STAT 685: Dr. Suojin Wang's Group

Modeling Seoul Bike Sharing Demand

Nam Tran, Bai Zou

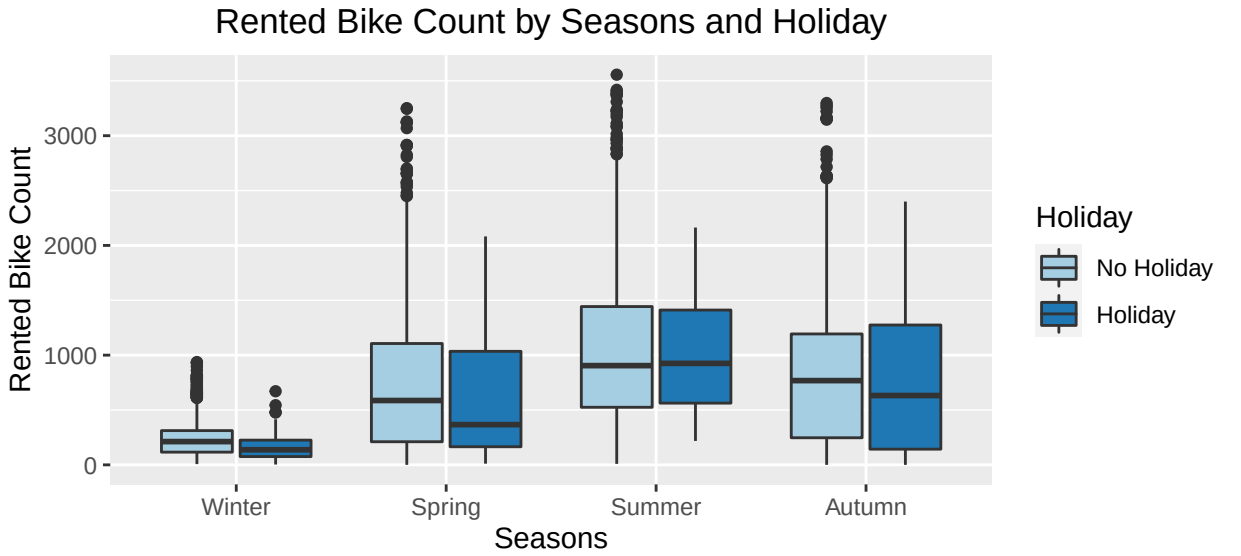
9/2/2020

Data Exploration

Qualtitative Variables

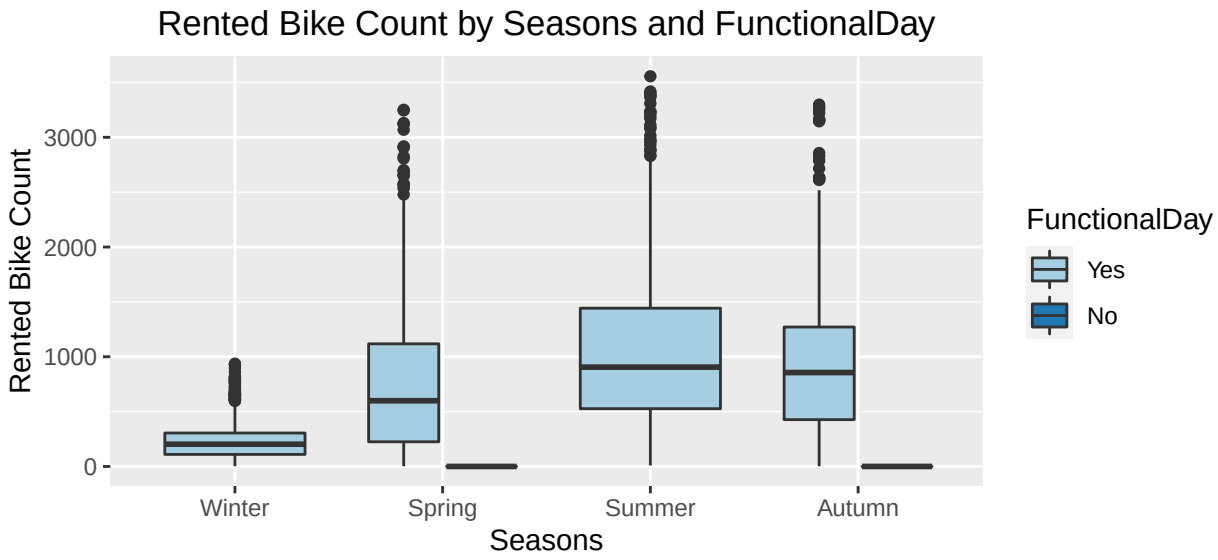
- The plots shows more rented bike count in non-holidays than holidays except for summer.
- If functional day is "no", there's no any bike rented.
- Day of week is not making significant difference in rented bike count.

```
# adding day of time attributes
dat$DayOfWeek <- weekdays(dat$Date)

# plotting by Seasons and Holiday
dat %>%
  ggplot(aes(x=Seasons, y=RentedBikeCount, fill=Holiday)) +
  geom_boxplot() +
  scale_fill_brewer(palette="Paired") +
  labs(y="Rented Bike Count", x="Seasons", title="Rented Bike Count by Seasons and Holiday")
```

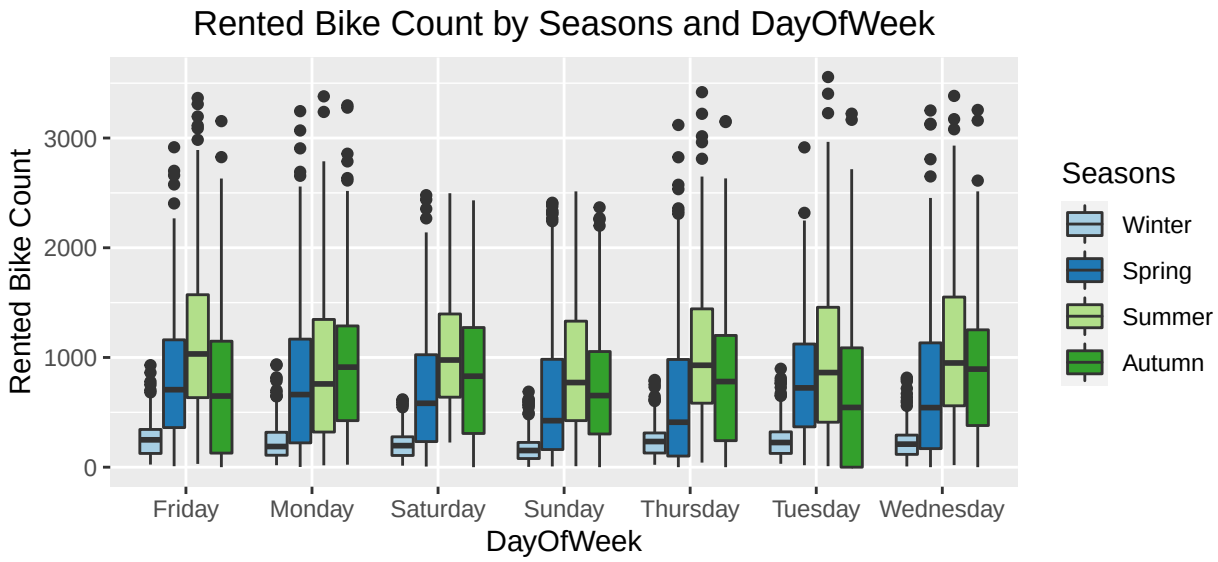



```
# plotting by Seasons and FunctionalDay
dat %>%
  ggplot(aes(x=Seasons, y=RentedBikeCount, fill=FunctionalDay)) +
  geom_boxplot() +
```

```
  scale_fill_brewer(palette="Paired") +
  labs(y="Rented Bike Count", x="Seasons", title="Rented Bike Count by Seasons and FunctionalDay")
```




```
# plotting by Seasons and DayOfWeek
dat %>%
  ggplot(aes(x=DayOfWeek, y=RentedBikeCount, fill=Seasons)) +
  geom_boxplot() +
  scale_fill_brewer(palette="Paired") +
  labs(y="Rented Bike Count", x="DayOfWeek", title="Rented Bike Count by Seasons and DayOfWeek")
```

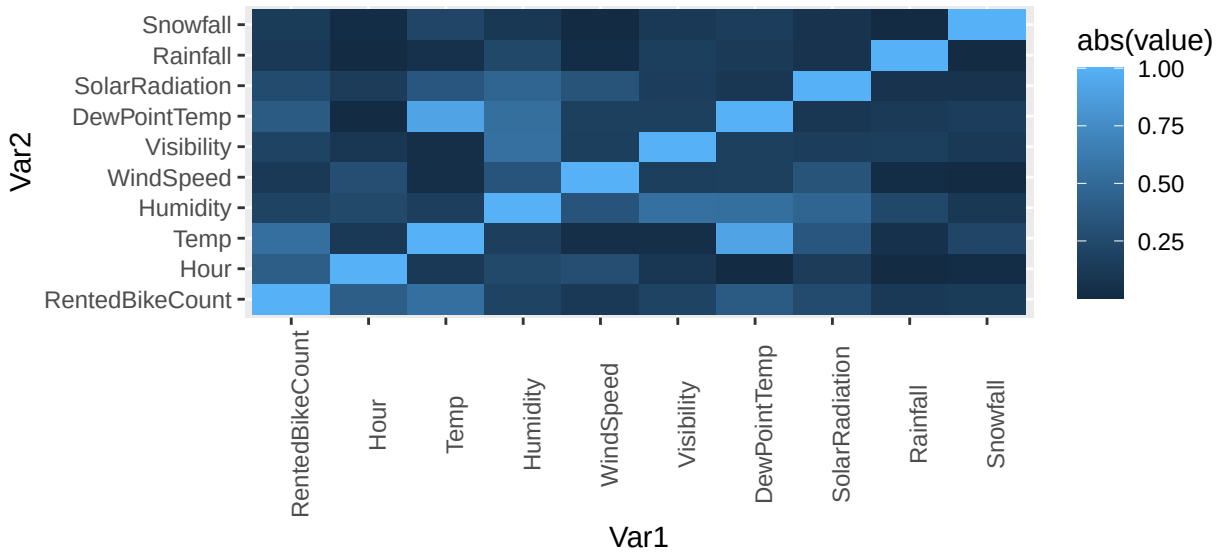


Quantitative Variables

- The covariance matrix shows Temp, Hour has relatively higher correlation with RentedBikeCount (>0.4).
- DewPointTemp and SolarRadiation have correlation greater than 0.2.
- Temp and DewPointTemp are highly correlated (0.9).

- **No clear linear relationship can be identified between response variable and quantitative Variables**

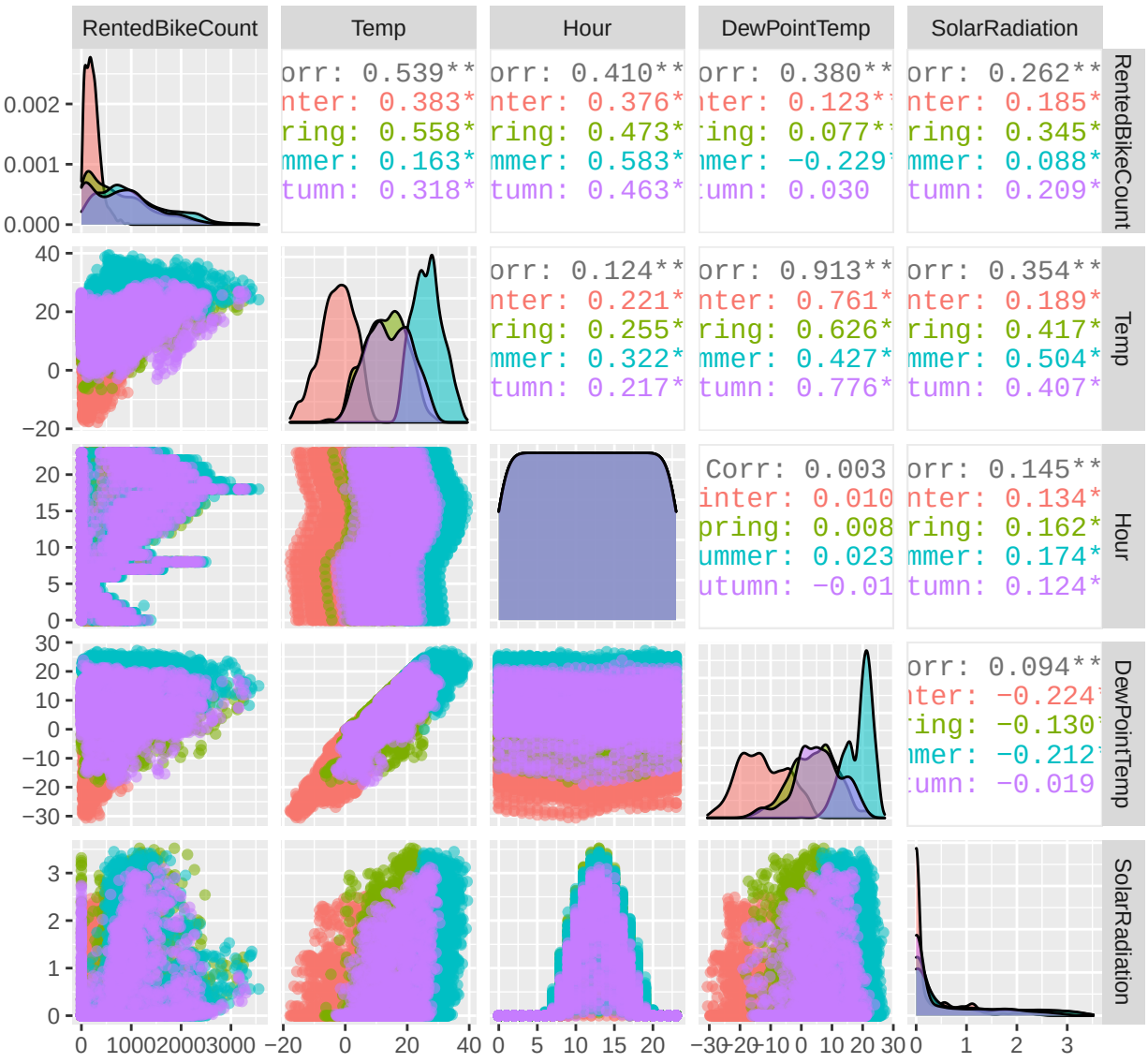
```
quantitative_var = c("Hour", "Temp", "Humidity", "WindSpeed", "Visibility", "DewPointTemp",
             "SolarRadiation", "Rainfall", "Snowfall")

# check covariance
cor_matrix = cor(dat[c("RentedBikeCount", quantitative_var)])
cor_matrix2 = melt(cor_matrix)

cor_matrix2 %>%
  ggplot(aes(x=Var1, y=Var2, fill=abs(value))) +
    geom_tile() +
    theme(axis.text.x = element_text(angle = 90))
```




```
cor_matrix = cor_matrix[order(abs(cor_matrix[,1]), decreasing=TRUE),]
data.frame(cor_matrix[2:10, 1])
```

```
##                 cor_matrix.2.10..1.
## Temp                      0.5385582
## Hour                      0.4102573
## DewPointTemp              0.3797881
## SolarRadiation            0.2618370
## Humidity                 -0.1997802
## Visibility                0.1992803
## Snowfall                 -0.1418036
## Rainfall                 -0.1230740
## WindSpeed                 0.1211084
```

```
# scatter plot matrix
select_var = rownames(cor_matrix)[1:5]
p = ggpairs(dat[select_var],
            aes(colour = dat$Seasons, alpha = 0.2))
show(p)
```



Simple Linear Regression Fit

- Hour is highly related to RentedBikeCount, but not linearly related. Using Hour as qualtitative variable improves 10% in R-squared.
- Adding second order for Temp doesn't bring significant improvement.
- Even with some modification on variables, the simple linear fit is still not very good with R-squared below 70%. The residual plots are showing 'v' shaped pattern that needs further investigation.

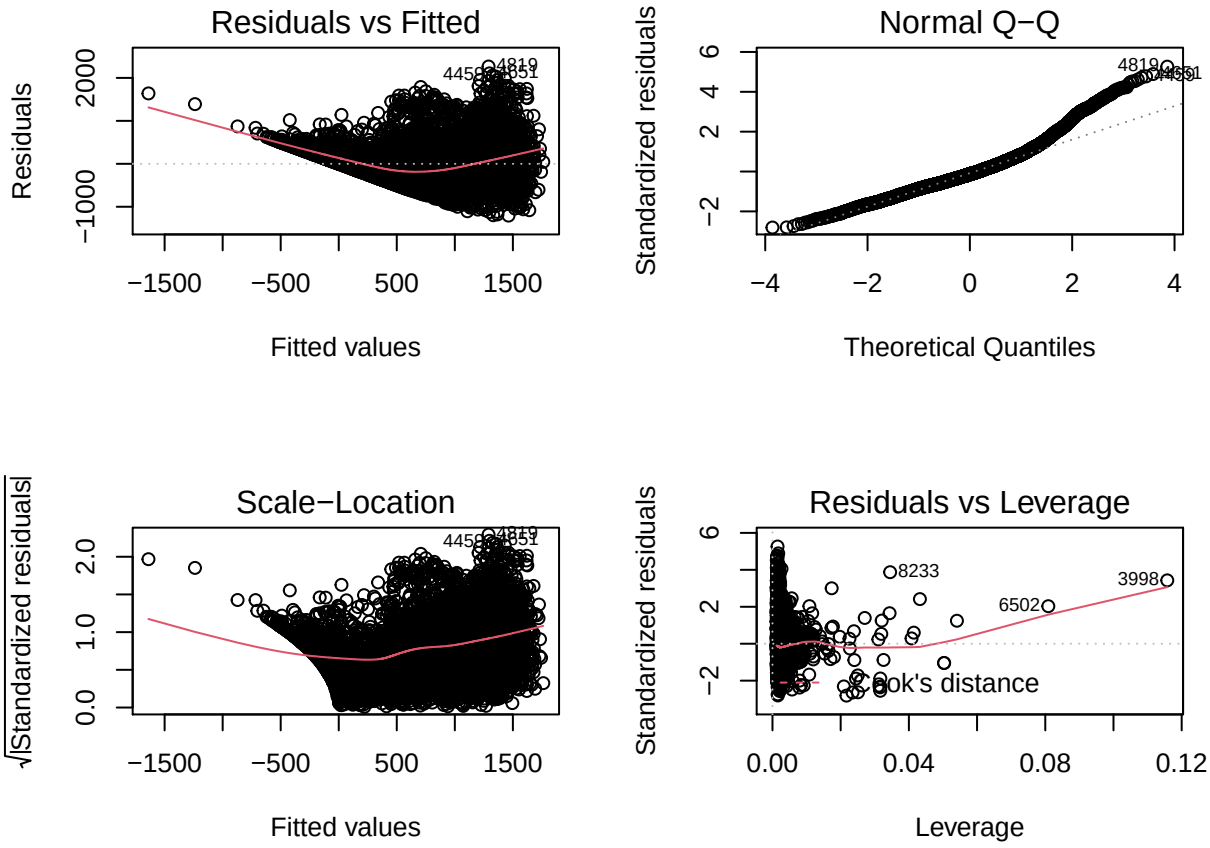
Baseline

```
fit_dat1 = dat[, -1]
# Fit the full model
full.model <- lm(RentedBikeCount ~., data = fit_dat1)
# Stepwise regression model
```

```r
step.model <- stepAIC(full.model, direction = "both",
                      trace = FALSE)
summary(step.model)
```

```
##
## Call:
## lm(formula = RentedBikeCount ~ Hour + Temp + Humidity + WindSpeed +
##     DewPointTemp + SolarRadiation + Rainfall + Snowfall + Seasons +
##     Holiday + FunctionalDay + DayOfWeek, data = fit_dat1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1205.75  -277.81   -56.72   210.17  2265.28
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         655.0045    88.4991   7.401 1.47e-13 ***
## Hour                 27.4944     0.7298  37.674  < 2e-16 ***
## Temp                 17.3718     3.6534   4.755 2.02e-06 ***
## Humidity            -10.7614     0.9957 -10.807  < 2e-16 ***
## WindSpeed            17.5471     5.0603   3.468 0.000528 ***
## DewPointTemp         10.0314     3.8268   2.621 0.008774 **
## SolarRadiation      -79.8386     7.4092 -10.776  < 2e-16 ***
## Rainfall            -58.7742     4.2505 -13.828  < 2e-16 ***
## Snowfall             31.1161    11.1685   2.786 0.005347 **
## SeasonsSpring       227.2609    18.5238  12.269  < 2e-16 ***
## SeasonsSummer       208.5637    27.6990   7.530 5.59e-14 ***
## SeasonsAutumn       366.4037    19.3114  18.973  < 2e-16 ***
## HolidayHoliday     -116.8251    21.5670  -5.417 6.23e-08 ***
## FunctionalDayNo    -945.7506    26.7366 -35.373  < 2e-16 ***
## DayOfWeekMonday     -55.3183    17.2383  -3.209 0.001337 **
## DayOfWeekSaturday   -64.8346    17.1965  -3.770 0.000164 ***
## DayOfWeekSunday    -138.4660    17.2069  -8.047 9.58e-16 ***
## DayOfWeekThursday   -33.3889    17.1986  -1.941 0.052246 .
## DayOfWeekTuesday    -29.6620    17.2584  -1.719 0.085705 .
## DayOfWeekWednesday   -5.9749    17.2201  -0.347 0.728619
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 430.8 on 8740 degrees of freedom
## Multiple R-squared:  0.5549, Adjusted R-squared:  0.5539
## F-statistic: 573.4 on 19 and 8740 DF,  p-value: < 2.2e-16
```

```r
# Residual plots
par(mfrow = c(2, 2))
plot(step.model)
```



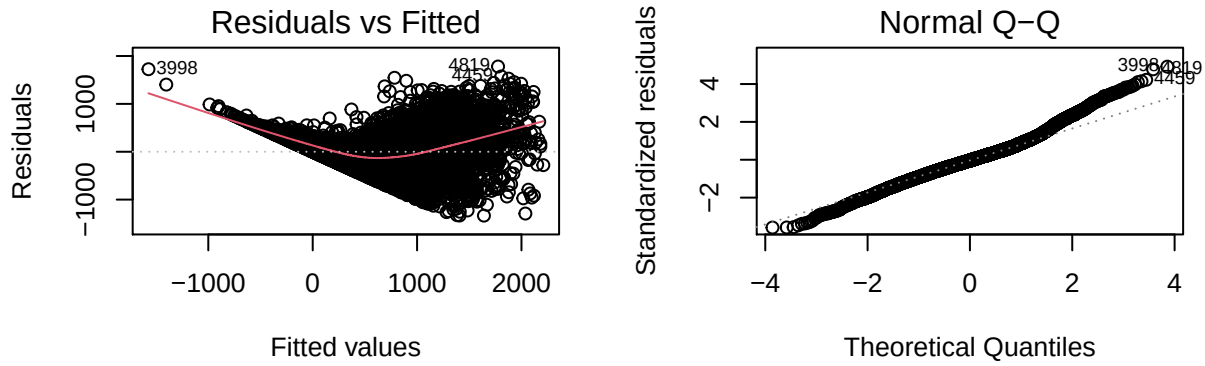
Treat Hour as Qualtitative Variables

```
fit_dat2 = dat[, -1]
fit_dat2$Hour = as.factor(fit_dat2$Hour)
# Fit the full model
full.model2 <- lm(RentedBikeCount ~., data = fit_dat2)
# Stepwise regression model
step.model2 <- stepAIC(full.model2, direction = "both",
                       trace = FALSE)
summary(step.model2)
```

```
##
## Call:
## lm(formula = RentedBikeCount ~ Hour + Temp + Humidity + DewPointTemp +
##     SolarRadiation + Rainfall + Snowfall + Seasons + Holiday +
##     FunctionalDay + DayOfWeek, data = fit_dat2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1334.56  -226.27   -13.77   198.26  1780.47
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          892.3762    79.2600  11.259  < 2e-16 ***
## Hour1               -104.4087    27.6473  -3.776 0.000160 ***
```

```
## Hour2              -211.3843    27.6570  -7.643 2.35e-14 ***
## Hour3              -294.8931    27.6690 -10.658  < 2e-16 ***
## Hour4              -355.1072    27.6831 -12.828  < 2e-16 ***
## Hour5              -342.2907    27.7068 -12.354  < 2e-16 ***
## Hour6              -179.6405    27.7229  -6.480 9.68e-11 ***
## Hour7               120.0883    27.7512   4.327 1.53e-05 ***
## Hour8               477.4402    27.9226  17.099  < 2e-16 ***
## Hour9                14.3974    28.5526   0.504 0.614105
## Hour10             -218.2549    29.6764  -7.355 2.09e-13 ***
## Hour11             -230.8062    30.8620  -7.479 8.24e-14 ***
## Hour12             -193.8176    31.7653  -6.102 1.10e-09 ***
## Hour13             -193.1546    32.0417  -6.028 1.72e-09 ***
## Hour14             -186.8716    31.5442  -5.924 3.26e-09 ***
## Hour15             -101.0473    30.7052  -3.291 0.001003 **
## Hour16               32.9042    29.6119   1.111 0.266520
## Hour17              308.3013    28.6545  10.759  < 2e-16 ***
## Hour18              754.3062    28.1149  26.829  < 2e-16 ***
## Hour19              504.4139    27.8750  18.096  < 2e-16 ***
## Hour20              434.1133    27.7884  15.622  < 2e-16 ***
## Hour21              427.2791    27.7140  15.417  < 2e-16 ***
## Hour22              333.5947    27.6660  12.058  < 2e-16 ***
## Hour23              102.9312    27.6476   3.723 0.000198 ***
## Temp                 10.9872     3.2510   3.380 0.000729 ***
## Humidity             -9.8092     0.8705 -11.268  < 2e-16 ***
## DewPointTemp         13.0888     3.3710   3.883 0.000104 ***
## SolarRadiation       80.6662     9.7970   8.234  < 2e-16 ***
## Rainfall            -58.3332     3.7159 -15.698  < 2e-16 ***
## Snowfall             26.5765     9.6961   2.741 0.006139 **
## SeasonsSpring       202.8230    16.4266  12.347  < 2e-16 ***
## SeasonsSummer       203.5704    24.5589   8.289  < 2e-16 ***
## SeasonsAutumn       364.8688    16.9177  21.567  < 2e-16 ***
## HolidayHoliday     -118.4943    18.7028  -6.336 2.48e-10 ***
## FunctionalDayNo    -948.1886    23.1712 -40.921  < 2e-16 ***
## DayOfWeekMonday     -51.0903    14.9420  -3.419 0.000631 ***
## DayOfWeekSaturday   -67.7648    14.9080  -4.546 5.55e-06 ***
## DayOfWeekSunday    -133.2836    14.9136  -8.937  < 2e-16 ***
## DayOfWeekThursday   -29.1534    14.9052  -1.956 0.050505 .
## DayOfWeekTuesday    -21.8918    14.9610  -1.463 0.143433
## DayOfWeekWednesday   -4.0060    14.9168  -0.269 0.788276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 373.4 on 8719 degrees of freedom
## Multiple R-squared:  0.6664, Adjusted R-squared:  0.6648
## F-statistic: 435.4 on 40 and 8719 DF,  p-value: < 2.2e-16
```

```
# Residual Plots
par(mfrow = c(2, 2))  # Split the plotting panel into a 2 x 2 grid
plot(step.model2)  # Plot the model information
```



Adding 2nd Order of Temp

```
fit_dat3 = dat[, -1]
fit_dat3$Hour = as.factor(fit_dat3$Hour)

# mean temp
fit_dat3$Temp2 = fit_dat3$Temp ^ 2

# Fit the full model
full.model3 <- lm(RentedBikeCount ~., data = fit_dat3)
# Stepwise regression model
step.model3 <- stepAIC(full.model3, direction = "both",
                      trace = FALSE)
summary(step.model3)
```

```
##
## Call:
## lm(formula = RentedBikeCount ~ Hour + Temp + Humidity + Visibility +
##     DewPointTemp + SolarRadiation + Rainfall + Snowfall + Seasons +
##     Holiday + FunctionalDay + DayOfWeek + Temp2, data = fit_dat3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1338.24  -228.46   -19.25   200.92  1756.82
##
```

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.332e+02  8.355e+01  11.169  < 2e-16 ***
## Hour1               -1.042e+02  2.758e+01  -3.779 0.000158 ***
## Hour2               -2.106e+02  2.759e+01  -7.633 2.54e-14 ***
## Hour3               -2.938e+02  2.760e+01 -10.646  < 2e-16 ***
## Hour4               -3.537e+02  2.761e+01 -12.810  < 2e-16 ***
## Hour5               -3.407e+02  2.764e+01 -12.328  < 2e-16 ***
## Hour6               -1.775e+02  2.766e+01  -6.417 1.47e-10 ***
## Hour7                1.222e+02  2.769e+01   4.413 1.03e-05 ***
## Hour8                4.777e+02  2.786e+01  17.148  < 2e-16 ***
## Hour9                9.916e+00  2.852e+01   0.348 0.728103
## Hour10              -2.295e+02  2.970e+01  -7.727 1.22e-14 ***
## Hour11              -2.466e+02  3.094e+01  -7.971 1.77e-15 ***
## Hour12              -2.115e+02  3.186e+01  -6.640 3.33e-11 ***
## Hour13              -2.097e+02  3.213e+01  -6.527 7.10e-11 ***
## Hour14              -2.004e+02  3.161e+01  -6.340 2.41e-10 ***
## Hour15              -1.111e+02  3.074e+01  -3.615 0.000303 ***
## Hour16               2.741e+01  2.962e+01   0.925 0.354829
## Hour17               3.072e+02  2.863e+01  10.730  < 2e-16 ***
## Hour18               7.564e+02  2.807e+01  26.946  < 2e-16 ***
## Hour19               5.066e+02  2.782e+01  18.213  < 2e-16 ***
## Hour20               4.352e+02  2.772e+01  15.698  < 2e-16 ***
## Hour21               4.275e+02  2.765e+01  15.464  < 2e-16 ***
## Hour22               3.335e+02  2.760e+01  12.083  < 2e-16 ***
## Hour23               1.030e+02  2.758e+01   3.735 0.000189 ***
## Temp                 1.468e+01  3.290e+00   4.461 8.25e-06 ***
## Humidity            -1.028e+01  9.005e-01 -11.421  < 2e-16 ***
## Visibility           1.789e-02  8.759e-03   2.043 0.041131 *
## DewPointTemp         1.522e+01  3.383e+00   4.501 6.86e-06 ***
## SolarRadiation       9.503e+01  1.001e+01   9.498  < 2e-16 ***
## Rainfall            -5.779e+01  3.708e+00 -15.584  < 2e-16 ***
## Snowfall             3.276e+01  9.716e+00   3.372 0.000751 ***
## SeasonsSpring        1.576e+02  1.770e+01   8.905  < 2e-16 ***
## SeasonsSummer        2.026e+02  2.474e+01   8.192 2.94e-16 ***
## SeasonsAutumn        3.189e+02  1.833e+01  17.400  < 2e-16 ***
## HolidayHoliday      -1.277e+02  1.871e+01  -6.825 9.37e-12 ***
## FunctionalDayNo     -9.533e+02  2.313e+01 -41.220  < 2e-16 ***
## DayOfWeekMonday     -5.232e+01  1.491e+01  -3.510 0.000451 ***
## DayOfWeekSaturday   -6.923e+01  1.487e+01  -4.655 3.29e-06 ***
## DayOfWeekSunday     -1.351e+02  1.488e+01  -9.083  < 2e-16 ***
## DayOfWeekThursday   -2.568e+01  1.489e+01  -1.725 0.084563 .
## DayOfWeekTuesday    -2.042e+01  1.494e+01  -1.367 0.171687
## DayOfWeekWednesday  -2.106e+00  1.491e+01  -0.141 0.887698
## Temp2               -2.496e-01  3.708e-02  -6.733 1.77e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 372.5 on 8717 degrees of freedom
## Multiple R-squared:  0.6681, Adjusted R-squared:  0.6665
## F-statistic: 417.8 on 42 and 8717 DF,  p-value: < 2.2e-16
```

```
# Residual Plots
par(mfrow = c(2, 2))  # Split the plotting panel into a 2 x 2 grid
```

```r
plot(step.model3)  # Plot the model information
```