2022

TORONTO BIKE SHARE DEMAND ESTIMATION



**EXECUTIVE SUMMARY**

**Problem Statement**: Bike Share Toronto is the most important mode of transportation in Toronto. For a sustainable future, cities must focus on transportation to reduce carbon emissions to zero. Road traffic accounts for over 80% of all carbon emissions from transportation in the urban environment. Because of this, bike share programs have been advocated as one of the numerous instruments that communities may use to reduce the number of vehicles on the road while also delivering a variety of additional advantages to people's mental and physical health. Since 2011, Toronto has operated a sizable bike share network. However, the city ranks at the bottom among Canadian cities in terms of the total percentage of the population that lives within the boundaries of bike-sharing service zones. Hence there is a need to estimate the bikeshare demand to supply the adequate number of bikes wherever required.

**Approach**: With the help of Toronto Bike Share data, the city can improve the bike share program to assist urban infrastructure. The riders are defined under two categories: **Yearly Members and Casual**. In this research paper, we analyzed the **impact of weather, gas prices, and traffic congestion** in Toronto on Bike Share Toronto from 2017 to 2022 on a weekly basis. We used real-time data for Bike Share Toronto and for the other variables as well.

There are **600 bike rider stations** across various intersections in Toronto, and we have segmented these stations based on high and low demand stations. Additionally, we have added a **high and low number of annual and casual members** to the segmentation. The model helps in understanding the bike demand at the weekly level and the factors affecting the demand for the station segments to maintain a steady supply of the rental bikes.

**Conclusion**:

The methodology established in this study and the resulting framework for expansion offer a data driven approach to guide the expansion of the Bike Share Toronto network. We could establish a clear picture that the bike demand is dependent on the temperature, cloud cover, gas price and traffic volume of which temperature is of the major driving factors followed by gas price. Created a polynomial model and normalized the data using Z-Score to lower the coefficient value and improve our model performance. We achieve **R square** values ranging between **65% - 75%** and RMSE score of **0.5**. To further improvise the model, we performed various regularization techniques of which we could see a significant improvement through ridge regression. After regularization, our R square values increased in the range of **71% - 80%** while the RMSE score dropped to 0.4.

**DATA PREPARATION**

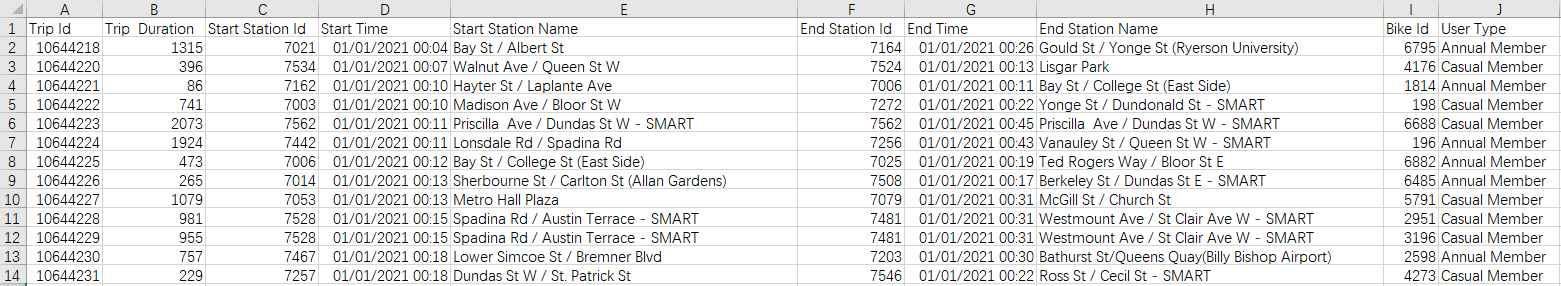
**A. Shared Bike Data**

**Summary:** Shared bike data was collected from Open Data Toronto which was recorded and updated by the operating company - Shared Bike Toronto monthly. The dataset recorded each bike trip’s detailed information including **Start Station & Time, End Station & Time, Trip duration and Type of users** (Annual Users which refers to annual pass holder and Casual Users which refers to single rider or day-pass rider).

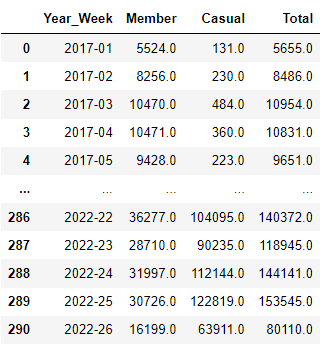
**Approach:** The goal was to build a weekly number of riders table, recording the number of the total number of riders on a weekly basis for further segmentation and modeling.

