# MSDS-601: LINEAR REGRESSION ANALYSIS

# FINAL PROJECT

Fall 2023, Professor Shan Wang
By: Shagun Kala, Varsha Moturi, Amadeo Cabanela


# BIKE RENTAL DEMAND PREDICTION

# DATASET DESCRIPTION

## Data Source:

https://www.kaggle.com/datasets/harbhajansingh21/bike-sharing-dataset/data

We perform our analysis on a bike sharing dataset available through the Kaggle platform. The dataset contains hourly bike rental information from the Capital Bikeshare system in Washington D.C. for the years 2011 and 2012 combined with corresponding weather and seasonal information extracted from Freemeteo. The dataset comes with 17,381 observations and 17 columns. The target variable is cnt, which is the number of bikes rented per hour.

## Data Dictionary:

| S.No. | Variable | Data Type | Description |
|---|---|---|---|
| 1. | instant | int64 | Record index |
| 2. | dteday | object | Date |
| 3. | season | object | Season category:<br>1: spring<br>2: summer<br>3: fall<br>4: winter |
| 4. | yr | object | Year:<br>0: 2011<br>1: 2012 |
| 5. | mnth | object | Month (ranges from 1 to 12) |
| 6. | hr | int64 | Hour (ranges from 0 to 23) |
| 7. | holiday | int64 | Whether day is a holiday or not |
| 8. | weekday | int64 | Day of the week |
| 9. | workingday | int64 | 1: Day is neither weekend or holiday<br>0: Otherwise |
| 10. | weathersit | object | Weather category:<br>1: Clear<br>2: Mist<br>3: Light Snow/Rain<br>4: Heavy Snow/Rain |
| 11. | temp | float64 | Normalized temperature in Celsius. The values are divided to 41 (max) |

| | | | |
|---|---|---|---|
| 12. | atemp | float64 | Normalized feeling temperature in Celsius. The values are divided to 50 (max) |
| 13. | hum | float64 | Normalized humidity. The values are divided to 100 (max) |
| 14. | windspeed | float64 | Normalized wind speed. The values are divided to 67 (max) |
| 15. | casual | int64 | Number of casual users |
| 16. | registered | int64 | Count of registered users |
| 17. | cnt | int64 | Count of total rental bikes per hour (including both casual and registered) |

'`cnt`' is the target variable in this dataset indicating the hourly bike rental count.

# ABSTRACT

Bike sharing is a shared transport service where the process of administering bike rentals including registration, payment, rentals and returns are automated. Given the benefits bike sharing systems offer to society, from reducing traffic congestion to reducing carbon emissions to promoting public health, it is important to analyze the valuable data collected to understand bike sharing patterns and make data-driven decisions to optimize bike sharing usage.

# RESEARCH PROBLEM

Problem statement: Optimize bike rental operations and maintenance within the Washington D.C. area by predicting the hourly number of bike rentals based on environmental and seasonal factors and identifying the factors having strong influence on the hourly bike count.

We conduct regression analysis to answer the following research questions:

1. Which factors significantly influence the hourly bike rental count?
2. Is there a linear relationship between bike rental factors and the hourly bike rental count?
3. How accurately can we predict the hourly bike rental count?

# APPROACH- SUMMARY OF METHODS

We performed multiple linear regression (MLR) analysis to fit models to the dataset for prediction and understand the impact of predictor variables on the response variable for inference. We determine the quality of fit of our models by performing model diagnostics and model assumption checks. Our analysis involves the following steps:

- Data Cleaning

- Exploratory Data Analysis
- Selected predictors of interest
- Model Diagnosis:
    - Multicollinearity check and remediation
    - Fitted initial linear regression model
    - Influential points analysis and remediation
    - Heteroscedasticity check and remediation
    - Normality check and remediation
- Fitted the final linear regression model on rectified data
- Checked feature significance on the fitted model using T-test, global F-test, ANOVA type 1 test and ANOVA type 2 test.
- Conducted model (feature) selection based on the best subset with adj-R-squared, AIC and BIC performance metrics.
- Fit the selected/final model using the following combination of predictors: humidity, apparent temperature, windspeed, weather, season, casual, hour, is_holiday, is_weekday, month, year.
- Compared the performance of the final linear regression model with the advanced non-linear ensemble models like Random Forest, Light GBM and XGBoost.
- Conclusion

# DATA CLEANING

We improved the dataset's comprehensibility and overall quality by renaming features. Using the data dictionary, we associated label encodings with their corresponding category names, enabling more insightful Exploratory Data Analysis (EDA). The dataset exhibited no missing values or unusual anomalies. After identifying and removing five duplicate rows, we omitted the `'instant'` variable as it merely served as an index for bookkeeping and held no relevance to our analysis. Additionally, we excluded the `'dteday'` variable, as our analysis focuses solely on linear regression rather than time series analysis.

# EXPLORATORY DATA ANALYSIS

Following the data cleaning process, we proceeded with exploratory data analysis to gain a deeper understanding of both the predictors and target variables within the dataset. Post-cleaning, the dataset now comprises 17,374 rows and 15 columns in terms of dimensions.

## Univariate Analysis

We initiated the analysis by creating histograms for all numerical variables and examining their distributions. Notably, in Figure 1a, it is evident that the distribution of the target variable `'count'` exhibits a rightward skew, displaying an average hourly count of 189.46 with a standard deviation of 181.38. Figure 1b illustrates that the majority of records indicate a weather condition of "Clear". This observation implies that the `'weather'` variable could potentially influence the count of hourly bike rentals.

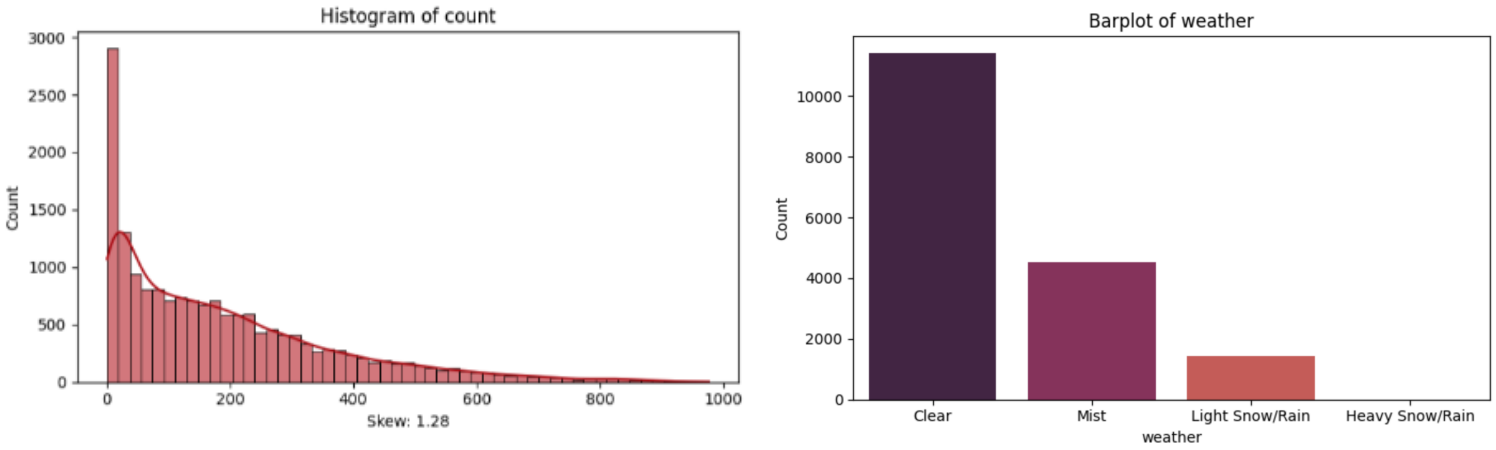



*Figure 1: Figure 1a on the left shows the distribution of the count variable. Figure 1b on the right shows the distribution of weather condition categories for the weather variable.*

## Bi-Variate Analysis

We also observed the behavior of two variables together. In particular, as shown in Figure 2, the bike rental count has increased from 2011 to 2012. The curve has flattened in the later year.

