# Linear + Regression + Subjective + Questions and Answers

**Assignment-based Subjective Questions**

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

**Season**:

There is demand for bikes usually high in the month of 'Summer' and 'Fall'

**Weathersit:**

Demand for bike is usually LOW when there are Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds. (Category 3 of Weathersit)

**Month:**

- There is significant growth in demand of bike from January to June
- High demand of bike from the month of June till September
- There is decreasing trend in demand from October till December

**Year:**

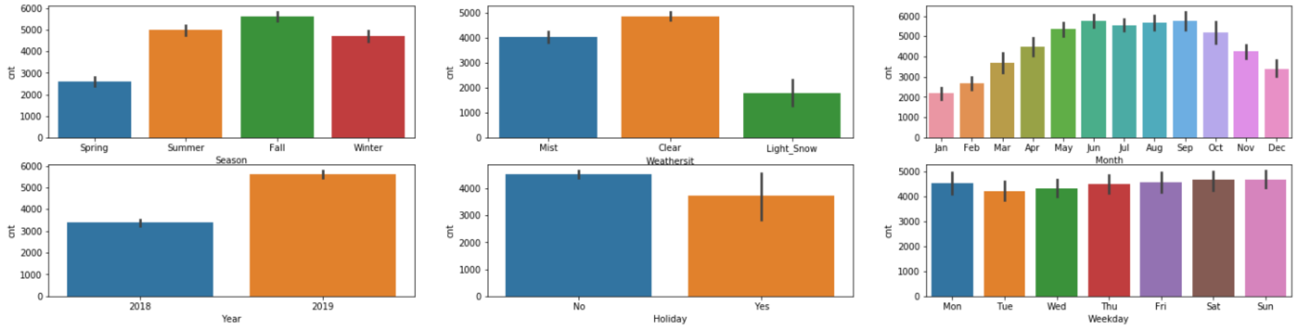
2019 has higher number of bookings compare to 2018

**Holiday:**

Demands of bike are usually high in working days

**Weekdays:**

There are no effect of weekdays on demand of bikes

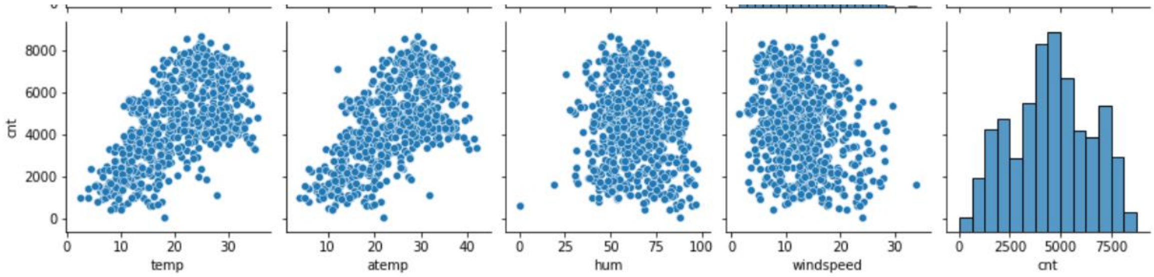


2. Why is it important to use **drop_first=True** during dummy variable creation?

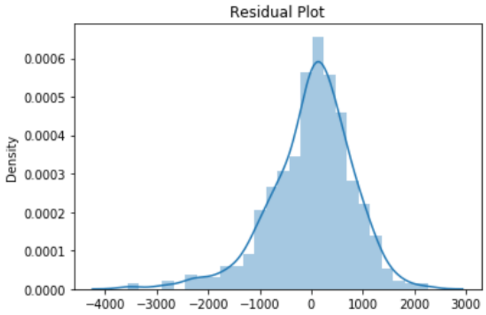
   If we keep all dummy variables then it will create problem of multicollinearity. This phenomenon is also called Dummy Variable Trap.
   Due to multicollinearity it would be very fifficult to interpretability of regression coefficients

