215.75 per hour, and the remaining 50% of the time are either below 35 rentals per hour or above 215.75 rentals per hour. One thing to be noticed in the table is that the median of the casual user rentals is closer to the first quantile. This may reflect that the data are skewing left for the casual user rentals, which might be caused by some extremely low hourly rentals.

Chart 1



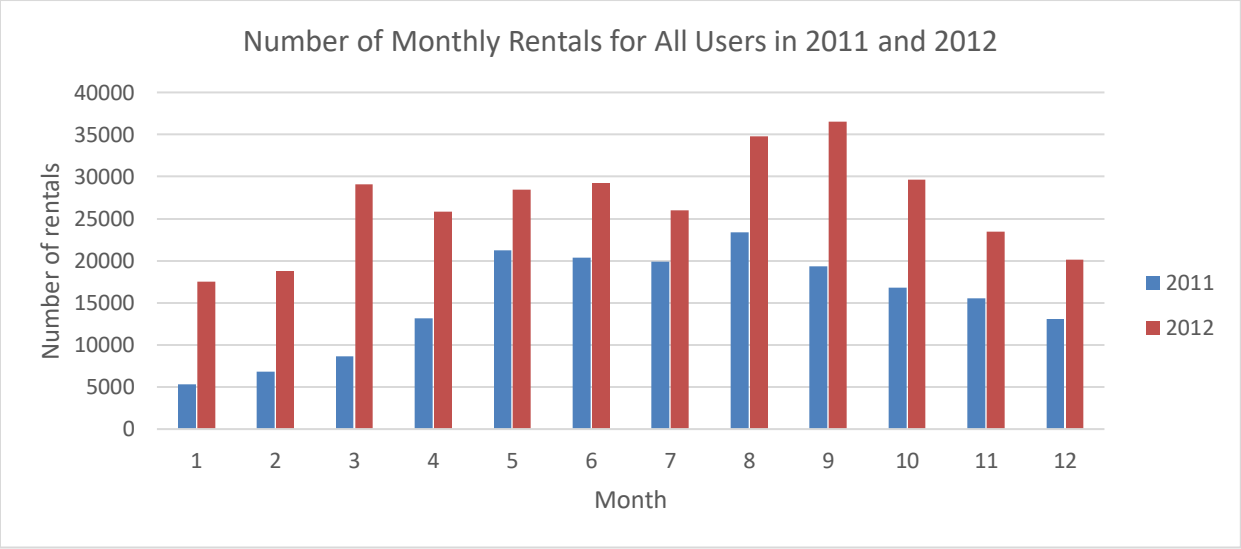
Above is the chart representing the proportion of casual users and registered users to all users. 81% of the users are registered users and 19% of the users are casual users. This might correspond to the result we found in the table 1 that there are more hourly rentals made in the registered users than the hourly rentals made in the casual users.