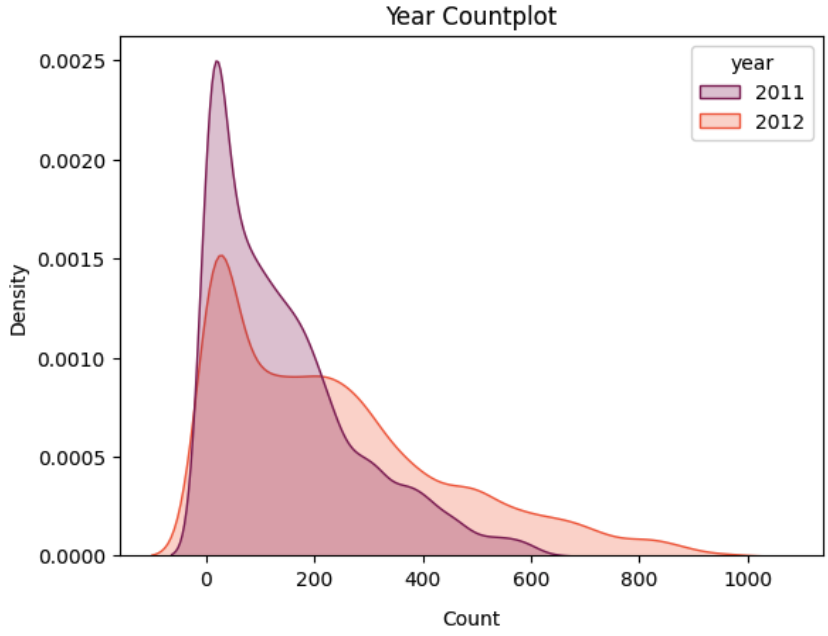



*Figure 2: Density plot of count for year 2011 and year 2012.*

Observing the higher frequency of "Clear" weather conditions in Figure 1b's histogram, we became interested in investigating a potential correlation between weather and rental counts. Figure 3's box plot reveals a trend: fewer bikes are rented during severe weather conditions, such as heavy snow/rain and light snow/rain, in contrast to favorable weather conditions like clear and mist.

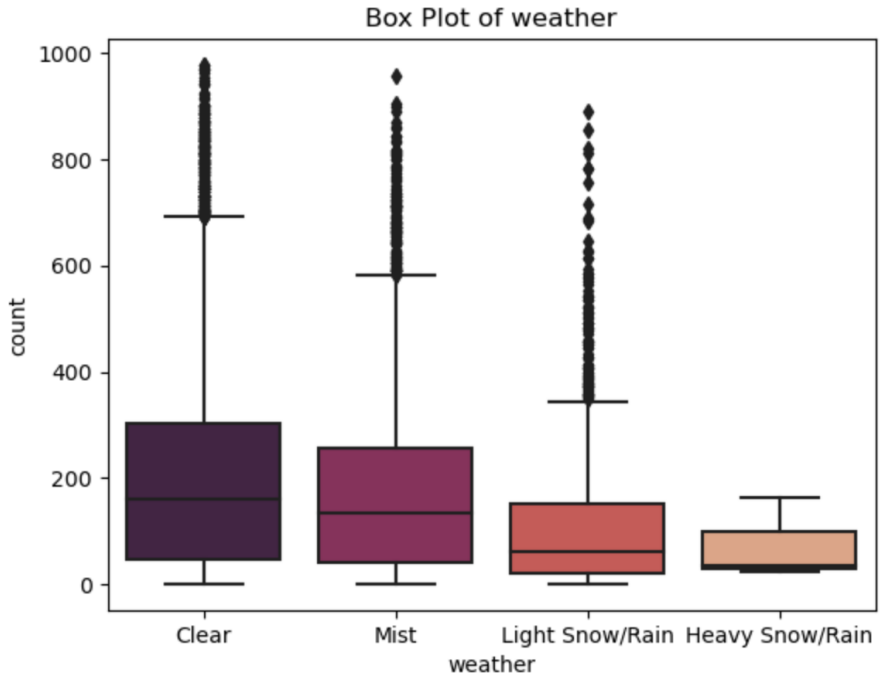



*Figure 3: Box plot of weather vs. count.*

## Multi–Variate Analysis

In Figure 4, it's evident that there's an uptick in bike rentals during typical office commute hours. Additionally, during weekends, there's a trend of higher bike rentals in the afternoon and lower rentals in the morning.

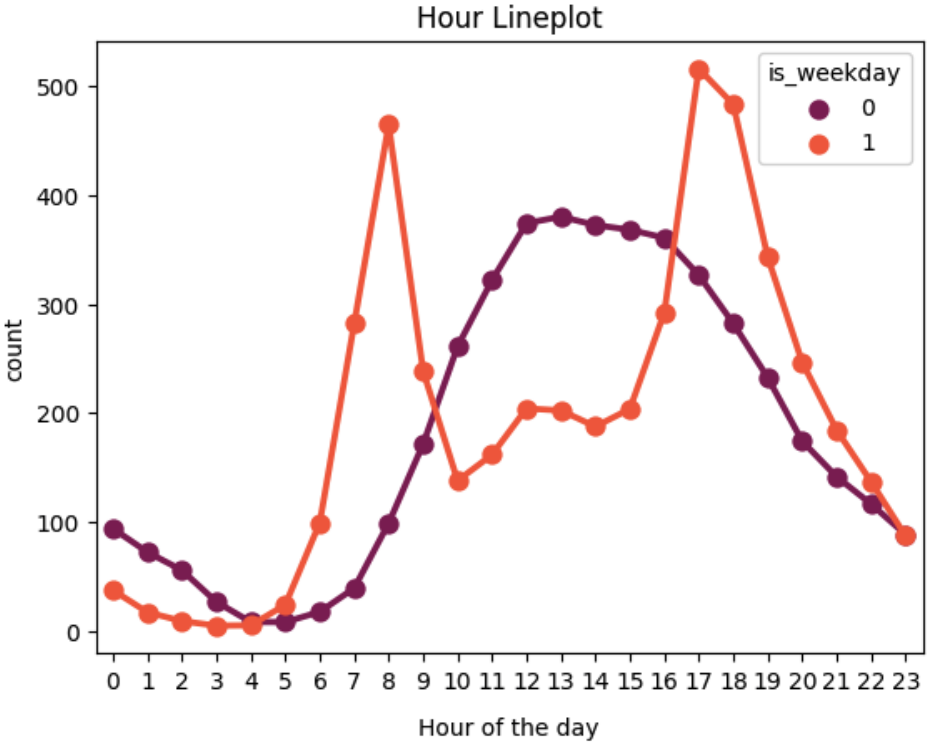



*Figure 4 Line Plot between hour of the day, count and is_weekday*

Interestingly, people tend to rent bikes much less during the spring season compared to other seasons (See Figure 5).

