Problem Set 2

Thao Duyen Tran & Uyen Ha

MISM6202: Foundations of Data Analysis

Professor Kate Ashley

Sunday, December 5, 2021

◆ ◆ ◆

Data Omission

When inspecting the data, we can see that there are two days in which there were no rides taken and no revenue generated. This is because those two days were federal holidays (Martin Luther King Day and President's Day), and CVE most likely suspended its services then. For the entirety of the problem set, we will eliminate those two days from our data, as they can skew the models created (since some users were still accessing and clicking on the app, despite not being able to book a ride.

◆ ◆ ◆

Regression Analysis

1. Because customers value flexibility in their commuting plans, CVE allows customers to cancel a booking without penalty up until the van they booked arrives at their chosen stop. As a result, not all ride bookings result in a ride actually taking place. Estimate a simple linear regression model to understand the relationship between daily bookings and daily completed rides. Report the estimated regression equation and R2 value and interpret them in words.

There are two ways to approach this analysis. We are can either: 1) use the rides column (number of rides completed) and regress it with booked.rides (number of bookings) or 2) create a new column with completed rides in app, by taking booked.rides and subtracting it by number of cancellations (trip.cancelled). The first method will provide a more comprehensive understanding of how many rides are booked on both the CVE website and app, while the second method will be more specific to the app.

When we regress number of completed rides (rides) on number of bookings (booked.ride), are given a regression output which includes:

Variable	Value	P-value
Intercept	364.837	<0.000
booked.ride	0.831	<0.000
Multiple R^2	0.939	
F-statistic	927.8 on 1 and 60 DF	<0.000

Regression Equation

$$Number\ of\ completed\ rides = 0.831(number\ of\ bookings) + 364.87$$

The booked.ride coefficient means that, for every 1 increase in booking, the average number of daily completed rides will increase by 0.83. The p-value of the booked.ride coefficient is also extremely low, indicating that this variable is statistically significant in predicting number of completed rides; there is less than a 0.0001% chance that the observed interaction between booked.ride and rides is the result of pure chance (which makes sense, considering that, in order to complete a ride, there must first be a ride booked). Also, an R^2 value of 0.939, meaning that 93.9% of the variation in our fit is explained by the regression model, which is extremely good.

Our intercept coefficient is rather interesting. The intercept is 364.837, meaning that the expected number of completed rides daily is around 365, holding the predictor (booked.ride, in this case) at 0. This intercept does not make any real-world sense and can be more-or-less ignored, especially since our predictor coefficients have extremely low p-values, indicating that they are strongly related to the outcome. That said, we should still include the intercept in our predictions, since omitting it would be disingenuous.

When we create a new variable for app completed rides (app.rides), which is booked.ride subtracted by trip.cancelled, we will acquire new values that are different from what was provided by the rides column. This is due to the fact that the rides column includes completed rides for both the website and app, while our new variable will only include completed rides for the app. The regression output values are as follow:

Variable	Value	P-value
Intercept	-117.143	0.143
booked.ride	0.762	<0.000
Multiple R^2	0.937	
F-statistic	892 on 1 and 60 DF	<0.000
Regression Equation		
$Number\ of\ completed\ app\ rides = 0.762(number\ of\ bookings) - 117.143$		

In this case, the coefficient for booked.ride is a little lower, so that the average number of completed app rides will only increase by 0.762 per point increase in app bookings. Like the model above, the p-value for this coefficient remains to be extremely low (it is less than <0.000, just like the previous coefficient), which indicates strong statistical significance. The R^2 value is also very similar, it is only 0.002 less than the previous model had (meaning that this model can still explain 93.7% of the variation in the data).

The intercept is where this model truly differs from the previous one, as it is a negative value this time. This implies that the value that our completed rides will take, given that booked.ride is 0, is -117. This, again, does not make any real-world sense. Furthermore, the p-value of our F-stat is <0.000, indicating that we can reject the null hypothesis on all general alpha level and conclude that our model is significant.

The residual standard error for our model is 90.88 on 60 DF, and it tells us that the regression model predicts the value of completed app rides with an average error of 90.88. Considering that values of app.rides range from 1530 to 2864 on 60 DF, the residual standard error of 90.88 is insignificant. We can reasonably believe that the regression model fits our dataset well.

The plot visualizing the relationship between completed rides and number of bookings can be found in Appendix A. While the plot for completed app rides and number of bookings can be found in Appendix B. Both models show that there is a positive linear relationship between the number of bookings and the number of completed rides, regardless of whether it is exclusively an in-app booking or not.

2. CVE would like to know if ride bookings through the mobile app can be predicted using the actions that an app user may perform prior to booking: namely, starting a session, tapping on the sidebar, tapping on a stop, and viewing van ETAs. Estimate a multiple regression model that uses the relevant variables to predict ride bookings. Multiple models involving these variables are possible; select the best model and explain your choice, citing specific numerical evidence from the regression output. Report the estimated regression equation and R^2 value and interpret them.

The regression output with ride bookings (booked.ride) as an outcome and the four aforementioned variables (starts.session, tapped.sidebar, tapped.on.stop, viewed.eta) as predictors includes:

Variable	Value	P-value
Intercept	314.780	0.026
starts.session	0.220	<0.000
tapped.sidebar	-0.565	<0.000
tapped.on.stop	0.019	0.680
viewed.eta	-0.161	<0.000
Multiple R^2	0.915	
Adjusted R^2	0.909	
F-statistic	153.5 on 4 and 57 DF	<0.000

Regression Equation

$$\boldsymbol{Number\ of\ bookings} = 0.22(starts.session) + 0.02(tapped.on.stop) - 0.57(tapped.sidebar) - 0.16(viewed.eta) + 314.78$$

Once again, the intercept coefficient for ride bookings tells us that the number of bookings daily would be 314 rides, given that all other predictors are 0. However, this, again, does not make sense since it should be impossible to book a ride through the app without starting a session on the app (so if starts.session is 0, then booked.ride will also be 0). Since our predictors are statistically significant, we can essentially ignore this intercept value during interpretation (but not during prediction). Also, since our predictor can never be 0 (because there always has to be a session started for a ride to be booked through the app), the intercept is more-or-less meaningless. If we were to run a regression with just booked.ride and starts.session, the p-value for the intercept will be 0.272, while the p-value for the booked.ride coefficient will be <0.000, implying that the intercept can be ignored in favor of the significance in predictor.

For our predictor coefficients, we can explain it as: every time a user starts a session in the app, or taps a specific stop, the average number of bookings is expected to increase by 0.22 and 0.02, respectively. Conversely, every time a user taps a sidebar, or views the ETA, the average number of bookings decreases by 0.57 and 0.16, respectively. With the exception of tapped.on.stop, all other predictor variables appear to be significant, with extremely low p-values.

The adjusted R^2 (R^2 that is modified in order to account for the increase in number of predictors) for this model is 0.96, which is quite impressive as it means that the regression model generated can explain 96% of the variance in the data.

While the model we had was quite nice, and has a fantastic adjusted R^2, we can still look to build an even better model by removing the variable tapped.on.stop (which was not statistically significant). The new regression model has:

Variable	Value	P-value
Intercept	315.694	0.246
starts.session	0.225	<0.000
tapped.sidebar	-0.534	<0.000
viewed.eta	-0.174	<0.000
Multiple R^2	0.915	
Adjusted R^2	0.910	
F-statistic	207.5 on 3 and 58 DF	<0.000

Regression Equation

$$\boldsymbol{Number\ of\ bookings} = 0.23(starts.session) - 0.53(tapped.sidebar) - 0.17(viewed.eta) + 315.69$$

We will not go into depth on the interpretation of these numbers, as the interpretation is the same as the previous model, but with different inputs. None of the predictor coefficients saw a drastic change in value or signage, and the adjusted R^2 is just a little bit higher (if we were to round the number to 2 significant figures, they would both be 0.91).

The second model is the best model to use, despite it only having a 0.001 higher adjusted R^2, since we also removed a variable that is not statistically significant from our model.

It is worth noting that both of these models have a relatively high F-statistic (165.5 on 1 and 60 DF for the first model and 207.5 on 3 and 58 DF for the second) and very low p-values (both are <0.000), which strongly suggests that we should reject the null hypothesis.

◆ ◆ ◆

Forecasting

3. Create a well-formatted and labeled scatter plot to visually inspect the 'rides' variable. Describe any trend and seasonality that appear to be present.