Chart 2



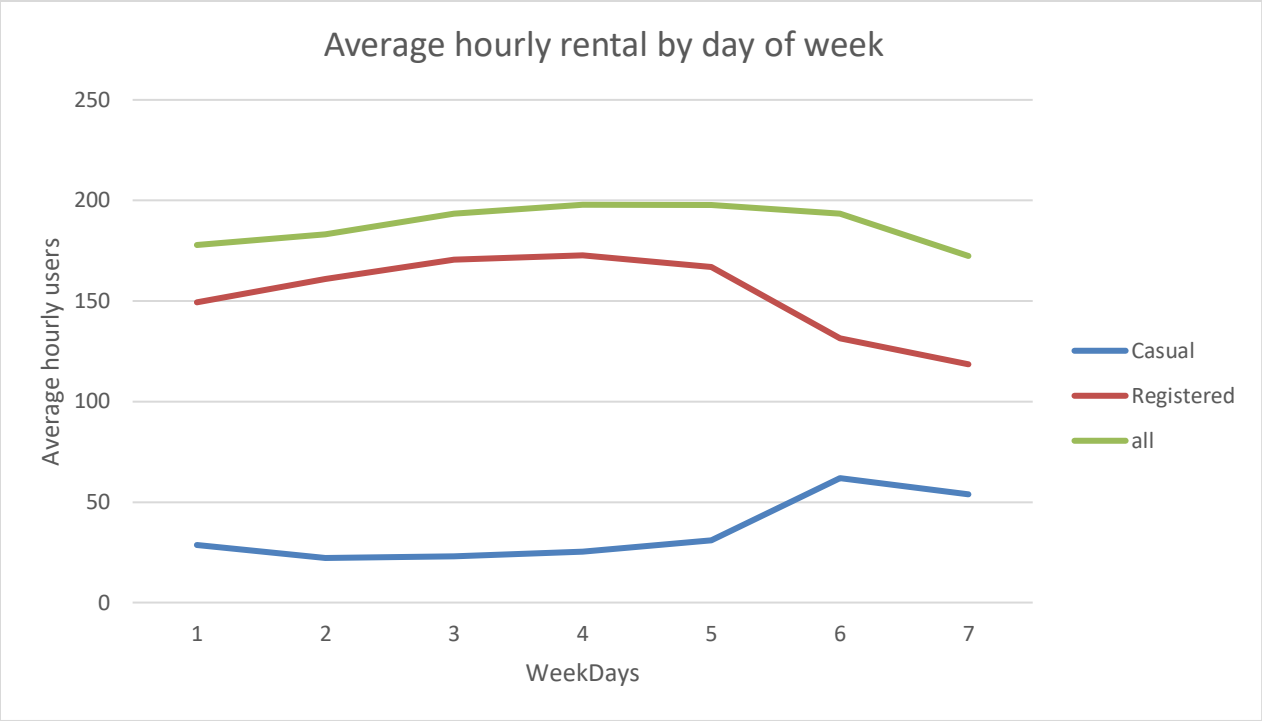
The chart 2 displays the hourly rental number in each season, and it is obvious that winter is the least popular season for people to rent bikes. The possible reason could be it is too cold in the winter so that less people want to go out and take a bike. On the opposite, since the weather in summer is usually sunny and warm, the hourly rental number in the summer is almost double times than the hourly rental number in the winter. And spring and fall have the almost equal amounts of rentals as summer.

Chart 3



According to chart 3 above, we can see apparent increase rental number in 2012 compared with 2011. However, the differences of the rentals numbers between these two years are the most obvious during the January, February, and March of these two years. It might be that the winter in 2012 was warmer than usual. However, further explorations for this situation will to be down in the future work. However, the overall trend of the rental numbers during these two years are similar.

Chart 4



In chart 4, one can see those registered users and all users have the similar trend, which consistently increase during the weekdays (Monday through Friday) but dropped dramatically by around 40% on the weekends. The reason that registered users and all users have the similar trend might be that registered users covered 81% of all users, which means that the variations in registered users partly represent the changes in all users. However, on the opposite, the average hourly rentals for casual users are consistently low during the weekdays, but have uptrend during the weekends, especially in Saturday. In addition, the differences between registered users and casual users are large especially during the weekdays.