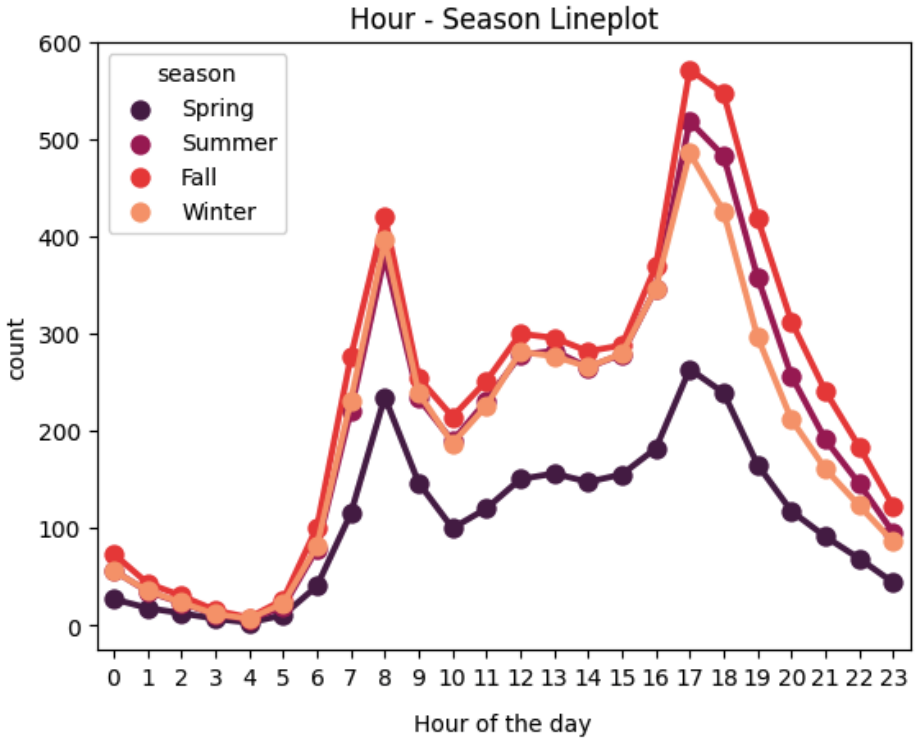



*Figure 5 Line Plot between hour of the day, count and season*

People tend to rent more bikes during warmer climates than colder climates. The number of bikes rented does not vary significantly on working days and holidays. (See Figure 6)

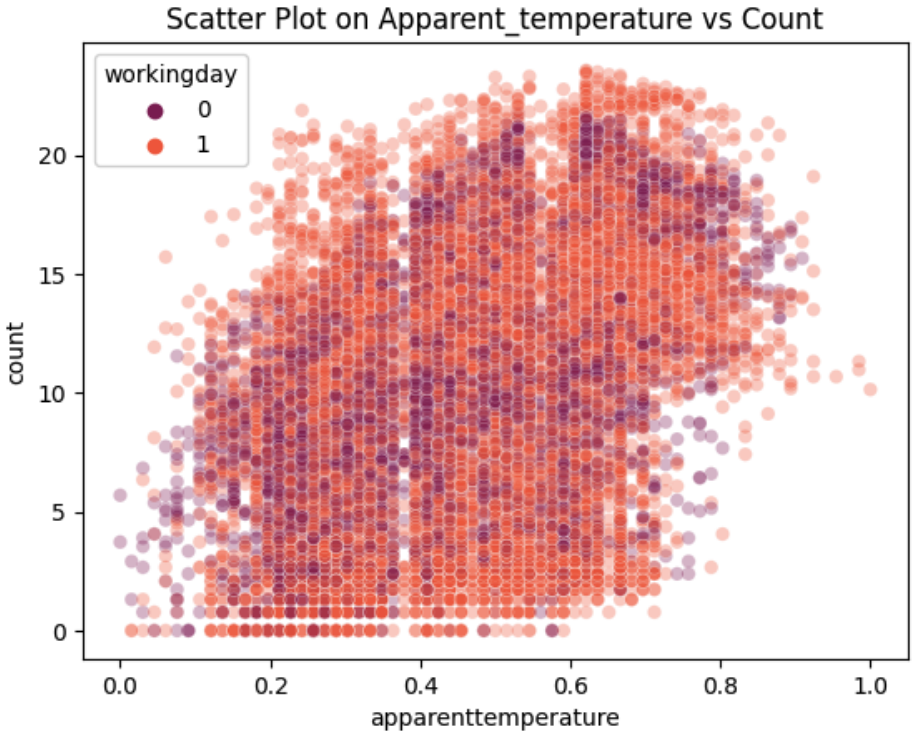



*Figure 6 Scatter Plot between apparent temperature, count and workingday*

# REGRESSION: MODEL DIAGNOSIS

## 1. Multicollinearity Check:

To evaluate multicollinearity among the numerical variables in our regression model, we undertook a comprehensive examination. This process involved the following steps:

**Correlation Analysis:**

We began by plotting a correlation heatmap, as shown in Figure 7, to visualize the correlation strengths between the numeric variables and to identify any potential multicollinearity issues.

The heatmap revealed several important insights:

1. `'Registered'` rentals displayed a high correlation with the total rentals count.
2. `'Casual'` rentals exhibited a moderate level of correlation with the total rentals count.
3. `'Casual'` rentals were also correlated with the temperature variables.
4. The temperature variables, `'temperature'` and `'apparenttemperature,'` displayed a high level of correlation with each other.

