# Building a Linear Regression Model to Predict Total Daily Rental Bike Demands

Helen Liang, Ivy Zhao, Xiaotong Zhao

# Dataset & Research Question

- The Bike Sharing Dataset (source: UCI Machine Learning Repository) includes the daily rental bike demands (casual and registered) from the Capital bikeshare system spanning from 2011 to 2012.
- This dataset contains 731 observations and 16 variables. We excluded time variables (e.g., date, month, etc.) and focused on using 9 of those 16 variables for our study (e.g., weather, season, normalized temperature, etc).
- Research Question: Can we use the Bike Sharing Dataset to create a linear regression model to predict the daily rental bike demands?
- Outcome: Daily rental bike demands.

https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset

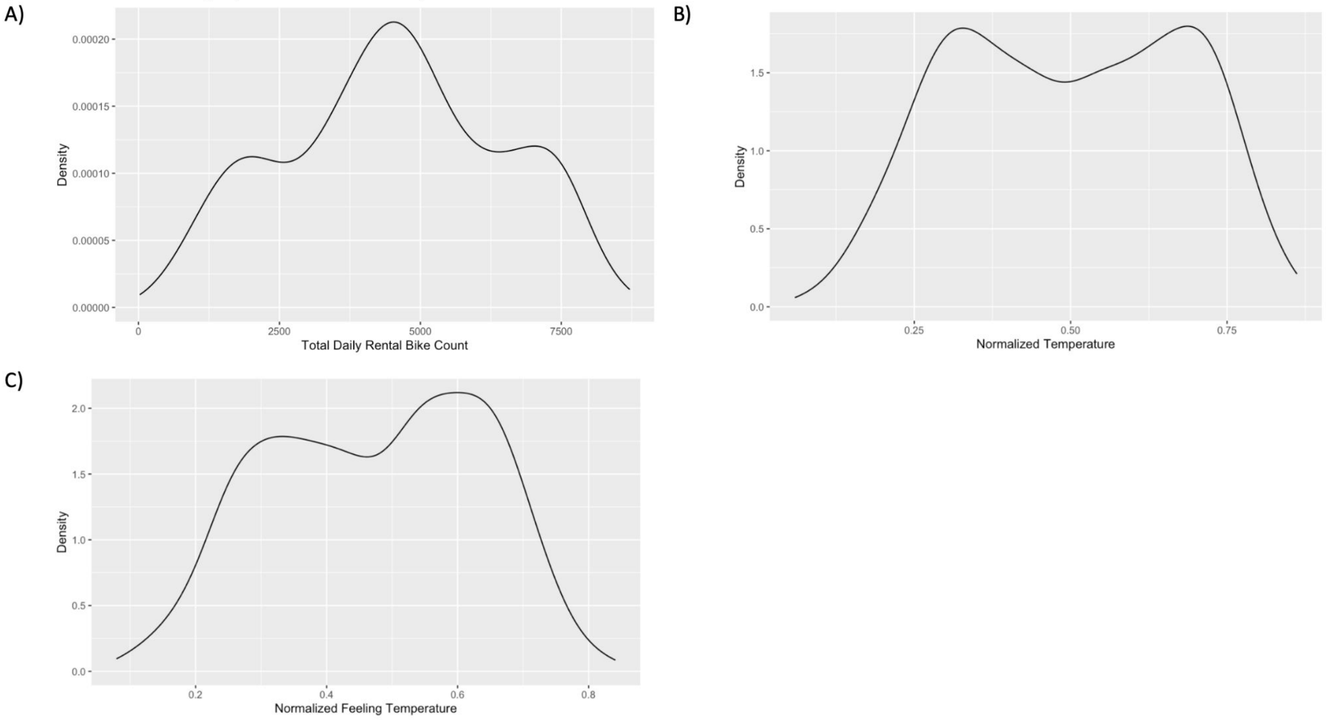
**Table 1. Summary of the Bike Sharing Dataset (n=731), 2011-2012.** Mean (SD), median (min, max) are reported for continuous variables. Frequencies (%) are reported for categorical variables.