We inspected the daily number of completed rides (rides) and its relationship with these predictors:

- Time period index (t), colored by day of the week
- Time (t), colored by month
- Revenue (revenue), colored by day of the week
- Completed rides in the app (app.rides), colored by day of the week

We also aggregated the daily number of completed rides into weeks (1-13) and summed up the numbers to get values for weekly completed rides (instead of daily). We then made scatterplots to see:

- The total number of completed rides per week
- The total number of completed rides and total revenue per week
- The total number of completed rides and revenue per ride (weekly)
    - This was also plotted again but laid out in a more chronological way (where x=time)

Some notable insights we found were:

- There is a general increase in number of bookings, and subsequently, number of completed rides as time progresses, implying that more people are using the shuttle service over time.
- Friday is the least popular day for ride completion, while Monday and Tuesday are the most popular.
    - There are many reasons why this could be the case, for instance, Mondays and Tuesdays might be when public transportation is delayed most often, or perhaps people want to get to work early at the start of the week and do not want to experience the commuter rush. However, these speculations are difficult to provide evidence for, since we are not given details about the social context.
- Weeks 3, 7, and 13 generated less revenue. However, this can be explained by the fact that all these weeks are missing a day (Weeks 3 and 7 both had a Monday with no revenue due to it being a federal holiday, while Week 13 does not have data for Friday).
- Weeks 1-3 actually generated the most revenue per ride, and there is a sharp decrease in revenue per ride in Week 4.
    - The revenue per ride will see a slight recovery on some weeks but will drop drastically again in Weeks 8 and 13.
    - This could be the result of returning customers choosing to purchase discount packages/coupons instead of paying full price per seat, which would explain the increase in number of completed rides and subsequent decrease in revenue per ride.

(These plots can be found in Appendices C to J)

◆ ◆ ◆

4. Construct a k-period simple moving average for the rides variable, where k is chosen based on your assessment of the seasonality patterns in the data. Explain your choice of k and report MSE, MAD, and MAPE for this forecasting model.

To find the appropriate k-period, we first plotted the completed rides variable (rides) against the time index variable (t), which is what was done for Appendices C and D. However, this time we will make the plot more geared towards representing a k-period trend. This plot can be found in Appendix K. This plot shows us that, after every 5 period indexes (5 days), the seasonality pattern will repeat in an almost identical order (where Monday is the most popular, with decreasing popularity every day, until Friday is the least).

With this information, we will construct our k-period simple moving average for rides with k=5. The table below contains the rides and 5-days average for our first 9 observations.

Date	Completed Rides (rides)	5-days Average
2016-01-04	2654	N/A
2016-01-05	2658	N/A
2016-01-06	2669	N/A
2016-01-07	2636	N/A
2016-01-08	2166	2556.6
2016-01-11	2763	2578.4
2016-01-12	2673	2581.4
2016-01-13	2657	2579.0
2016-01-14	2601	2572.0

We can then calculate the mean squared error (MSE), mean absolute deviation (MAD) and mean absolute percentage error (MAPE) to evaluate the model performance. The results are as follow:

Metric	Value
MSE	104401.3
MAD	260.079
MAPE	0.095 (9.5%)

Considering that the actual number of daily rides ranges from 2113 to 3562, our MSE, MAD and MAPE values demonstrate that our prediction model has high accuracy and low error. We also re-plotted the graph we made earlier, but with a 5-day simple moving average line, which can be found in Appendix L.

◆ ◆ ◆

5. Estimate a linear trend model for the 'rides' variable. Report the estimated linear trend equation and the R^2 of the model and interpret both the equation and the R^2 in words.

Our linear trend model for rides (plotted against the time series index) yielded the following results:

Variable	Value	P-value
Intercept	2526.171	<0.000
Time series index (t)	11.607	<0.000
Multiple R^2	0.309	
F-statistic	26.83 on 1 and 60 DF	<0.000

Regression Equation

$$\boldsymbol{Number\ of\ completed\ rides = 11.61(time\ index) + 2526.17}$$

Similar to the previous models, we can ignore this intercept value, as it does not make any real-world sense, and we cannot reasonably expect our time coefficient to be 0 (as its p-value is very low, which points towards a strong relationship between the predictor and outcome). While our intercept value does not make sense in terms of interpretation, it is still very important that we include it in our predictions, as it is an integral part of the regression equation. As for our t coefficient, the value tells us that, for every point increase in t, we should expect to see a subsequent average increase of 11.61 for completed rides.

Since this is a linear trend model and only includes one predictor, we will be looking at the multiple R^2, as opposed to the adjusted R^2. Our R^2 value is, unfortunately, rather low; a value of 0.309 means that the model generated can explain about 31% of the variance in the data. Our R^2 is quite low because it cannot accurately capture the unusually low number of rides happening during Fridays, as well as the differences in ride numbers between the other weekdays (which can be seen more clearly in Appendix C).

◆ ◆ ◆

6. Estimate a linear trend model with day-of-week dummy variables for the 'rides' variable. Interpret both the estimated regression equation and the R^2 in words, and comment on the magnitude of the adjusted R^2 relative to the adjusted R^2 from the regression you performed in (5).

We transformed our dayofweek into dummy variables in a way that:

- Monday is represented as d1
- Tuesday is represented as d2
- Wednesday is represented as d3
- Thursday is represented as d4
- Friday will not have a dummy variable since it is our reference category

When we run our linear trend model, regressing rides with the dummy variables and time index (t), we are given this output:

Variable	Value	P-value
Intercept (b0)	1940.184	<0.000
Monday (d1)	827.433	<0.000
Tuesday (d2)	841.474	<0.000
Wednesday (d3)	693.834	<0.000
Thursday (d4)	554.039	<0.000
Time series index (t)	11.641	<0.000
Adjusted R^2	0.923	
F-statistic	146.2 on 5 and 56 DF	<0.000

Regression Equation

$$\boldsymbol{Overall\ trend =}$$
$$\boldsymbol{827.43(d1) + 841.47(d2) + 693.83(d3) + 554.04(d4) + 11.64(time) + 1940.18 + error}$$

$$\boldsymbol{Day\ of\ week\ trend =}$$
$$\boldsymbol{(1940.18 + dx) + 11.64(time)}$$

While the time index coefficient remained the same (at 11.64), we can see stark differences in values for the intercept coefficient, adjusted R^2, and F-statistic. With the introduction of dummy variables, our intercept value has dropped, from 2526 to 1940, however, this still is not very realistic to interpret, as all our predictor variables are statistically significant. Again, we do not have to worry about the intercept since, looking at the p-value of the F-statistic (which is lower than any general alpha level), we can reject the null hypothesis and conclude that the prediction model is meaningful. In other words, our predictors will not be equal to zero and the intercept would not be significant independently.

For our predictor coefficients, we now have 4 dummy variables to represent the days from Monday to Thursday. The coefficients imply that, when it is either a Monday, Tuesday, Wednesday, or Thursday, the expected number of completed rides will increase by either 827.43, 841.47, 693.83, or 554.03, respectively. Since Friday is our reference variable, it does not have an assigned coefficient (its value is 0). If we wanted to forecast Friday's completed rides, we would simply do $1940.18 + 11.64(time)$. The coefficient of time period (t) tells us that, for every 1 unit increase in time period, an increase of 11.64 in the number of daily completed rides will follow.

Our adjusted R^2, which has increased drastically, indicates that our model is much more reliable. An adjusted R^2 of 0.923 means that our model is able to now capture 92.3% of the variance in the data, which is a big step up to the previous 30.9%.

◆ ◆ ◆

7a. Use the estimated regression equation from (6) to calculate a forecast of 'rides' for each day in your dataset. Calculate MSE, MAD, and MAPE for this forecast. Comment on which of the two forecasts you have calculated in this problem set (from Q4 and this question) performs the best and why that method is best suited to this data.

The table below contains our reported MSE, MAD, and MAPE for the forecast of 'rides' variable (Q6), compared with the metric values obtained from Question 4 (Q4).

Metric	Value Q6	Value Q4
MSE	10691.97	104401.3
MAD	78.51485	260.079
MAPE	0.0281 (2.81%)	0.095 (9.5%)

Compared with the Question 4 metrics, we can clearly see that the values of MSE, MAD and MAPE of the Q6 model much smaller. These number indicate that errors made by the prediction model in Q6 are rather insignificant, and so it performs better predictions for the outcome variable. In conclusion, we are confident that the prediction using seasonal dummy variables performs the best in forecasting daily completed rides as opposed to the model using 5-day moving average in Q4. This is the case as there is a very prominent seasonal trend that can be captured rather well by seasonal dummy variables, but not by 5-day moving average.

7b. Use the estimated regression equation from (7a) to forecast daily completed rides for each weekday in the next month (April 1-April 29). Optional: Also forecast revenues for each day.

To tackle this question, we first created a new dataset ("newdata") that includes t, dayofweek, rides, revenue and seasonal dummy variables (d1, d2, d3, d4) from our original dataset. Since there are 29 days from April 1 to April 29, and we skip Saturdays and Sundays for the entire month, the new data set will have 21 rows. Since, in our original dataset, March 31st is a Thursday, we can assume that the new data month will start on Friday.

