Part 2: Linear regression models

Charles Julien, Chike Odenigbo, Atul Sharma, Gabriel Jobert

10/20/2023

Contents

Introduction

Unlocking the Wheels of Urban Mobility: A Data-Driven Analysis of BIXI

In the fast-paced, ever-evolving landscape of urban transportation, the quest to create efficient and sustainable solutions for city dwellers continues to be a paramount concern. Amidst the diverse array of options that have emerged in recent years, the BIXI public cycling service stands as a beacon of sustainable urban mobility. Offering an accessible, convenient, and eco-friendly mode of transportation, BIXI has transformed the way people navigate and experience cities.

As part of our commitment to understanding and improving urban transportation systems, our consultant team has embarked on an in-depth exploration of BIXI's operational data. The objective of this report is to provide a comprehensive analysis of the data collected from the BIXI service. By leveraging statistical and data analysis techniques, we aim to uncover valuable insights into the usage patterns, financial dynamics, and various factors affecting BIXI's performance. Our study covers an extensive range of factors, including ridership trends, environmental conditions, user classifications, and more.

One of the central questions addressed in this report is whether revenue generated by the BIXI service and trip duration significantly varies during weekends compared to weekdays. we also delve into the other factors affecting the duration of trips and the revenue generated by non-members. Our methodology combines data analysis, data visualization, and statistical modeling, with a primary focus on using R, a powerful statistical tool, to extract meaningful information from the BIXI dataset.

By analyzing this data, we aim to assist BIXI in making data-informed decisions to enhance the efficiency and quality of their services, ultimately contributing to the betterment of urban living. We believe that the findings and recommendations presented in this report will not only provide valuable insights to BIXI but also serve as a valuable reference for urban planners, researchers, and policymakers who are dedicated to creating more sustainable, convenient, and enjoyable urban environments.

The following sections of the report will delve into the specifics of our data analysis, share our findings, and provide recommendations based on the insights gathered during this project.

Business/Research questions

- Research Question 1: How do seasonal factors impact trip revenue for BIXI Montréal?
- Research Question 2: How do daily and weekly patterns impact trip durations for BIXI Montréal?
- Research Question 3: What variables impact the average bixi trip duration?

Pre-processing

Imputation

```
imputation_model <- lm(rev ~ dur + avg + n_tot , data = df_main)
df_main$rev_pred = predict(imputation_model, df_main)
```

To impute revenue for members, we consider the same formula of revenue used for non-members. This deterministic function is a linear combination of `dur`, `avg` and `n_tot`. The imputation model has an r-squared of 1.

Outlier detection

```
model <- lm(n_AM_PM_delta ~ long_wknd_ind + season + rain_ind + mem, data = df_main) # Goal is to look
summary(model)
```

```
##
## Call:
## lm(formula = n_AM_PM_delta ~ long_wknd_ind + season + rain_ind +
##     mem, data = df_main)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -117.382   -2.083    1.433    4.928   24.147
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -4.1407     0.8188  -5.057 4.33e-07 ***
## long_wknd_indWeekday    1.7079     0.8182   2.087   0.0369 *
## long_wknd_indWeekend    0.4048     0.8293   0.488   0.6255
## seasonSpring           -0.1763     0.2692  -0.655   0.5126
## seasonSummer           -1.5288     0.2199  -6.951 3.86e-12 ***
## rain_indRain            1.6899     0.2006   8.424  < 2e-16 ***
## mem1                   -8.1851     0.1951 -41.950  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.726 on 9993 degrees of freedom
## Multiple R-squared:  0.1587, Adjusted R-squared:  0.1582
## F-statistic: 314.2 on 6 and 9993 DF,  p-value: < 2.2e-16
```

```
model.diag.metrics <- augment(model)
head(model.diag.metrics)
```

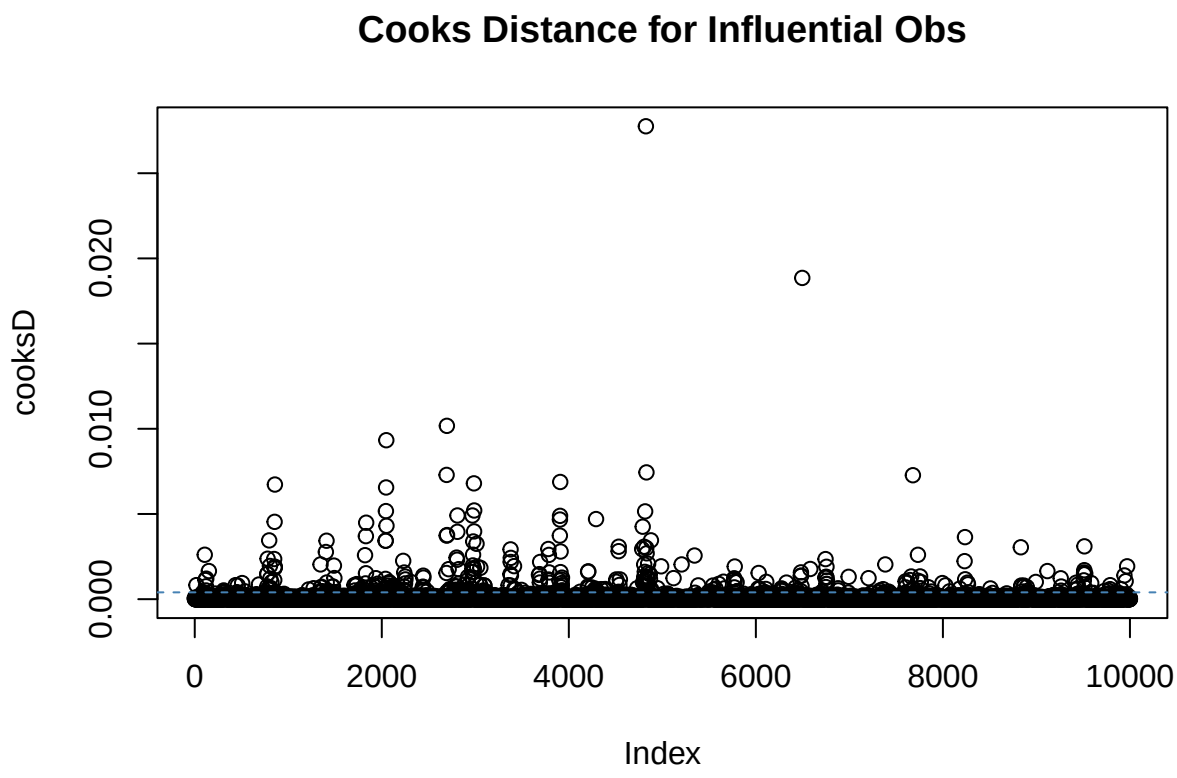
```
## # A tibble: 6 x 11
##   n_AM_PM_delta long_wknd_ind season rain_ind mem   .fitted .resid     .hat
##           <int> <fct>         <chr>  <fct>    <fct>   <dbl>  <dbl>    <dbl>
## 1            -1 Weekday       Spring Rain     1      -9.10    8.10 0.000798
## 2            -7 Weekday       Spring NoRain   1     -10.8     3.79 0.000673
```

```
## 3            -5 Weekend       Spring NoRain   0      -3.91   -1.09 0.000874
## 4            -2 Weekday       Spring Rain     0      -0.919  -1.08 0.000875
## 5            -6 Weekend       Spring Rain     1     -10.4     4.41 0.000979
## 6             0 Weekday       Summer NoRain   0      -3.96    3.96 0.000441
## # i 3 more variables: .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>
```

```
# OUTLIERS  WITH COOKS DISTANCE
model.diag.metrics %>%
  top_n(3, wt = .cooksd)
```

```
## # A tibble: 3 x 11
##   n_AM_PM_delta long_wknd_ind season rain_ind mem   .fitted .resid     .hat
##           <int> <fct>         <chr>  <fct>    <fct>   <dbl>  <dbl>    <dbl>
## 1          -128 Weekday       Fall   NoRain   1      -10.6  -117.  0.000488
## 2           -54 Long Weekend  Spring NoRain   0       -4.32  -49.7 0.00734
## 3           -54 Long Weekend  Fall   NoRain   1      -12.3   -41.7 0.00709
## # i 3 more variables: .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>
```

```
n <- nrow(df_main)
cooksD <- cooks.distance(model)
plot(cooksD, main = "Cooks Distance for Influential Obs")
abline(h = 4/n, lty = 2, col = "steelblue") # add cutoff line
```



Research Question 1: How do seasonal factors impact trip revenue for BIXI Montréal?

Objective of Analysis: This regression model is examining the impact of the month (`mm`), average daily temperature (`temp`), and total amount of rainfall (`rain`) and membershipt (`mem`) on the revenue (`rev_pred`) generated by trips leaving from a specified station.

Model

```
##
## Call:
## lm(formula = rev_pred ~ mm + temp + rain + mem, data = df_main)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -116.63  -31.67  -10.15   16.08  797.15
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13.9996     3.0451  -4.597 4.33e-06 ***
## mm5          22.9154     3.0485   7.517 6.09e-14 ***
## mm6          29.2179     3.4973   8.354  < 2e-16 ***
## mm7          34.4310     3.3877  10.164  < 2e-16 ***
## mm8          31.3085     3.7026   8.456  < 2e-16 ***
## mm9          41.5894     3.1402  13.244  < 2e-16 ***
## mm10         20.5731     2.8750   7.156 8.89e-13 ***
## mm11         13.1843     3.4067   3.870 0.000109 ***
## temp          1.0229     0.1758   5.818 6.14e-09 ***
## rain         -1.2649     0.1242 -10.182  < 2e-16 ***
## mem1         72.8839     1.2766  57.093  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.6 on 9989 degrees of freedom
## Multiple R-squared:  0.2782, Adjusted R-squared:  0.2775
## F-statistic:   385 on 10 and 9989 DF,  p-value: < 2.2e-16
```

Interpretation

Intercept: The intercept is -13.9996. This means that when all other predictors are zero (for non member on april), the predicted revenue is -$13.99. However, this should be treated with caution since an intercept in this context doesn't have a clear business interpretation.