| | **Overall (n=731)** |
|---|---|
| **Total Daily Rental Bike Count** | |
| Mean (SD) | 4504 (1937.21) |
| Median [Min, Max] | 4548 [22, 8714] |
| **Year** | |
| 2011 | 365 (49.93%) |
| 2012 | 366 (50.07%) |
| **Season** | |
| Spring | 181 (24.76%) |
| Summer | 184 (25.17%) |
| Fall | 188 (25.72%) |
| Winter | 178 (24.35%) |
| **Holiday** | |
| No | 710 (97.13%) |
| Yes | 21 (2.87%) |
| **Workday** | |
| No | 231 (31.60%) |
| Yes | 500 (68.40%) |
| **Weather** | |
| Clear, Few Clouds, Partly Cloudy | 463 (63.34%) |
| Mist + Cloudy, Mist + Broken Clouds, Mist + Few Clouds, Mist | 247 (33.79%) |
| Light Snow, Light Rain, Thunderstorm + Scattered Clouds, Light Rain + Scattered Clouds | 21 (2.87%) |
| Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog | 0 |
| **Normalized Temperature** | |
| Mean (SD) | 0.50 (1.83) |
| Median [Min, Max] | 0.50 [0.06, 0.86] |
| **Normalized Feeling Temperature** | |
| Mean (SD) | 0.47 (1.63) |
| Median [Min, Max] | 0.49 [0.08, 0.84] |
| **Normalized Humidity** | |
| Mean (SD) | 0.63 (1.42) |
| Median [Min, Max] | 0.63 [0, 0.97] |
| **Normalized Windspeed** | |
| Mean (SD) | 0.19 (7.75) |
| Median [Min, Max] | 0.18 [0.02, 0.51] |

Acryonym: SD (standard deviation)

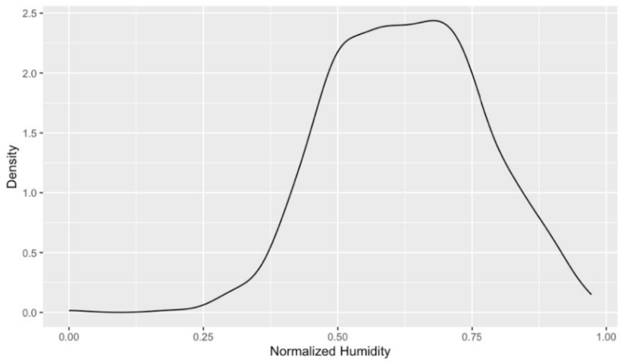
# Univariable Plots

**Figure 1.** Univariable plots for continuous variables. A) Bike Count. B) Normalized Temperature. C) Normalized Feeling Temperature. D) Normalized Humidity. E) Normalized Wind Speed.

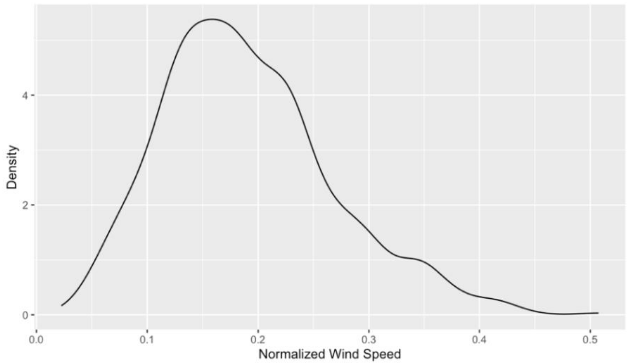


# Univariable Plots (cont.)

D)