Chart 5

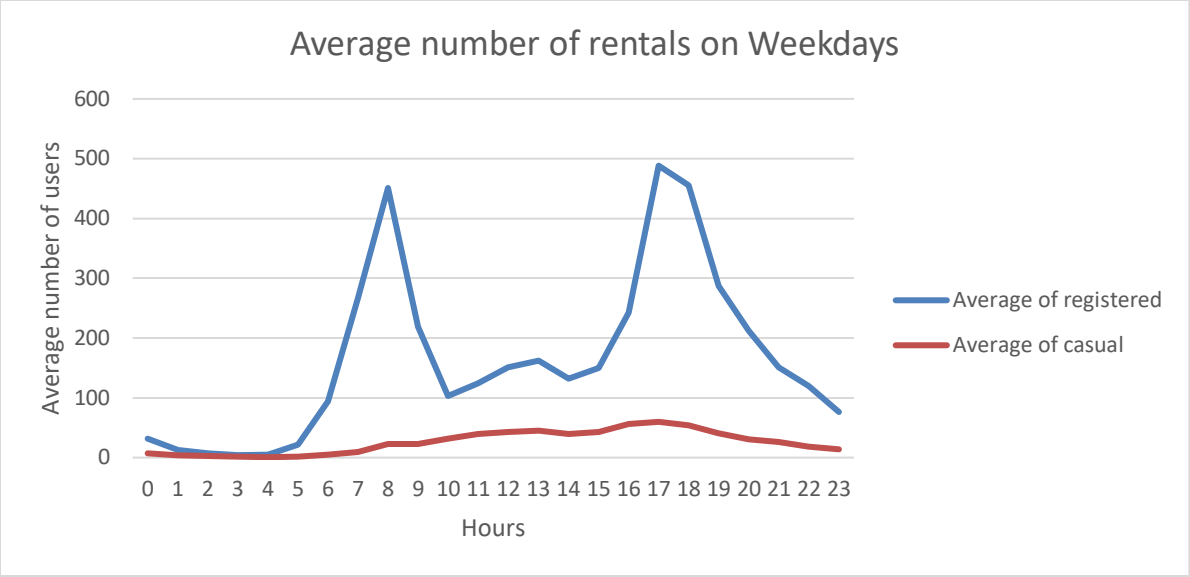


Chart 6

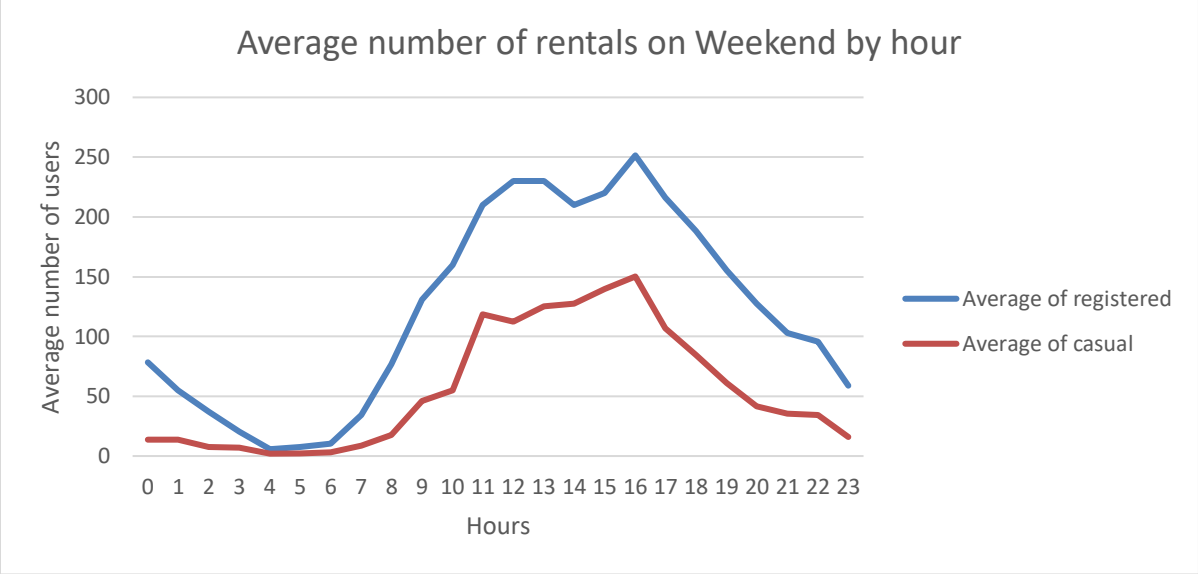


Chart 5 and 6 display the average number of rentals for both registered users and casual users in each hour on weekdays and weekends respectively. More specifically, the average rental numbers reach the highest at 8 am and 17 am, which correspond to the peak hours of the weekdays. One possible explanation would be that the registered user rent the bikes as their transportation to go to or go from their workplaces. On the other hand, the overall trend of registered users and casual users are similar each other on the weekends, which the rental hours focused from noon to 17 pm. In other words, people usually rent bikes in the afternoon on the weekends.