Seasonality (Month): - The revenue seems to have a seasonal pattern. Compared to April (reference month), the model suggests that there's an increase in predicted revenue from May (mm5) through September (mm9). The highest increase in predicted revenue relative to April is in September (mm9) with an increase of $41.58 on average. After September, there is a decline in October (mm10) and November (mm11), suggesting that as we move into cooler months, the revenue decreases, but it's still higher than in April.

Temperature (`temp`): For every degree Celsius increase in temperature, the predicted revenue increases by approximately $1.02 on average. This suggests that warmer days tend to generate more revenue.

Rainfall (`rain`): - For every additional mm of rainfall, the revenue decreases by approximately $1.26 on average. This is intuitive as rainy days likely deter users from renting bikes, leading to reduced revenue.

Membership: If a rider is a member (mem1), the predicted revenue increases by $72.88 on average compared to non-member. This means that members contribute significantly more to the revenue compared to non-members.

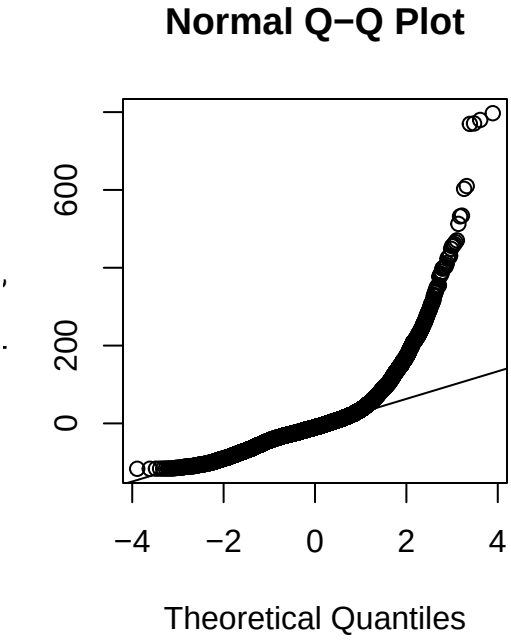
Model Fit: - The model has an R-squared value of 0.2782, indicating that approximately 27.82% of the variation in predicted revenue is explained by the model. - The F-statistic and its associated p-value (which is very close to zero) suggest that the model is statistically significant, and at least one of the predictors is useful in predicting revenue.

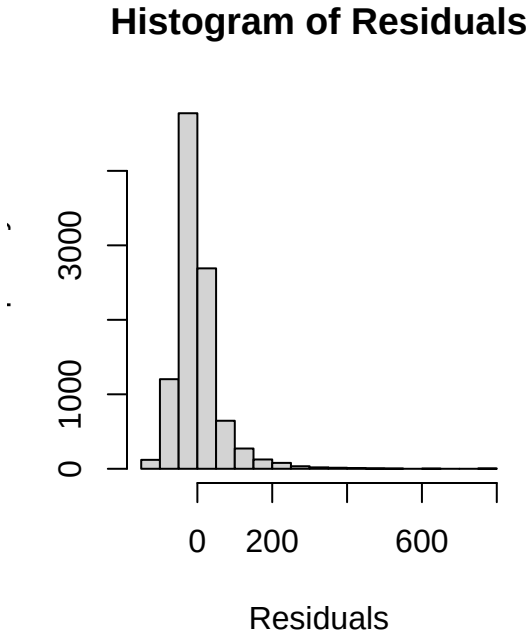
Business implications

1. **Operational Adjustments:** Given that revenue is higher in warmer months, consider optimizing operations for this period. This might involve higher staffing, more promotional activities, or ensuring optimal equipment availability.
2. **Rainy Day Strategies:** Since rainfall seems to negatively impact revenue, consider implementing strategies to mitigate this. For instance, promotional offers or special activities/events for rainy days might help attract customers.
3. **Membership** : While temperature and rain have expected impacts on bike rentals, it's notable how significant the role of membership is in driving revenue. Bixi should consider incentivizing memberships, given the clear revenue increase associated with members.

Verification of Assumptions

Normality of residuals





1. **Histogram of Residuals**: These histograms have a clear right-skew with a peak close to zero and a long tail towards the right. This suggests that most residuals are clustered around zero, but there are a few larger positive residuals. This is an indication that the normality assumption of the residuals may be violated.

2. **Normal Q-Q Plot**: Most of the points are close to the line, which is a good sign. However, there's a clear deviation from the line on the top right corner, suggesting the presence of larger residuals that are not explained by a normal distribution. This reiterates the presence of the right skew seen in these histograms.

Overall Interpretation: The residuals are not perfectly normal. They show a positive skewness, indicating there might be some observations with higher residuals (perhaps outliers or instances where the model systematically underpredicts). The deviation from normality might not be a problem because the sample is large enough.

Research Question 2: How do daily and weekly patterns impact trip durations for BIXI Montréal?

Objective of Analysis: This regression model is examining the impact of the day of the month (`dd`), day of the week (`wday`), and holidays (`holiday`) on the revenue (`rev`) generated by trips leaving from a specified station.

Model

```
##
## Call:
## lm(formula = dur ~ dd + wday + holiday, data = df_main)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -330.0 -201.4 -101.5   96.4 3953.1
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   284.97626    9.71890  29.322  < 2e-16 ***
## dd              0.03792    0.35263   0.108   0.9144
## wdayMonday    -49.00428   11.63373  -4.212 2.55e-05 ***
## wdaySaturday   15.05883   11.40702   1.320   0.1868
## wdaySunday     -6.50623   11.48715  -0.566   0.5711
## wdayThursday  -16.94467   11.56156  -1.466   0.1428
## wdayTuesday   -34.48313   11.61775  -2.968   0.0030 **
## wdayWednesday -20.56897   11.53032  -1.784   0.0745 .
## holiday1       68.77802   21.23112   3.239   0.0012 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 307 on 9991 degrees of freedom
## Multiple R-squared:  0.00463,    Adjusted R-squared:  0.003833
## F-statistic: 5.809 on 8 and 9991 DF,  p-value: 2.016e-07
```

Interpretation

Overall Model : The model explains about 0,5% of the variability in total rental durations. The F-statistic and its associated p-value confirm that the model is statistically significant and that at least some of the predictors have significant effects.

Intercept (284.98)*: - On an average day (specifically, a non-holiday), the expected rental duration is approximately 4h45. - This value is statistically significant (*** p-value < 2e-16), which indicates strong evidence against the null hypothesis.

Day of the Month (`dd`): - For each additional day in the month, the rental duration increases by an average of 0.04 minutes. - This effect is not statistically significant (p-value = 0.9144), suggesting the day of the month might not be a meaningful predictor for the duration of BIXI bike rentals.

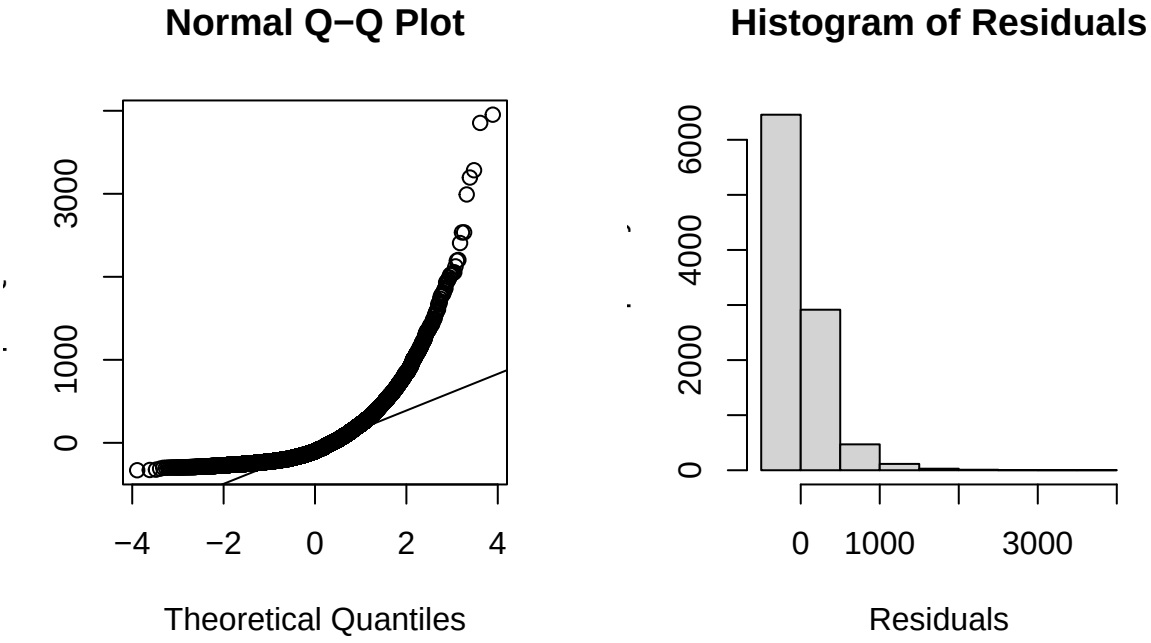
Day of the Week (`wday`): - Compared to Fridays: - Rentals on **Mondays** are, on average, **49 minutes shorter**. This is statistically significant (*** p-value = 2.55e-05). - Rentals on **Saturdays** are about **15.05 minutes longer** on average, but this is not statistically significant (p-value = 0.1868). - Rentals on **Sundays** are about **6.51 minutes shorter** on average, but this is also not statistically significant (p-value = 0.5711). - Rentals on **Thursdays** are **16.95 minutes shorter** on average, but this isn't statistically significant either (p-value = 0.1428). - Rentals on **Tuesdays** are **34.48 minutes shorter** on average, and this is statistically significant (** p-value = 0.0030). - Rentals on **Wednesdays** are **20.57 minutes shorter** on average. This result is at the borderline of significance (p-value = 0.0745).

Holiday (`holiday1`): - On holidays, bike rentals are, on average, **1 hour and 8 minutes longer** compared to non-holidays. This is statistically significant (** p-value = 0.0012), suggesting that holidays have a meaningful impact on the duration of bike rentals.

Business Implications: BIXI bike rentals tend to be shorter on Mondays and Tuesdays compared to Fridays, and rentals on holidays are significantly longer than on non-holidays. Planning for resource allocation, marketing strategies, or promotional campaigns should consider these patterns to optimize business operations. Keeping in mind the higher trip duration during holidays, adequate steps should be taken to ensure Bixi availability at stations.