**Data Cleaning Process:** The data was stored in different datasets with similar but different file names from 2017 to June 2022. For the cleaning process, we followed the steps below:

* Used **for()** loop to load all datasets into Jupyter Notebook as we had more than 5 files to be loaded
* Changed column names in all the tables to make them consistent across tables.
* Concatenate all datasets into one data frame.
* Loaded Station Information dataset to fill the missing Station Id value and missing Station Name value (e.g if Station Name is missing, use Station Id’s corresponding Station Name in Station information table to fill)
* Since the missing data was **less than 5%**, we dropped all rows with missing Station Id or Station Name, Bike ID, None or “0” Trip Duration,
* Used **pd.to\_datetime** and datetime function to make the format of Start Time consistent (time from early six months data was recorded in day-month-year but others were in month-day-year format).
* Changed “Annual User” to “Annual” and “Casual User” to “Casual” in the User Type column to ensure consistency.
* Used **groupby()** function on Station Id, Year\_Week to get the number of Casual riders, Member riders and Total riders of each station during each week.



***Figure: Main Dataframe for ridership data***

******

***Figure: Final output for Weekly Traffic Data***

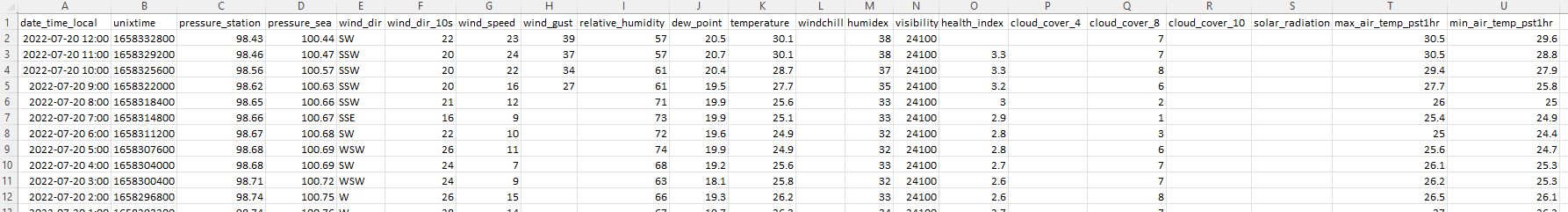
**B. Weather Dataset**

**Summary**: Toronto weather data is collected on hourly basis by Toronto weather stats for a variety of purposes. It records parameters like pressure, wind, humidity, temperature, visibility, and cloud cover.

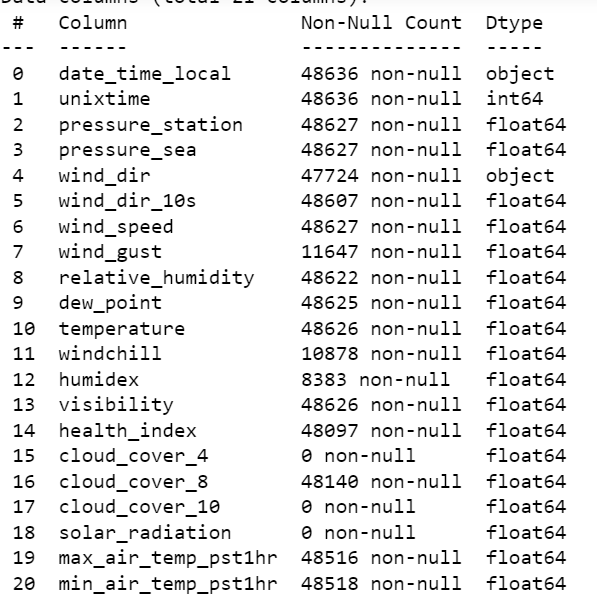
**Approach:** Clean the weather data on the hourly basis so that it can be combined with the ridership data.

**Data Cleaning Process:**

***Figure: Main Dataframe for Weather***

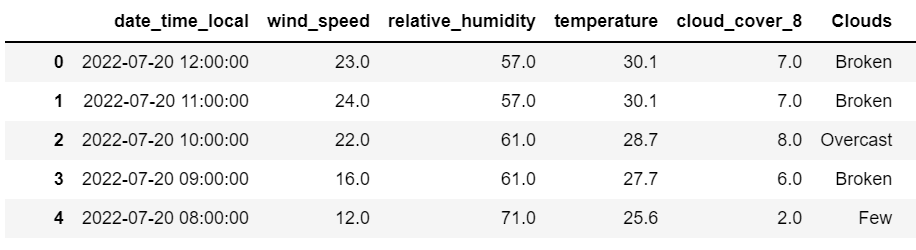
****

1. Identify the relevant columns that would affect the bike ridership and remove the ones where more than 50% of the data is missing (marked in red below)