After calculate prediction for the number of completed rides of each day in the month of April based on the model built in Q6, we can then visualize the results in 2 scatterplots, which are found in Appendices N and O. The predicted values fall in a consistent uptrend with previous values. We are confident that the prediction of completed rides made by this model is highly reliable.

To predict revenue, we ran a linear regression of revenue along with time period index (t), seasonal dummy variables, and number of completed rides. We then ran another regression with just time period t and 'rides' (as these are the significant predictors according to its p-value) so that we can see which model we should use for our prediction. The outputs of both models are below:

Output 1 – Using dummy variables

Variable	Value	P-value
Intercept	1111.293	0.013
rides	3.493	<0.000
Monday (d1)	294.011	0.143
Tuesday (d2)	371.564	0.068
Wednesday (d3)	311.212	0.072
Thursday (d4)	257.900	0.075
Time series index (t)	-6.136	0.036
Adjusted R^2	0.984	
F-statistic	625 on 6 and 55 DF	<0.000

Output 2 – No dummy variables

Variable	Value	P-value
Intercept	434.295	0.023
rides	3.860	<0.000
Time series index (t)	-10.396	<0.000
Adjusted R^2	0.9838	
F-statistic	1856 on 2 and 59 DF	<0.000

Since the R-squared of these two models are almost the same, we decide to keep going with the first model, as including the dummy variables will give us a more comprehensive understanding of revenue.

After we predicted revenue for April, we also plotted the results along with our predicted rides variable. The plots can be found in Appendices P and Q. The data distribution for revenue is almost identical to the distribution of completed rides (which makes sense, as more completed rides will result in more revenue); we can clearly see that, just like completed rides, predicted values of revenue consistently fall on the uptrend.

◆ ◆ ◆

8. Write a concise but thorough 1-2 paragraph summary of the forecasting analysis you performed in this problem set, focusing on the most important findings. In other words, think about the work you did for Q3-Q7 and summarize what you would communicate to CVE to help them better understand their ridership data.

Summary:

The data presented by CVE shows several clear seasonality trends in terms of ridership and revenue, primarily:

- Day of the week differences in completed rides – Friday is the least popular day for rides (by a big margin), while Monday and Tuesday are the most popular days for rides.
- Revenue per ride differences between weeks – While there is a general increase in ridership and total revenue over time, there is also a drop in revenue generated per ride (appendix J)

Since there are two separate time trends occurring in this data (which are illustrated in appendices C and J), it can be difficult to make accurate predictions for upcoming weeks. To make the best possible time trend predictions, we will need to create a model that can capture both the differences between days, as well as the differences over time. With our current models, we are able to visualize the overall daily trend, where Friday has the lowest ridership, and Monday/Tuesday have the highest. The model is also not able to currently capture the changes in ridership if there is a holiday (which would result in 0 rides and 0 revenue generated). The revenue per ride prediction cannot be captured in the previous plots and will require its own descriptive and predictive models in order to visualize properly.

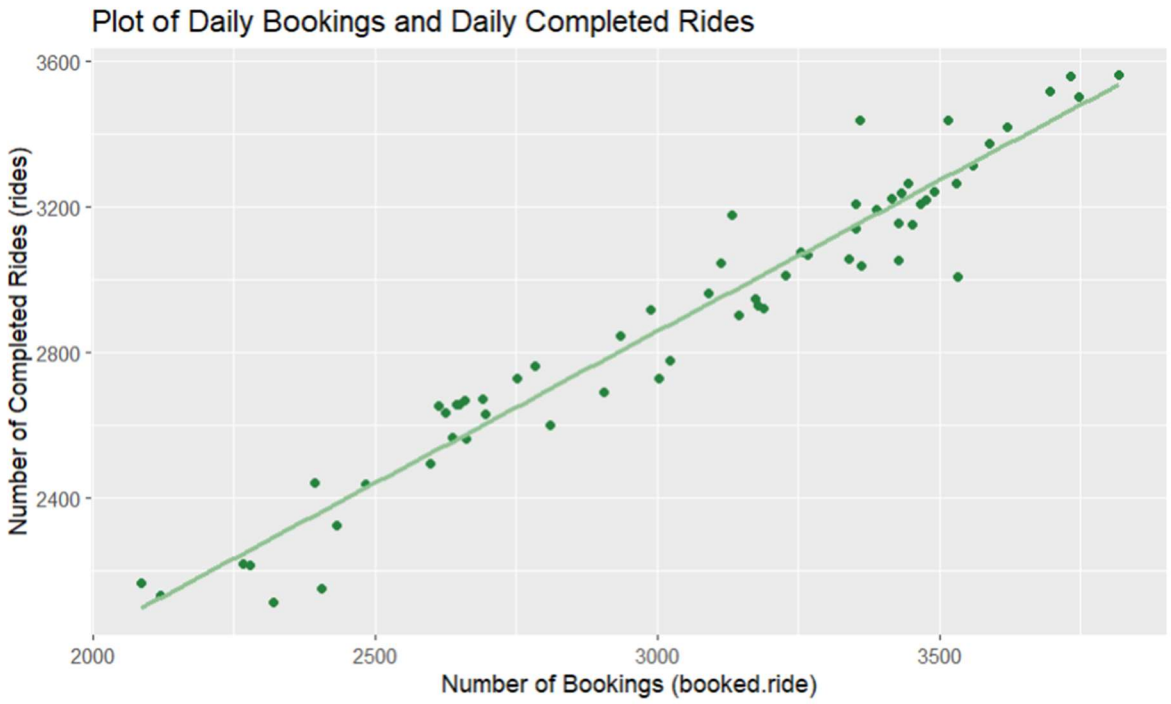
Moreover, in order to make more accurate and actionable predictive models, we will need more data from CVE, either for longer periods of time, or for time periods where there are no federal holidays which can skew the data (because we would have to omit it, impute it, or leave it at 0).

Recommendations:

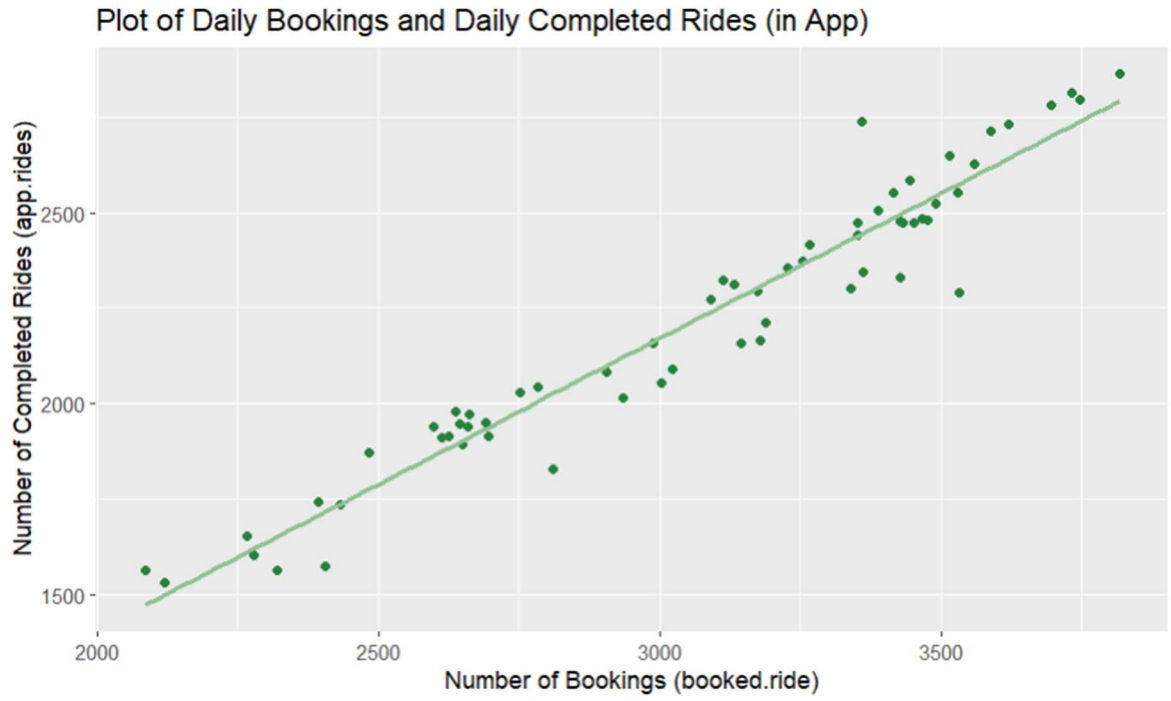
From these insights, CVE can look to focus their resources on these primary areas:

- Implementing weekly/monthly subscriptions, as opposed to discount packages. This can allow CVE to be less affected by ridership fluctuations within the week, instead of allowing customers to use their discount rides on high-traffic days (i.e., Monday/Tuesday).
- Change ride schedules for Friday. Since Friday has substantially less traffic than other days, CVE can look to maximize profits by changing ride availability and schedule on that day (e.g., have the interval between rides longer than high-traffic days).