Verification of assumptions

Normality of residuals



Normal Q-Q Plot: From the given Q-Q plot, the points deviate significantly from the diagonal line, especially in the tails. This suggests that the residuals are not normally distributed. The heavy tails (points deviating from the line at both ends) suggest the presence of potential outliers or extreme values in the residuals.

Histogram of Residuals: The histogram shows that the majority of residuals are clustered around zero, but there are some extreme positive values. This is consistent with the observation from the Q-Q plot and indicates a possible right-skewed distribution of residuals. In the context of regression, the CLT means that even if the residuals aren't perfectly normally distributed in the population, the sampling distribution of the regression coefficients will be approximately normal if the sample size is large enough.

Implications:

- While large sample sizes can make the assumption of normality less crucial, it doesn't mean analysts should ignore violations of other assumptions or entirely disregard the distribution of residuals. Diagnostics and plots (like Q-Q plots) still provide valuable information about potential model misspecifications or the presence of influential outliers.
- Moreover, a large sample size can sometimes detect statistically significant relationships even when they are practically insignificant. So, while p-values might be small, the effect sizes or coefficients might not be practically meaningful.

Research Question 3: What variables impact the average bixi trip duration?

The idea is to identify the driving factors of a bixi trip length when we control for most of the variables. Trip length is one of the three important variables that drives revenue, the other ones being the number of trips and the pricing scheme. Keep in mind that increasing the trip length does not necessarily increase revenues since an unwanted increase in trip length may discourage users from using bixi's system and result in a decrease in trip number.

Variables Selection

Our goal is to incorporate most of the important variables in order to increase our chance of respecting the assumption of E(e)=0 and thus making our model more telling.

Variables that make business sense to include:

From our seasonality analysis we identified:

- Season (`season`)
- Temperature in degrees celcius (`temp`)
- Rainfall in mm (`rain`)

From our daily and weekly pattern analysis we identified:

- Part of theweek i.e. weekend or weekday (`wknd_ind`)
- If it is a holiday (`holiday`)

Some other variables that are interesting:

- If the user is a member(`mem`)
- Location of the bixi station compared to Parc Lafontaine (`North_South`) and (`West_East`)
- Proportion of trips in the morning versus the whole day (`percent_AM`)
- Total number of trips (`n_tot`)
- If the station is a metro station (`Metro_ind`)
- If we are at the beginning of month or the end of the month (`PartOfMonth`)

Interactions: In our EDA we observed a different week day usage of the member and non members, thus an interaction term between members and day of week would be interesting. (`wday*mem`).

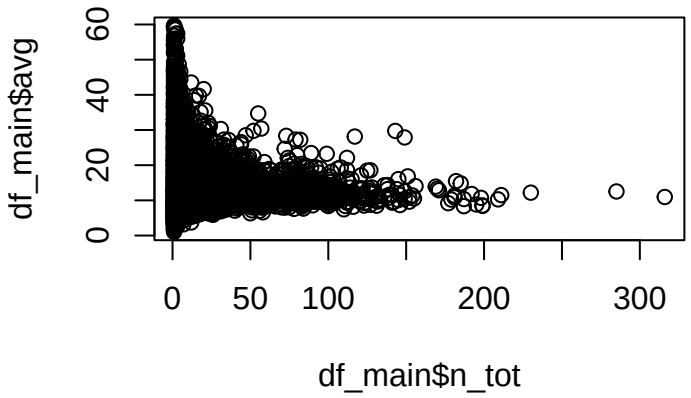
Correlation:

Let's take a quick look at the correlation between our numerical variables to estimate the effect of collinearity.

```
##                  avg        temp        rain        n_tot   percent_AM
## avg         1.00000000  0.09639054 -0.10619900 -0.215866274 -0.107387372
## temp        0.09639054  1.00000000 -0.02794911  0.139997362 -0.078110564
## rain       -0.10619900 -0.02794911  1.00000000 -0.054717667  0.013211523
## n_tot      -0.21586627  0.13999736 -0.05471767  1.000000000 -0.008953075
## percent_AM -0.10738737 -0.07811056  0.01321152 -0.008953075  1.000000000
```

We see very low correlation between the Xs which means we should not get any problems with collinearity between our numerical variables.

After the assumptions verification we chose to exclude `n_tot` :



The variable total number of trip (`n_tot`) has been removed from the regression because it did not pass the assumption of constant variance, making the model not correctly specified. This makes intuitive sense since as the number of trip increases, the average trip duration should converge towards the true mean.

Model

```
##
## Call:
## lm(formula = avg ~ season + temp + rain + wknd_ind * mem + holiday +
##     North_South + West_East + percent_AM + PartOfMonth + Metro_ind,
##     data = df_main)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.525  -3.568  -1.159   2.085  43.513
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          14.00789    0.26879  52.114  < 2e-16 ***
## seasonSpring          2.71442    0.17663  15.368  < 2e-16 ***
## seasonSummer          0.52048    0.18461   2.819  0.00482 **
## temp                  0.11474    0.01351   8.494  < 2e-16 ***
## rain                 -0.10242    0.01219  -8.402  < 2e-16 ***
## wknd_indWeekend       2.47108    0.20000  12.356  < 2e-16 ***
## mem1                 -1.83782    0.15034 -12.224  < 2e-16 ***
## holiday1              1.06532    0.42213   2.524  0.01163 *
## North_SouthSouth      0.08859    0.12686   0.698  0.48497
## West_EastWest        -0.26904    0.13389  -2.009  0.04451 *
## percent_AM           -2.03011    0.31203  -6.506 8.08e-11 ***
## PartOfMonthEOM       -0.22545    0.12776  -1.765  0.07766 .
```

```
## Metro_ind1           -0.74882    0.23050  -3.249  0.00116 **
## wknd_indWeekend:mem1 -1.54124    0.27642  -5.576 2.53e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.263 on 9986 degrees of freedom
## Multiple R-squared:  0.09851,    Adjusted R-squared:  0.09733
## F-statistic: 83.94 on 13 and 9986 DF,  p-value: < 2.2e-16
```

Interpretation

Overall Model - The model explains approximately 12.52% of the variation in the average trip duration. which means that other factors are also at play and not included in the model.

Intercept - The interpretation of the intercept does not make sense in this case since the number of trips would have to be zero.

Coefficients - Season: The reference level is fall. We can see that on average trip duration during spring and summer are respectively 2.33 and 0.38 minutes longer than in fall holding everything else constant.

- **Temperature**: The coefficient of temperature is 0.14 which means that an increase in temperature of 1 degree celcius corresponds to an increase of average trip duration of 0.14 minutes on average holding all else constant.
- **Rainfall**: The coefficient for rain is -0.12 which means that an increase in rainfall of 1 mm corresponds to a decrease of average trip duration of 0.12 minutes on average holding all else constant.
- **Effect of Weekend Indicator and membership**: Since there exists an interaction between both variables, it is no longer possible to interpret one without the other. This implies that the relation between average trip duration and membership is different depending on the moment of the week. The opposite is also true, the relation between average trip duration and the moment of the week is different depending on the membership status.
  - **Weekend indicator's** coefficient 2.638923 is the average difference between average trip duration during weekend and weekday for non-members. In other words, for non-members, average trip duration is higher on average than for members holding all else constant.
  - **Membership's** coefficient -0.385385 is the average difference between average trip duration for members and non-members for weekdays. In other words, during weekdays, the average trip duration is shorter on average for members than for non-members holding all else constant.
  - **Interaction term's** coefficient -1.868383 is ...
- **Holiday**: The coefficient for holiday is 1.069557 which means that during holidays average trip duration is 1.069 minutes higher on average than during non-holidays, holding all else constant.
- **North_South and West_East**: Their coefficients are 0.35 and -0.23 which means that on average the average trip duration for trips starting at a station South of Parc Lafontaine or West is 0.35 and -0.23 minutes different from their counter parts respectively, holding all else constant.
- **Total number of trips**: The coefficient is -0.055315 which means that on average as number of trips increase, the average trip length in minutes decreases, holding all else constant.
- **Percent AM**: The magnitude of the coefficient -2.351044 is less important than its sign for our interpretation. What it means is that as the proportion of trips in the morning increases, the average trip duration generally decreases when holding all else constant. This hints that trips in the morning might be shorter on average than trip in the afternoon, hence bring in less revenue.

Business Implications:

1. **Promotion and Marketing**: For the same temperature, average trip length tends to be the longest in spring. This indicates that users are eager to use bikes after winter. This insight could be used for promotion purposes.

2. **Resource Allocation**: Expect longer trips when it is hot and non-rainy outside. Even more if it is a weekend or holiday. Also, bikes tend to be borrowed longer during the afternoon than in the morning. Stations south of Parc Lafontaine have on average longer trip duration, which may suggests that stations are further from one another. There might be some space for additional stations.

3. **Pricing Strategy**: The usage that is associated with the longest trip length based on our interaction term is for non-members during the weekend. Charging a heftier price for these people at that time may increase profit margins significantly.

4. **Operational Strategy**: It is important to keep in mind the tradeoff between the number of trips and the average trip length. Indeed, as the number of trips increase for a given day, the average trip length decreases. This may suggest that the additional trips during those days are short haul.