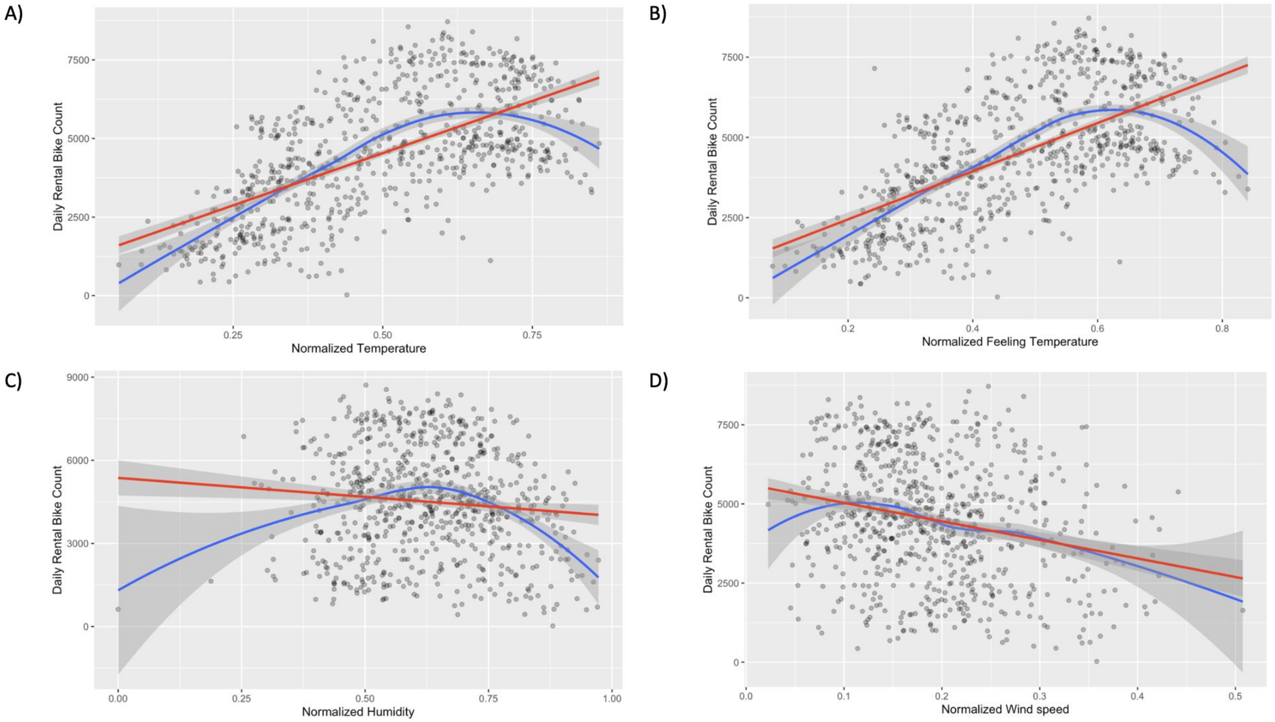



E)



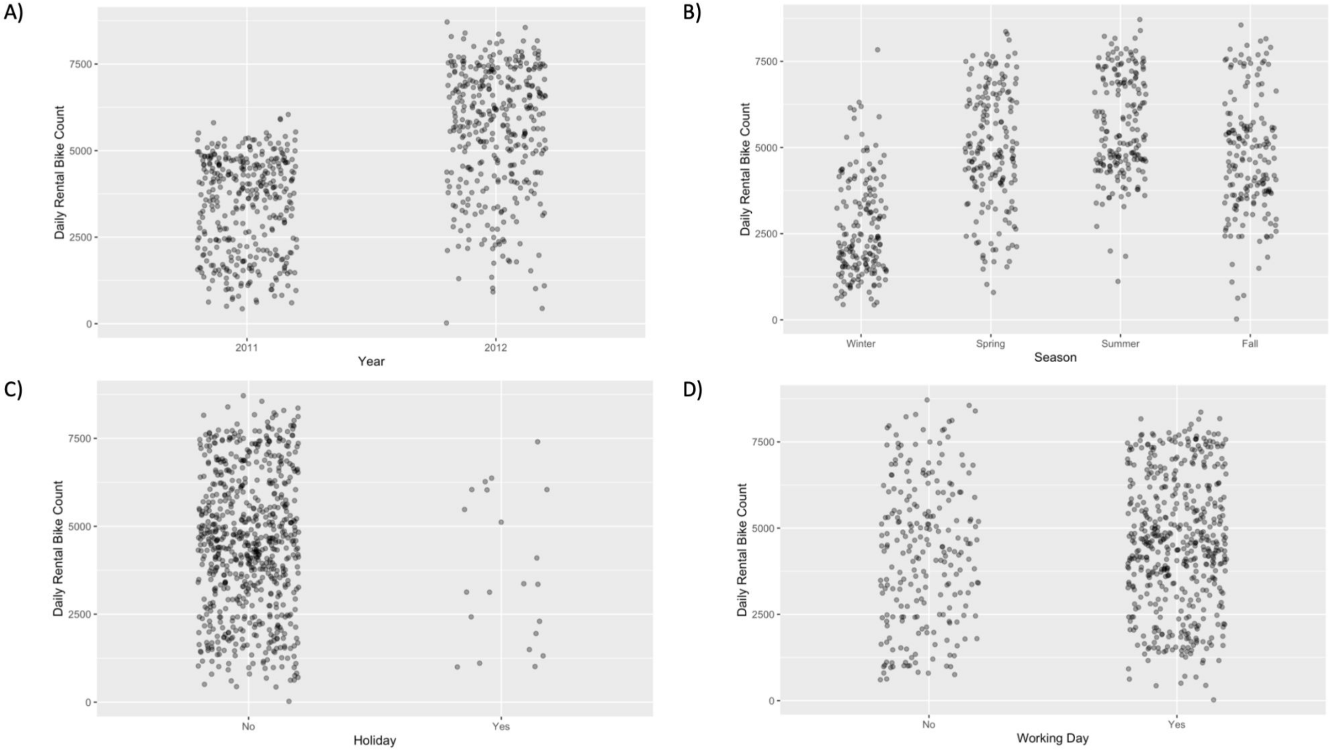
# Bivariable Plots

**Figure 2.** Bivariable plots for continuous variables. A) Normalized Temperature. B) Normalized Feeling Temperature. C) Normalized Humidity. D) Normalized Wind Speed.