1. Final columns used - 'date\_time\_local', 'wind\_speed', 'relative\_humidity','temperature', 'cloud\_cover\_8'.
2. Converted 'date\_time\_local' to datetime format using function “to\_datetime”
3. Filled the missing values in 'wind\_speed', 'relative\_humidity', 'temperature', using **Interpolate()** function.
   1. **Why Interpolation** - Since this data is hourly, any change (if any) will take place gradually instead of suddenly. Hence missing data for one particular hour can be filled as average of the previous and the next hour.
4. Missing data for cloud\_cover\_8 is categorical; hence it cannot be interpolated. Used **ffil** method to fill the missing values.
5. Created bins for the cloud cover as below:
   1. **0-1 :**  Clear
   2. **1-3**: Few
   3. **3-5:** Scattered
   4. **5-7:** Broken
   5. **7-8:** Overcast

***Figure: Final output for Hourly Weather Data***

******

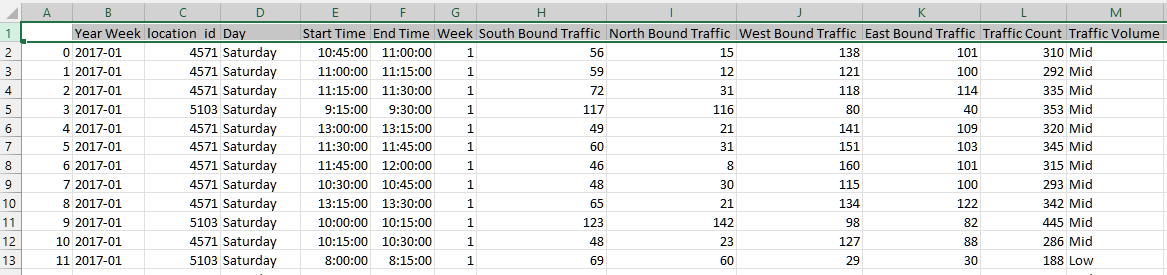
**C. Traffic Data**

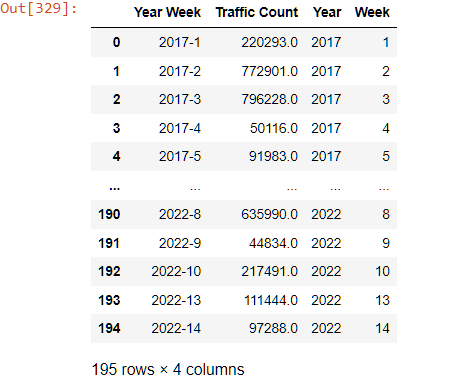
**Summary**: Traffic data is collected in real-time by the City's Transportation Services division for a variety of purposes. The Automatic Traffic Recorder system is used for every car on a certain street and their turning maneuvers. Each hour, 15-minute intervals were used to build the data collection. The data set was divided into sections like "Northbound," "Southbound," "Westbound," and "Eastbound." Additionally, the data set indicated whether or not the vehicle made a right turn or a left turn.

**Approach:** Our ultimate goal in the data cleaning process was to produce three separate data sets showing weekly traffic volumes of high, medium, and low for each of the various sites.

**Data Cleaning Processing**: We had two different data sets. One had the data from the year 2010-2017 and the other from 2018-2022. For the cleaning process, we followed the steps below:

* Concated both of the datasets into one data frame.
* Filtered the dates by “pd.date >= 2017”. Now our data set shows the dates from 2017-2022.
* Extracted the year and week from both the 'Start Time’ and the 'End Time’ using panda's “**to\_datetime()**” function.
* Converted the ‘Date’ column to date format. Then using the “pd.df[].dt.year/week” created a separate column for both the year and week.
* Identified the missing fields and used the moving averages rolling() function to fill them.
* Using group by () on "Year Week" summed up all the vehicles from each intersection into one.
* Dropped the columns that were not required for our final data set.
* Segmented the locations into High Traffic, Medium Traffic and Low Traffic using 25th and 50th percentile limits.

***Figure: Main Data Frame for Traffic Data***



***Figure: Final output for High Traffic Data***

**D. Gas Price Data**

**Summary**: The raw gas price data was sourced from the calibrate retail analytics platform in form of weekly price averages. The initial raw data included prices for three different types of gasoline namely regular, midgrade and premium as well as diesel of which we selected regular grade for our dataset.