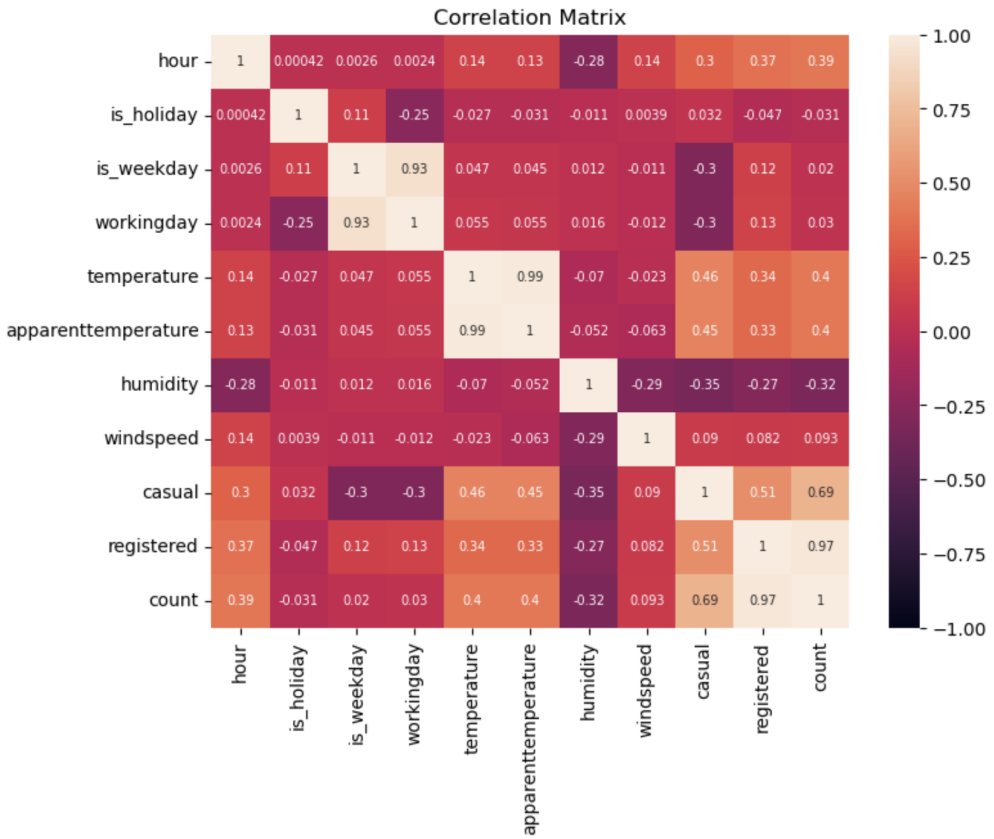



*Figure 7: Correlation heatmap of numeric variables.*

**Multicollinearity Remediation Steps:**

To mitigate the multicollinearity issue and enhance the stability of our regression model, we implemented the following corrective actions:

1. We eliminated the predictor variable `'registered'` due to its high correlation with our target variable `'count`.'
2. Similarly, we opted to retain `'apparenttemperature'` while excluding `'temperature'` from the model, as they exhibited a high level of correlation. This decision was made to preserve crucial weather-related information while mitigating multicollinearity.
3. We also considered the Variance Inflation Factor (VIF) scores for our variables. Notably, we observed that the temperature variables `'temperature'` and `'apparenttemperature'` displayed notably elevated VIF scores of 53 and 46, respectively. This further affirmed our choice to retain only `'apparenttemperature'` in order to address the multicollinearity issue.
4. Additionally, we computed the VIF score for the categorical variables `'is_holiday`,' `'is_working`,' and `'workingday`.' These variables yielded infinite (INF) VIF scores, indicating a severe multicollinearity problem. Consequently, we decided to exclude the `'workingday'` column from our analysis.

**Before treating multicollinearity**

| | VIF Factor | features |
|---|---|---|
| 0 | 118.739069 | Intercept |
| 1 | 1.003148 | C(weather)[T.Heavy Snow/Rain] |
| 2 | 1.272442 | C(weather)[T.Light Snow/Rain] |
| 3 | 1.165111 | C(weather)[T.Mist] |
| 4 | 10.228692 | C(season)[T.Spring] |
| 5 | 7.868323 | C(season)[T.Summer] |
| 6 | 8.234225 | C(season)[T.Winter] |
| 7 | 5.483144 | C(month)[T.Aug] |
| 8 | 5.302250 | C(month)[T.Dec] |
| 9 | 4.849149 | C(month)[T.Feb] |
| 10 | 5.325437 | C(month)[T.Jan] |
| 11 | 5.660041 | C(month)[T.Jul] |
| 12 | 2.627121 | C(month)[T.Jun] |
| 13 | 3.195975 | C(month)[T.Mar] |
| 14 | 2.093734 | C(month)[T.May] |
| 15 | 6.192519 | C(month)[T.Nov] |
| 16 | 6.086167 | C(month)[T.Oct] |
| 17 | 4.755080 | C(month)[T.Sept] |
| 18 | 1.098596 | C(year)[T.2012] |
| 19 | 1.779642 | humidity |
| 20 | 53.679940 | temperature |
| 21 | 46.262115 | apparenttemperature |
| 22 | 1.241649 | windspeed |
| 23 | 2.280382 | casual |
| 24 | 1.799971 | registered |
| 25 | 1.301207 | hour |
| 26 | inf | is_holiday |
| 27 | inf | is_weekday |
| 28 | inf | workingday |

**After treating multicollinearity**

| | VIF Factor | features |
|---|---|---|
| 0 | 118.320737 | Intercept |
| 1 | 1.003129 | C(weather)[T.Heavy Snow/Rain] |
| 2 | 1.263581 | C(weather)[T.Light Snow/Rain] |
| 3 | 1.162367 | C(weather)[T.Mist] |
| 4 | 10.206236 | C(season)[T.Spring] |
| 5 | 7.842814 | C(season)[T.Summer] |
| 6 | 8.210231 | C(season)[T.Winter] |
| 7 | 5.284272 | C(month)[T.Aug] |
| 8 | 5.279579 | C(month)[T.Dec] |
| 9 | 4.823006 | C(month)[T.Feb] |
| 10 | 5.289964 | C(month)[T.Jan] |
| 11 | 5.507640 | C(month)[T.Jul] |
| 12 | 2.490613 | C(month)[T.Jun] |
| 13 | 3.192335 | C(month)[T.Mar] |
| 14 | 2.040214 | C(month)[T.May] |
| 15 | 6.176367 | C(month)[T.Nov] |
| 16 | 6.085931 | C(month)[T.Oct] |
| 17 | 4.693312 | C(month)[T.Sept] |
| 18 | 1.035109 | C(year)[T.2012] |
| 19 | 1.764634 | humidity |
| 20 | 4.675204 | apparenttemperature |
| 21 | 1.161717 | windspeed |
| 22 | 1.918959 | casual |
| 23 | 1.212039 | hour |
| 24 | 1.045143 | is_holiday |
| 25 | 1.227410 | is_weekday |

*Figure 8: VIF Score of all predictors*

Multicollinearity has been effectively resolved, as all variables now exhibit VIF values below the threshold of 10.