# Bivariable Plots (cont.)

**Figure 3.** Bivariable plots for categorical variables. A) Year. B) Season. C) Holiday. D) Working Day.



# Initial Model

- Our adjusted $R^2$ for the initial model is **0.7914**.
- We will use diagnostic plots to check for model assumptions (e.g., normality, homoscedasticity, influential outliers, etc.)

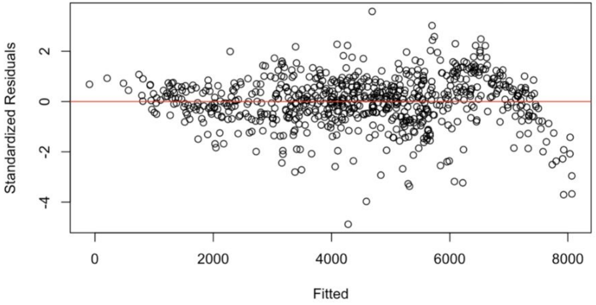
```
Call:
lm(formula = cnt ~ ., data = bikeDf)

Residuals:
    Min      1Q  Median      3Q     Max 
-4258.3  -458.1    75.1   539.6  3149.2 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1730.19     235.59   7.344 5.62e-13 ***
yr           2034.07      66.09  30.778  < 2e-16 ***
season        409.40      32.20  12.714  < 2e-16 ***
holiday      -621.06     202.71  -3.064 0.002267 ** 
workingday    124.12      73.02   1.700 0.089582 .  
weathersit   -580.37      79.20  -7.328 6.28e-13 ***
temp         2326.31    1421.46   1.637 0.102158    
atemp        3312.12    1609.35   2.058 0.039945 *  
hum         -1202.79     315.67  -3.810 0.000151 ***
windspeed   -2582.41     462.54  -5.583 3.35e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 884.7 on 721 degrees of freedom
Multiple R-squared:  0.794,	Adjusted R-squared:  0.7914 
F-statistic: 308.8 on 9 and 721 DF,  p-value: < 2.2e-16
```

**cnt = 1730.19 + 2034.07 (yr) + 409.40 (season) − 621.06 (holiday) + 124.12 (workingday) − 580.37 (weathersit) + 2326.31 (temp) + 3312.12 (atemp) − 1202.79 (hum) − 2582.41 (windspeed)**

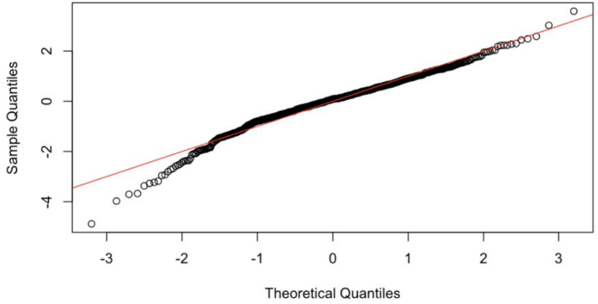
# Initial Model Plots

**Figure 4.** Plots for the initial model. A) Residual vs. Fitted Plot. B) Quantile-Quantile Plot. C) Histogram. D) Fitted vs. Square-Root of Standardized Residuals. E) Residuals vs. Leverages. F) Half-Norm Plot for Leverages. G) Half-Norm Plot for Cook's Distances.

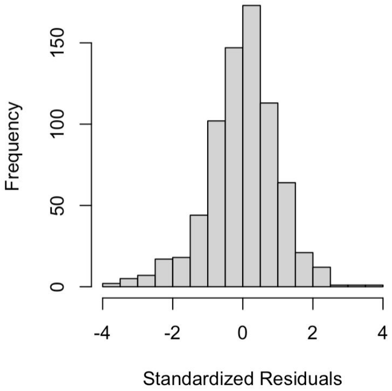
A)




B)

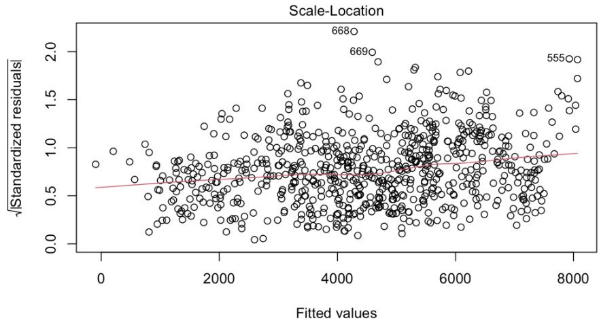



C)