Verification of assumptions and collinearity

Variance Inflation Factor

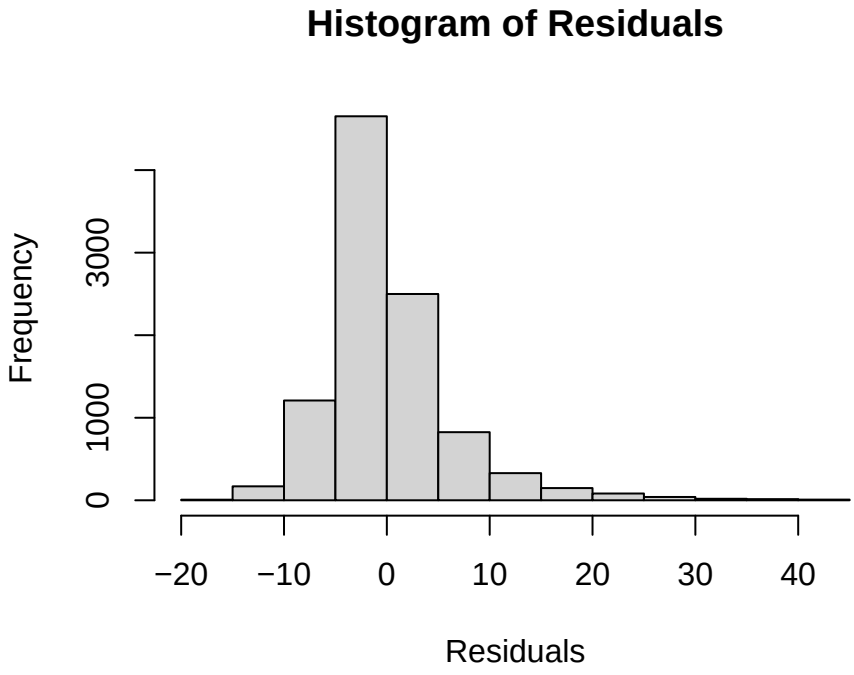
Let's use the variance inflation factor to verify for collinearity, we will use a standard threshold of 5.

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## season       1.952937  2        1.182149
## temp         1.885069  1        1.372978
## rain         1.030170  1        1.014973
## wknd_ind     2.101725  1        1.449733
## mem          1.436479  1        1.198532
## holiday      1.016486  1        1.008209
## North_South  1.010347  1        1.005160
## West_East    1.002827  1        1.001413
## percent_AM   1.045308  1        1.022403
## PartOfMonth  1.039869  1        1.019740
## Metro_ind    1.003732  1        1.001864
## wknd_ind:mem 2.472760  1        1.572501
```

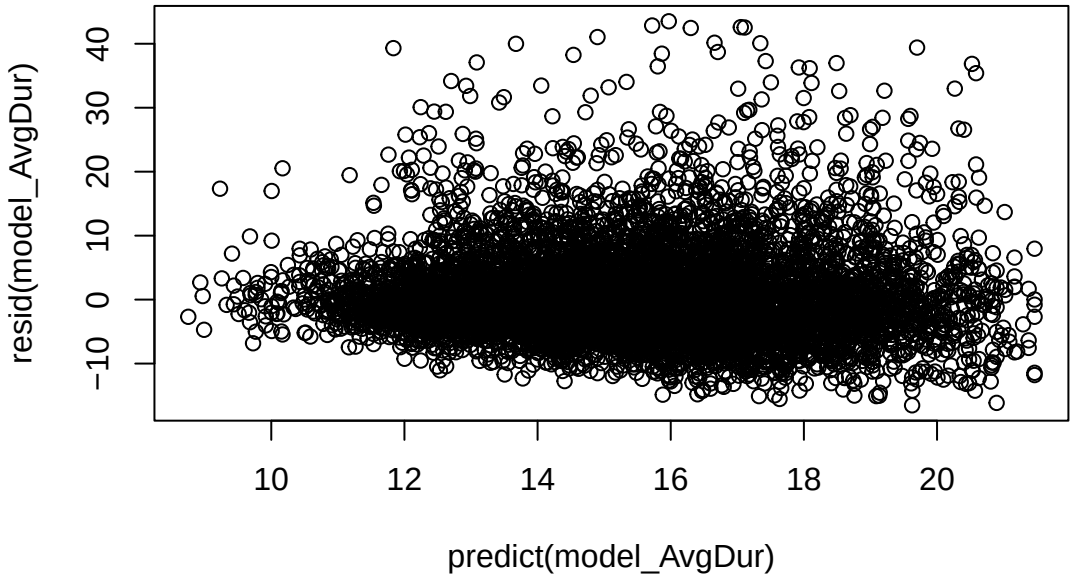
No major problem is detected, since the global vifs are all relatively low.

Verification of Normality of Residuals



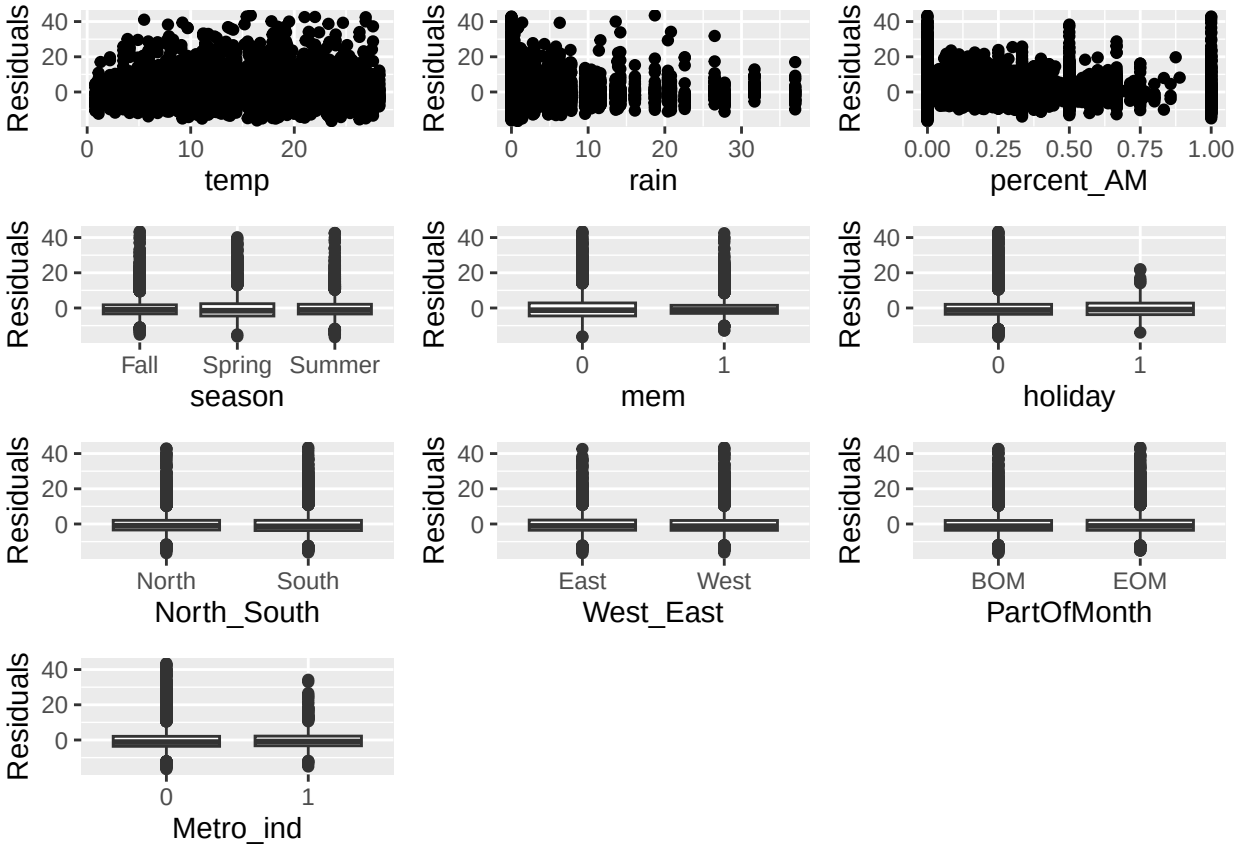
No problem here, residuals are normally distributed.

Model correctly specified



The model seems to be correctly specified.

Verificaiton of Heteroscedasticity



No major problem of heteroscedasticity were detected. The variable `n_tot` has been removed as stated earlier.

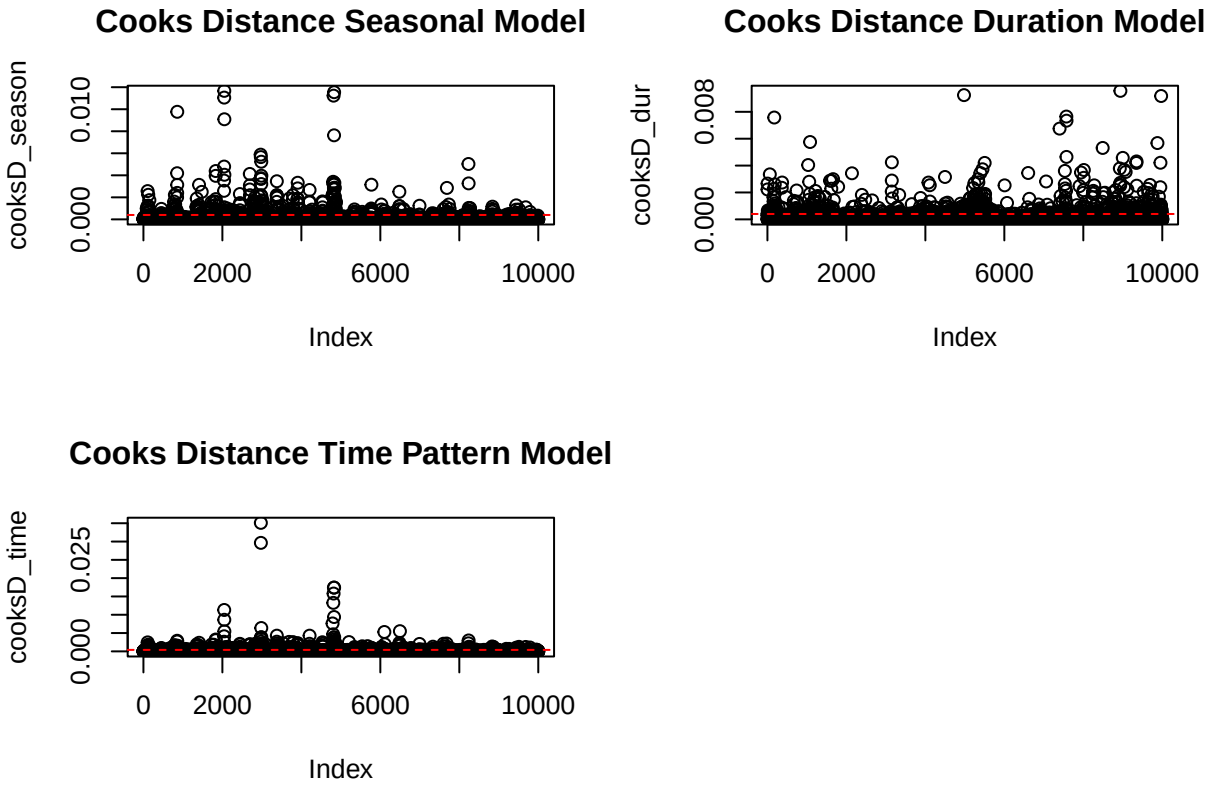
Influential Observations

In order to confirm the validity of our previous analysis, we explored the impact of influential observations on our model parameters and statistical measures. Our methodology was to use Cook's distance to identify and remove highly influential explanatory features and to remove them to ensure accuracy in our coefficients when retraining. Cook's distance identifies observations that cause the largest change in fitted values when the observation is deleted. The limitation of retraining this way is that we could be losing information from highly influential data points such as if those observations are highly correlated to a subset of the data such as period of extreme rain.

As can be seen by the plots below, each of the aforementioned models have a number of observations with a large Cook's distance.

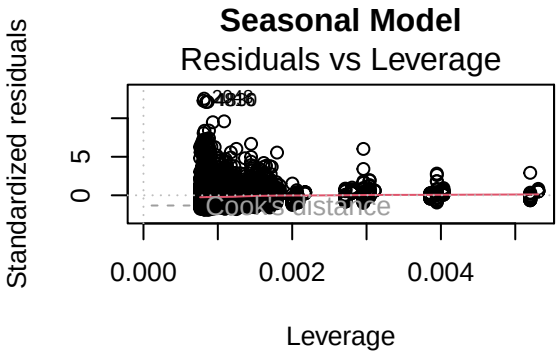
```
layout_matrix_1 <- matrix(1:4, ncol = 2)
layout(layout_matrix_1)
n <- nrow(df_main)
cooksD_season <- cooks.distance(seasonal_effect_rev_model)
plot(cooksD_season, main = "Cooks Distance Seasonal Model")
abline(h = 4/n, lty = 2, col = "red") # add cutoff line
```

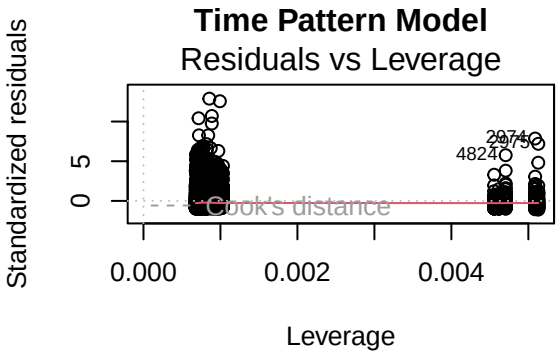
```
cooksD_time <- cooks.distance(time_pattern_dur_model)
plot(cooksD_time, main = "Cooks Distance Time Pattern Model")
abline(h = 4/n, lty = 2, col = "red") # add cutoff line