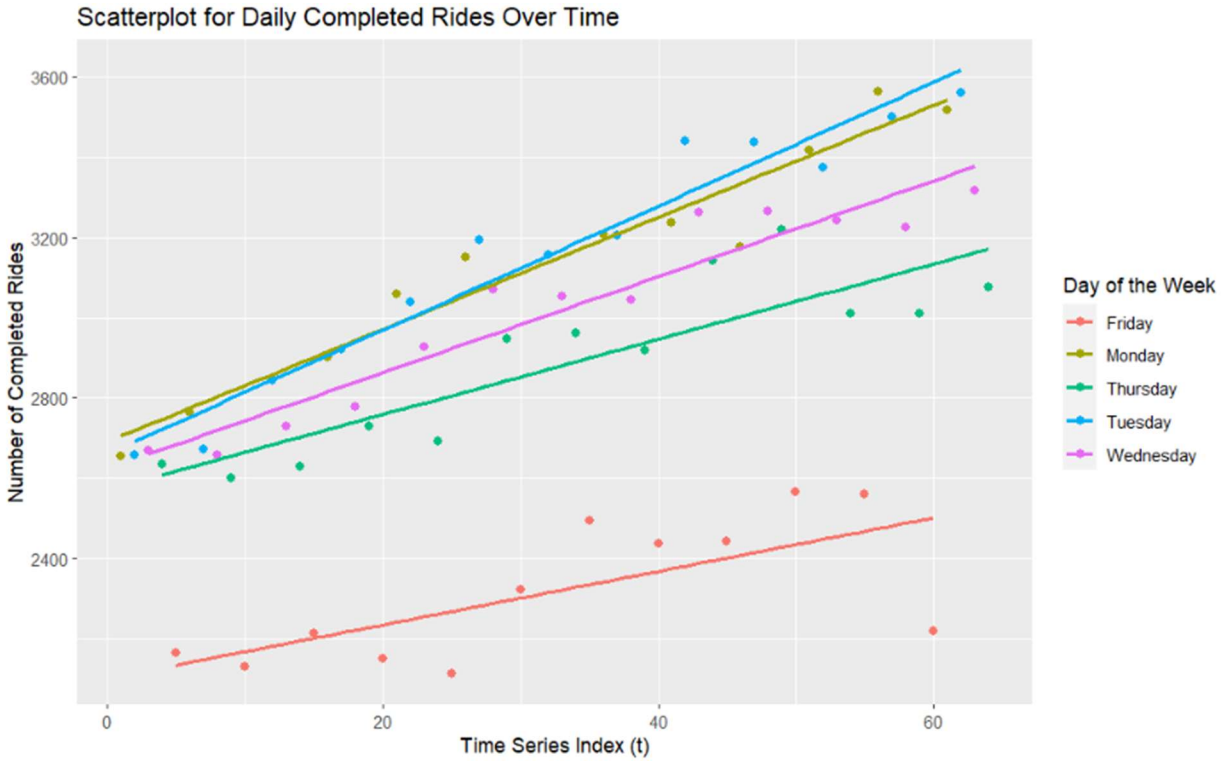
Appendix A



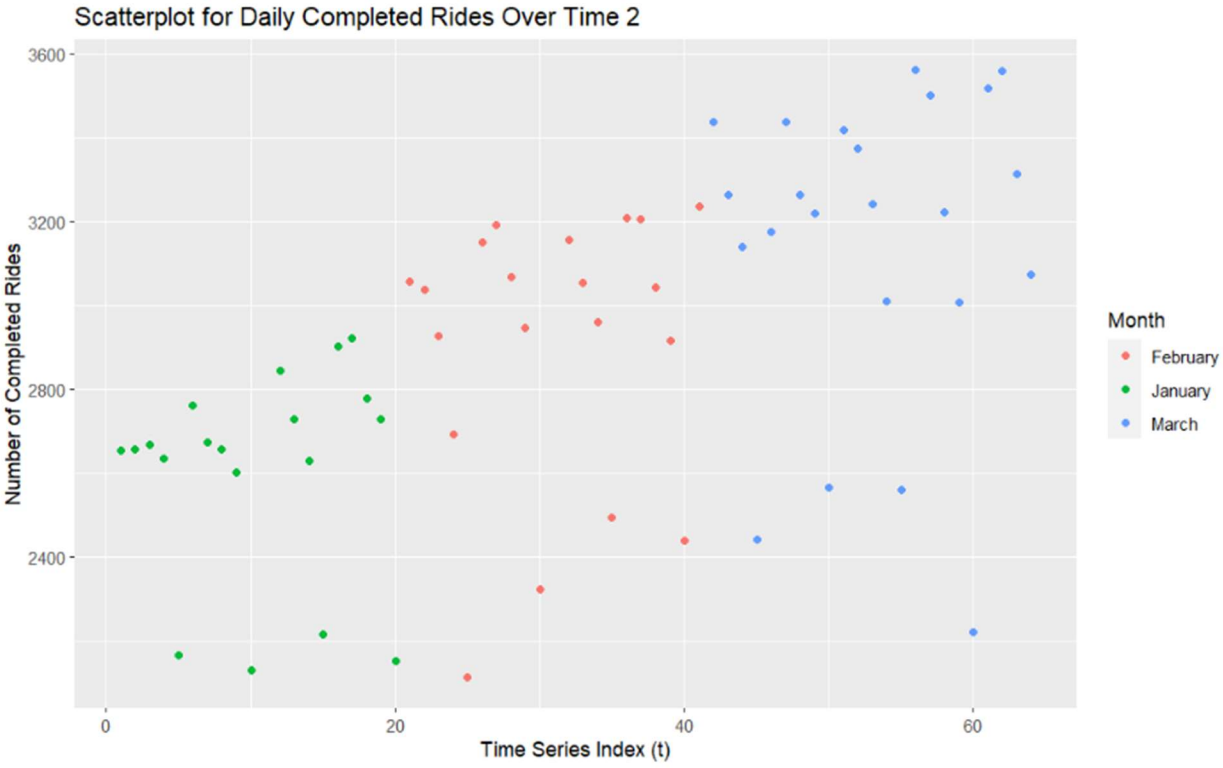
Appendix B



Appendix C

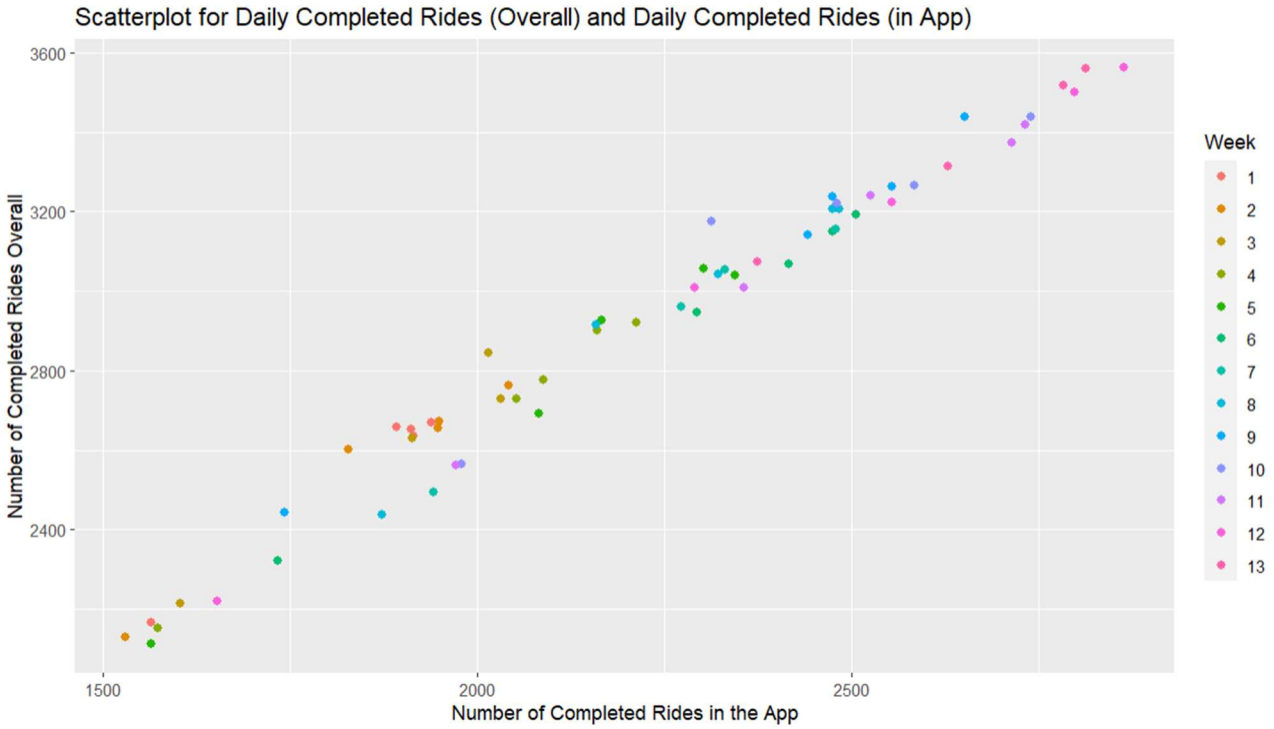


Appendix D