**Approach:** The overall goal was to create a table of weekly average prices of gas for the city of Toronto over the period of June 2017 to present and combine it with other datasets.

**Data Cleaning Process:** The dataset was extracted from the Kalibrate platform in the form of individual annual data sets from January 2017 to present. The following steps were involved in the cleaning process to create a desired dataset:

* Concated all the individual annual data sets into a single data frame.
* Converted the ‘Date’ column to date format. Then using the “pd.df[].dt.year/week” created a separate column for both the year and week.
* No missing data found
* Using group by () on "Year Week" and averaged the gas price at weekly level

Table

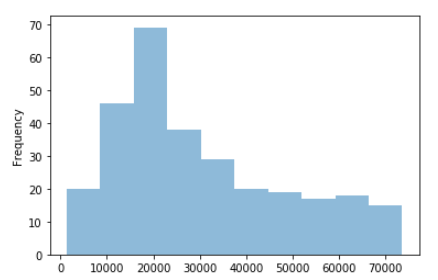
Description automatically generated

***Figure: Main Data Frame for Gas price data*** ***Figure: Final output for Gas price data***

***Now, combine all the datasets to Toronto Ridership data***

**Data Exploration and Segmentation**

1. **Total bike rider stations -** We have 600 different stations across various intersections.
2. **Casual and Member Distribution**: It was found that the distribution of the member riders and casual riders varies (figures on the next page). We can speculate that it may differ depending on the distribution of riders at different bike stations. For example, there may be more member riders than casual riders at some bike stations, while other bike stations may have more casual riders than member riders.

 ***Chart, histogram

Description automatically generated***

***Figure: Distribution of Member Riders Figure: Distribution of Casual Riders***

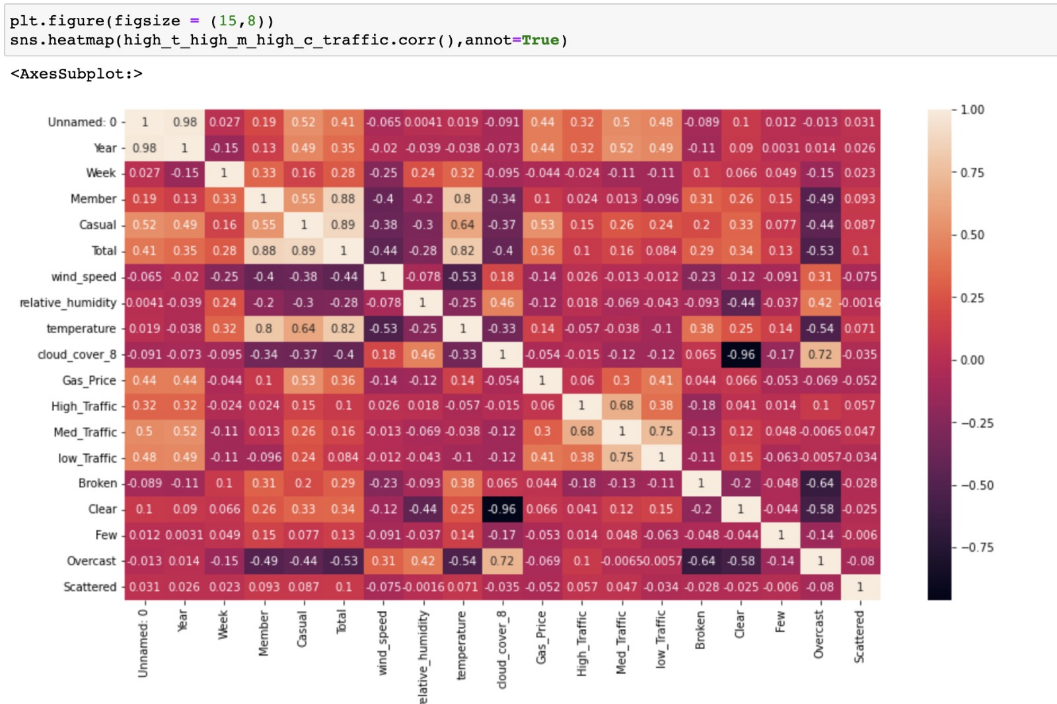
1. **Segmentation**: Since we have 600 stations, it is not possible for us to create 600 models, hence we segment the bike stations. It is done using a combination of Total riders, Casual riders, and Annual Member riders. Used median() to get the 50th percentile of all the three categories and created the following segments in the dataset:

Table

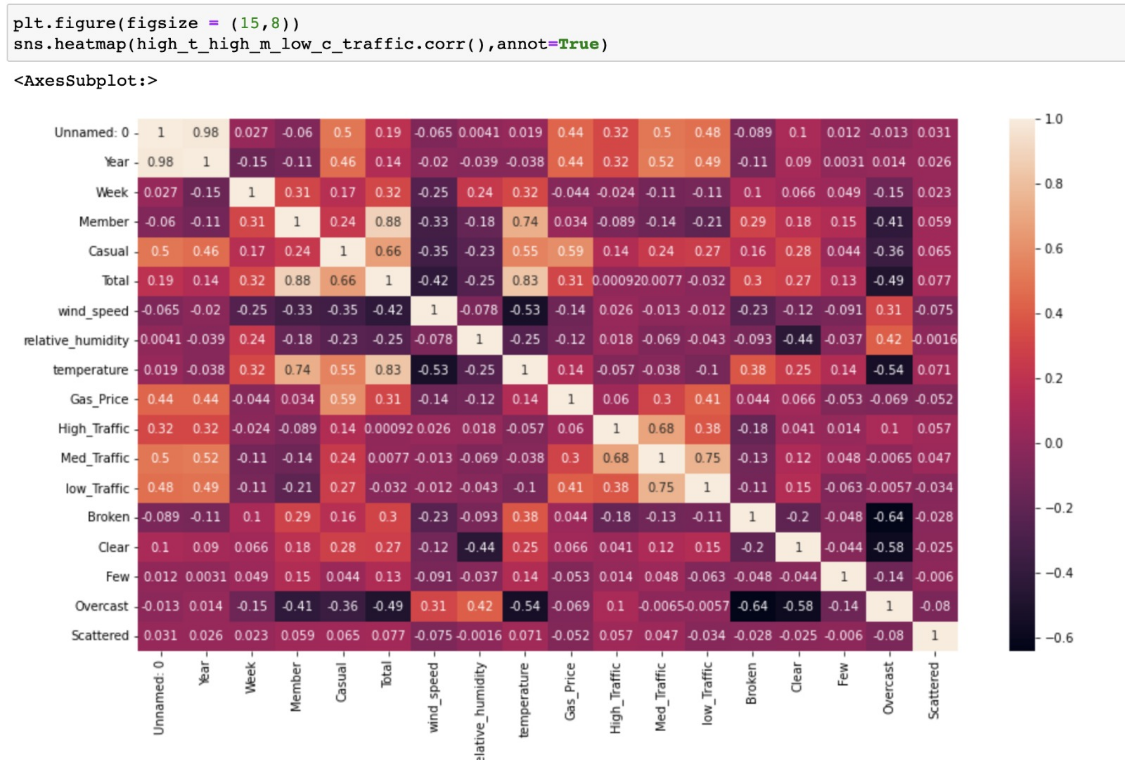
Description automatically generated

1. **Correlation Analysis:** The below correlation matrix shows that temperature has a high correlation with wind speed while cloud cover has a high correlation with relative humidity. For such variables that have high correlation, the one with the lower predictability (R-square) is removed. This helps in maintaining a balance amongst non-collinear variables and not loose on the predictability of the mode.

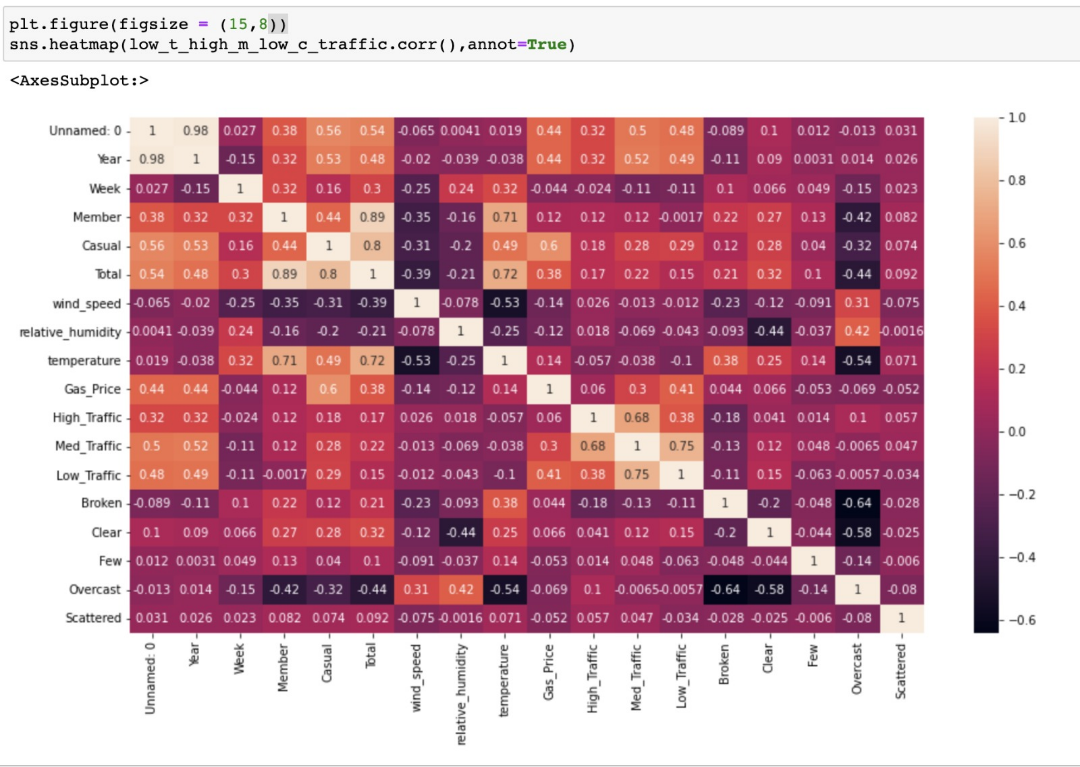
**High Total, High Member and High Casual**