cooksD_dur <- cooks.distance(model_AvgDur)
plot(cooksD_dur, main = "Cooks Distance Duration Model")
abline(h = 4/n, lty = 2, col = "red") # add cutoff line
```

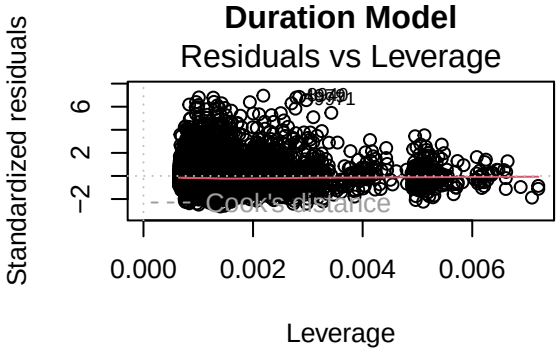


To confirm our finding on influential observations, we looked at leverage residuals plots. Leverage represents how much a coefficient would change if the observation were removed from the dataset. Observations in the charts below that fall above the red line can be considered influential observations. The spread of the residuals shouldn't change much as a function of leverage which would indicate heteroskedasticity. In this case, all three model seem to display some form of heteroskedasticity particularly the seasonal model.

```
par(mfrow = c(2, 2))
plot(seasonal_effect_rev_model, main = "Seasonal Model",5)
plot(time_pattern_dur_model, main = "Time Pattern Model",5)
plot(model_AvgDur, main = "Duration Model",5)
```








```
influ_season<- cooksD_season[(cooksD_season > (3 * mean(cooksD_season, na.rm = TRUE)))]
names_season <- names(influ_season)
outliers_season <- df_main[names_season,]
df_main_season <- df_main %>% anti_join(outliers_season)
```

```
## Joining with `by = join_by(station, mm, dd, wday, mem, holiday, dur, avg, rev,
## n_AM, n_PM, n_tot, temp, rain, name, latitude, longitude, wday_ordered, season,
## rain_ind, North_South, West_East, PartOfMonth, Metro_ind, lntot, rev_per_min,
## rev_per_trip, percent_AM, percent_PM, percent_AM_PM_delta, n_AM_PM_delta,
## am_pm_ind, wknd_ind, long_wknd_ind, year, date, week_num, rev_pred)`
```

```
model_season <- lm(rev_pred ~ mm + temp + rain + mem, data = df_main_season)
```

```
#df_main_time
influ_time<- cooksD_time[(cooksD_time > (3 * mean(cooksD_time, na.rm = TRUE)))]
names_time <- names(influ_time)
outliers_time <- df_main[names_time,]
df_main_time <- df_main %>% anti_join(outliers_time)
```

```
## Joining with `by = join_by(station, mm, dd, wday, mem, holiday, dur, avg, rev,
## n_AM, n_PM, n_tot, temp, rain, name, latitude, longitude, wday_ordered, season,
## rain_ind, North_South, West_East, PartOfMonth, Metro_ind, lntot, rev_per_min,
## rev_per_trip, percent_AM, percent_PM, percent_AM_PM_delta, n_AM_PM_delta,
## am_pm_ind, wknd_ind, long_wknd_ind, year, date, week_num, rev_pred)`
```

```
model_time <- lm(dur ~ dd + wday + holiday, data = df_main_time)
```

```
#df_main_dur
influ_dur<- cooksD_dur[(cooksD_dur > (3 * mean(cooksD_dur, na.rm = TRUE)))]
names_dur <- names(influ_dur)
outliers_dur <- df_main[names_dur,]
df_main_dur <- df_main %>% anti_join(outliers_dur)
```

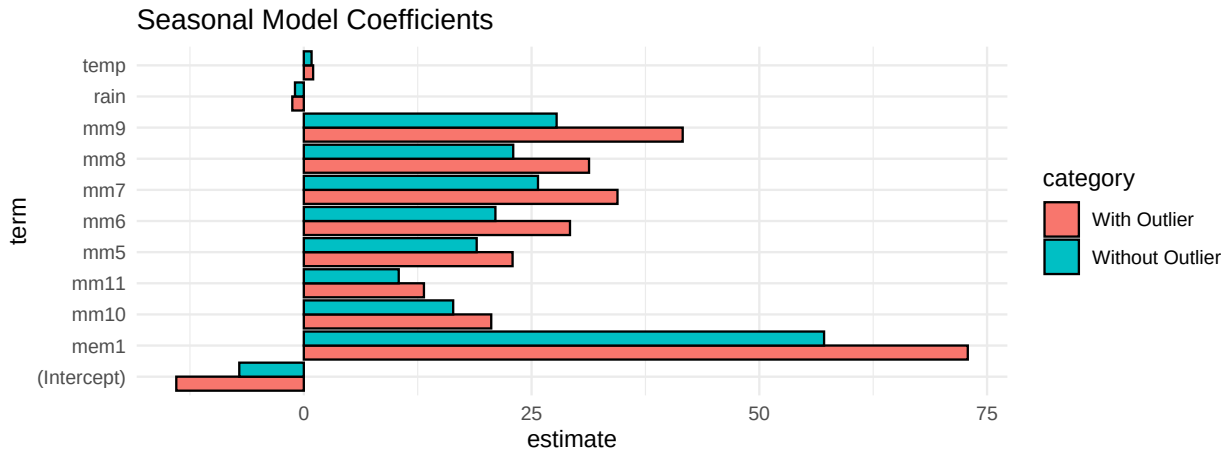
```
## Joining with `by = join_by(station, mm, dd, wday, mem, holiday, dur, avg, rev,
## n_AM, n_PM, n_tot, temp, rain, name, latitude, longitude, wday_ordered, season,
## rain_ind, North_South, West_East, PartOfMonth, Metro_ind, lntot, rev_per_min,
## rev_per_trip, percent_AM, percent_PM, percent_AM_PM_delta, n_AM_PM_delta,
## am_pm_ind, wknd_ind, long_wknd_ind, year, date, week_num, rev_pred)`
```