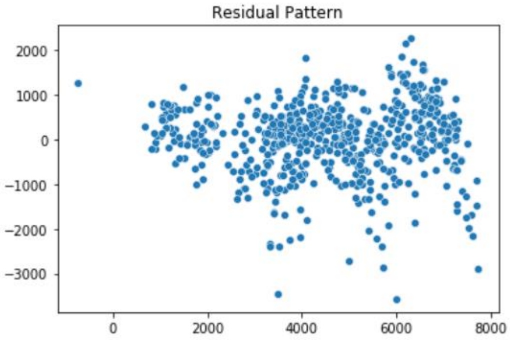
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

   Temperature variable (temp and atemp) is showing highest Correlation with Target Variable



4. How did you validate the assumptions of Linear Regression after building the model on the training set?
   After building the model we validate the model by checking
   a) Residual Plot: It should be Normally distributed
   b) Heteroscedasticity: Residual plot should not show any pattern

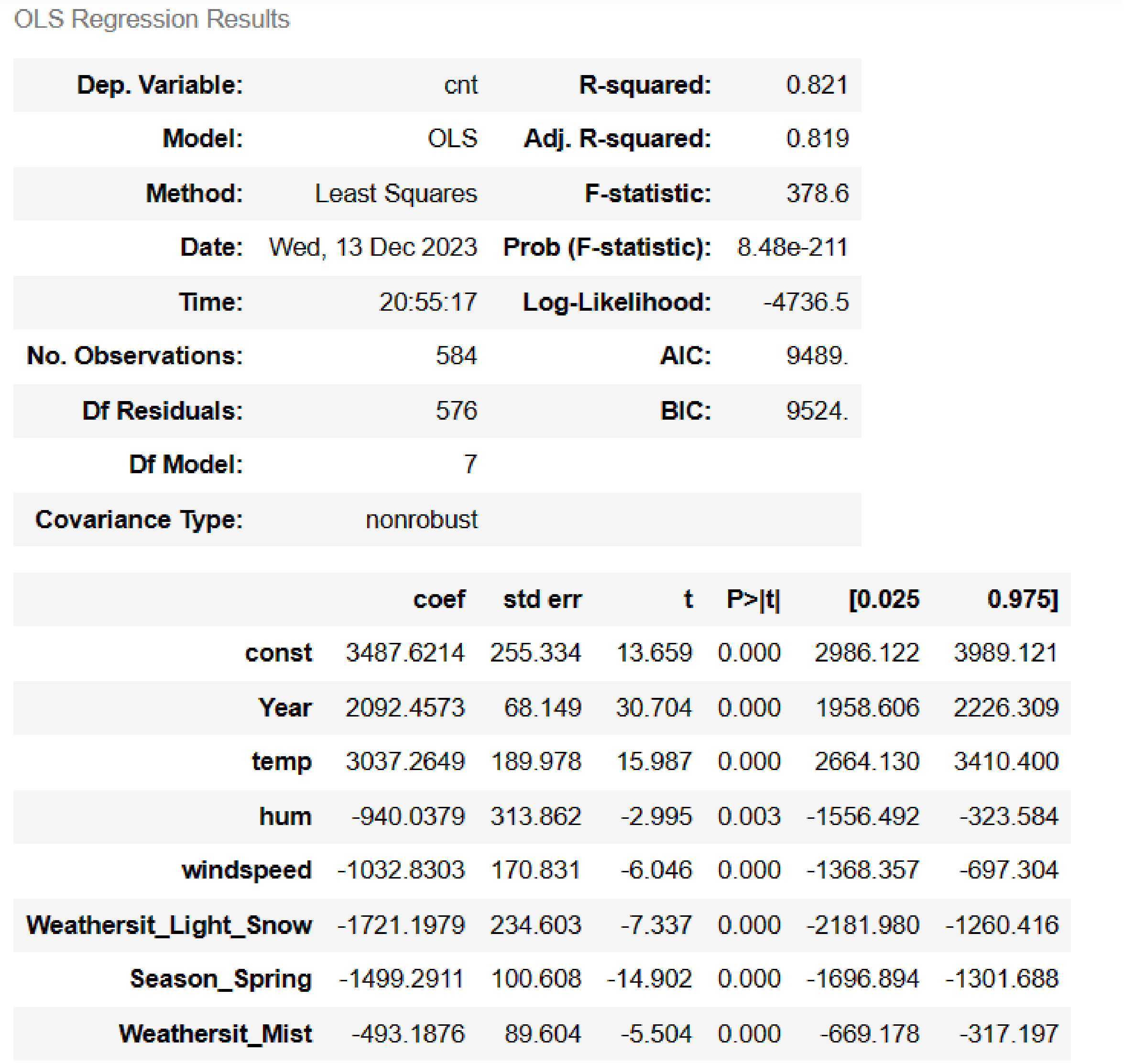




5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

   In the final model, top three features contributing in predicting Target variables are:

   1. Year
   2. Weather_Light_Snow
   3. Season_Spring



OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | cnt | **R-squared:** | 0.821 |
| **Model:** | OLS | **Adj. R-squared:** | 0.819 |
| **Method:** | Least Squares | **F-statistic:** | 378.6 |
| **Date:** | Wed, 13 Dec 2023 | **Prob (F-statistic):** | 8.48e-211 |
| **Time:** | 20:55:17 | **Log-Likelihood:** | -4736.5 |
| **No. Observations:** | 584 | **AIC:** | 9489. |
| **Df Residuals:** | 576 | **BIC:** | 9524. |
| **Df Model:** | 7 | | |
| **Covariance Type:** | nonrobust | | |

| | **coef** | **std err** | **t** | **P>\|t\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **const** | 3487.6214 | 255.334 | 13.659 | 0.000 | 2986.122 | 3989.121 |
| **Year** | 2092.4573 | 68.149 | 30.704 | 0.000 | 1958.606 | 2226.309 |
| **temp** | 3037.2649 | 189.978 | 15.987 | 0.000 | 2664.130 | 3410.400 |
| **hum** | -940.0379 | 313.862 | -2.995 | 0.003 | -1556.492 | -323.584 |