Chart 7

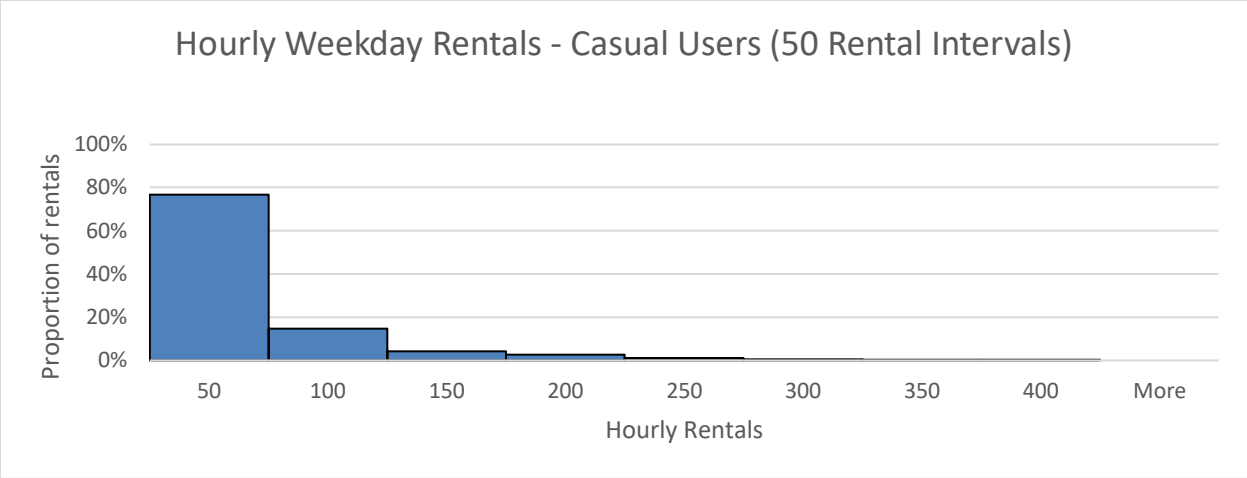
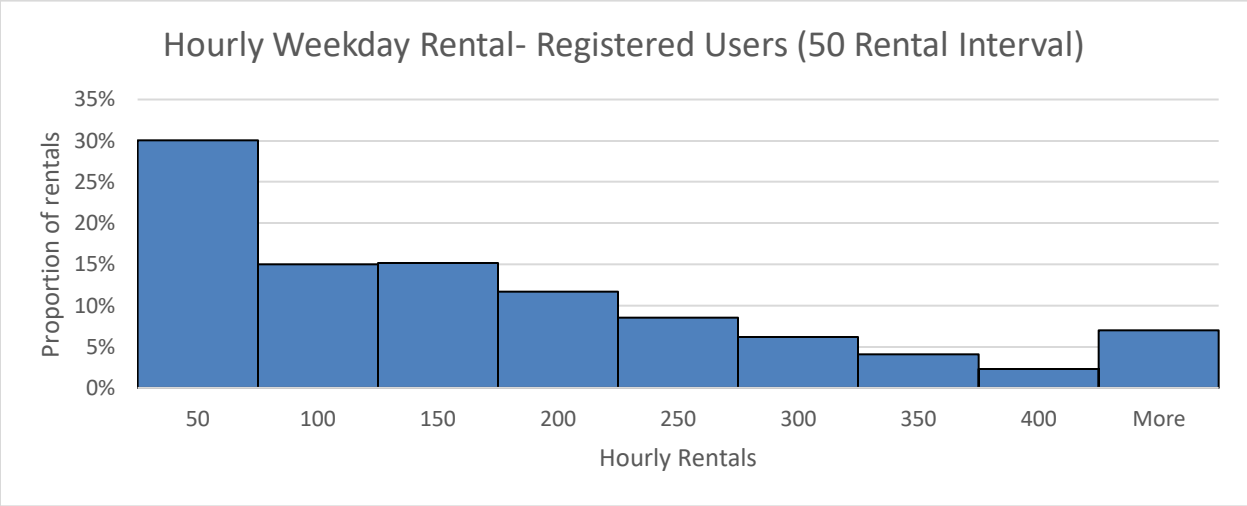
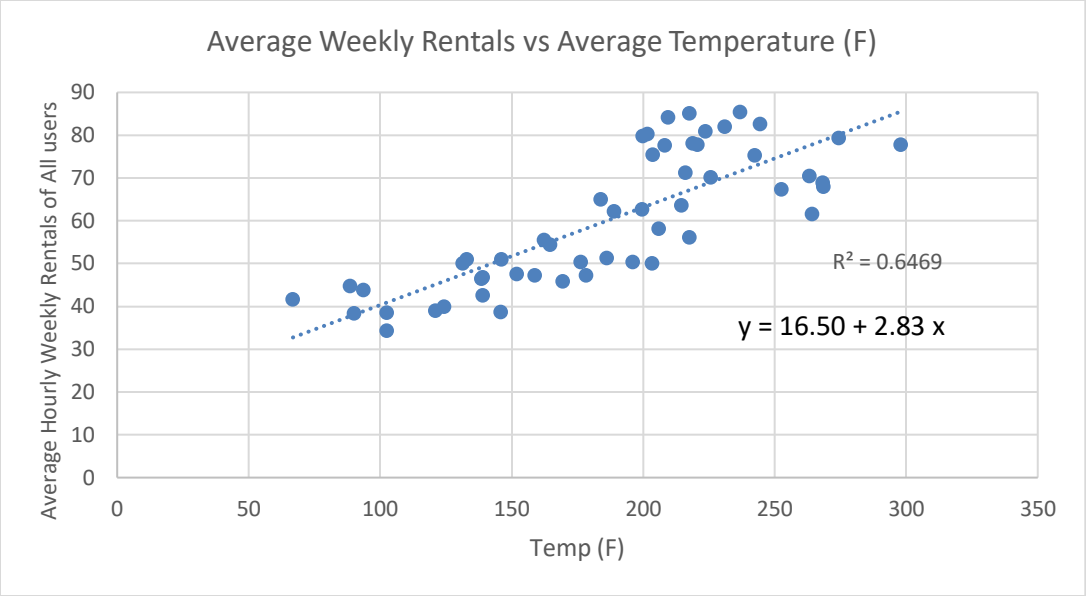