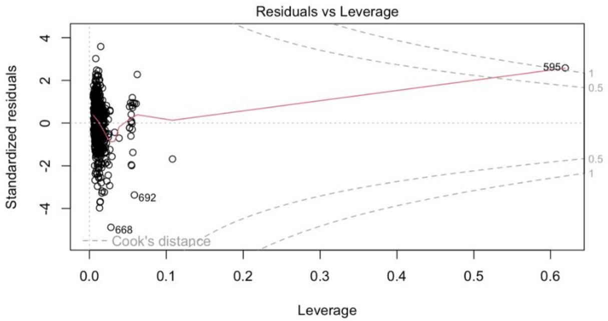


# Initial Model Plots (cont.)

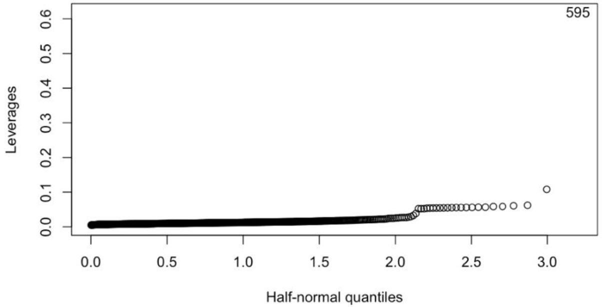
D)



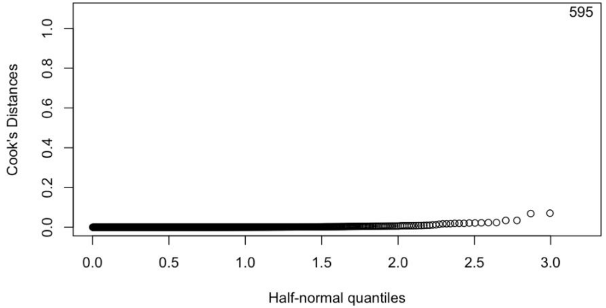

E)




F)




G)




Obs #595 has a high leverage and cook's distance and therefore is an influential outlier and is removed from the initial model as the first step to improve our model.

# Initial Model

## (influential outlier removed)

- After removing the influential outlier Obs #595, Our model slightly improved according to the adjusted $R^2$. **(From 0.7914 to 0.7999)**
- Our next step is to check for Variance Inflation Factors (VIF) and see if any predictor(s) can be dropped to further improve our model.

```
Call:
lm(formula = cnt ~ ., data = bikeDf_new)

Residuals:
    Min      1Q  Median      3Q     Max 
-3234.7  -467.2    61.0   523.3  3149.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1551.82     231.19   6.712 3.89e-11 ***
yr           2060.51      64.51  31.942  < 2e-16 ***
season        426.24      31.52  13.524  < 2e-16 ***
holiday      -621.54     197.36  -3.149  0.00170 ** 
workingday    136.59      71.24   1.917  0.05560 .  
weathersit   -555.18      77.28  -7.184 1.69e-12 ***
temp         2070.38    1385.96   1.494  0.13566    
atemp        3531.90    1569.18   2.251  0.02470 *  
hum         -1103.53     307.81  -3.585  0.00036 ***
windspeed   -2312.29     452.66  -5.108 4.17e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 861.2 on 718 degrees of freedom
Multiple R-squared:  0.8024,	Adjusted R-squared:  0.7999 
F-statistic:   324 on 9 and 718 DF,  p-value: < 2.2e-16
```

**cnt = 1730.19 + 2034.07 (yr) + 409.40 (season) − 621.06 (holiday) + 124.12 (workingday) − 580.37 (weathersit) + 2326.31 (temp) + 3312.12 (atemp) − 1202.79 (hum) − 2582.41 (windspeed)**

# Multicollinearity

- By looking at the VIFs of the initial model, we can see that both “Normalized Temperature” and “Normalized Feeling Temperature” have large VIFs, indicating that multicollinearity exists in our model.
- We’re going to see if removing “Normalized Feeling Temperature”, the predictor with the highest VIF, will reduce our issue with multicollinearity and further improve our model.

**Table 2. Variance Inflation Factors (VIF) for the initial model.** This table shows the VIFs calculated for each predictor in the initial model after removal of leverages/outliers/influential points.

| **Variable** | Year | Season | Holiday | Workday | Weather | Normalized Temperature | Normalized Feeling Temperature | Normalized Humidity | Normalized Windspeed |
|---|---|---|---|---|---|---|---|---|---|
| **VIF** | 1.02 | 1.19 | 1.07 | 1.08 | 1.72 | 63.20 | 64.22 | 1.87 | 1.20 |

# Fixing Multicollinearity

- We calculated the VIFs for each predictor for our new model and each VIF is close to 1, indicating that removing “Normalized Feeling Temperature” did solve our issue with multicollinearity.