| **windspeed** | -1032.8303 | 170.831 | -6.046 | 0.000 | -1368.357 | -697.304 |
| **Weathersit_Light_Snow** | -1721.1979 | 234.603 | -7.337 | 0.000 | -2181.980 | -1260.416 |
| **Season_Spring** | -1499.2911 | 100.608 | -14.902 | 0.000 | -1696.894 | -1301.688 |
| **Weathersit_Mist** | -493.1876 | 89.604 | -5.504 | 0.000 | -669.178 | -317.197 |

# General Subjective Questions

1. **Explain the linear regression algorithm in detail**
   It is a type of supervised learning technique, In Linear regression model the Target variable is a continuous number.
   E.g.: Total Sales / Salary / Age / Area of House / Population etc

   In regression model primary goal is to achieve the ***best fit line***. A best fit line is that line where the *overall error is minimum.

   ***Error= 1/n * (Actual value - Predicted value)^2**

   A regression line is called Best fit line if the line touches all(~near to) actual points.
   This best fit line represents the linear relationship between dependent variable(Y) and set of Independent variables(x1, x2,…xn) .

   Equation of multiple linear regression is :

   Y= Bo + B1X1 + B2X2 + …+ BnXn

   **Y**=Prediction point
   **B0**= Beta Coefficient for Constant (Intercept)
   **B1**= Beta Coefficient of first Independent Variable (x1)
   **x1**= First Independent Variable (x1)

2. **Explain the Anscombe’s quartet in detail**
   Anscombe’s Quartets is tells about corelation, according to Anscombe’s Quartet analysis just looking at Corelation statistics to find association between two feature is not enough. We also need to inspect visually the relationship and this will be a supplement for Correlation coefficient.