Chart 8



In chart 7, we can see, for casual users in weekdays, around 80% of all hours had 50 or less rentals and the maximum rentals is 300 per hour. Compared with casual users, the registered users have overall larger number of rentals per hour. In detail, 80% of all hours had 300 or less rentals for registered users and the maximum rentals number are beyond 400. In this case, PhillyCycle could take advantages on these charts, as the charts tell us about the potential max demand for bikes each hour. In our cases, the 350 bikes should be sufficed for 80% of the time. However, during the peak times we mentioned in the previous charts, more bikes may be needed. My suggestion would be preparing at least 600 bikes for supply during the peak period of weekdays.

Chart 9



As we can see in the scatterplot above, the average of total users has a relatively strong and positive linear relationship with temperature(F). The particular reason would be that the regression line showed in the graph is upward sloping. In addition, the R-squared value of 0.64 referring that around 64% of the variation in average hourly weekly rentals of all users is explained by the variation in temperature. In the equation displayed in the regression graph, the y stands for the average hourly weekly rentals of all users and x stand for the temperature. In this case, the equation is telling us that with 1-degree increasement in temperature, there are 2.83 increase in the weekly rentals.

Table 2

	Weekday sample mean	Weekend sample mean	Lower bound	Upper bound	Significant or not
All	189.85	183.62	-8.58	21.03	No
Casual	25.89	58.18	-37.59	-26.99	Yes
Registered	163.96	125.44	27.91	49.12	Yes

Before going into the detail conclusion of the above table, we need to clarify the two hypothesis tests for this case. The null hypothesis would be that the average number of rentals by users in the user group on weekends is not significantly different from the rentals by these users on weekdays. And the alternative hypothesis would be there is significant difference of average rental number for weekdays versus weekends.

Above is the table that contains the information of the sample means for both weekdays and weekends as well as the lower and upper bound of 95% confidence intervals for users in each group (all users, casual users, and registered users). In general, the average number in weekday for all users is 189.85 rentals per hour and 183.62 in weekend. Since 0 falls between the 95% confidence interval of [-8.58, 21.03], it is fair to say that the difference in average hourly rentals between weekdays and weekends is not significant. In other words, there is no evidence to

support the null hypothesis that there is a difference between the number of rentals on weekdays versus weekends.