## 2. Fitting the initial model

In the regression model, we've fitted the variable `'count'` as the dependent variable, and the following independent variables:

- `Humidity`
- `Apparent Temperature`
- `Wind Speed`
- Categorical variable `'Weather'` (treated as categorical)
- Categorical variable `'Season'` (treated as categorical)
- `Casual`
- `Hour`
- `Is_Holiday`
- `Is_Weekday`
- Categorical variable `'Month'` (treated as categorical)
- Categorical variable `'Year'` (treated as categorical)

This yields the following model:

*count ~ humidity + apparenttemperature + windspeed + C(weather) + C(season) + casual + hour + is_holiday + is_weekday + C(month) + C(year)*

Our selected model was fitted using the least squares method, which is a common approach for linear regression. We can see the initial performance of this model in the summary table in Figure 9 below:

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | count | **R-squared:** | 0.613 |
| **Model:** | OLS | **Adj. R-squared:** | 0.613 |
| **Method:** | Least Squares | **F-statistic:** | 1100. |
| **Date:** | Sat, 07 Oct 2023 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 16:13:58 | **Log-Likelihood:** | -1.0676e+05 |
| **No. Observations:** | 17374 | **AIC:** | 2.136e+05 |
| **Df Residuals:** | 17348 | **BIC:** | 2.138e+05 |
| **Df Model:** | 25 | | |
| **Covariance Type:** | nonrobust | | |

*Figure 9: Summary table of OLS regression results for the initial linear regression model.*

**Model Performance:**

The R-squared and Adjusted R-squared values are essential indicators of the model's goodness of fit. From Figure 9 we can see the values are:

- R-squared: 0.613
- Adjusted R-squared: 0.613

These values suggest that the model explains approximately 61.3% of the variance in the dependent variable `'count`.' This indicates a moderate level of predictive power, meaning that the included independent variables collectively account for a substantial portion of the variability in `'count`.'

# 3. Influential Points Analysis:

In our efforts to assess and enhance the robustness of our regression model, we conducted an influential points analysis. This analysis aimed to identify and subsequently remove potential outliers that could significantly impact the model's stability and performance.

Two common criteria were employed to identify influential points:

a. **External Studentized Residuals:** To identify potential outliers, we calculated the external studentized residuals. A threshold was set using the t-distribution at a significance level of alpha=0.05/2 (corresponding to a two-tailed test), and the degrees of freedom (Df) were adjusted to Df=n-1-p, where 'n' represents the number of data points, and 'p' denotes the number of predictors in the model. Data points with external studentized residuals exceeding this threshold were flagged as potential outliers.

b. **Cook's Distance:** Cook's distance was employed as a measure of the influence of individual data points on the regression coefficients. A threshold, commonly set at 4/n, where 'n' represents the total number of data points, was used to identify data points with substantial influence on the model.

**Influential Points Remediation:**

Following the application of these refined criteria, we identified 814 data points that met both conditions. Consequently, these data points were systematically removed from the dataset. This process aimed to bolster the model's reliability by mitigating the potential negative effects of outliers on model performance, parameter estimates, and the overall predictive accuracy.

# 4. Heteroscedasticity Check

As part of our regression model diagnostics, we examined the assumption of constant variance, also known as homoscedasticity, for the model's residuals. This assumption implies that the spread or variance of the residuals should remain relatively consistent across different values of the independent variables.

### a. Visual Inspection:

Initially, we visually inspected the relationship between the fitted values and the residuals by plotting a graph of fitted values against residuals. The resulting plot revealed a characteristic "<" shape, indicating that the variance of the residuals exhibited a distinct pattern. Specifically, the residuals displayed a smaller variance in the beginning and progressively increased as the fitted values increased. This pattern strongly suggests the presence of heteroscedasticity.

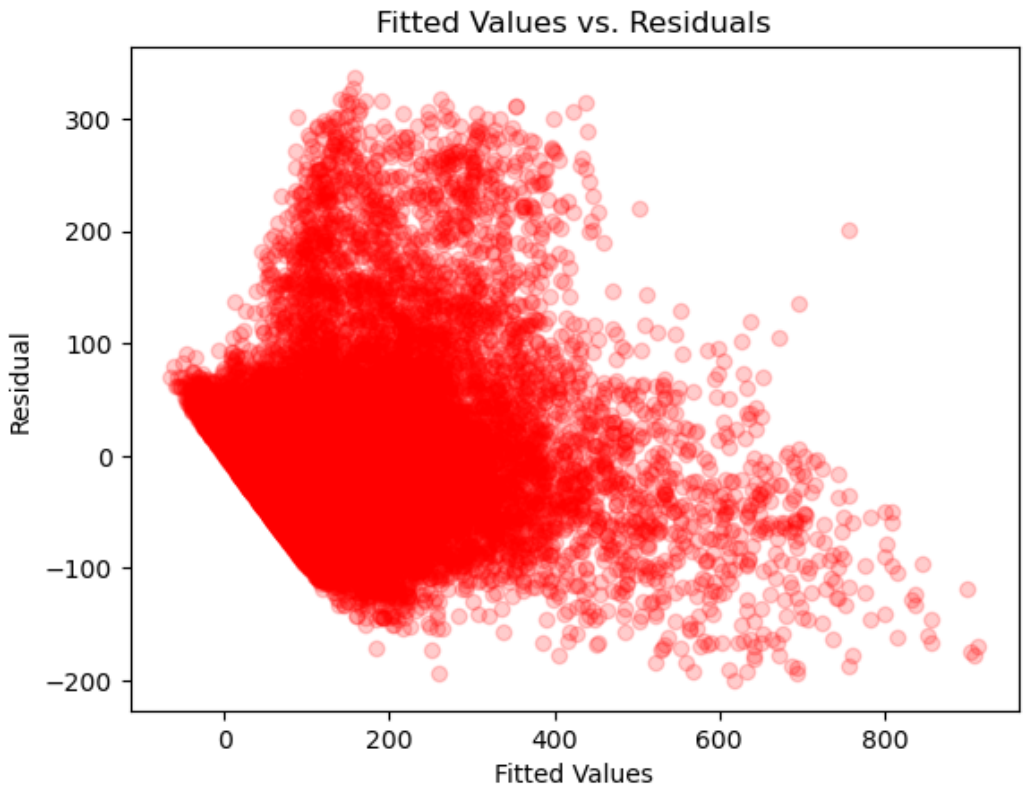



*Figure 10: Fitted Values vs. Residuals Plot*

### b. Breusch-Pagan Test:

To substantiate our visual observations, we conducted a formal statistical test known as the Breusch-Pagan test. The Breusch-Pagan test assesses whether the variance of the residuals is dependent on the values of the independent variables.

The results of the Breusch-Pagan test are as follows:

- 'BP Statistic': 1320.76
- 'BP-Test p-value': 3.16e-263 (approximately)

The remarkably high BP Statistic and the extremely low p-value (approximately 3.16e-263) provide strong statistical evidence for the presence of heteroscedasticity in the data. The high BP Statistic indicates a significant departure from homoscedasticity, and the low p-value suggests that the null hypothesis of constant variance is highly unlikely to be true.

We will apply Box Cox transformation on the target variable to stabilize the variance and make it more homoscedastic.

## 5. Check for Normality

In our regression analysis, we made a critical examination of the normality assumption, which is one of the key assumptions underlying linear regression. This assumption suggests that the residuals or errors should be normally distributed for reliable model results.

### a. QQ Plot

To assess the normality of our data, we created a Quantile-Quantile (Q-Q) plot. This plot visualizes the distribution of residuals against the expected quantiles of a normal distribution. In our Q-Q plot, we observed a curved line, indicating a significant departure from the normal distribution. The curvature suggests that the data does not adhere to a perfectly normal distribution.