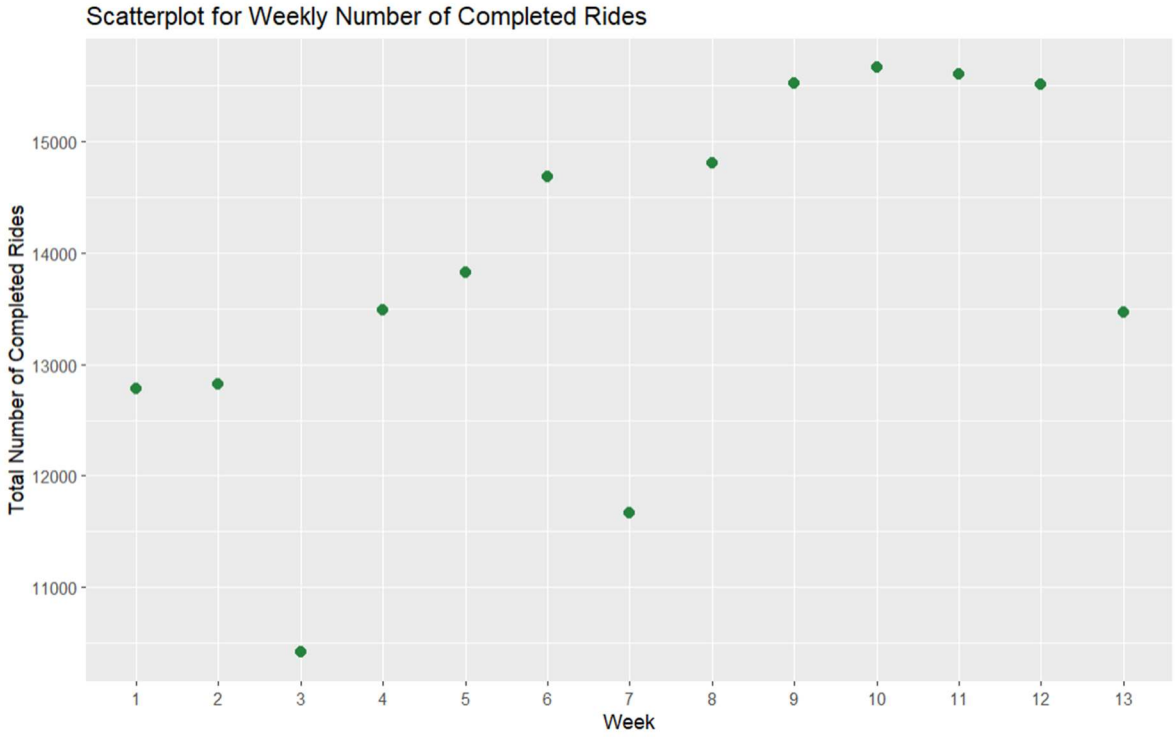


Appendix E

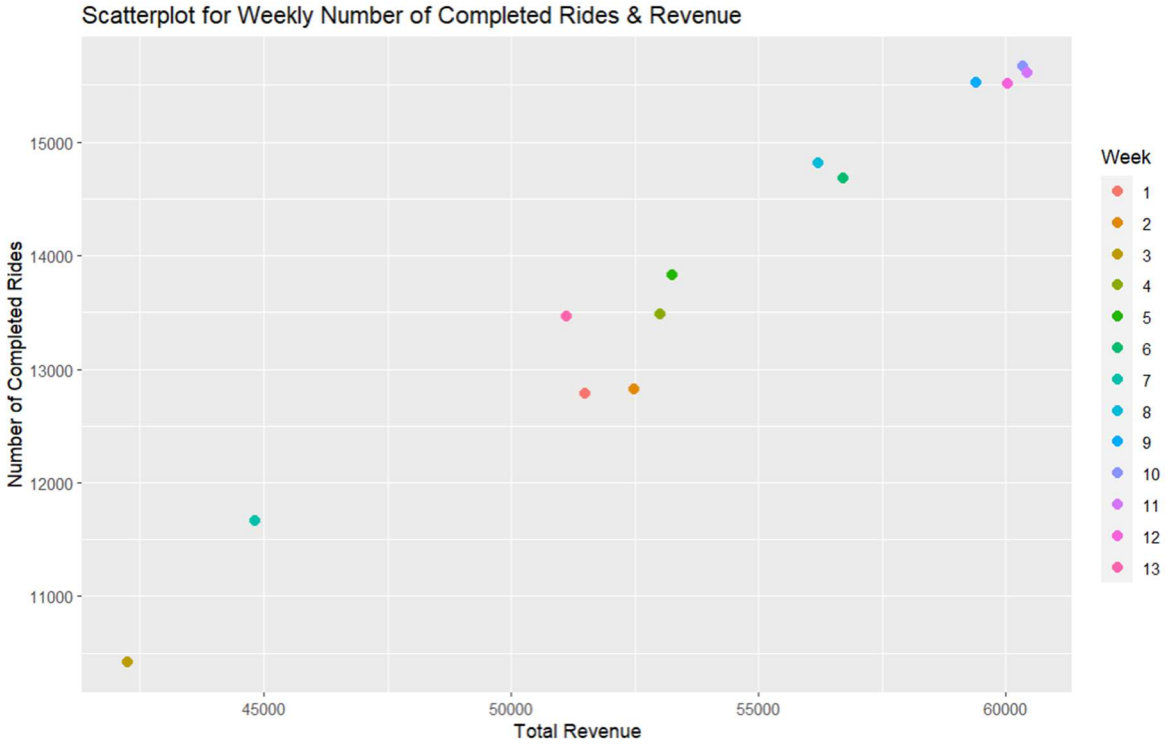
Appendix F



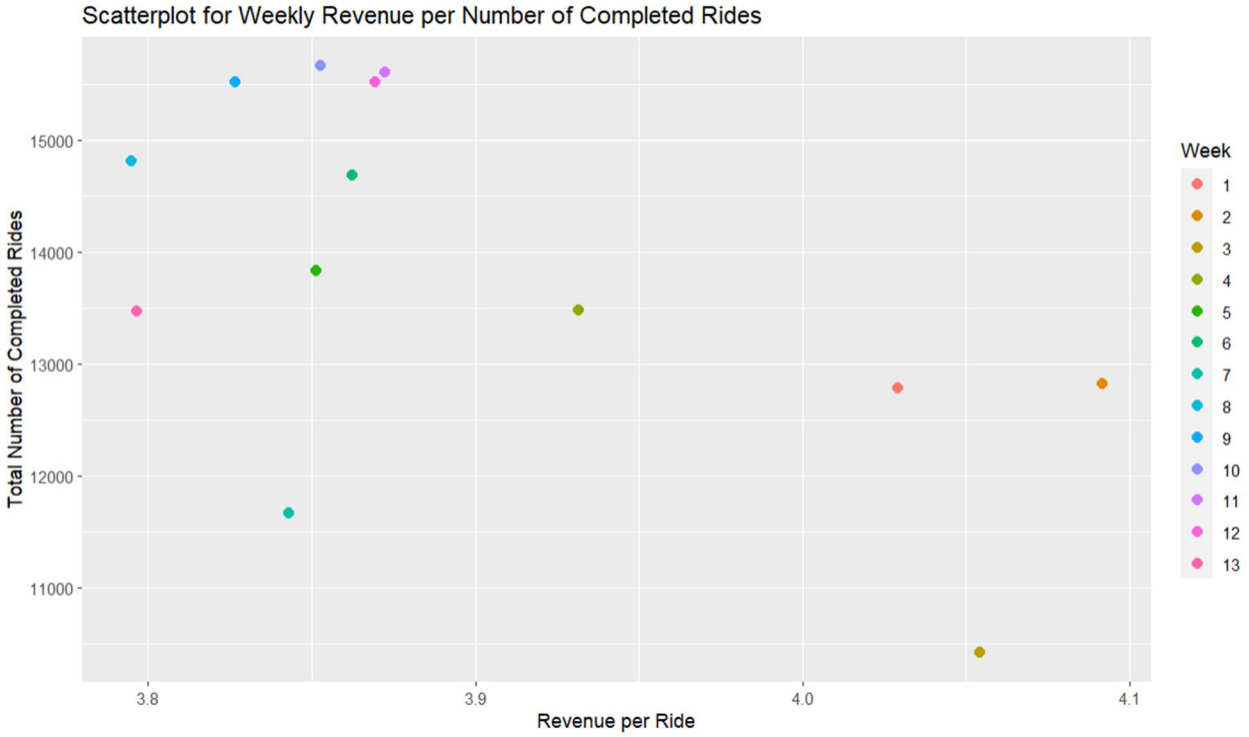
Appendix G



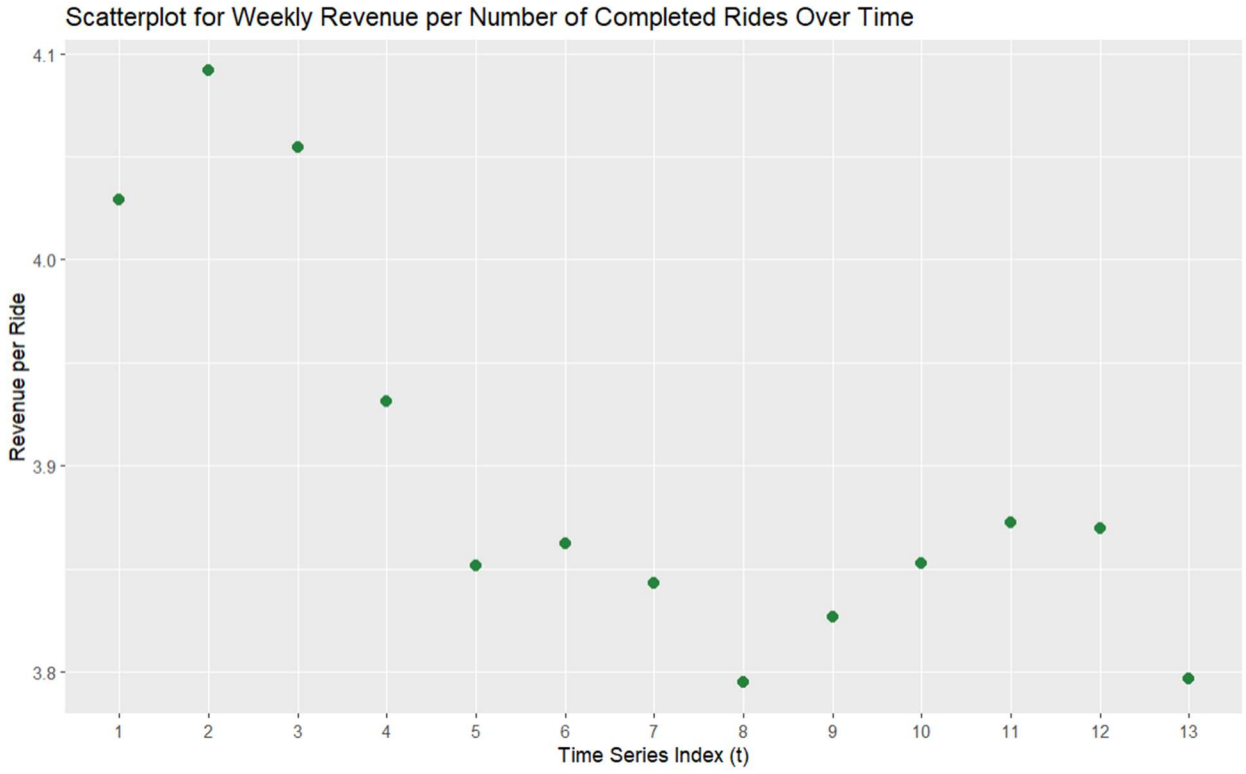