```
model_dur <- lm(avg ~ season + temp + rain + wknd_ind*mem + holiday  + North_South + West_East + percen
```

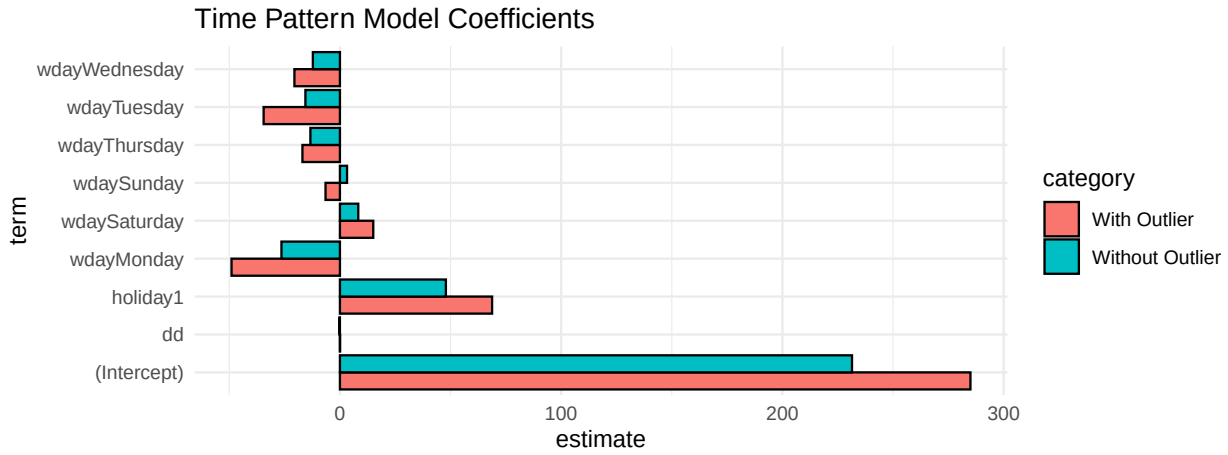
```
summary_season = tidy(seasonal_effect_rev_model)
summary_season$category = 'With Outlier'
summary_season_outlier = tidy(model_season)
summary_season_outlier$category = 'Without Outlier'
season_summary_combined <- rbind(summary_season, summary_season_outlier)
season_summary_combined$significance <- ifelse(season_summary_combined$p.value<0.05, 'Significant Featu

summary_dur = tidy(model_AvgDur)
summary_dur$category = 'With Outlier'
summary_dur_outlier = tidy(model_dur)
summary_dur_outlier$category = 'Without Outlier'
dur_summary_combined <- rbind(summary_dur, summary_dur_outlier)
dur_summary_combined$significance <- ifelse(dur_summary_combined$p.value<0.05, 'Significant_Feature', '

summary_time = tidy(time_pattern_dur_model)
summary_time$category = 'With Outlier'
summary_time_outlier = tidy(model_time)
summary_time_outlier$category = 'Without Outlier'
time_summary_combined <- rbind(summary_time, summary_time_outlier)
time_summary_combined$significance <- ifelse(time_summary_combined$p.value<0.05, 'Significant_Feature',


p_coef1 <- ggplot(season_summary_combined, aes(fill=category, y=term, x=estimate)) +
  geom_bar(colour="black",position='dodge', stat='identity')  + ggtitle ("Seasonal Model Coefficients")

p_coef2 <- ggplot(time_summary_combined, aes(fill=category, y=term, x=estimate)) +
  geom_bar(colour="black",position='dodge', stat='identity')  + ggtitle ("Time Pattern Model Coefficien

p_coef3 <- ggplot(dur_summary_combined, aes(fill=category, y=term, x=estimate)) +
  geom_bar(colour="black",position='dodge', stat='identity')  + ggtitle ("Average Duration Model Coeffi

#grid.arrange(p_coef1, p_coef2, p_coef3, nrow = 3, ncol = 1,heights = c(5,5,5))
```

Upon retraining each model without the influential, we can observe that there was a slight change in the coefficients, which increases our confidence in the initial models that we trained. Furthermore, the number of significant vs insignificant features at a 5% threshold did not change much as well.

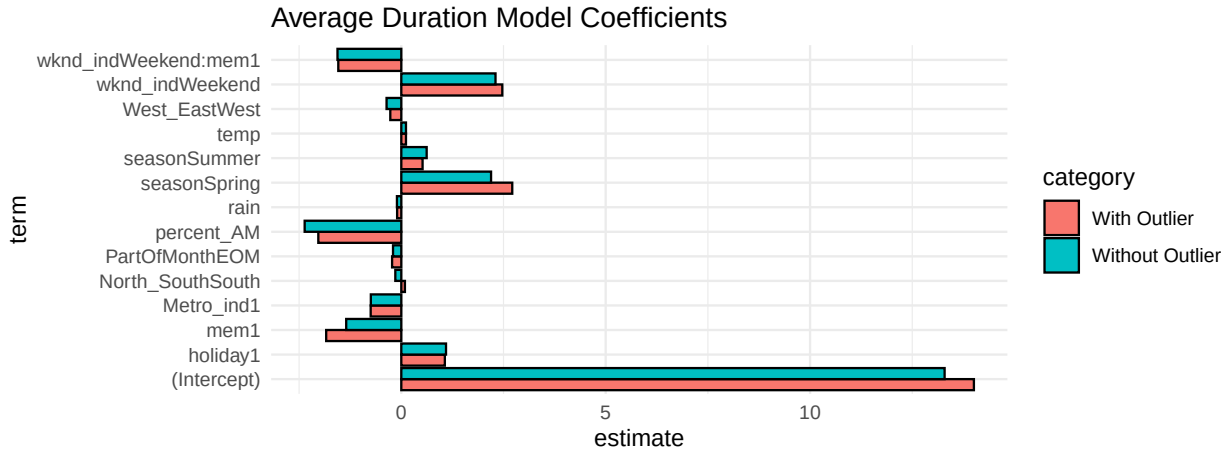
```
par(mfrow = c(3, 1))
p_coef1
```




```
p_coef2
```




```
p_coef3
```



```
dur_heatmap = dur_summary_combined %>%
    tabyl(category, significance)%>%
   as.data.frame()

season_heatmap = season_summary_combined %>%
    tabyl(category, significance)%>%
   as.data.frame()

time_heatmap = time_summary_combined %>%
    tabyl(category, significance)%>%
   as.data.frame()



knitr::kable(dur_heatmap, caption = "Average Duration Model")
```

Table 1: Average Duration Model

category	Not_Significant	Significant_Feature
With Outlier	2	12
Without Outlier	1	13

```
knitr::kable(season_heatmap, caption = "Seasonal Model")
```

Table 2: Seasonal Model

category	Significant Feature
With Outlier	11
Without Outlier	11

```
knitr::kable(time_heatmap, caption = "Time Pattern Model")
```

Table 3: Time Pattern Model

category	Not_Significant	Significant_Feature
With Outlier	5	4
Without Outlier	5	4

Ultimately, the R Squared of the models increased by a large amount in the duration and seasonal models, and decreased for the time pattern model. Of the 10,000 observations in the original dataset, less than 700 observations in each model training set was considered influential.

```
model_dur_outlier_rsquared = summary(model_dur)$r.squared
model_time_outlier_rsquared = summary(model_time)$r.squared
model_season_outlier_rsquared = summary(model_season)$r.squared

model_dur_rsquared = summary(model_AvgDur)$r.squared
```

```
model_time_rsquared = summary(time_pattern_dur_model)$r.squared
model_season_rsquared = summary(seasonal_effect_rev_model)$r.squared

dur_noninfl_obs =nrow(df_main_dur)
time_noninfl_obs =nrow(df_main_time)
season_noninfl_obs = nrow(df_main_season)
tot_obs = nrow(df_main)

rsquared_df <- tribble(
~Model, ~With_Outliers, ~Without_Outliers, ~Num_Influential_Obs,
'Duration',model_dur_rsquared, model_dur_outlier_rsquared, tot_obs - dur_noninfl_obs,
'Time Pattern', model_time_rsquared, model_time_outlier_rsquared,tot_obs - time_noninfl_obs,
'Seasonal', model_season_rsquared, model_season_outlier_rsquared, tot_obs - season_noninfl_obs
)
knitr::kable(rsquared_df, caption = "R Squared Changes")
```

Table 4: R Squared Changes

Model	With_Outliers	Without_Outliers	Num_Influential_Obs
Duration	0.0985064	0.1571804	669
Time Pattern	0.0046296	0.0039334	571
Seasonal	0.2781768	0.3839293	524

```
#strwrap(print(paste0("The Average Duration Model's R Squared went from ", model_dur_rsquared," to " ,m

#strwrap(print(paste0("The Time Pattern Model's R Squared went from ", model_time_rsquared," to " ,mode

