Seoul Bike Share

A Regression Analysis on Environmental Variables

Group 1: Snap Storm Bikers

Anish Singla - 1005801401 Malhar Pandya - 1005893008 Tan Lin -

06/12/2021

Background

Rental bikes serve as the backbone of affordable transportation for densely populated urban cities like Seoul, South Korea as they make mobility easier and faster. These bikes are very important as they can influence the population to lead a healthier lifestyle and a greener one (Nikitas, 2019). In fact, there are many countries that have more bikes than cars, to name some would be: Denmark, China, and the Netherlands (2011).

The widespread availability of bikes can reduce car emissions (Nikitas, 2019) as well as it becomes much easier to commute through an urban city. We know that at 2013 there were about 52 countries that have a bike-sharing program (Midgley, 2015), these bike-sharing programs have had a real impact on the majority of these cities as they have also helped reduce traffic congestion along with the healthier and greener lifestyle (Ngo, 2021) I mentioned before. Seoul is known to be a great city to bike through as recently they have built massive cycling infrastructure (Jensen, 2019) which helps promote the use of bikes over other types of travel.

A crucial part of providing such an infrastructure involves in maintaining a sufficient supply of rental bikes in the city to ensure the availability of the service to the public.

One way to predict the required supply would be to identify patterns between rental bike usage and environmental factors such as weather and time. Not only will this help us directly infer the rental bike usage, but also assist in identifying maintenance schedules that cause the least hindrance to people, which will further improve the service.

This leads us to the question, can we accurately predict the number of rented bikes for a given time based on the environmental conditions? The goal of this study is to answer this question and determine how weather and time affect Seoul's rental bike usage.

Exploratory Data Analysis

Data Size and Validation

After loading the data set and modifying the column names, we first identify if the data set has missing entries and perform remediation if needed.

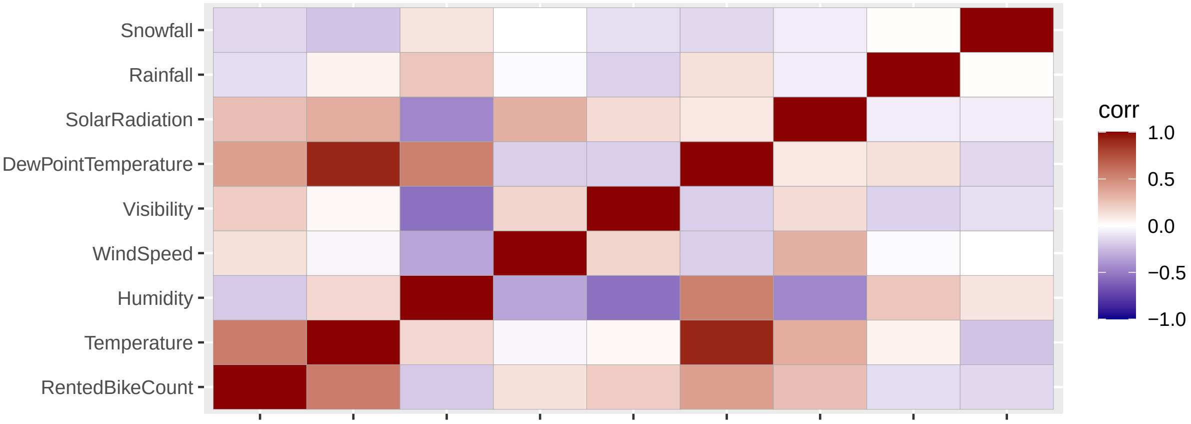
```
Missing entries: 0 , Number of rows: 8760
```

Data Cleaning and Transformation

We can now filter out unwanted data based on how it affects our research question.

1. We can clearly see that the entries in the data for which the bike share is non-functional bears no inferential value for predicting bike usage, and can therefore remove those entries from the data set. Once done, the "Functioning Day" covariate serves no purpose so we remove the column entirely

2. Now we identify redundancies in the data by checking the correlation between the quantitative covariates in the data set. We can see that "Temperature" and "Dew Point Temperature" have a significant correlation, and of the two, "Temperature" has a higher correlation with "Rented Bike Count". Therefore, in order to preserve the independence constraint among covariates, we remove the "Dew Point Temperature" covariate.