**Table 3. Variance Inflation Factors (VIF) of the new model after removing "Normalized Feeling Temperature".** This table shows the VIFs calculated for each predictor in the new model after removing "Normalized Feeling Temperature" from the model.

| **Variable** | Year | Season | Holiday | Workday | Weather | Normalized Temperature | Normalized Humidity | Normalized Windspeed |
|---|---|---|---|---|---|---|---|---|
| **VIF** | 1.02 | 1.19 | 1.07 | 1.08 | 1.71 | 1.20 | 1.86 | 1.17 |

# New Model a.k.a Model 2

- It seems that removing "Normalized Feeling Temperature" slightly decreased the adjusted $R^2$ of our model **(From 0.7999 to 0.7988)** .
- However, "Normalized Temperature" became a significant predictor of the model ($p < 0.01$) while the overall p-value of the model stayed the same ($p < 0.01$).
- Our next step is to use the Akaike Information Criterion (AIC)-based model selection to select the best model.

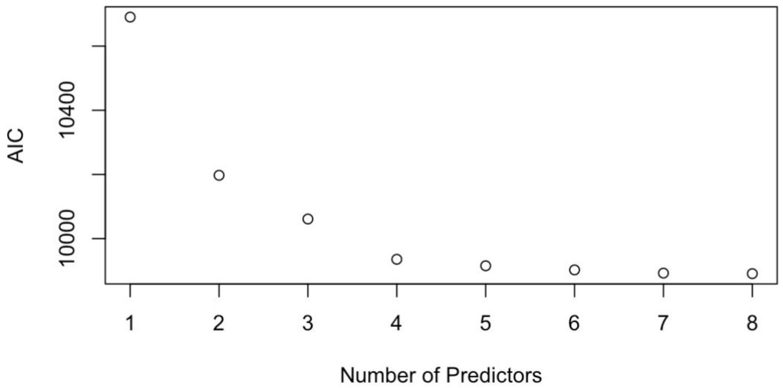
```
Call:
lm(formula = cnt ~ yr + season + holiday + workingday + weathersit +
    temp + hum + windspeed, data = bikeDf_new)

Residuals:
    Min      1Q  Median      3Q     Max
-3169.9  -472.9    54.1   519.3  3190.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1701.99     221.98   7.667 5.72e-14 ***
yr            2060.78      64.69  31.856  < 2e-16 ***
season         429.52      31.57  13.604  < 2e-16 ***
holiday       -637.14     197.79  -3.221 0.001334 **
workingday     135.76      71.44   1.900 0.057806 .
weathersit    -567.47      77.30  -7.341 5.75e-13 ***
temp          5160.16     191.46  26.952  < 2e-16 ***
hum          -1044.52     307.56  -3.396 0.000721 ***
windspeed    -2484.42     447.41  -5.553 3.96e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 863.7 on 719 degrees of freedom
Multiple R-squared:  0.801,	Adjusted R-squared:  0.7988
F-statistic: 361.8 on 8 and 719 DF,  p-value: < 2.2e-16
```

**cnt = 1701.99 + 2060.78 (yr) + 429.52 (season) − 637.14 (holiday) + 135.76 (workingday) − 567.47 (weathersit) + 5160.16 (temp) − 1044.52 (hum) − 2284.42 (windspeed)**

# AIC-Based Model Selection

- We used the AIC-based model selection to see if dropping any other predictor(s) will further improve our model by selecting the model with the lowest AIC value and we decided to keep Model 2 since it has the lowest AIC value.

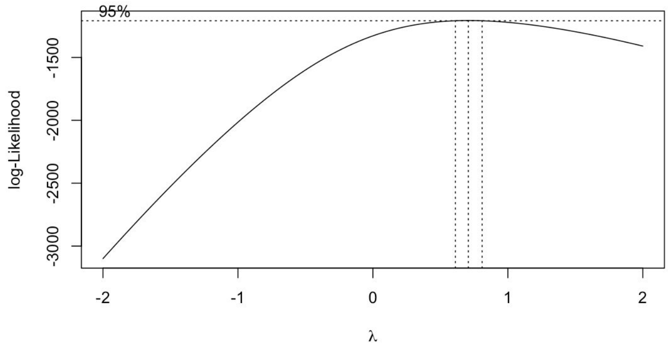
```
  (Intercept)    yr  season holiday workingday weathersit temp   hum windspeed
1        TRUE FALSE  FALSE   FALSE      FALSE      FALSE TRUE FALSE     FALSE
2        TRUE  TRUE  FALSE   FALSE      FALSE      FALSE TRUE FALSE     FALSE
3        TRUE  TRUE   TRUE   FALSE      FALSE      FALSE TRUE FALSE     FALSE
4        TRUE  TRUE   TRUE   FALSE      FALSE       TRUE TRUE FALSE     FALSE
5        TRUE  TRUE   TRUE   FALSE      FALSE       TRUE TRUE FALSE      TRUE
6        TRUE  TRUE   TRUE    TRUE      FALSE       TRUE TRUE FALSE      TRUE
7        TRUE  TRUE   TRUE    TRUE      FALSE       TRUE TRUE  TRUE      TRUE
8        TRUE  TRUE   TRUE    TRUE       TRUE       TRUE TRUE  TRUE      TRUE
[1] 10690.446 10197.308 10061.100  9935.528  9915.260  9902.437  9892.431
[8]  9890.770
```