Appendix H



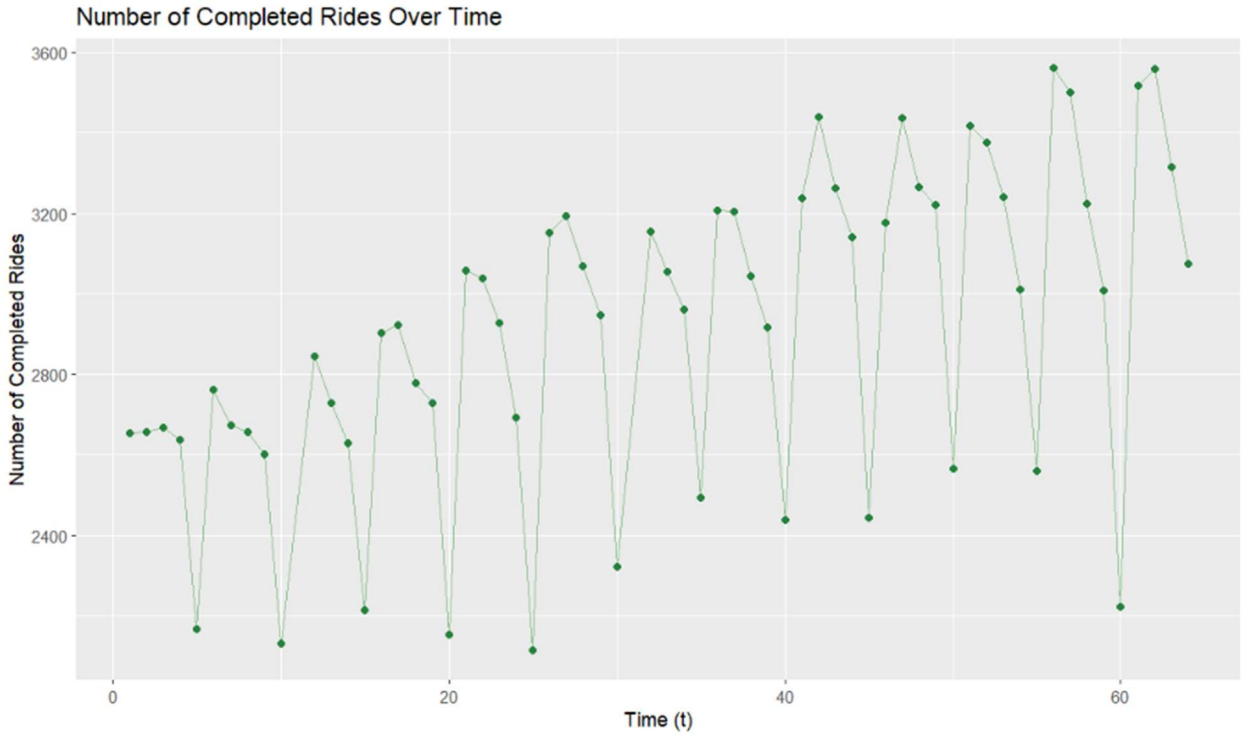
Appendix I



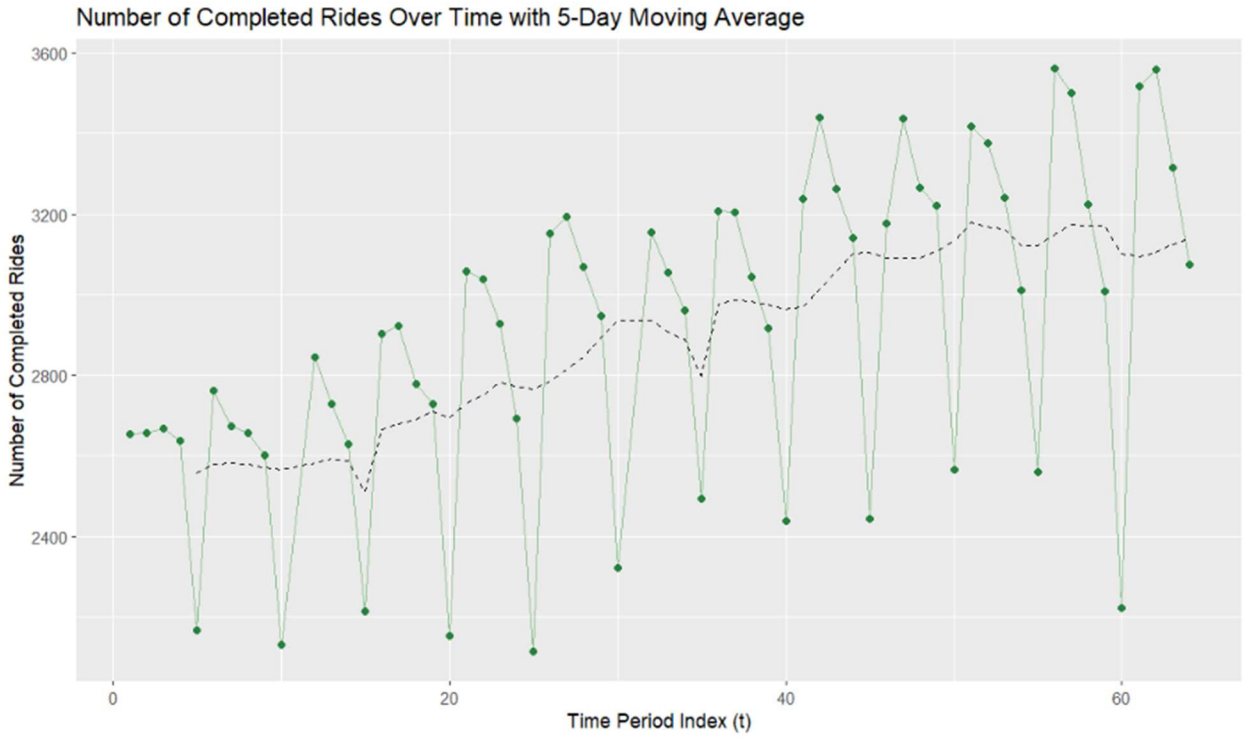
Appendix J



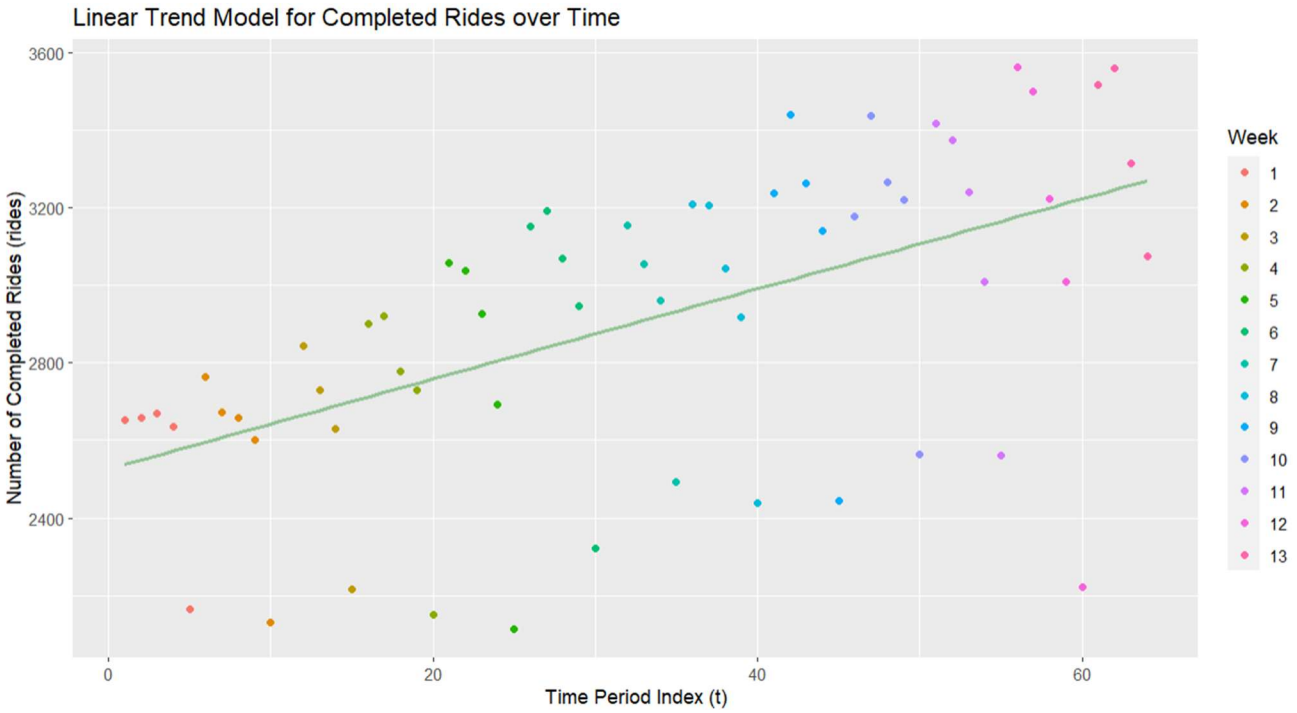
Appendix K



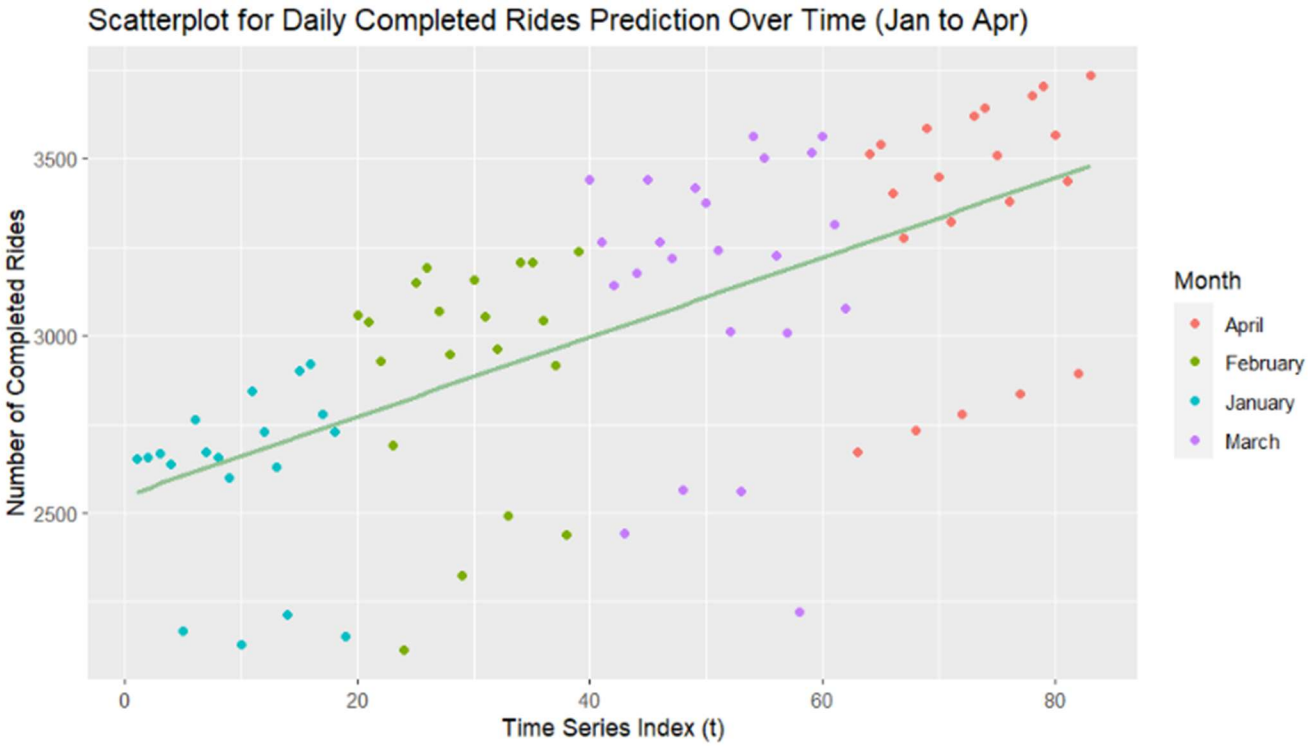