****

**High Total, High Member and Low Casual**

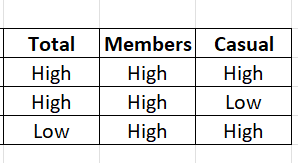
****

**Low Total, High Member and Low Casual**

****

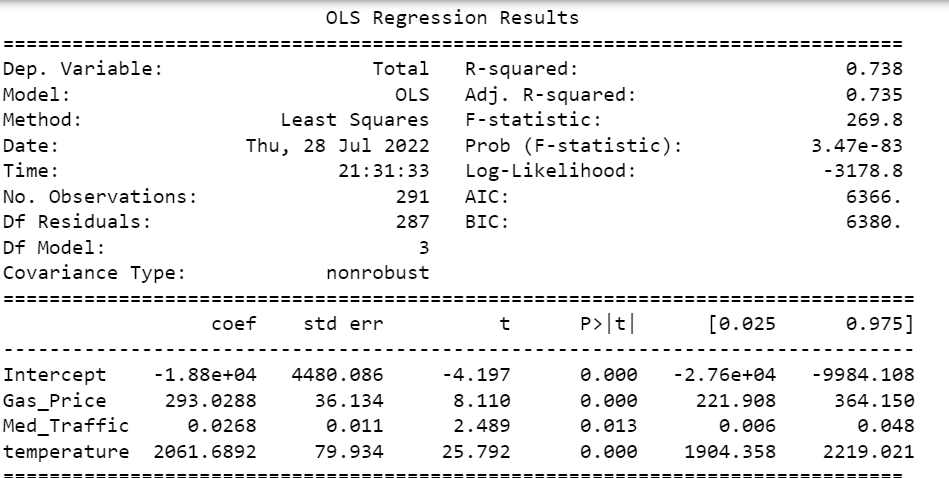
We can conclude the following from above correlation analysis:

* **Variation in trends across segments**: Trend and regularity in different segmentations would vary that the factors which influenced the number of riders in high member volume stations may not have the same impact on stations with lower member volume.
* **High Member Demand Stations**: The independent variables (weather, gas price and traffic) had a strong relationship with the total number of riders in high member volume stations.
* **Low Member Demand Stations**: The chosen variables do not significantly decide the number of the total riders. An unsuitable temperature, cheap gas price and fewer traffic on road may lead to fewer casual rider bikes but it was not certain that suitable temperature, high gas price and overcrowded road would result in more casual riders. The variables that would influence the behavior of casual riders are more complicated.
* **Final segments to be used in this research**: Since we don't have more information on what would affect the casual riders’ behavior, we would mainly focus on 3 high member volume tables (we excluded low total volume\_high member volume\_high casual volume table because it only included one station which may have an overfitting issue).