One the opposite with all users, both casual users and registered users shows different results. More specifically, the sample means of casual users on weekdays and weekends are 25.89 and 58.18 respectively. And the corresponding 95% confidence interval is [-37.59, -26.99], which does not contain 0. In this case, we can say that 95% of the data shows that there are significant differences between the two sample means. Similarly, for registered users, the sample means for weekdays and weekends are 163.96 and 125.44 respectively. And 0 does not fall between the confidence interval of [27.91, 49.12]. Thus, same as casual users, we can say that we are 95% confidence that there are significant differences between the sample mean of weekdays and the sample mean of weekends.

To sum up, 95% of the data tell us that the difference between the average rental number for all users on weekdays and weekends is not significant. On the other hand, we are 95% confidence with the result that the differences between the average rental number for both casual and registered users are significant.

Regression Analysis

Table 3

Regression Statistics	
Multiple R	0.57759606
R Square	0.33361721
Adjusted R Square	0.33111671
Standard Error	149.212242
Observations	2676

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-15.73	19.51	-0.81	0.42	-54.00	22.53
humidity	-2.62	0.18	-14.53	0.00	-2.97	-2.27
windspeed	0.64	0.37	1.71	0.09	-0.09	1.37
weekend	2.56	6.40	0.40	0.69	-9.98	15.10
mist	16.11	7.04	2.29	0.02	2.30	29.93
precipitation	3.07	11.70	0.26	0.79	-19.88	26.02
spring	-8.21	10.35	-0.79	0.43	-28.49	12.08
summer	-60.78	13.14	-4.63	0.00	-86.54	-35.01
fall	50.08	8.99	5.57	0.00	32.44	67.72
year 2012	86.21	5.80	14.86	0.00	74.84	97.59
Temp	5.27	0.28	18.87	0.00	4.72	5.81

Table 3 displays the result for a multiple regression analysis. The predictor in this multiple regression model is the hourly rental number for all users (all_users) and the estimators are temperature (F), humidity, windspeed (significant at 10% level), dummy variables for seasons (Spring, Summer, and Fall), year(year_2012), weekend, and weather (mist and precipitation). The R-squared value is 0.33, which means that 33% of the variation in the average hourly rentals is explained in this model. The equation for the regression model is as following:

Y (all users) = -15.75 – 2.62*humidity + 0.64*windspeed + 2.56*weekend + 16.11*mist + 3.07*precipitation – 8.21*spring – 60.78*summer + 50.08*fall + 86.21*year 2012 + 5.27*Temp

To interpret the model, we can need to first take a look at p value for each variable. By only looking at the p value in the table, we found that humidity(significant at 1% level), windspeed (significant at 10% level), mist (significant at 5% level), summer(significant at 1% level), fall(significant at 1% level), year 2012(significant at 1% level), and temp(significant at 1% level), have significant effect on number of hourly rentals made. Since the other few variables are not statistically importance, we will not consider their coefficients while interpreting the regression model. Holding everything constant, for every 1 unit increase in humidity, the average hourly rental decreases 2.62; for every 1 unit increase in windspeed, the average hourly rental increase 0.64; for every 1 degree increase in temperature(F), the average hourly rental increase 5.27. In addition, holding everything else constant, if it is a mist day, there is a 16.11 increase in the average hourly rentals; if it is a summer day, there is a 60.78 decrease in the average hourly rentals; it is a fall day, there is a 50.08 increase in the average hourly rentals; if it is in year 2012, there is an 86.21 increase in the average hourly rentals.

One thing to be noticed in this regression model is that, unlike what we expected, the coefficient for summer is negative, which means that the model is telling us that there is less bike rental in summer. However, if we take a closer look, we may find that temperature partly explains the variations in summer. In this case, the possible reason that summer has a small p value could be that the p value for temperature is small, and summer may not be a fit estimator in this regression model while temperature is in the model at the same time. Another way to explain it is that when holding all other variables constant, the increase in number of rentals due to temperature is larger than the decrease in number of rentals due to summer. For example, if it is a 90-degree (F) summer day, there will be 90*5.27 – 60.78 = 413.52 increases in the average hourly bike rentals, instead of a decrease in rental number.