Appendix L



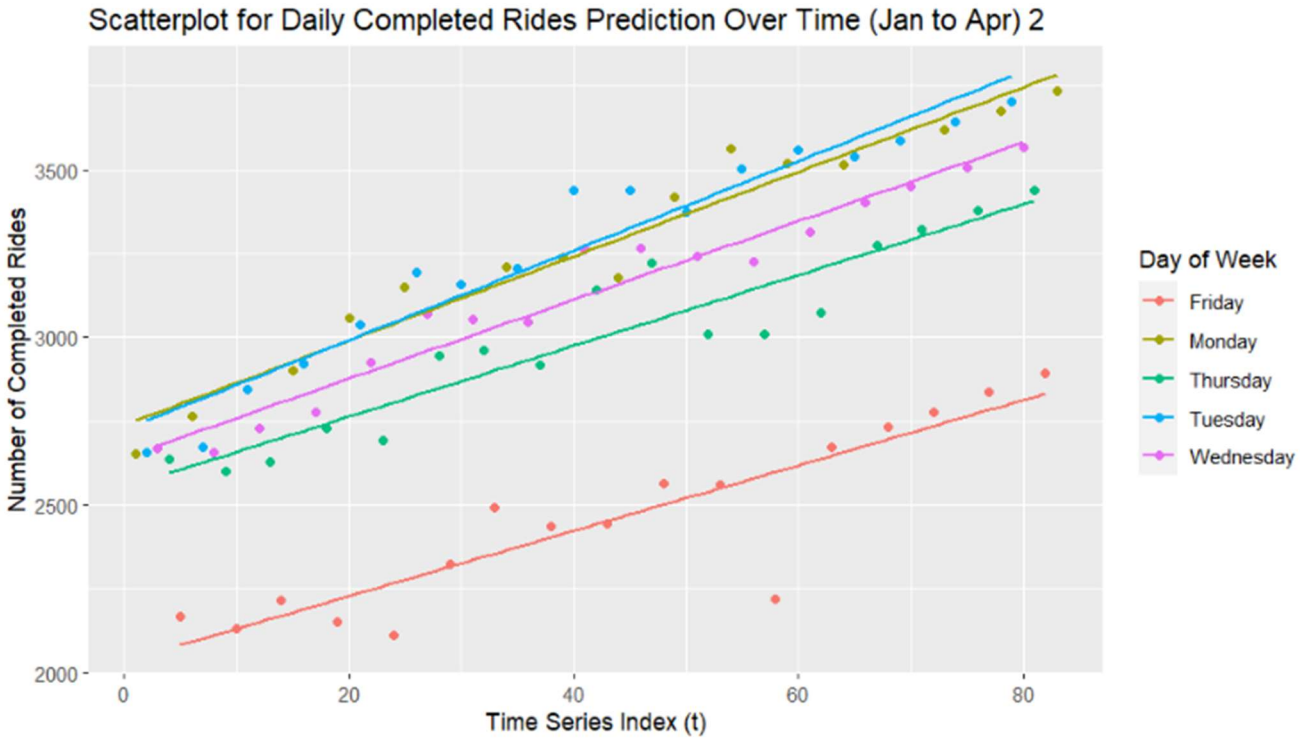
Appendix M



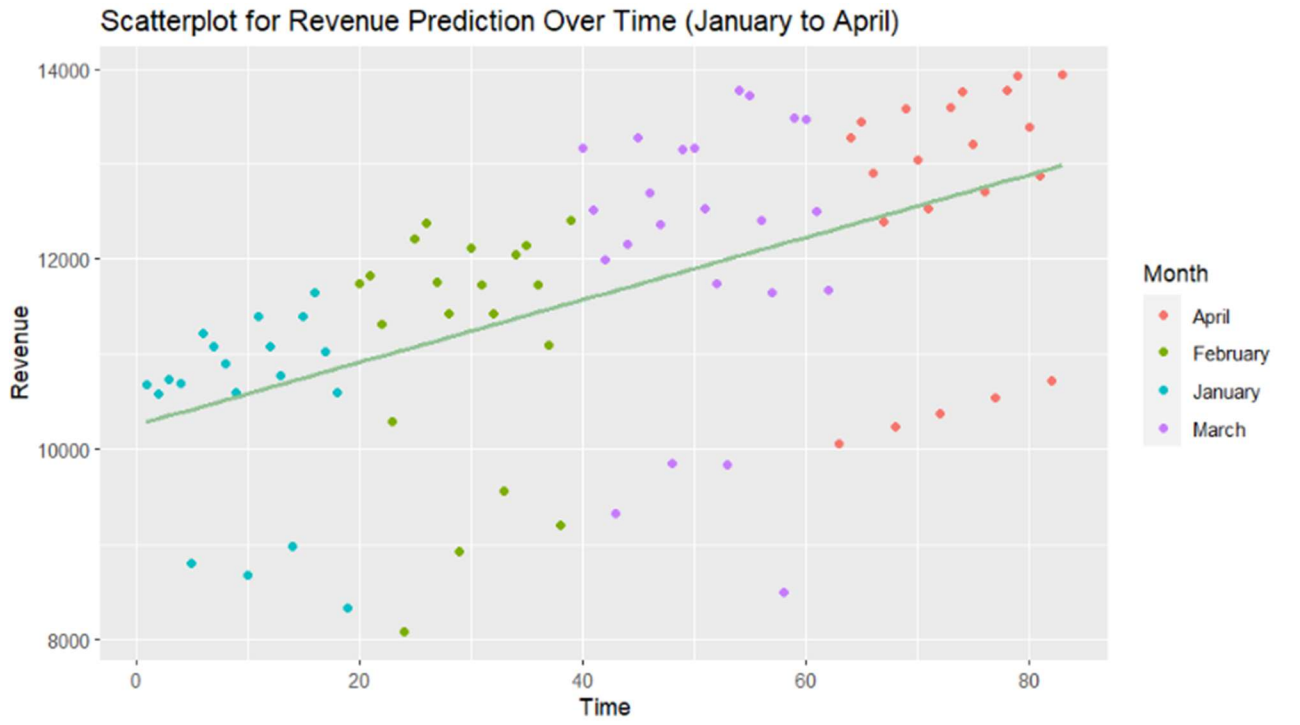
Appendix N



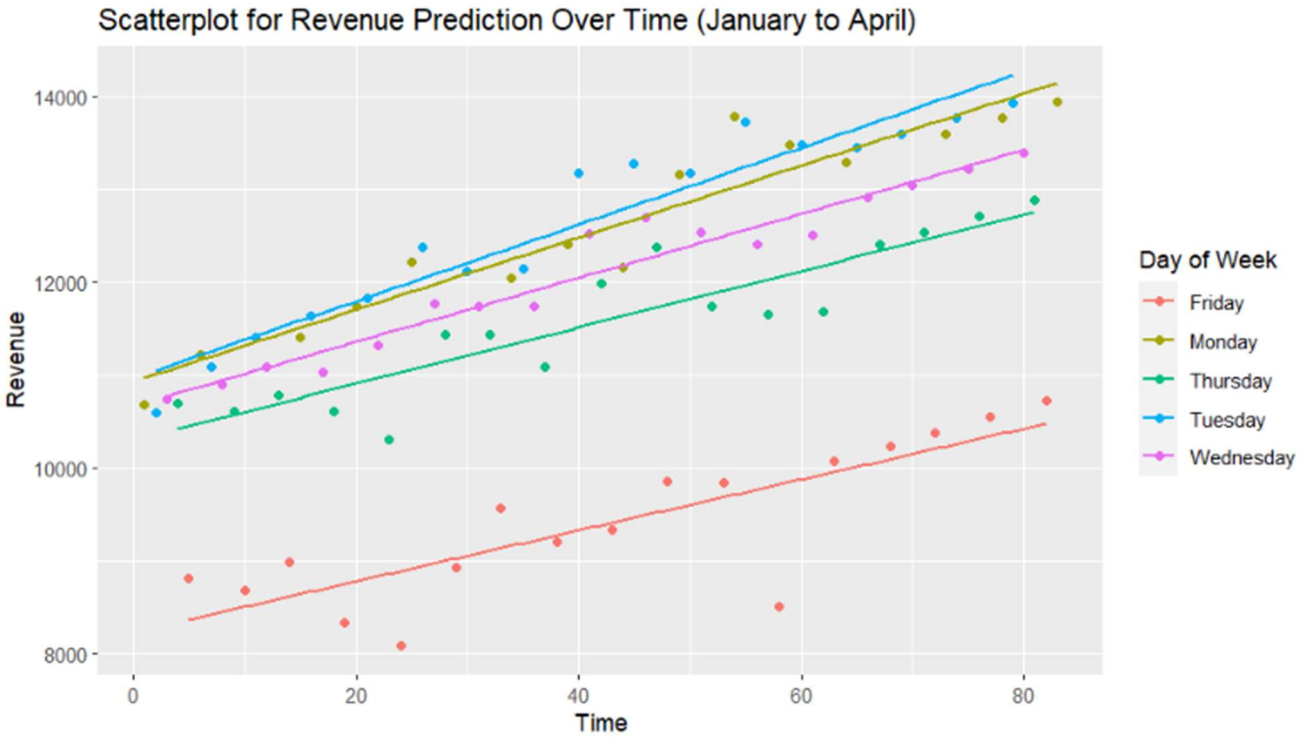
Appendix O



Appendix P



Appendix Q


Scatterplot for Revenue Prediction Over Time (January to April)
Revenue
14000
12000
10000
8000
0
20
40
60
80
Time
Day of Week
Friday
Monday
Thursday
Tuesday
Wednesday