3. Next, we observe that "Rainfall" and "Snowfall" have quite a low correlation with "Rented Bike Count". Our primary hypothesis is that this is due to the large number of entries for which these covariates have a value of 0.

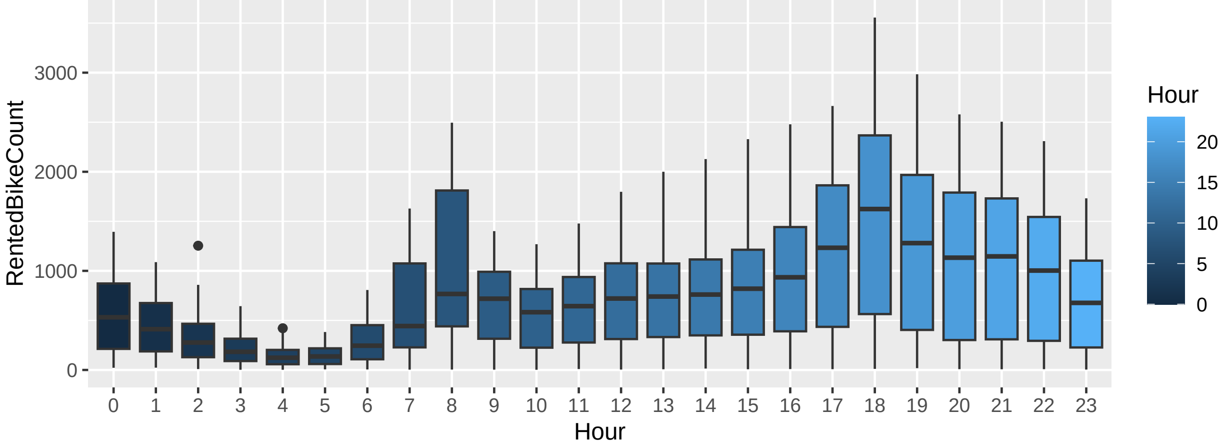
```
The number of entries with no rainfall: 7949
```

```
The number of entries with no snowfall: 8022
```

As predicted, a majority of the entries are like this. We remedy this by converting both these covariates into booleans, based on whether their values are 0 or not.

To clean the data further, we need to split the "Date" variable into days and months.

4. We now attempt to identify when rental bikes are being used the most. We hypothesize that they are used mainly for work or school transportation, and bike usage will peak when work or school begins and ends. We visualize the bike usage per hour as a boxplot



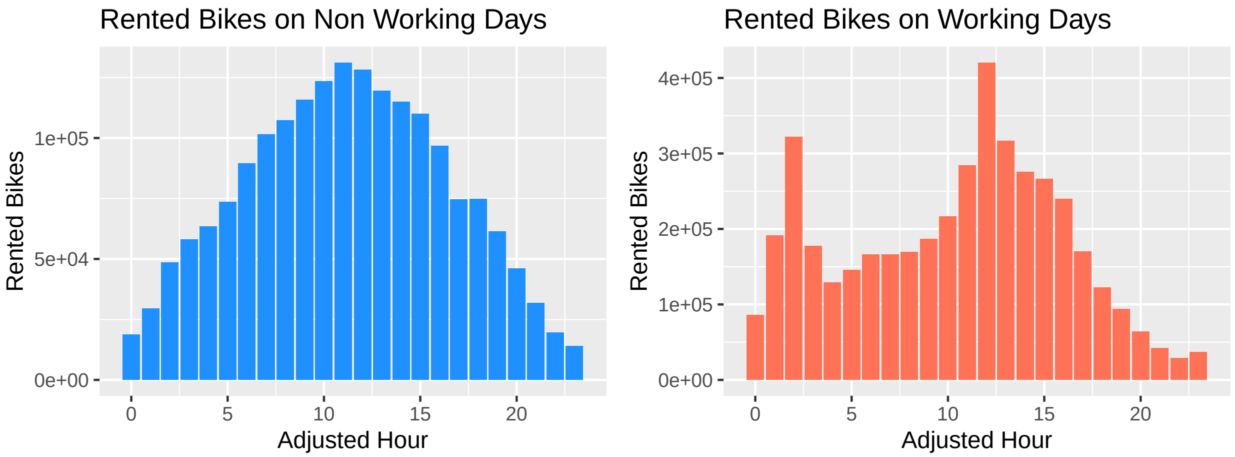
We can see peaks happening at the 8 and 18 hour marks, which we believe are the rush hour times for Seoul (Ryerson, 2018).