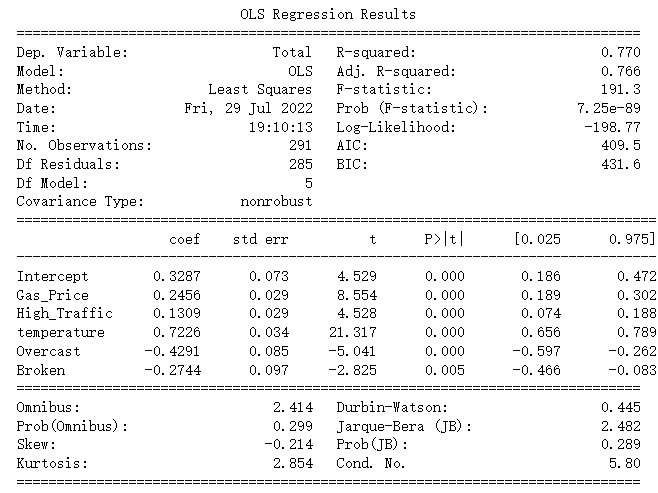
**MODEL DEVELOPMENT: POLYNOMIAL MODEL**

1. **Stepwise regression and p-value significance**: First we run stepwise regression for every variable to check for their R square value and remove the ones that have high correlation with another feature but have low r squared value. In this case, we drop wind speed and relative humidity as they have very low r squared value.
2. **Final Features:** Temperature, Cloud Cover, Gas price, Traffic volume
3. **High total\_high members\_high casual Segment:** Below is the ANOVA table for our first iteration

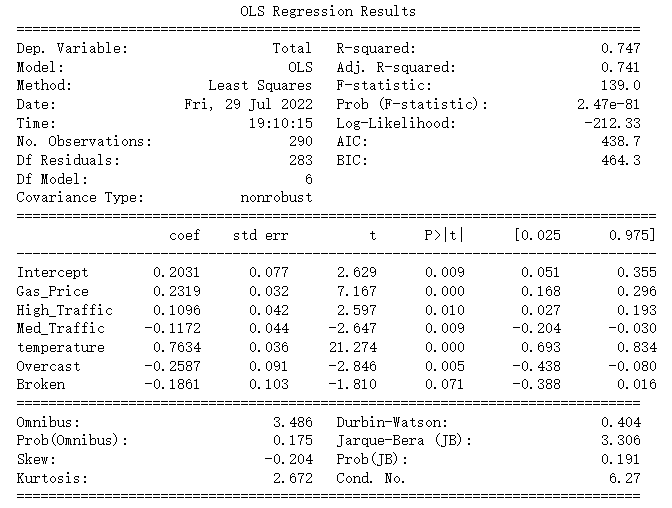


**Interpretation:** Though our R squared value is good (74%) with all the p-values below 0.05, we see that the temperature coefficient is **too high** - 2061. Hence to reduce the temperature coefficient, we need to normalize the data

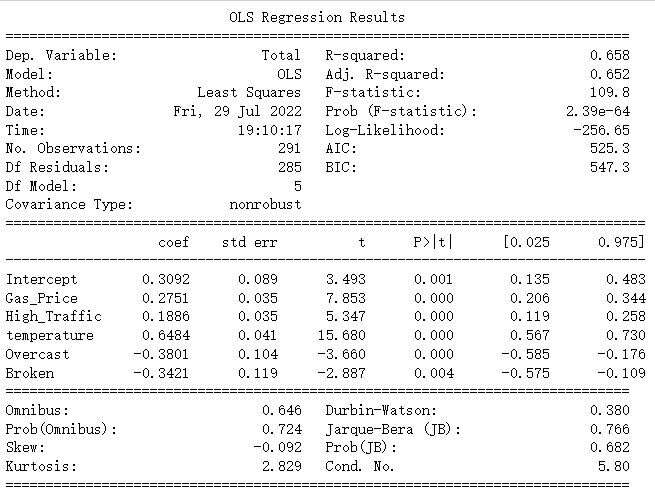
1. **Need for Normalization:** Since the coefficient for temperature is very high, we tried only normalizing the temperature column, however we didn't see any drop in the coefficient value. On a deeper analysis, we identified that **great variability** in the numbers in our features. For example: the temperature range was from -25 to 35, street traffic volume may reach thousands of hundreds, the gas price varied from 70 to 210. Hence it was necessary to normalize all the variables before modeling to know which variable drove the number of total riders. In this project, **Z-score** was used to normalize the dataset.
2. **Performance Analysis after Normalization:** Below are the ANOVA tables of the three models.  
   It can be concluded from the tables that the temperature was the biggest drive among all independent variables. Three tables’ R-squared values all indicated a strong predicting ability. The difference between three models was that in the High Total \_High Member\_Low Casual model, Med\_Traffic value also played an important role while not in the other two models. The adjusted R-squared value did not drop significantly with more variables being input into the model which also indicated every independent variable was meaningful and useful.