#strwrap(print(cat(paste0("The Seasonal Model's R Squared went from ", model_season_rsquared," to " ,mo
```

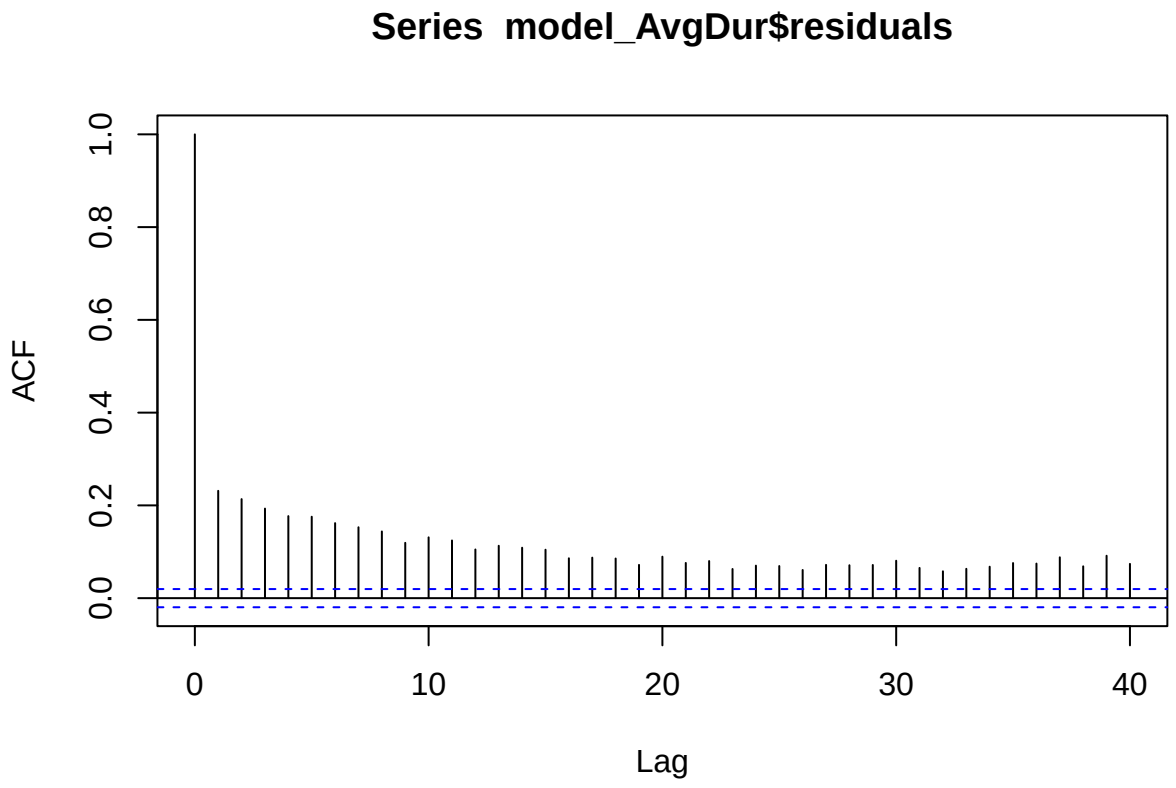
Autocorrelation Analysis

To further validate our analysis, we checked for autocorrelation in the data using ACF residual plots and the Durbin-Watson Test. The result of these checks could help us determine if we should pursue linear mixed models as potential next steps.

For ACF plots, we can observe whether or not there is correlation among the residuals at different lag intervals. For there to be no autocorrelation, we would expect the residuals to be close to 0. Using this method, it seems that each of the models have some level of autocorrelation, with the strongest autocorrelation being observed in the time pattern and the seasonal models.

The other method we explored was the Durbin-Watson test. The goal of this statistical test is to detect the presence of autocorrelation at using a lag period of 1. The null hypothesis is that there is no positive first order serial correlation in the data (i.e. correlation at 1 lag is less than or equal to 0) and the alternative hypothesis is that there is a positive first order serial correlation in the data (i.e. correlation at lag 1 is greater than 0). Given the p values and the autocorrelation statistics below, we must reject the null hypothesis using a threshold of 5% and conclude that each model has some level of autocorrelation. As a result, we suggest using linear mixed models moving forward.

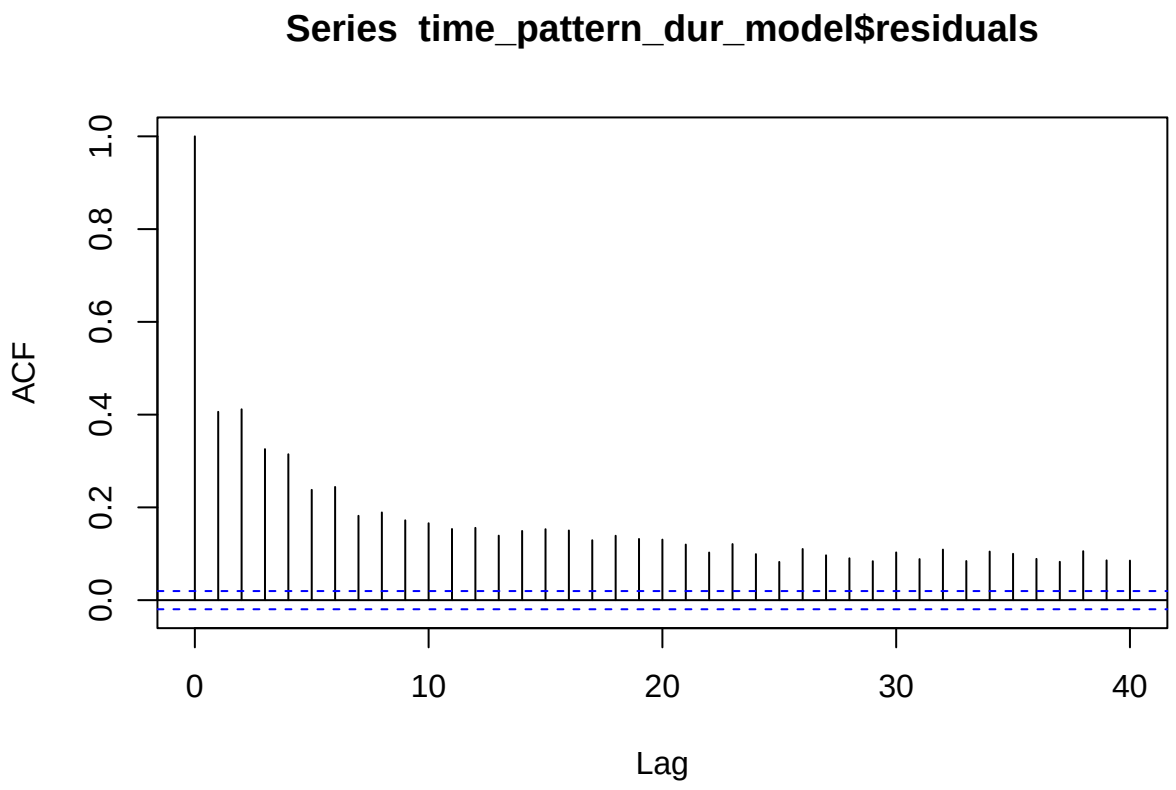
```
acf(model_AvgDur$residuals, type = "correlation")
```




```
durbinWatsonTest(model_AvgDur)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.2314049      1.537042       0
##  Alternative hypothesis: rho != 0
```

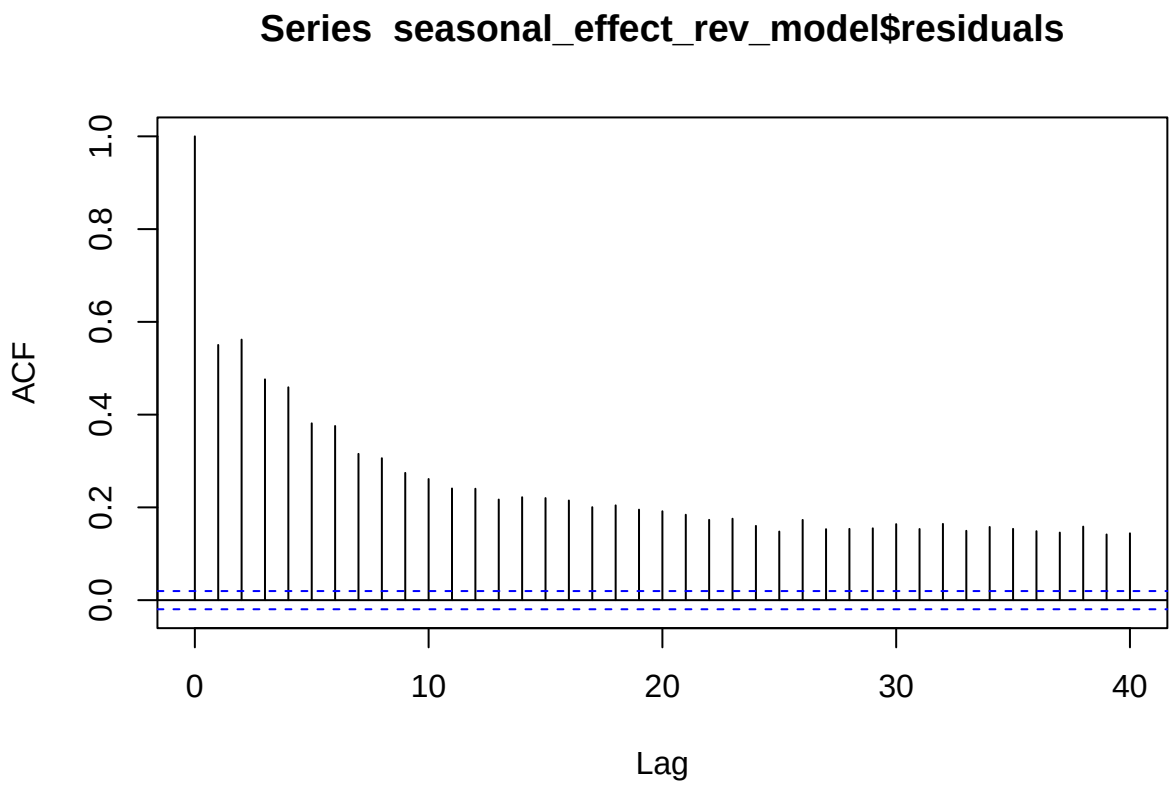
```
acf(time_pattern_dur_model$residuals, type = "correlation")
```




```
durbinWatsonTest(time_pattern_dur_model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.4060858      1.187774       0
##  Alternative hypothesis: rho != 0
```

```
acf(seasonal_effect_rev_model$residuals, type = "correlation")
```




```
durbinWatsonTest(seasonal_effect_rev_model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.5502328      0.899486       0
##  Alternative hypothesis: rho != 0
```

Exploratory Regression

(I would only keep the ones that bring new information and that answer business question) ## AM/PM Delta

```
model <- lm(n_AM_PM_delta ~ long_wknd_ind + season + rain_ind + mem, data = df_main) # Goal is to look
summary(model)
```

```
##
## Call:
## lm(formula = n_AM_PM_delta ~ long_wknd_ind + season + rain_ind +
##     mem, data = df_main)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -117.382   -2.083    1.433    4.928   24.147
##
```

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -4.1407     0.8188  -5.057 4.33e-07 ***
## long_wknd_indWeekday    1.7079     0.8182   2.087   0.0369 *
## long_wknd_indWeekend    0.4048     0.8293   0.488   0.6255
## seasonSpring           -0.1763     0.2692  -0.655   0.5126
## seasonSummer           -1.5288     0.2199  -6.951 3.86e-12 ***
## rain_indRain            1.6899     0.2006   8.424  < 2e-16 ***
## mem1                   -8.1851     0.1951 -41.950  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.726 on 9993 degrees of freedom
## Multiple R-squared:  0.1587, Adjusted R-squared:  0.1582
## F-statistic: 314.2 on 6 and 9993 DF,  p-value: < 2.2e-16
```

Trip length wkday/wknd

```
#df_main
# SHORTER TRIPS ON WEEKDAYS THAN WEEKENDS
model <- lm(avg ~ long_wknd_ind + season + rain_ind + mem, data = df_main)
summary(model)
```

```
##
## Call:
## lm(formula = avg ~ long_wknd_ind + season + rain_ind + mem, data = df_main)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.872  -3.611  -1.152   2.113  43.681
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            16.1900     0.5321  30.429  < 2e-16 ***
## long_wknd_indWeekday   -1.0186     0.5317  -1.916   0.0554 .
## long_wknd_indWeekend    0.6199     0.5389   1.150   0.2501
## seasonSpring            2.8607     0.1749  16.353  < 2e-16 ***
## seasonSummer            1.6624     0.1429  11.631  < 2e-16 ***
## rain_indRain           -1.0075     0.1304  -7.728  1.2e-14 ***
## mem1                   -2.4147     0.1268 -19.044  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.321 on 9993 degrees of freedom
## Multiple R-squared:  0.0812, Adjusted R-squared:  0.08064
## F-statistic: 147.2 on 6 and 9993 DF,  p-value: < 2.2e-16
```

trip length for members/non-members

```r
# MEMBERS TAKE SHORTER TRIPS
# MEMBERS TAKE LONGER TRIPS IN THE RAIN
model <- lm(avg ~ (rain_ind *mem) + long_wknd_ind, data = df_main)
summary(model)
```

```
##
## Call:
## lm(formula = avg ~ (rain_ind * mem) + long_wknd_ind, data = df_main)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.676  -3.701  -1.221   2.310  43.073
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           17.2772     0.5396  32.021  < 2e-16 ***
## rain_indRain          -1.4170     0.1938  -7.312 2.84e-13 ***
## mem1                  -2.5621     0.1633 -15.693  < 2e-16 ***
## long_wknd_indWeekday  -0.7002     0.5362  -1.306   0.1916
## long_wknd_indWeekend   0.9991     0.5438   1.837   0.0662 .
## rain_indRain:mem1      0.5474     0.2647   2.068   0.0387 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.411 on 9994 degrees of freedom
## Multiple R-squared:  0.05475,    Adjusted R-squared:  0.05428
## F-statistic: 115.8 on 5 and 9994 DF,  p-value: < 2.2e-16
```