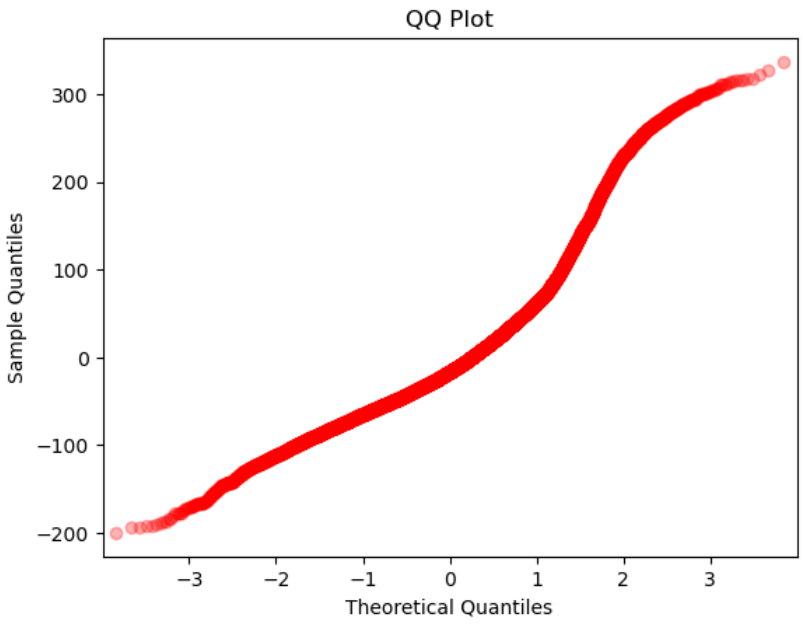



*Figure 11: QQ Plot on the fitted model*

### b. Jarque-Bera Test

To further quantify the departure from normality, we conducted the Jarque-Bera test, which is a statistical test used to assess the normality of data by examining skewness and kurtosis. The test results were as follows:

- 'Jarque-Bera Statistic': 8018.91
- 'p-value': 0.0

The Jarque-Bera statistic exhibited an exceptionally high value, and the p-value was substantially below the common significance level of 0.05. This provides strong evidence that the data significantly deviates from a normal distribution.

### c. Skew and kurtosis

| | | | |
|---|---|---|---|
| **Omnibus:** | 6247.750 | **Durbin-Watson:** | 0.687 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 23139.511 |
| **Skew:** | 1.801 | **Prob(JB):** | 0.00 |
| **Kurtosis:** | 7.358 | **Cond. No.** | 4.69e+03 |

*Figure 11a: Summary table of OLS regression results for the initial linear regression model*

Furthermore, the model summary revealed significant skewness (1.801) and kurtosis (7.358), providing additional confirmation of the non-normality issue within the dataset. This observation further emphasized the need to address the data's departure from normal distribution for more robust and reliable modeling results.

**Remediation of Normality:**

In response to the observed departure from normality in our data, we took a proactive step to address this issue. Specifically, we applied a Box-Cox transformation to the target feature, `'count.'` The Box-Cox transformation is a well-established and effective method for mitigating non-normality in data by reshaping its distribution to more closely resemble a normal distribution. By performing this transformation, we aimed to rectify the non-normality issue and enhance the reliability of our regression model's results.

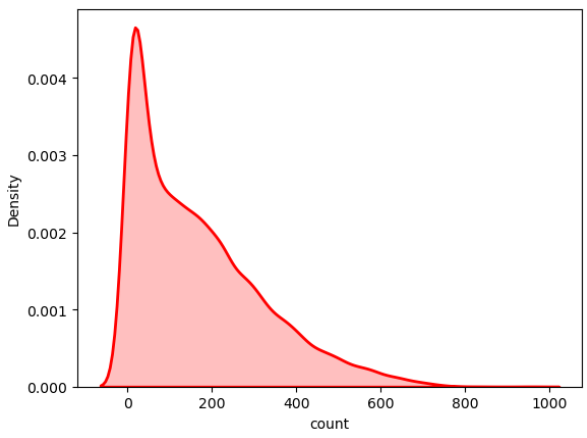
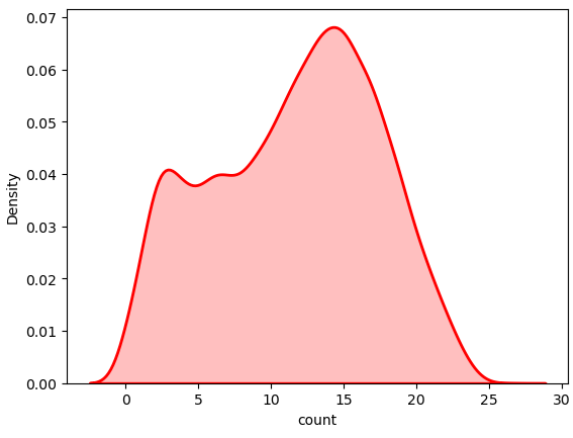
| **'Count' Variable before treating Normality** | **'Count' Variable after treating Normality** |
|---|---|
|  |  |



*Figure 12: Before and after applying Box Cox transformation on 'Count' Variable.*

The effectiveness of the Box-Cox transformation is evident when examining the distribution of the target variable. After applying the transformation, we can observe a notable improvement in the alignment of the `'count'` variable with the characteristics of a normal distribution, as visually depicted in the histogram plot.

# 6. Fitting the final linear regression model on rectified data

After conducting a comprehensive series of analyses, including the identification and removal of influential points, addressing heteroscedasticity, and rectifying non-normality through a Box-Cox transformation, we proceeded to train our regression model on the refined data. The variables used in the model remained consistent with our previous specifications.

**OLS Regression Results:**

The R-squared and Adjusted R-squared values have seen a notable improvement. We progressed from an initial R-squared of 0.613 and Adjusted R-squared of 0.613 to an enhanced R-squared of 0.678 and an Adjusted R-squared of 0.677. These higher values indicate that the model explains a greater proportion of the variance in the dependent variable `'count'`,' demonstrating improved predictive power and model fit.

OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | count | **R-squared:** | 0.678 |
| **Model:** | OLS | **Adj. R-squared:** | 0.677 |
| **Method:** | Least Squares | **F-statistic:** | 1391. |
| **Date:** | Sat, 07 Oct 2023 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 16:20:23 | **Log-Likelihood:** | -42987. |
| **No. Observations:** | 16560 | **AIC:** | 8.603e+04 |
| **Df Residuals:** | 16534 | **BIC:** | 8.623e+04 |
| **Df Model:** | 25 | | |
| **Covariance Type:** | nonrobust | | |

*Figure 13: Summary table of OLS regression results for the final linear regression model*

**Global F-Test:**

The global F-test is a statistical test that assesses the overall significance of the model. It evaluates whether the model, as a whole, significantly explains the variance in the dependent variable `'count'`.' The F-statistic measures the ratio of the explained variance to the unexplained variance in the model.

- F-statistic: In this case, the F-statistic is 1391, which is a large and positive value.

- Prob (F-statistic): The associated p-value for the F-statistic is extremely close to zero (0.00).

**Interpretation of global F-test:**

The F-statistic tests the null hypothesis that all the regression coefficients are equal to zero, meaning that none of the predictor variables in the model are significant. A low p-value, close to zero, suggests that we should reject this null hypothesis.