***Figure: ANOVA of High Total Volume\_High Member Volume\_High Casual Volume***

******

***Figure: ANOVA of High Total Volume\_High Member Volume\_Low Casual Volume***

******

***Figure: ANOVA of Low Total Volume\_High Member Volume\_Low Casual Volume***

|  |  |  |
| --- | --- | --- |
| **Datasets** | **R^2 Score** | **RMSE** |
| High Total High Member High Casual | 0.763685036 | 0.510380263 |
| High Total High Member Low Casual | 0.679329341 | 0.515434988 |
| Low Total High Member Low Casual | 0.655088664 | 0.586285495 |

***Figure: R^2 Value and RMSE before regularization***

After implementing the gradient Descent, the performance of the model did not improve and even dropped, which means Gradient Descent may not be a proper way to do so.

**REGULARIZATION TO IMPROVE R-SQUARED VALUE**: To further increase the R squared value and decrease the RMSE, we decided to implement regularization.

1. **Gradient Descent:**

|  |  |  |
| --- | --- | --- |
| **Datasets** | **R^2 Score** | **RMSE** |
| High Total High Member High Casual | 0.753207287 | 0.521572188 |
| High Total High Member Low Casual | 0.669842802 | 0.523003588 |
| Low Total High Member Low Casual | 0.6490671 | 0.591381124 |

***Figure: R^2 Value and RMSE after Gradient Descent***

1. **Ridge and Elastic Net Regression**: Lasso would shrink the coefficient to 0 and in our normalized model, it was not appropriate and instead of using Lasso, Ridge and Elastic Net Regression were implemented to improve the performance of the model. As a result, Elastic significantly decreased the performance of 3 models with higher RMSE and lower R^2 indicating the unsuitability for our model. Ridge improved the model performance especially at 2 degree polynomial features.

|  |  |  |
| --- | --- | --- |
| **Datasets** | **R^2 score** | **RMSE** |
| High Total High Member High Casual | 0.484163944 | 0.718217276 |
| High Total High Member Low Casual | 0.45769577 | 0.736413084 |
| Low Total High Member Low Casual | 0.403001892 | 0.772656527 |

***Figure: R^2 Value and RMSE after Elastic Net Regression***

|  |  |  |
| --- | --- | --- |
| **Datasets** | **R^2 score** | **RMSE** |
| High Total High Member High Casual | 0.808836682 | 0.437222275 |
| High Total High Member Low Casual | 0.77667779 | 0.472569794 |
| Low Total High Member Low Casual | 0.709155517 | 0.539299993 |

***Figure: R^2 Value and RMSE after Ridge***

**CONCLUSION**

As per our analysis, we can establish a clear picture that the bike demand is dependent on the temperature, cloud cover, gas price and traffic volume of which temperature is of the major driving factors followed by gas price. We segmented the stations into 8 categories based on total riders, annual riders, and casual riders and chose to work with the categories which includes high members as the chosen variables have a higher impact on their decision as compared to the low member segments. Created a polynomial model and normalized the data using Z-Score to lower the coefficient value and improve our model performance. Below is the summary of our model performance along with the equations that can be used to make the predictions:

|  |  |  |  |
| --- | --- | --- | --- |
| **Before Regularization** | | | |
| **Datasets** | **R^2 score** | **RMSE** | **Predicting equation** |
| **High Total High Member High Casual** | 0.7636 | 0.5103 | Total = 0.33 + (0.25\*gas price) + (0.13\* high traffic) + (0.72\* temperature) – (0.43\* overcast) -  (0.27 \* broken) |
| **High Total High Member Low Casual** | 0.6793 | 0.5154 | Total = 0.20 + (0.23\*gas price) + (0.11\* high traffic) – (0.12 \* low traffic) + (0.76\* temperature) – (0.25\* overcast) -  (0.18 \* broken) |
| **Low Total High Member Low Casual** | 0.6550 | 0.586285 | Total = 0.20 + (0.23\*gas price) + (0.11\* high traffic) + (0.72\* temperature) – (0.43\* overcast) -  (0.27 \* broken) |

To further improvise the model, we performed various regularization techniques of which we could see a significant improvement through ridge regression.

|  |  |  |
| --- | --- | --- |
| **After Regularization - Ridge Regression** | | |
| **Datasets** | **R^2 score** | **RMSE** |
| **High Total High Member High Casual** | 0.8088 | 0.4372 |
| **High Total High Member Low Casual** | 0.7766 | 0.4725 |
| **Low Total High Member Low Casual** | 0.7091 | 0.5392 |

**RECOMMENDATION:**

1. It is recommended that model results be regenerated after Phase 1 performance results become available. Results from the new stations will be instrumental in refining the predictive capacity of the model and more accurately delineating the areas of least risk for future expansion.
2. We could not implement our model to the low member stations as of now, hence that can be the next step to identify what motivates them to ride bikes