Revenue per trip: seasonal effect

```r
#HIGHER REVENUE PER TRIP IN SPRING AND SUMMER THAN WINTER
model <- lm(rev_per_trip ~ long_wknd_ind + season + rain_ind, data = df_main)
summary(model)
```

```
##
## Call:
## lm(formula = rev_per_trip ~ long_wknd_ind + season + rain_ind,
##     data = df_main)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5592 -0.6838 -0.1753  0.4379  6.5634
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.60460    0.13423  26.854  < 2e-16 ***
## long_wknd_indWeekday -0.16215    0.13525  -1.199    0.231
## long_wknd_indWeekend  0.19319    0.13702   1.410    0.159
## seasonSpring          0.58644    0.04544  12.905  < 2e-16 ***
## seasonSummer          0.32748    0.03601   9.094  < 2e-16 ***
```

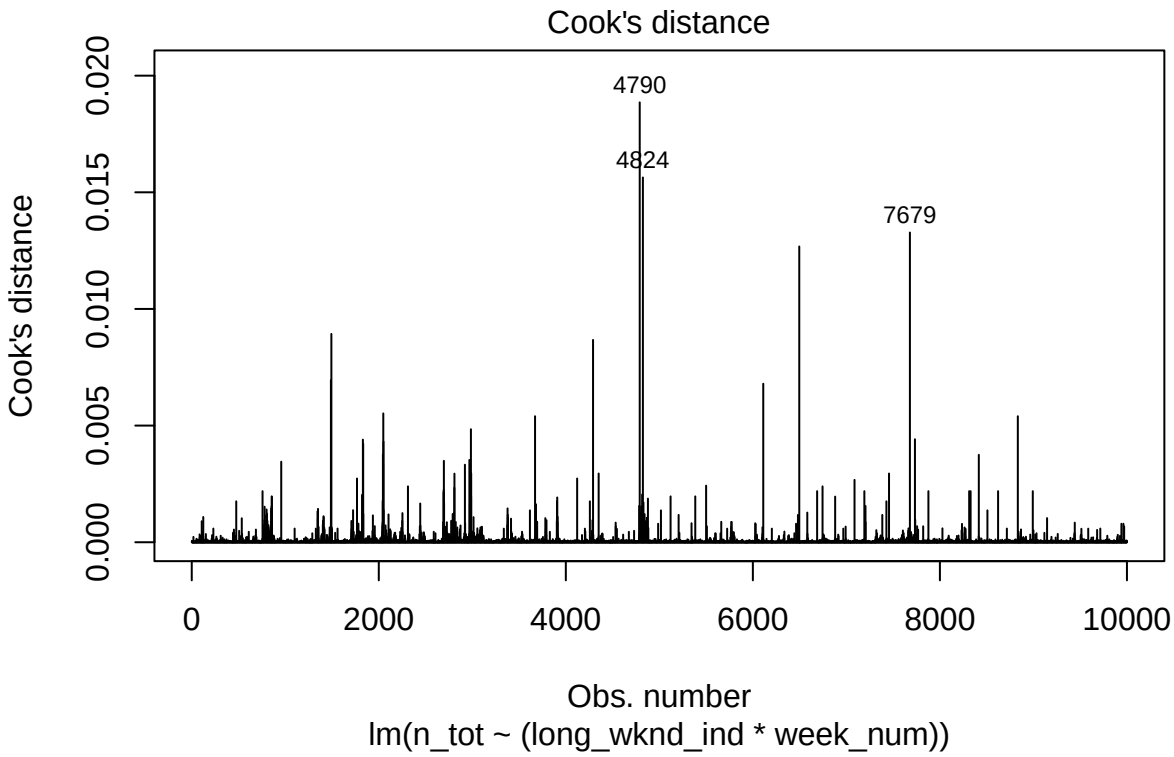
```
## rain_indRain         -0.18870    0.03346  -5.639 1.81e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.105 on 4728 degrees of freedom
##   (5266 observations deleted due to missingness)
## Multiple R-squared:  0.06704,    Adjusted R-squared:  0.06605
## F-statistic: 67.95 on 5 and 4728 DF,  p-value: < 2.2e-16
```

Number of trips as season advances

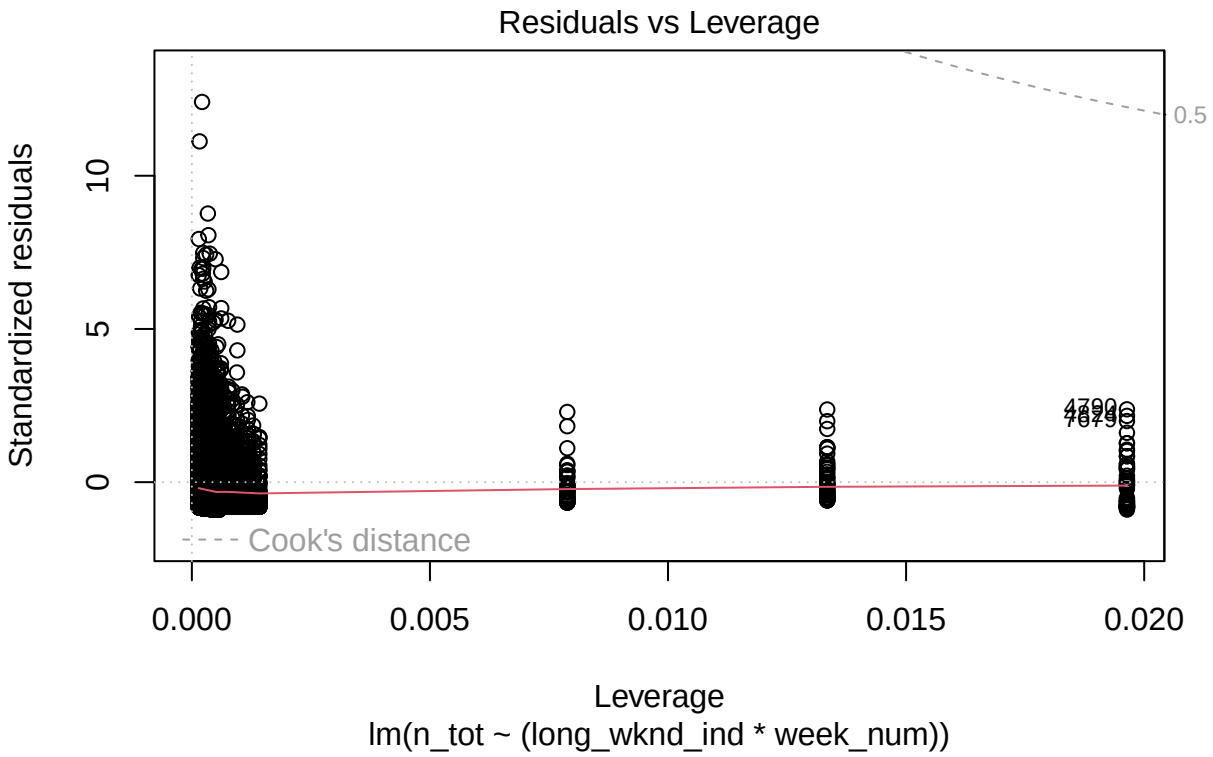
```
#df_main
# AS BIXI SEASON GOES ON, WEEKDAY NUMBER OF TRIPS GO UP WITH A STATISTICAL SIGNIFICANCE
model <- lm(n_tot ~ (long_wknd_ind*week_num), data = df_main)
summary(model)
```

```
##
## Call:
## lm(formula = n_tot ~ (long_wknd_ind * week_num), data = df_main)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.290 -15.548  -8.577   6.966 294.938
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    29.4754     7.7757   3.791 0.000151 ***
## long_wknd_indWeekday          -13.4649     7.8444  -1.716 0.086103 .
## long_wknd_indWeekend           -8.7238     7.9243  -1.101 0.270971
## week_num                       -0.3531     0.2302  -1.534 0.125152
## long_wknd_indWeekday:week_num   0.4896     0.2325   2.106 0.035231 *
## long_wknd_indWeekend:week_num   0.3068     0.2353   1.304 0.192407
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.77 on 9994 degrees of freedom
## Multiple R-squared:  0.002469,   Adjusted R-squared:  0.001969
## F-statistic: 4.946 on 5 and 9994 DF,  p-value: 0.0001588
```

```
plot(model,4)
```


Cook's distance
4790
4824
7679
Cook's distance
0.000
0.005
0.010
0.015
0.020
0
2000
4000
6000
8000
10000
Obs. number
lm(n_tot ~ (long_wknd_ind * week_num))

```
plot(model,5)
```



Limitations and shortcomings

- Causation vs. Correlation: The regression model captures relationships but does not establish causation.
- Data Exclusions: The data only considers trips under 60 minutes, which might exclude a segment of users who use BIXI for longer journeys.
- Other External Factors: Events, road conditions, or public transportation disruptions can affect BIXI usage but are not captured in the dataset.
- Mention Auto-correlation (Chike)

Conclusion

(Review the research questions that were answered)

Contribution

Charles Julien :

Gabriel Jobert :

Chike Odenigbo:

Atul Sharma: