# **Introduction**

Capital Bike Share operates a bike-sharing program in Washington D.C., where customers rent bikes from automated kiosks located throughout the city. Having the ability to accurately predict demand for bike rentals is crucial for ensuring that bikes are available when and where they are needed, thus enhancing customer satisfaction and operational efficiency. As Capital Bike Share’s consultant, our goal is to develop a predictive model that identifies the factors that affect demand and forecast total bike rentals.

## Data Set

The analysis uses the Bike Sharing Systems data set, which includes hourly bike rentals for the years 2011 and 2012, and covers the first 19 days of each month. Variables included in the analysis are noted below.

**Dependent Variable:**

* **Total Count:** The total number of bike rentals (i.e., the sum of registered and casual users).

**Independent Variables:**

* **Seasonal Factors:** Season, Year, Month, Hour, Weekday, Holiday, Working Day
* **Weather Conditions:** Weather Category, Temperature, ‘Feels Like’ Temperature, Humidity, Windspeed

By leveraging these variables, we aim to build a predictive model that will enable Capital Bike Share to meet customer demand for bike rentals.

# **Exploratory Data Analysis**

## Total Count (Total Bike Rentals)

Data is skewed right. In order to find a linear model to the data, a transformation will likely be needed to make a normal distribution.

## Correlation Analysis

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **Transformed Total Count** | | | |
| **Variable** | **Unadjusted** | **log()** | **Sq Rt** | **Cube Rt** | **Reciprocal** |
| Temperature | 0.40 | 0.39 | 0.42 | 0.42 | -0.23 |
| ‘Feels Like' Temp | 0.40 | 0.38 | 0.42 | 0.41 | -0.23 |
| Windspeed | 0.09 | 0.11 | 0.11 | 0.11 | -0.05 |
| Humidity | -0.32 | -0.34 | -0.35 | -0.35 | 0.15 |

Based on correlation coefficients, **Temperature** and **‘Feels Like’ Temperature** show evidence of a moderately positive relationship to Total Rentals – meaning that an increase to those variables suggest an increase to total bike rentals. Windspeed does show a positive correlation, but the low value indicates that the relationship is weak. **Humidity** has a moderately negative correlation – suggesting that an increase to humidity results in a decrease in rentals.

As previously mentioned, the positively skewed distribution of Total Rentals suggests that the variable may need to be transformed to improve the linear model. Correlation coefficients have also been calculated for transformed versions of Total Rentals. Except for the log and reciprocal transformation, the correlation coefficients saw some modest improvements.

## Multicollinearity

Multicollinearity occurs when two or more independent variables are highly correlated. This means they provide redundant information about the outcome and affects the parameter estimates and predictions in the model.

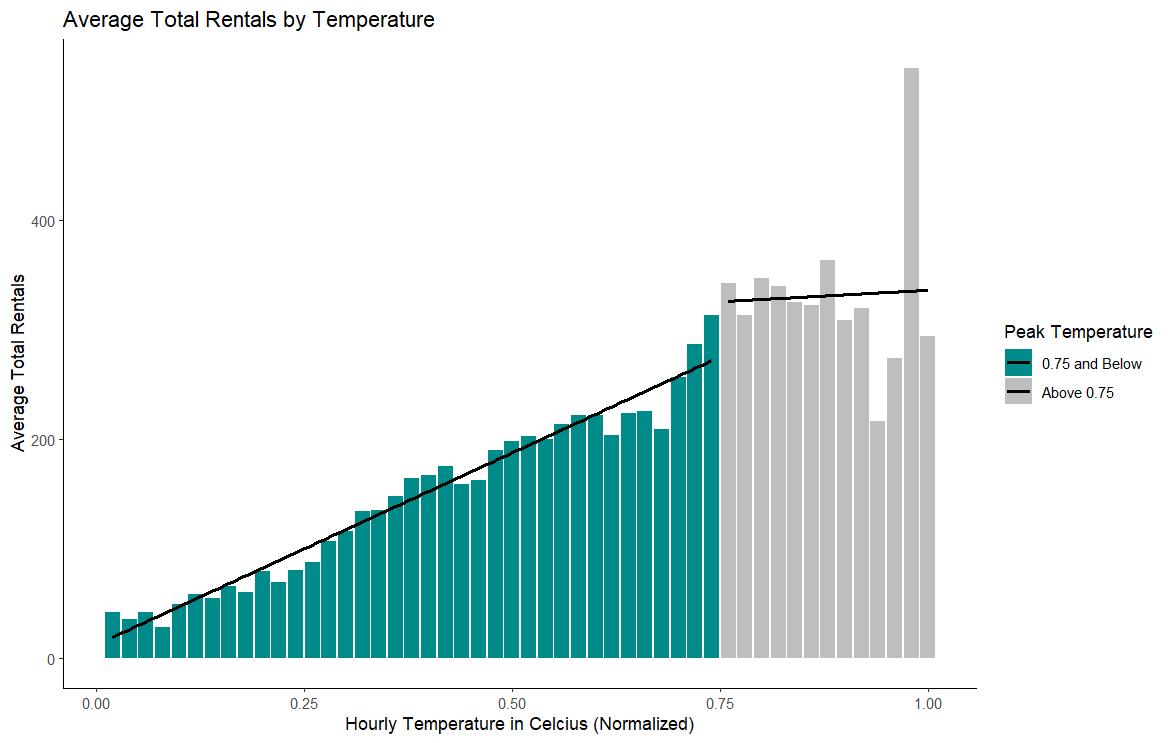
Correlation coefficients between independent variables were measured, and the following variables were identified as showing evidence of multicollinearity were Temperature and ‘Feels Like’ Temperature (r = 0.99).

To address multicollinearity, one of the variables will be excluded from the model. In these cases, the regression model will include Temperature the model and exclude ‘Feels Like’ Temperature since they are redundant to predicting Total Rentals.

## Additional Observations

### Temperature

Temperature has a linear relation with Average Total Rentals up to normalized value of 0.75. Points after start to level out as seen in the visual below.



To account for the change in the relationship, a dummy variable was added to the data. In this case, the variable evaluates if the normalized temperature is greater than 0.75 (True, False)

# **Regression Model**

The final model selected had an adjusted R-Squared of 0.868 on training data – meaning that 86.6% of the variance in the dependent variable is explained by the independent variable. The model versus a testing set of data resulted in a reduced adjusted R-Squared of 0.678. Variables included in the model and any transformations are noted below.

**Dependent Variable:** Cube Root Total Count

As described in the exploratory analysis section, the distribution for Total Count is not normally distributed. A review of the distributions for transformed Total Count showed that the Cube Root transformation was the best candidate as the dependent variable.

**Independent Variables:**

|  |  |  |
| --- | --- | --- |
| Variable | Transformations | Notes |
| Season | Converted to factor. | While presented as numerical in the dataset. The nature of the variable is categorical and was converted to ensure proper treatment in the model. |
| Year | Converted to factor. | See note above. |
| Hour | Converted to factor. | See note above. |
| Temperature | Added Peak Temperature variable | Linear up to 0.75 normalized value. |
| Peak Temperature | Conditional if normalized temperature is greater than 0.75. | Accounts for rentals not increasing at the same rate one above 0.75 |
| Cube Root Total Count Lag | Lagged Dependent Variable | Used to address autocorrelation present with time series data. |

Any variables not included in the model were not statistically significant or showed evidence of multicollinearity. The only exception being **Holiday**, which was manually removed due to the signs of leverage on the model that could lead to misleading results.

# **Interpretation of Results**

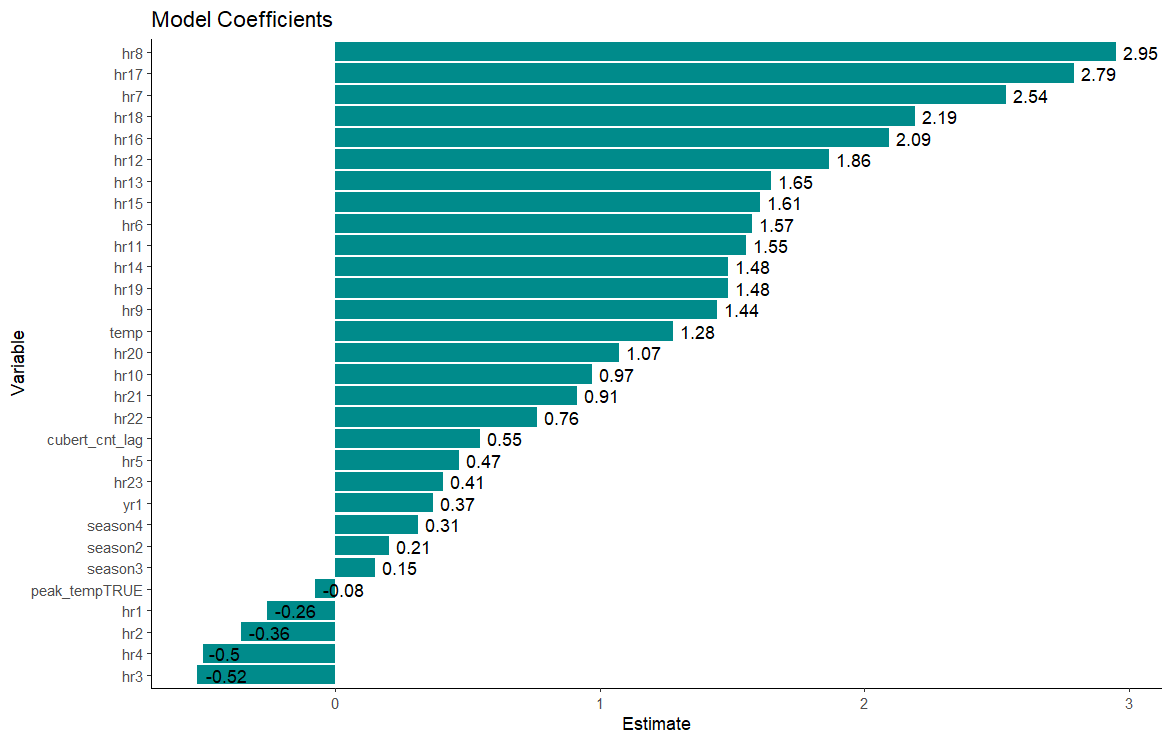
## Independent Variables of focus

The implications of the model include:

* **Time of Day**: Have the largest positive marginal effect on cube root rentals.
* **Temperature**: Warmer temperatures have a positive effect on cube root rentals.

## Marginal Effects

The visual below shows the marginal effects of changes to cube root total rentals. Of particular interest, the model shows that rush hour times (7AM – 8AM and 4PM – 6PM) show the most increase in rentals.



# **Limitations and Assumptions**

## Model Performance: Training vs Test

The model shows evidence of heteroscedasticity and autocorrelation, thus likely contributing differing model performance in the training and test sets.

* **Heteroscedasticity**: Refers to the non-constant variance of residuals in a model. A Breuch-Pagan Test was used and showed evidence. Transformations of independent variables such as temperature were explored but did not result in improvement.
* **Autocorrelation**: Occurs when the residuals from a regression model are correlated with each other and are common with time-series data. A Durbin-Waton test was used to identify autocorrelation. A lag variable was added to address the correlation. While there was improvement in the DW test, there is still evidence of autocorrelation in the model.

In addition, Q-Q diagnostic plot indicates that the data is not normally distributed with the deviations shown from the reference line, despite the transformations made.

## Limitations

The points made above mean that there are limitations to the model. The evidence of Heteroscedasticity and a non-normal distribution are violations of the assumptions of linear regression. In addition, there are likely omitted variables that could affect the accuracy of the model. For example, price and kiosk location are independent variables not in the data that could be predictors of demand.

# **Recommendations and Conclusions**

## Model Refinement

Given the limitations of the model, further transformations should be explored to address heteroscedasticity and non-normal distributions. In addition, alternative machine learning models should be explored.

## Ethical Considerations

Any machine learning model should consider the following:

* **Potential Misuse** – Can the model be misused?
* **Potential Persecution** – Can this model be used to persecute people?
* **Potential Inherent Bias** – Is the data inherently biased to a certain group of people?
* **Security and Credibility** –Data is required for machine learning. Is the data used collected ethically?
* **Data Sensitivity, Ownership, and Consent** – Does the data contain any personal, sensitive, or financial information?

In the context of the model developed, it is important to ensure that the data collected complies with privacy laws. Any personal information collected should be anonymized where possible to protect the customer’s identity. Also, underrepresented groups should be considered in the model to address potential inherent bias.

## Actionable Insights

Hours and Temperature should be used in operational planning such as adjusting bike availably during peak times and warmer periods. In addition, marketing promotions could be targeted for warmer periods throughout the year.

# **Appendix**

## Summary Statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Min | Q1 | Median | Mean | Q3 | Max | Sd |
| Temperature | 0.02 | 0.34 | 0.5 | 0.496987 | 0.66 | 1 | 0.192556 |
| Temperature 'Feels Like' | 0 | 0.3333 | 0.4848 | 0.475775 | 0.6212 | 1 | 0.17185 |
| Humidity | 0 | 0.48 | 0.63 | 0.627229 | 0.78 | 1 | 0.19293 |
| Windspeed | 0 | 0.1045 | 0.194 | 0.190098 | 0.2537 | 0.8507 | 0.12234 |
| Total Count | 1 | 40 | 142 | 189.4631 | 281 | 977 | 181.3876 |

## Model Summary

**Residual Standard Error:** 0.0.7379 **Adjusted R-Squared:** 0.868

**Breusch-Pagan Test:** BP = 1955.9, df = 30, p-value < 2.2e-16

**Durbin-Watson Test:** DW = 1.4179, p-value < 2.2e-16

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| term | estimate | std.error | | | statistic | p.value |
| (Intercept) | 0.12689904 | 0.042674247 | | | 2.973668 | 2.95E-03 |
| season2 | 0.20554811 | 0.024511038 | | | 8.385941 | 5.59E-17 |
| season3 | 0.15027876 | 0.031207353 | | | 4.815492 | 1.49E-06 |
| season4 | 0.31317252 | 0.021199758 | | | 14.772457 | 5.87E-49 |
| yr1 | 0.37087101 | 0.014435598 | | | 25.69142 | 8.68E-142 |
| hr1 | -0.25871185 | 0.046285154 | | | -5.589521 | 2.33E-08 |
| hr2 | -0.35683162 | 0.047407644 | | | -7.526879 | 5.56E-14 |
| hr3 | -0.52019293 | 0.048230667 | | | -10.785522 | 5.34E-27 |
| hr4 | -0.49858175 | 0.049342519 | | | -10.104505 | 6.56E-24 |
| hr5 | 0.46810936 | 0.049152621 | | | 9.523589 | 1.99E-21 |
| hr6 | 1.57380123 | 0.04820005 | | | 32.651444 | 3.27E-224 |
| hr7 | 2.53607394 | 0.047028004 | | | 53.926889 | 0.00E+00 |
| hr8 | 2.95198573 | 0.046314569 | | | 63.737735 | 0.00E+00 |
| hr9 | 1.44136097 | 0.047668639 | | | 30.237091 | 1.07E-193 |
| hr10 | 0.97019281 | 0.047317774 | | | 20.503771 | 7.13E-92 |
| hr11 | 1.55186852 | 0.047011313 | | | 33.010533 | 6.43E-229 |
| hr12 | 1.86486212 | 0.047019864 | | | 39.661155 | 1.976263e-323 |
| hr13 | 1.64799496 | 0.04751951 | | | 34.680386 | 2.31E-251 |
| hr14 | 1.48418172 | | 0.047964378 | 30.943416 | | 2.03E-202 |
| hr15 | 1.60684287 | | 0.047926057 | 33.527541 | | 9.05E-236 | |
| hr16 | 2.09196644 | | 0.047759326 | 43.80226 | | 0.00E+00 | |
| hr17 | 2.79297609 | | 0.048281185 | 57.848126 | | 0.00E+00 | |
| hr18 | 2.18929403 | | 0.0488157 | 44.848154 | | 0.00E+00 | |
| hr19 | 1.48300978 | | 0.048977313 | 30.279525 | | 3.23E-194 | |
| hr20 | 1.07198598 | | 0.047956059 | 22.353504 | | 1.70E-108 | |
| hr21 | 0.91265814 | | 0.047329352 | 19.283132 | | 1.23E-81 | |
| hr22 | 0.76242248 | | 0.047075677 | 16.195677 | | 2.21E-58 | |
| hr23 | 0.40661359 | | 0.04639567 | 8.764042 | | 2.13E-18 | |
| temp | 1.27630712 | | 0.066706189 | 19.133264 | | 2.03E-80 | |
| peak\_tempTRUE | -0.07724553 | | 0.028724745 | -2.689163 | | 7.17E-03 | |
| cubert\_cnt\_lag | 0.54599298 | | 0.006409086 | 85.190464 | | 0.00E+00 | |

Generalized Variance Inflation Factor (GVIF)

## Model Diagnostic Plots