In our context, the F-statistic of 1391 and the tiny p-value indicate that the model as a whole is highly significant. This implies that at least one predictor variable in the model has a significant effect on `'count,'` and the model provides a statistically significant explanation for the variation in bike rentals. In other words, the predictor variables collectively contribute to predicting the number of bike rentals, and the model is a valuable tool for understanding and forecasting this relationship.

**T-Test:**

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 3.3644 | 0.275 | 12.221 | 0.000 | 2.825 | 3.904 |
| **C(weather)[T.Heavy Snow/Rain]** | 1.1825 | 1.878 | 0.630 | 0.529 | -2.498 | 4.863 |
| **C(weather)[T.Light Snow/Rain]** | -0.8993 | 0.103 | -8.721 | 0.000 | -1.101 | -0.697 |
| **C(weather)[T.Mist]** | 0.4331 | 0.062 | 6.998 | 0.000 | 0.312 | 0.554 |
| **C(season)[T.Spring]** | -1.1022 | 0.188 | -5.852 | 0.000 | -1.471 | -0.733 |
| **C(season)[T.Summer]** | -0.3015 | 0.164 | -1.837 | 0.066 | -0.623 | 0.020 |
| **C(season)[T.Winter]** | 1.0381 | 0.170 | 6.104 | 0.000 | 0.705 | 1.371 |
| **C(month)[T.Aug]** | -0.0933 | 0.210 | -0.445 | 0.657 | -0.504 | 0.318 |
| **C(month)[T.Dec]** | 0.7117 | 0.207 | 3.434 | 0.001 | 0.306 | 1.118 |
| **C(month)[T.Feb]** | 1.2263 | 0.207 | 5.924 | 0.000 | 0.821 | 1.632 |
| **C(month)[T.Jan]** | 1.0577 | 0.210 | 5.026 | 0.000 | 0.645 | 1.470 |
| **C(month)[T.Jul]** | -0.4779 | 0.213 | -2.242 | 0.025 | -0.896 | -0.060 |
| **C(month)[T.Jun]** | -0.1461 | 0.146 | -1.003 | 0.316 | -0.432 | 0.139 |
| **C(month)[T.Mar]** | 0.4994 | 0.162 | 3.087 | 0.002 | 0.182 | 0.816 |
| **C(month)[T.May]** | 0.4781 | 0.129 | 3.702 | 0.000 | 0.225 | 0.731 |
| **C(month)[T.Nov]** | 0.1908 | 0.228 | 0.837 | 0.403 | -0.256 | 0.638 |
| **C(month)[T.Oct]** | -0.0333 | 0.225 | -0.147 | 0.883 | -0.475 | 0.409 |
| **C(month)[T.Sept]** | 0.1558 | 0.200 | 0.779 | 0.436 | -0.236 | 0.548 |
| **C(year)[T.2012]** | 1.0169 | 0.051 | 19.840 | 0.000 | 0.916 | 1.117 |
| **humidity** | -2.7779 | 0.173 | -16.015 | 0.000 | -3.118 | -2.438 |
| **apparenttemperature** | 5.2312 | 0.318 | 16.463 | 0.000 | 4.608 | 5.854 |
| **windspeed** | 1.4297 | 0.222 | 6.432 | 0.000 | 0.994 | 1.865 |
| **casual** | 0.0601 | 0.001 | 83.920 | 0.000 | 0.059 | 0.062 |
| **hour** | 0.2963 | 0.004 | 75.004 | 0.000 | 0.289 | 0.304 |
| **is_holiday** | -1.8677 | 0.152 | -12.251 | 0.000 | -2.166 | -1.569 |
| **is_weekday** | 1.4693 | 0.061 | 24.007 | 0.000 | 1.349 | 1.589 |

*Figure 14: Summary table of T-test results for the final linear regression model*

**Categorical Weather Variables:** As shown in Figure 14, the significance of the various weather categories is determined by their respective coefficients and t-statistics:

1. **Heavy Snow/Rain:** This category is not statistically significant (p-value = 0.529 > 0.05), it suggests that '`Heavy Snow/Rain`' may not have a substantial impact on the response variable '`count`.'
2. **Light Snow/Rain:** The '`Light Snow/Rain`' category is highly significant, with a t-statistic of -8.721 and a p-value of 0.000. This indicates a strong and negative impact on '`count`.' The coefficient of -0.8993 suggests that this weather condition is associated with a decrease in bike rentals.
3. **Mist:** The '`Mist`' category is also not statistically significant (p-value = 0.358 > 0.05). The coefficient for '`Mist`' is 0.4331, suggesting that this weather condition may not have a substantial effect on '`count`.'
4. Since at least one category is statistically significant, the whole '`weather`' variable is still statistically significant.

**Categorical Season Variables:** The seasonal categories' significance is determined similarly:

1. '`Spring`': '`Spring`' is highly significant (p-value of 0.000) with a negative coefficient of -1.1022.
2. '`Summer`': '`Summer`' is somewhat borderline in terms of significance with a p-value of 0.066 and a coefficient of -0.3015, but not significant.
3. '`Winter`': '`Winter`' is highly significant with a positive coefficient of 1.0381 and a very low p-value (0.000).
4. Since at least one category is statistically significant, the whole '`season`' variable is still statistically significant.

**Categorical Month Variables:** The month categories' significance is assessed in the same manner:

1. Some months, like '`Feb`', '`Jan`', '`Dec`', '`Mar`' and '`May`' are statistically significant with low p-values, indicating they impact 'count.'
2. Other months, like '`Jun`', '`Aug`', '`Oct`', '`Nov`' and '`Sept`' are not statistically significant (p-values > 0.05).
3. Since at least one category is statistically significant, the whole '`month`' variable is still statistically significant.

**Categorical Year Variable:** The year category '`2012`' is highly significant (p < 0.001) with a coefficient of 1.0169, indicating its strong influence on '`count`' than year '`2011`'

**Other Numeric Predictors:**

The predictors '`humidity`', '`apparenttemperature`', '`windspeed`', '`casual`', '`hour`', '`is_holiday`', and '`is_weekday`' are all highly significant, as their associated

t-test p-values are very close to 0.000. This indicates that these variables have a substantial impact on `'count'`.'

In summary, the results of the t-tests indicate that nearly all of the predictors in the regression model are statistically significant. This suggests that the majority of the included variables have a noteworthy impact on the dependent variable, `'count'`.' These findings underscore the importance of these predictors in explaining the variability in rental counts and emphasize their value in the predictive capability of the model.

**Improvement in Normality:**

The assessment of the data's skewness and kurtosis also yields positive results. The data now exhibits a skewness of 0.239 and a kurtosis of 2.957, which suggest a marked improvement in the normality of the data distribution. This development aligns the data more closely with the normality assumption, further enhancing the model's reliability.

| | | | |
|---|---|---|---|
| **Omnibus:** | 154.455 | **Durbin-Watson:** | 0.608 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 158.610 |
| **Skew:** | 0.239 | **Prob(JB):** | 3.62e-35 |
| **Kurtosis:** | 2.957 | **Cond. No.** | 4.49e+03 |

*Figure 15: Summary table of OLS regression results for the final linear regression model*

In summary, the combined efforts to enhance data quality and address assumptions have significantly improved the performance and reliability of our regression model. These positive outcomes underscore the effectiveness of the analytical steps taken and support our ability to make more accurate and meaningful predictions based on the model's results.