5. We also transform the hours so that they begin in the morning (6 a.m.) rather than midnight.
6. Additionally, we predict that the residents of Seoul use bikes to travel to work and/or school. We can check this by comparing the aggregated number of rented bikes for workdays and non-workdays (weekends and holidays).

```
Number of bikes rented on a working day: 4319348

Number of bikes rented on non-working days: 1852966
```

To include this information we can introduce a new boolean value to identify working days.

```
Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine
```



7. After observing the graphs, we identified two major points of interest; Rush hour and Working Times. We did some research and discovered the average Seoul resident's work times (Ryerson, 2018) and the rush hour time (Ae, 2021).

```
#Create Rush Hour
data$RushHour = 0
for(i in seq(1:length(data$WeekDay))){
  if((data[i,"WorkDay"] == 1) & (data[i,"Hour"] %in% c(0,1,2,11,12))){
    data[i,"RushHour"] = 1
  }
}
#Create Working
data$Working = 0
for(i in seq(1:length(data$WeekDay))){
  if((data[i,"WorkDay"] == 1) & (data[i,"Hour"] > 2) & (data[i,"Hour"] < 11))
  {
    data[i,"Working"] = 1
  }
}
```

Model Generation and Fitting

8. Before investigating possible models, we split the data into a training and a testing data set.

```
set.seed(1005102871)
sample = sample(1:length(data[,1]),size=round(length(data[,1])*0.8,0))
data.train = data[sample,]
data.test = data[-sample,]
```

9. We decided to build our model manually. We started by using hour and hourˆ2 as our starting model. We chose this because we thought that the overall shape of Rented Bikes would be well explained with a quadratic. This explained the Rental Bike Activity during holidays and weekends well, but fell short during the weekdays.

10. Building on top of this model, we used rush hour and working times to better explain the variations that occur during the work days.

11. Adding Temperature, Humidity, Rainfall, and Snowfall, the new model could explain the differences in Rental Bike Count that occurred through seasonal and daily weather patterns.

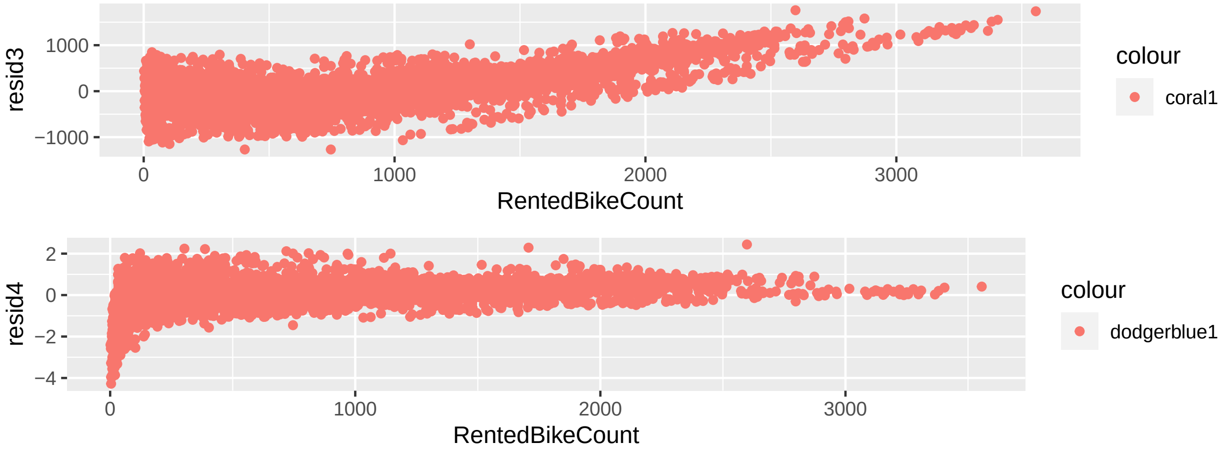
12. Plotting the residuals for fit3, we noticed that the residuals seem to sloping upwards. To address this, we decided to transform RentedBikeCount using log.

```
fit4 = lm(log(RentedBikeCount) ~ Hour + I(Hour^2) + Working + RushHour + Temperature + I(Temperature^2)
summary(fit4)
```

```
Call:
lm(formula = log(RentedBikeCount) ~ Hour + I(Hour^2) + Working + 
    RushHour + Temperature + I(Temperature^2) + Humidity + Rainfall + 
    Snowfall + Snowfall + WindSpeed, data = data.train)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-4.2773 -0.3120  0.0498  0.3584  2.4394 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       5.580e+00  4.135e-02 134.943  < 2e-16 ***
Hour              1.969e-01  4.923e-03  40.009  < 2e-16 ***
I(Hour^2)        -9.571e-03  2.106e-04 -45.454  < 2e-16 ***
Working          -1.742e-01  2.194e-02  -7.941 2.33e-15 ***
RushHour          6.550e-01  2.539e-02  25.796  < 2e-16 ***
Temperature       8.215e-02  1.509e-03  54.445  < 2e-16 ***
I(Temperature^2) -1.226e-03  5.275e-05 -23.232  < 2e-16 ***
Humidity         -9.118e-03  4.751e-04 -19.193  < 2e-16 ***
Rainfall         -1.836e+00  3.423e-02 -53.653  < 2e-16 ***
Snowfall          5.192e-02  3.694e-02   1.406     0.16    
WindSpeed        -6.918e-02  8.240e-03  -8.395  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6132 on 6761 degrees of freedom
Multiple R-squared:  0.7221,	Adjusted R-squared:  0.7217 
F-statistic:  1757 on 10 and 6761 DF,  p-value: < 2.2e-16
```

```r
#Residual Plots
resid3 = resid(fit3)
resid4 = resid(fit4)
p1 = ggplot(data.train) + geom_point(mapping=aes(x=RentedBikeCount,y=resid3,color="coral1"))
p2 = ggplot(data.train) + geom_point(mapping=aes(x=RentedBikeCount,y=resid4,color="dodgerblue1"))
grid.arrange(p1,p2)
```