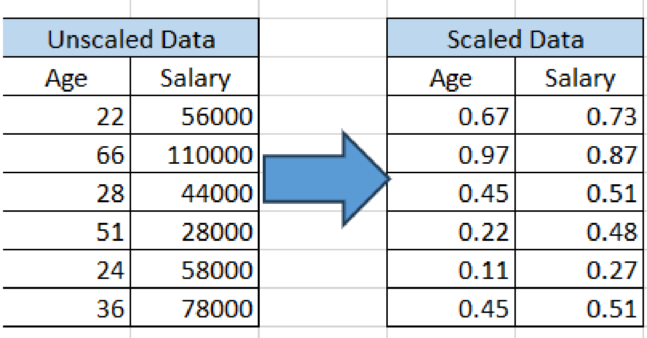
   Anscombe’s Quartet could be well-defined as a cluster of four data sets which are almost same in simple descriptive statistics, but there are some individualities in the dataset that fools the regression mode

3. **What is Pearson’s R?**
   It is a type of Correlation coefficient used to find the degree of strength that how two or more variables are associated (linear relationship) with each other.
   Correlation is always checked between Dependent Variable and Independent variables.

   The value of correlation ranges from -1 to +1.

| Corr Value | Description |
|---|---|
| -1 | Strong **Negative** relationship between X and Y |
| 1 | Strong **Positive** relationship between X and Y |
| 0 | **No** linear relationship between X and Y |

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
Scaling is a method to bring data from different scale to same scale.



| Unscaled Data |  |  | Scaled Data |  |
|---|---|---|---|---|
| Age | Salary |  | Age | Salary |
| 22 | 56000 |  | 0.67 | 0.73 |
| 66 | 110000 |  | 0.97 | 0.87 |
| 28 | 44000 |  | 0.45 | 0.51 |
| 51 | 28000 |  | 0.22 | 0.48 |
| 24 | 58000 |  | 0.11 | 0.27 |
| 36 | 78000 |  | 0.45 | 0.51 |

Majorly there are two types of Scaling:

1) **Standardization**:
    - Z - Score Normalization

2) **Normalization:**
    - Min Max Scaler
    - Mean Normalization
    - MaxAbs Scaling
    - Robust Scaling

One of the major characteristics of scaling is that it does not change the distribution of data, it just changed the scale of data.

Scaling is very important preprocessing if we are working on below algorithms:

- K Nearest Nabour
- Principal Component Analysis (PCA)
- Artificial Neural Network (ANN)

Scaling not required on below algorithms:

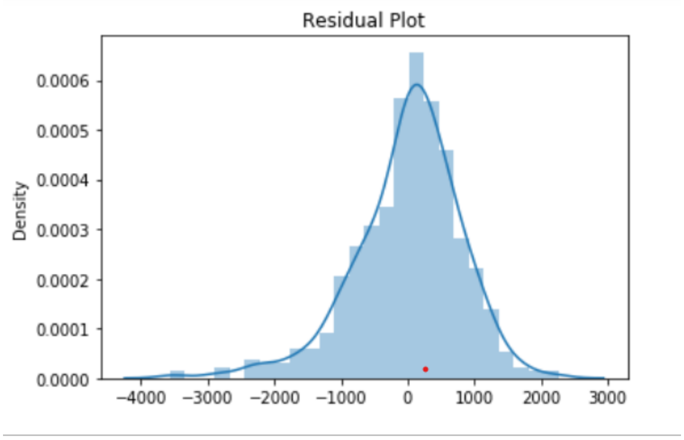
- Tree based algorithm (Decision Tree / Random Forest)
- Boosting algorithm (Gradient and Extreme Gradient boosting)

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   In case of **perfect multicollinearity** value of VIF turns to be infinite

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression**
   QQ (Quantile-Quantile) plot are used to graphically determine if two different dataset have same distribution or not.

   Also, if distribution of the data is **Normally Distributed** or not.

   In Linear Regression QQ plot is used to check the assumptions post building the dataset, If the **residuals**(errors) are Normally distributed or not.


Residual Plot
Density