**ANOVA Type-1 F-Test:**

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| **C(weather)** | 3.0 | 12018.667925 | 4006.222642 | 380.016624 | 1.045577e-238 |
| **C(season)** | 3.0 | 31314.535101 | 10438.178367 | 990.130021 | 0.000000e+00 |
| **C(month)** | 11.0 | 5376.312308 | 488.755664 | 46.361696 | 9.829755e-101 |
| **C(year)** | 1.0 | 14383.215345 | 14383.215345 | 1364.342781 | 4.765986e-287 |
| **humidity** | 1.0 | 84434.055987 | 84434.055987 | 8009.126752 | 0.000000e+00 |
| **apparenttemperature** | 1.0 | 51522.699200 | 51522.699200 | 4887.267627 | 0.000000e+00 |
| **windspeed** | 1.0 | 1324.281547 | 1324.281547 | 125.616834 | 4.746914e-29 |
| **casual** | 1.0 | 97138.928273 | 97138.928273 | 9214.267631 | 0.000000e+00 |
| **hour** | 1.0 | 62211.912946 | 62211.912946 | 5901.210009 | 0.000000e+00 |
| **is_holiday** | 1.0 | 807.881790 | 807.881790 | 76.632913 | 2.254533e-18 |
| **is_weekday** | 1.0 | 6076.112647 | 6076.112647 | 576.359335 | 3.202004e-125 |
| **Residual** | 16534.0 | 174305.230142 | 10.542230 | NaN | NaN |

*Figure 16: Summary table of ANOVA Type 1 test results for the final linear regression model*

- **Categorical Variables (Weather, Season, Month, Year):** The F-test results for these categorical variables demonstrate highly significant effects on 'count.' The extremely low p-values (close to zero) indicate that each of these variables significantly contributes to explaining the variance in `'count`.'
- **Numeric Predictors (Humidity, Apparent Temperature, Windspeed, Casual, Hour, Is Holiday, Is Weekday):** Similarly, the F-test results for these numeric predictors indicate highly significant effects. The low p-values and large F-statistics highlight the substantial influence of these variables on `'count`.'

**ANOVA Type-2 F-Test:**

| | sum_sq | df | F | PR(>F) |
|---|---|---|---|---|
| **C(weather)** | 1835.986342 | 3.0 | 58.051802 | 2.529531e-37 |
| **C(season)** | 1998.046579 | 3.0 | 63.175963 | 1.315922e-40 |
| **C(month)** | 1061.869046 | 11.0 | 9.156843 | 1.456486e-16 |
| **C(year)** | 4149.612355 | 1.0 | 393.618084 | 1.370485e-86 |
| **humidity** | 2703.802366 | 1.0 | 256.473476 | 2.717080e-57 |
| **apparenttemperature** | 2857.174628 | 1.0 | 271.021846 | 2.055242e-60 |
| **windspeed** | 436.081990 | 1.0 | 41.365251 | 1.297470e-10 |
| **casual** | 74244.115779 | 1.0 | 7042.543757 | 0.000000e+00 |
| **hour** | 59305.897998 | 1.0 | 5625.555336 | 0.000000e+00 |
| **is_holiday** | 1582.348442 | 1.0 | 150.096180 | 2.327791e-34 |
| **is_weekday** | 6076.112647 | 1.0 | 576.359335 | 3.202004e-125 |
| **Residual** | 174305.230142 | 16534.0 | NaN | NaN |

*Figure 17: Summary table of ANOVA Type 2 test results for the final linear regression model*

In this analysis, we are evaluating the individual significance of each predictor variable, considering the presence of all other predictors in the model.

- **Categorical Variables (Weather, Season, Month, Year):** The F-test results for these categorical variables are highly significant. Each of them has a substantial impact on the variance in `'count`,' as evidenced by their notably low p-values and high F-statistics.
- **Numeric Predictors (Humidity, Apparent Temperature, Windspeed, Casual, Hour, Is Holiday, Is Weekday):** Similar to the categorical variables, the F-test results for these numeric predictors are also highly significant. They exert significant influence on the variation in `'count`,' as indicated by the low p-values and large F-statistics.

**Overall Observations:**

The combined F-test results underscore the robustness and validity of our regression model. All the variables, following the order of [weather -> season -> humidity -> apparent temperature -> windspeed -> casual -> hour], have been assessed and found to be statistically significant within the context of our regression model. This means that each variable holds its own significance while considering the presence of all other variables in the model. In other words, their individual importance in explaining the variation in the dependent variable `'count'` remains robust and valid, even when accounting for the combined influence of all the other variables. This collective significance of each variable underscores their valuable contributions to the model's predictive power and reinforces the integrity of our analysis.

# MODEL SELECTION

**Best Linear Regression Model:**

As an integral part of our model selection process, we embarked on an exhaustive exploration, fitting all possible models with different combinations of predictors. Our primary aim was to identify the model that exhibits the most favorable combination of statistical performance indicators.After extensive model exploration and evaluation, we have reached a conclusive decision. The full model, described by the equation:

**count ~ humidity + apparent temperature + windspeed + C(weather) + C(season) + casual + hour + is holiday + is weekday + C(month) + C(year)**

stands out as the best model. It excels in meeting our selection criteria, boasting high R-squared and adjusted R-squared values while simultaneously achieving low AIC and BIC values.

**Potential data problem:**

However, since this data is having a non-linear relationship between the predictors and the assumptions getting violated. We have tried fitting this dataset with other advanced models starting with models close to linear regression like lasso and ridge regression and then fitting other ensemble models like Decision Tree, LGBM, Random Forest, XGBoost which can fit on the non-linear dataset well. We have also fitted the dummy regressor model which generates random values for prediction just to compare other models with a random predictor.

**Comparing Best Linear Regression Model performance with other advanced models:**

| Model | R2 Score |
|---|---|
| Dummy Regressor | -0.0154 |
| Linear Regression | 0.681 |
| Lasso Regression | -0.0 |
| Ridge Regression | 0.6814 |
| Decision Tree | 0.9293 |
| LGBM | 0.9694 |
| Random Forest | 0.9508 |
| XGBoost | 0.9701 |

We observed that XGBoost performed best on this dataset with R2 Score 0.97 explaining 97% of variance in the target variable. So, **XGBoost** is our best selected model.

# CONCLUSION

In summary, the analysis demonstrates that several factors have a significant impact on the hourly bike rental count. These factors encompass humidity, apparent temperature, windspeed, categorical variables (weather, season, month, and year), as well as casual usage, hour, holiday, and weekday. Contrary to the initial assumption, the relationship between these factors and the hourly bike rental count may not adhere strictly to a linear pattern. The linear regression model offers moderate predictability (with an R-squared value of 0.68), whereas the XGBoost model performs exceptionally well with a high R-squared of 0.970, indicating substantial predictive capability. This underscores that the model's efficacy in explaining the variance in the hourly bike rental count is heavily influenced by the chosen modeling technique. Further exploration might uncover more intricate relationships between these factors and bike rental counts.