13. We enhanced our model by factoring in the fact that most people bike during the day.

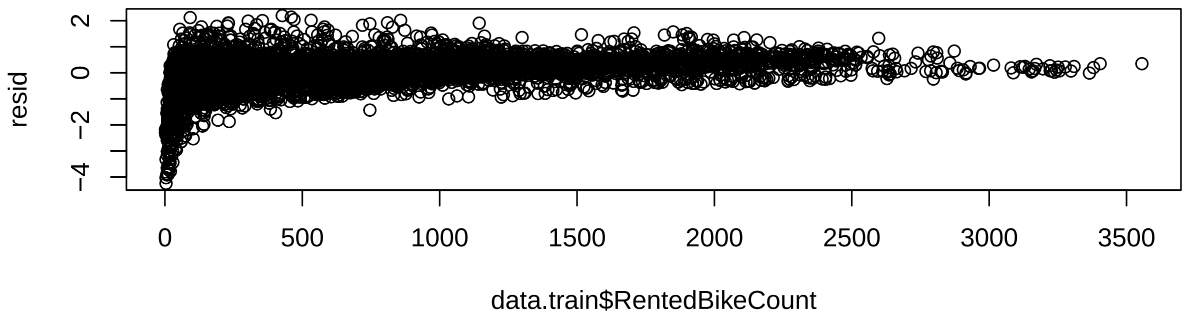
```r
fit5 = lm(log(RentedBikeCount) ~ Hour + I(Hour^2) + Working + RushHour + Temperature + I(Temperature^2)
summary(fit5)
```