# More Steps in Determining Final Model

**Interaction Term**: We checked for interaction terms using ANOVA. 12 out of 28 interaction terms are significant ($p < 0.05$) and the addition of these significant interaction terms improved our model (adjusted $R^2$ = **0.8252**). We'll call this new model Model 3.

```
           yr:season          yr:holiday       yr:workingday       yr:weathersit
        3.085459e-03        5.122350e-02        3.055526e-02        3.422734e-02
             yr:temp              yr:hum        yr:windspeed      season:holiday
        2.798929e-05        1.826086e-01        2.627452e-01        6.682483e-01
   season:workingday    season:weathersit         season:temp          season:hum
        4.355133e-01        4.814197e-01        1.663141e-11        2.890197e-02
    season:windspeed  holiday:workingday   holiday:weathersit        holiday:temp
        2.185703e-01                  NA        1.421502e-01        3.553888e-02
         holiday:hum   holiday:windspeed workingday:weathersit    workingday:temp
        3.061417e-01        5.745760e-01        4.600119e-01        2.708902e-03
      workingday:hum workingday:windspeed    weathersit:temp      weathersit:hum
        7.897260e-02        8.647682e-03        6.931291e-02        1.966592e-03
weathersit:windspeed            temp:hum      temp:windspeed       hum:windspeed
        2.559463e-02        6.153892e-01        4.435588e-02        3.389265e-01
```




**Transformation**: We used the Box-Cox method to transform Model 3 to find λ that maximizes the likelihood (λ = 0.7071) and created a new model with λ which we'll call it Transformed Model 3. The transformation improved our model (adjusted $R^2$ = **0.8328**).

# Final Model

## (Transformed Model 3)

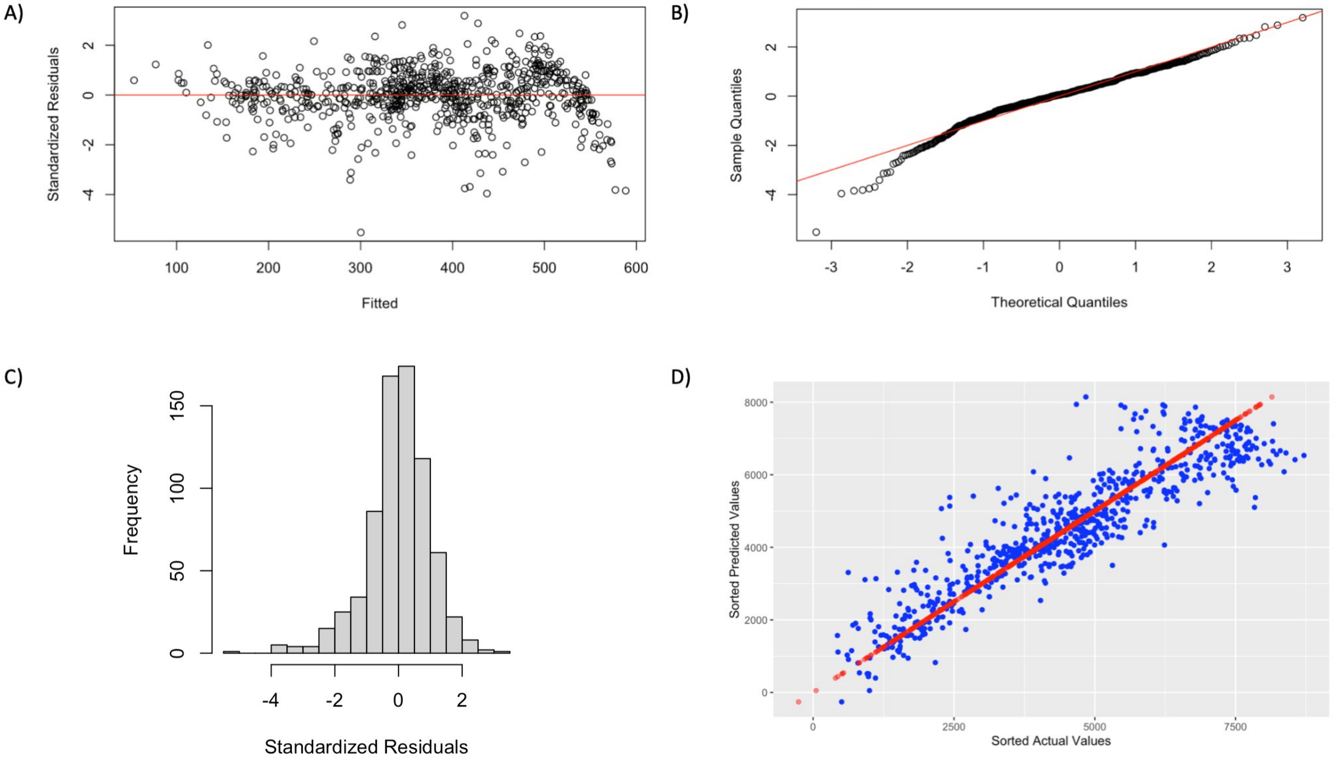
- Our final model for predicting total daily rental bike demands includes the following 8 predictors and 12 interaction terms, with the final adjusted $R^2$ = **0.8328**.
- <u>Note:</u> `trans` is the transformed outcome variable.