Conclusion

In this case, we aim to analyze the data provided by PhillyCycle, a bike sharing company, and provide corresponding interpretation for the analysis in order to help the company understand the customer behaviors so that they could make better decisions on the expansion strategies for bike rentals. According to our analysis above, the registered users cover most of bike rentals, so PhillyCycle should encourage users to register and become a member. One possible way to attract more people to become registered users is to reduce the payment of each bike rental for the registered users, instead, the company could charge a monthly or annually membership fee. In addition, we found that there are more rentals in demand at 8 am and 5 pm on weekdays and

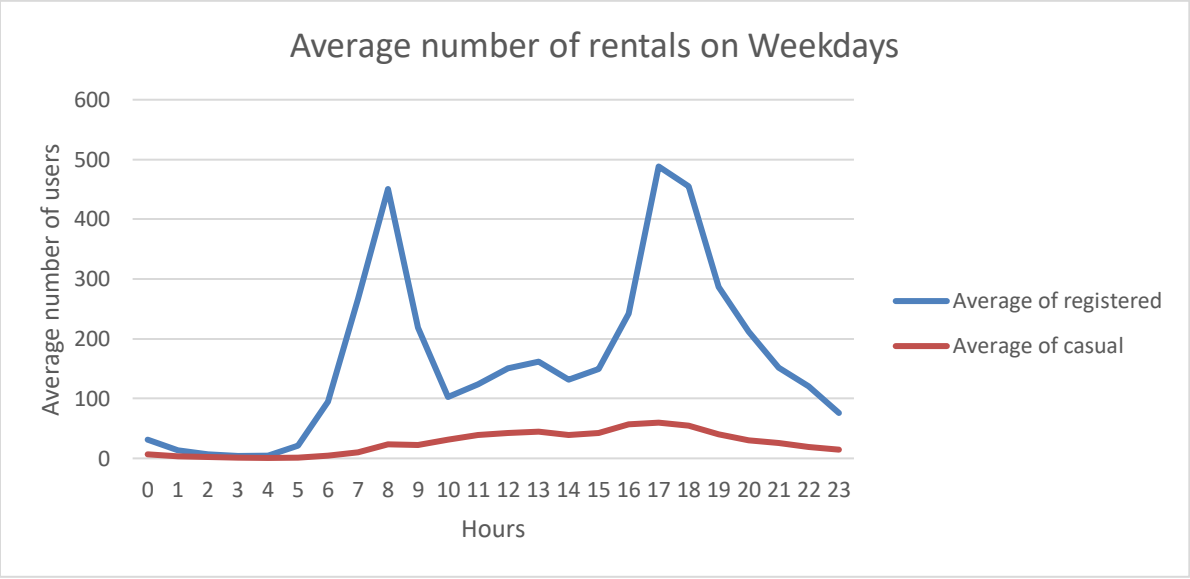
in the afternoon on the weekends, which could be the time to slightly increase the rental price. Also, weather is an important factor in our analysis. Thus, I recommend forecast the weather on each Sunday to see the overall weather for the next week and then vary the price based on the weather. For example, if it is a mist day, then the company should slightly increase the price since there might be more rentals. Another suggestion for the long-term profit is to open new stores in the warm place with low humidity such as California.