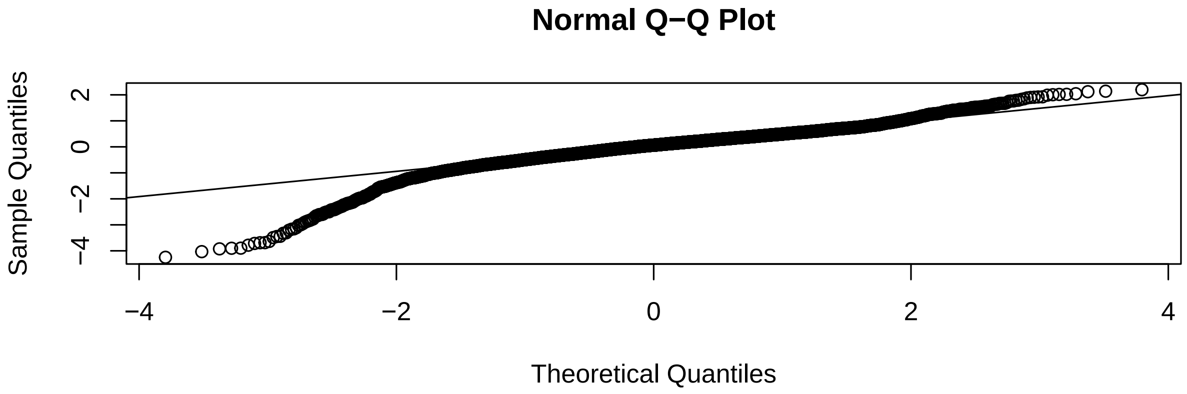
```
Call:
lm(formula = log(RentedBikeCount) ~ Hour + I(Hour^2) + Working +
    RushHour + Temperature + I(Temperature^2) + Humidity + Rainfall +
    Snowfall + SolarRadiation + Humidity:Rainfall, data = data.train)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2487 -0.2988  0.0652  0.3560  2.1968

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.624e+00  4.357e-02 129.081  < 2e-16 ***
Hour               1.845e-01  4.613e-03  40.006  < 2e-16 ***
I(Hour^2)         -9.258e-03  1.966e-04 -47.101  < 2e-16 ***
Working           -9.942e-02  2.267e-02  -4.385 1.17e-05 ***
RushHour           6.103e-01  2.528e-02  24.142  < 2e-16 ***
Temperature        8.688e-02  1.514e-03  57.399  < 2e-16 ***
I(Temperature^2)  -1.258e-03  5.158e-05 -24.382  < 2e-16 ***
Humidity          -1.001e-02  5.200e-04 -19.250  < 2e-16 ***
Rainfall           1.961e+00  2.793e-01   7.023 2.38e-12 ***
Snowfall           4.365e-02  3.645e-02   1.197    0.231
SolarRadiation    -1.443e-01  1.348e-02 -10.706  < 2e-16 ***
Humidity:Rainfall -4.270e-02  3.106e-03 -13.747  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.6024 on 6760 degrees of freedom
Multiple R-squared:  0.7319,    Adjusted R-squared:  0.7314
F-statistic:  1677 on 11 and 6760 DF,  p-value: < 2.2e-16
```