**trans = -26.816 + 102.404 (yr) + 71.918 (season) - 70.071 (holiday) + 10.215 (workingday) + 64.941 (weathersit) + 588.851 (temp) + 62.654 (hum) + 33.738 (windspeed) + 2.352 (yr*season) + 13.147 (yr*workingday) - 11.910 (yr*weathersit) + 41.648 (yr*temp) - 106.206 (season*temp) - 4.571 (season*hum) + 61.627 (holiday*temp) - 55.069 (workingday*temp) + 99.090 (workingday*windspeed) - 79.309 (weathersit*hum) - 202.108 (weathersit*windspeed) + 75.596 (temp*windspeed)**

```
Call:
lm(formula = trans ~ yr + season + holiday + workingday + weathersit +
    temp + hum + windspeed + yr * season + yr * workingday +
    yr * weathersit + yr * temp + season * temp + season * hum +
    holiday * temp + workingday * temp + workingday * windspeed +
    weathersit * hum + weathersit * windspeed + temp * windspeed,
    data = bikeDf_new)

Residuals:
     Min       1Q   Median       3Q      Max
-205.608  -23.250    2.409   28.506  153.965

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           -26.816     43.100  -0.622 0.534020
yr                    102.404     16.721   6.124 1.51e-09 ***
season                 71.918      9.106   7.898 1.08e-14 ***
holiday               -70.071     29.168  -2.402 0.016549 *
workingday             10.215     16.756   0.610 0.542322
weathersit             64.941     19.974   3.251 0.001204 **
temp                  588.851     50.934  11.561  < 2e-16 ***
hum                    62.654     47.376   1.322 0.186441
windspeed              33.738    103.229   0.327 0.743895
yr:season               2.352      3.546   0.663 0.507272
yr:workingday          13.147      7.930   1.658 0.097784 .
yr:weathersit         -11.910      7.057  -1.688 0.091931 .
yr:temp                41.648     21.594   1.929 0.054175 .
season:temp          -106.206     12.961  -8.194 1.18e-15 ***
season:hum             -4.571     12.416  -0.368 0.712867
holiday:temp           61.627     57.622   1.070 0.285206
workingday:temp       -55.069     22.398  -2.459 0.014183 *
workingday:windspeed   99.090     51.798   1.913 0.056152 .
weathersit:hum        -79.309     23.099  -3.433 0.000631 ***
weathersit:windspeed -202.108     43.214  -4.677 3.49e-06 ***
temp:windspeed         75.596    151.753   0.498 0.618533
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.15 on 707 degrees of freedom
Multiple R-squared:  0.8374,     Adjusted R-squared:  0.8328
F-statistic: 182.1 on 20 and 707 DF,  p-value: < 2.2e-16
```

# Final Model Plots

**Figure 6.** Plots for the final model. A) Residual vs. Fitted Plot. B) Quantile-Quantile Plot. C) Histogram. D) Actual vs. Predicted Values.



# Future Directions

Analysis/Study Design Improvements: Using machine learning techniques to conduct analyses on the Bike Sharing Dataset can potentially improve accuracy and robustness of our predictions.

Examples: Stepwise Selection Methods, KNN, Random Forests, Time Series Analysis

Advantages: Reduce overfitting/underfitting, increase flexibility

Thanks for watching!!

Any questions?

# References


- Fanaee-T, Hadi. (2013). Bike Sharing. UCI Machine Learning Repository. https://doi.org/10.24432/C5W894.