Appendix

Notes on Data Preparation

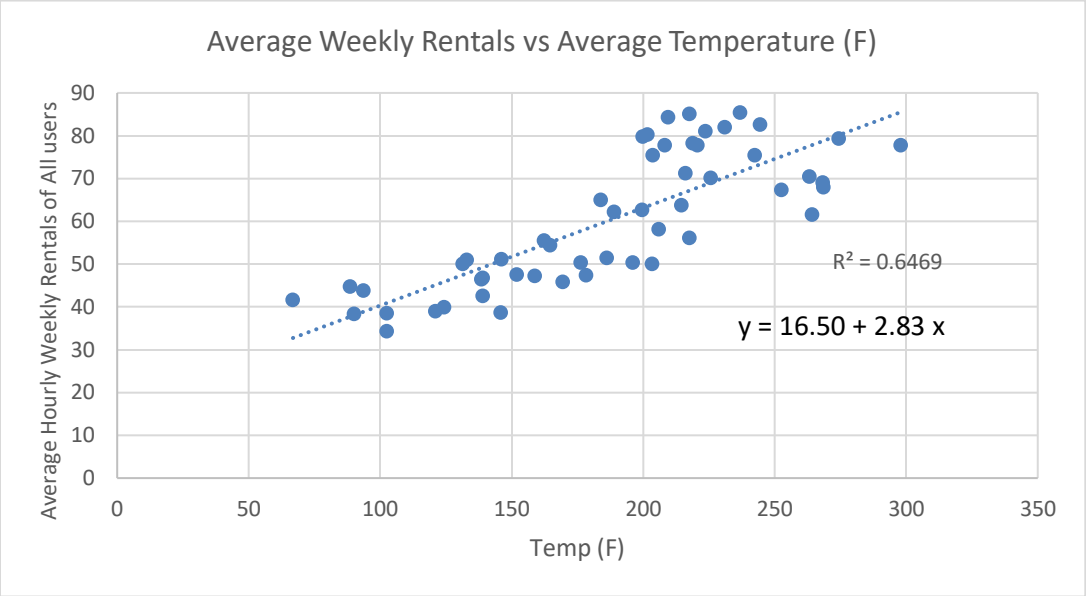
The sample data was provided by PhillyCycle. Each entry in the dataset contains information about the total number of rentals in the given hour by type of user from 2011 to 2012, which are our major predictors. Other variables include season, weather, temperature, humidity, wind speed, and weekday, which are all considered as our estimators. We also created some new variables based on existed variables. More specifically, we first create variable year, month, day of week, hour, and week number based on the date time in the origin dataset. Then we added dummy variables for the categories in season, year, and weather to help with our later regression analysis. At last, we created a new temperature variable by converting the original data from Celsius to Fahrenheit. After delete all the duplicate entries and any entries with temperature equal or higher than 122 degrees, we had 2676 remaining entries. The cleaned and reorganized data provided us a clear and complete overview of what happened in 2011 and 2012, which allowed us to have a smooth process with our analysis.

Elevator Charts



This chart is important as it tells us, on weekdays, at what hours of the day that registered users rent bikes. As we can see that 8 am and 5 pm are two peak time for registered users since the average hourly rentals number reaches the highest. This information helps the executers of PhillyCycle to find suitable customers and to encourage them become the registered users. For

example, those who leave a little bit far from work and do not have a car would be possible target customers.



This is a scatterplot with an upward sloping trend line. It is useful since it tells us that temperature and average hourly weekly rentals of all users are positively related with each other. In other words, the higher the temperature (within 122 Fahrenheit), the more rentals will be made. More specifically, regardless of all other factors, with 1 degree increase in temperature (F), there is 2.82 increase in average bike rental.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-15.73	19.51	-0.81	0.42	-54.00	22.53
humidity	-2.62	0.18	-14.53	0.00	-2.97	-2.27
windspeed	0.64	0.37	1.71	0.09	-0.09	1.37
weekend	2.56	6.40	0.40	0.69	-9.98	15.10
mist	16.11	7.04	2.29	0.02	2.30	29.93
precipitation	3.07	11.70	0.26	0.79	-19.88	26.02
spring	-8.21	10.35	-0.79	0.43	-28.49	12.08
summer	-60.78	13.14	-4.63	0.00	-86.54	-35.01
fall	50.08	8.99	5.57	0.00	32.44	67.72
year 2012	86.21	5.80	14.86	0.00	74.84	97.59
converting	5.27	0.28	18.87	0.00	4.72	5.81

This is the result table for our multiple regression analysis. We only focused on the features that has small p value, which were all highlighted above. This is an important table as it displays all the variables that could have possible influence on average hourly rentals both positively and negatively. It also provides the PhillyCycle company the information of the places that are suitable for them to expand their company. For example, a city that has long warm weather and low humidity would be a great place to develop their company.