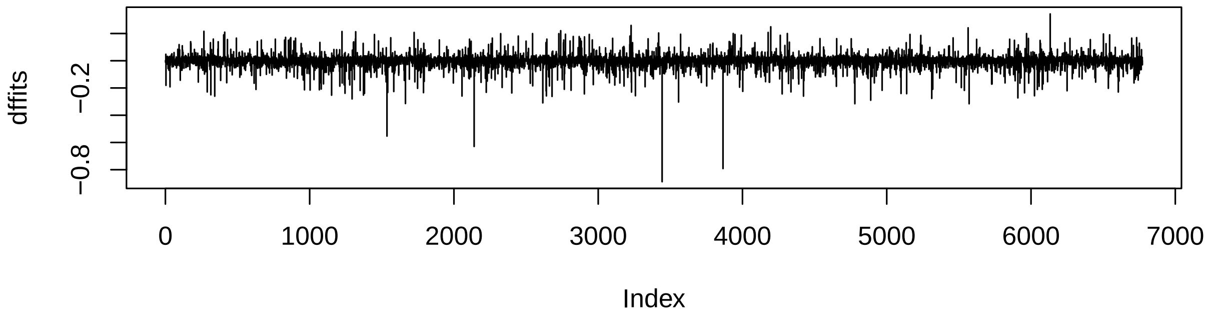
Model Validation and Analysis

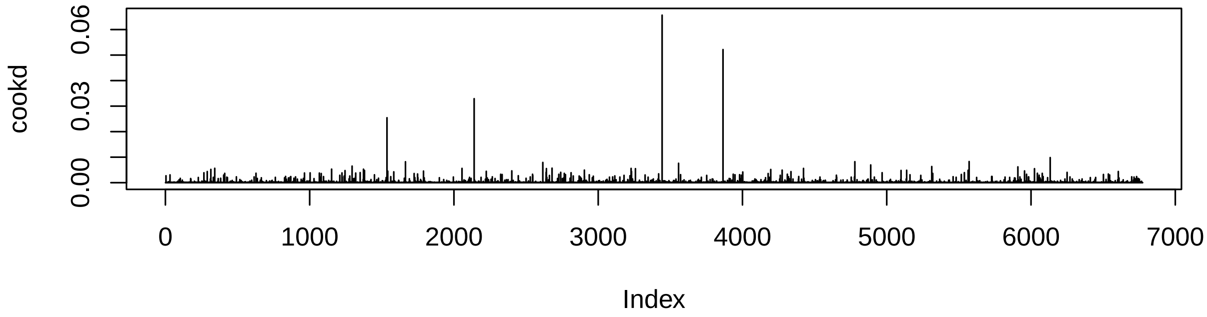
14. We created a function to evaluate our model.

```r
#Model Evaluation Function
evaluate_fit = function(fit){
  resid = residuals(fit)
  #Residual Plot
  plot(data.train$RentedBikeCount,resid)

  #QQ Plot
  qqnorm(resid)
  qqline(resid)

  #DFFITS
  dffits = dffits(fit)
  plot(dffits, type = 'h')

  #Cook's Distance
  cookd = cooks.distance(fit)
  plot(cookd,type='h')

  #DFBETAS
  dfbetas = dfbetas(fit)
  plot(dfbetas,type='h')

  #Variance Inflation Factor
  vif(fit)

}
evaluate_fit(fit5)
```

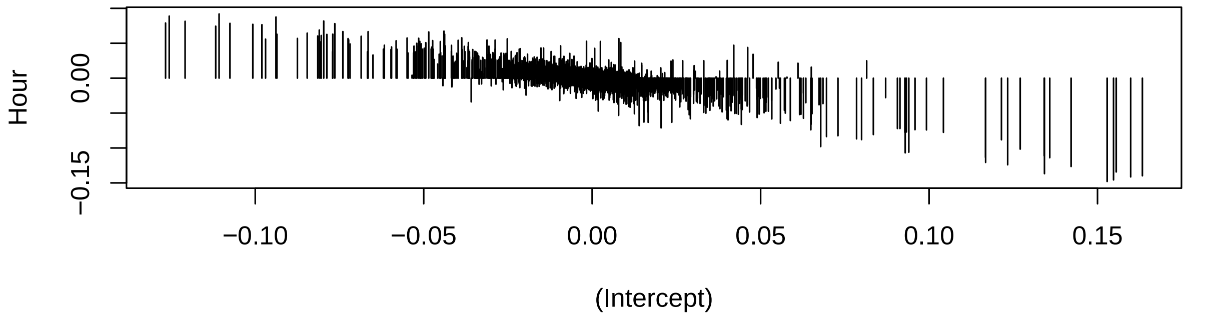








```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
```



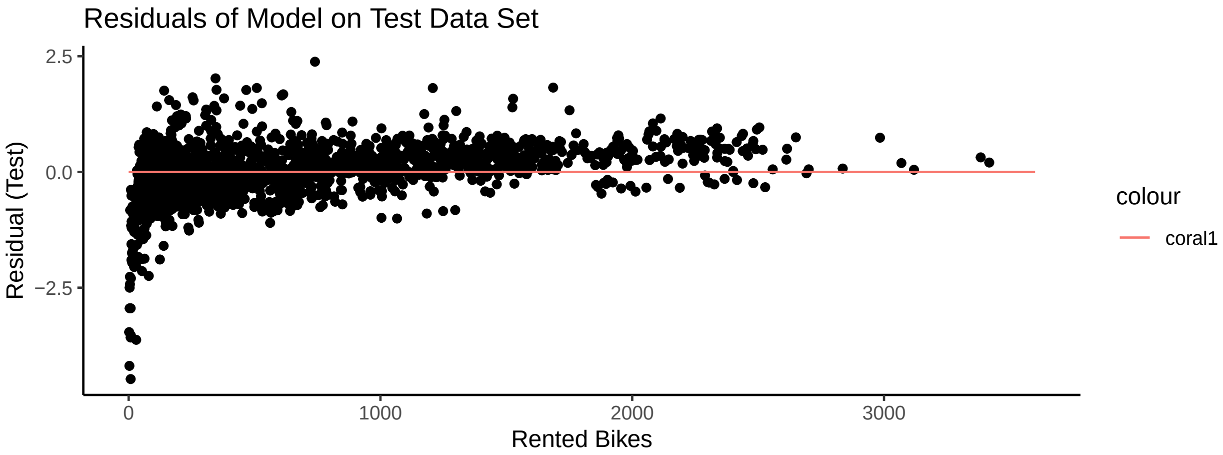
```
             Hour         I(Hour^2)           Working          RushHour
        18.928914         19.545567          1.672031          1.447443
      Temperature  I(Temperature^2)          Humidity          Rainfall
         6.280434          5.079018          2.123341         84.117232
         Snowfall    SolarRadiation Humidity:Rainfall
         1.274514          2.561192         85.169688
```

15. We then validated our model using the test data that we withheld.From the residual plot, it appears that our model works similarly on both the training data and the testing data. Our adjR2 is lower on the testing data compared to the training data, but that was expected.

```
predictions = predict(fit5, data.test)
data.frame( R2 = R2(predictions, data.test$RentedBikeCount),
            RMSE = RMSE(predictions, data.test$RentedBikeCount),
            MAE = MAE(predictions, data.test$RentedBikeCount))
```

```
         R2     RMSE      MAE
1 0.5855095 973.0091 731.5079
```

```
resid.test = log(data.test$RentedBikeCount) - predictions
ggplot(mapping=aes(x=data.test$RentedBikeCount,y=resid.test)) + geom_point() + geom_line(mapping = aes(
```



“‘ ## Conclusion Model Inferences: * People like to bike the higher the temperature is, but there is a limit * People do not like to bike in the rain, though a light drizzle has a much smaller impact than a storm * People are renting fewer bikes when they are working whereas they are renting more during rush hour * The more humid it is, the less likely people want to rent bikes Impact: The Seoul Bike System can ensure the availability of rental bikes by anticipating rises in demand caused by weather. Hot and dry summers would lead to a rise in demand and with out findings, they can better prepare for this surge. Limitations: Our model does not accurate predict bike rentals when they are low. The residuals are also not randomly scattered; Our model does a better job at predicting rental bike usage when the number is high. This is likely due to the fact that we transformed the number of rented bikes using log. Future Research: To more accurately maintain the availability of bikes, we would need information on the location of different bike locations. With this information, we would be able to better predict when and where demand for bikes will be high. This would allow the Seoul Bike Sharing System to utilize their existing bikes more efficiently. ## References 1. Jensen, E. (2019, January 1). Cycling in South Korea: 5 best paths, when to go, and travel tips. Retrieved December 06, 2021, from https://www.bookmundi.com/t/cycling-in-south-korea-5-best-paths 2. Nikitas, A. (2019, October 16). The global bike sharing boom – why cities love a cycling scheme. Retrieved December 06, 2021, from https://theconversation.com/the-global-bike-sharing-boom-why-cities-love-a-cycling-scheme-53895 3. Top 10 countries with most bicycles per capita: Top 10 hell. (2011, March 14). Retrieved December 06, 2021, from http://top10hell.com/top-10-countries-with-most-bicycles-per-capita/ 4. Midgley, P. (2015, February 22). Bike Sharing Systems - United Nations. Retrieved December 6, 2021, from https://sustainabledevelopment.un.org/content/documents/4803Bike%20Sharing%20UN%20DESA.pdf 5. Ngo, H. (2021, January 12). Why some bike shares work and others don’t. Retrieved December 06, 2021, from https://www.bbc.com/future/article/20210112-the-vast-bicycle-graveyards-of-china 6. Ryerson, L. (2018, May 15). What an average work day looks like in 18 countries around the world. Retrieved December 07, 2021, from https://www.insider.com/office-work-day-around-the-world-2018-5 7. Ae, C. (2021, August 28). What time is rush hour in Seoul? Retrieved December 07, 2021, from https://thekoreanguide.com/what-time-is-rush-hour-in-seoul/ Footer © 2023 GitHub, Inc. Footer navigation Terms Privacy Security Status Docs Contact GitHub Pricing API Training Blog About c67-case-study/CaseStudy_Final.Rmd at main · anisin22/c67-case-study