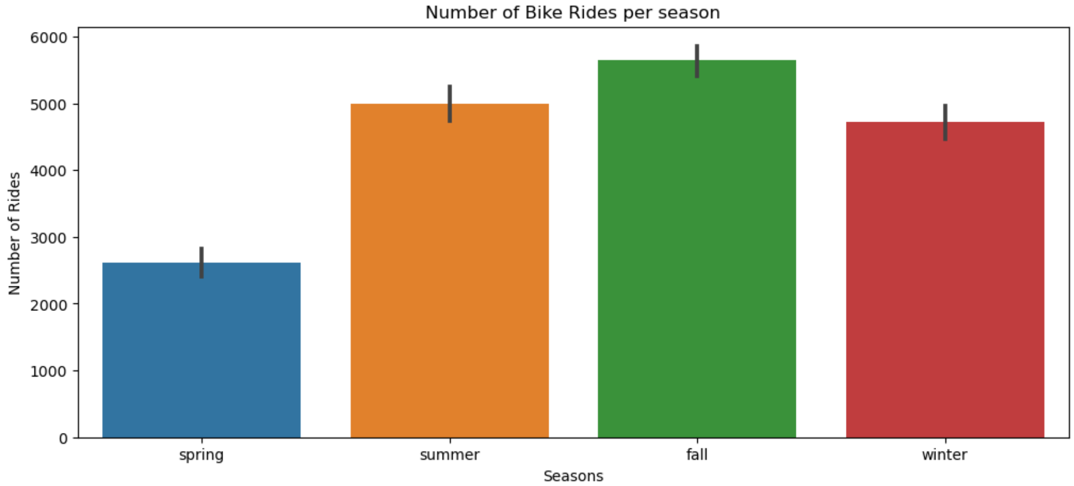
# Assignment-based Subjective Questions

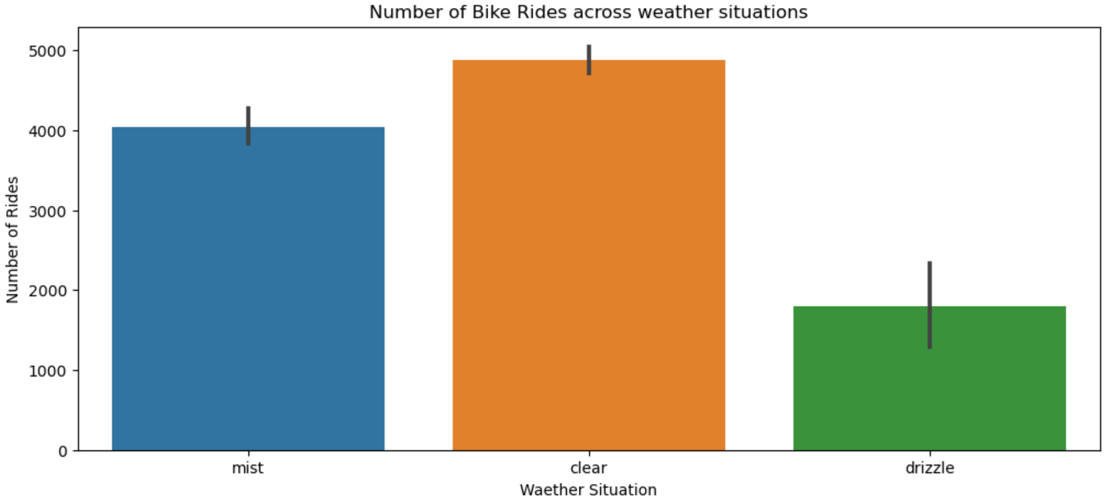
**1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?  (3 marks)**

**Answer**:
Categorical variables in the bike sharing dataset are as below:

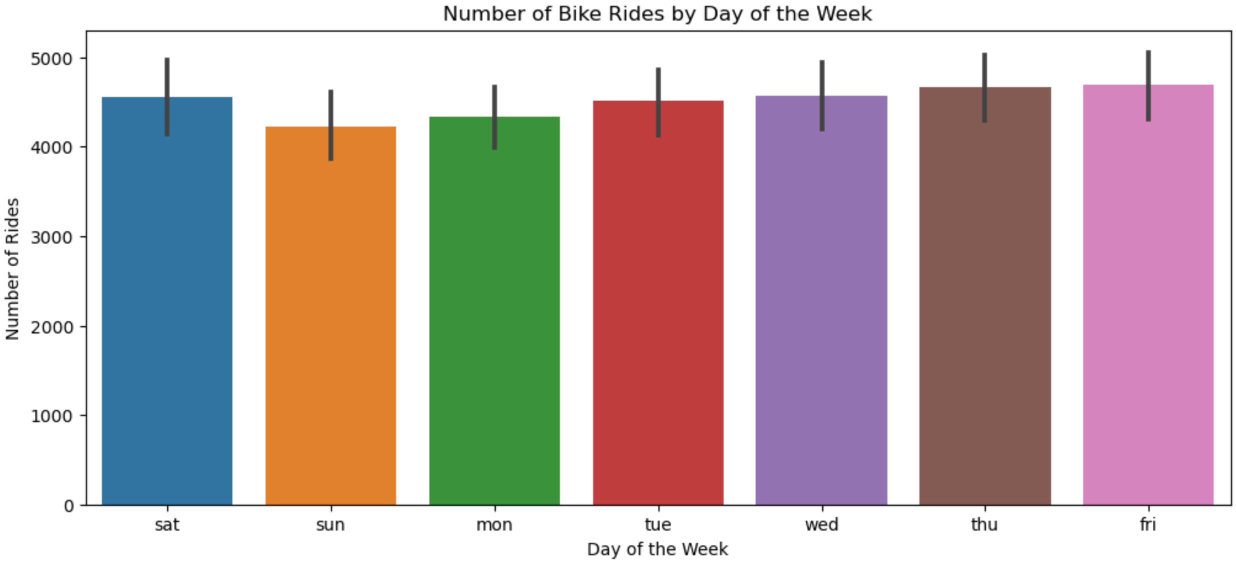
- Season:
  On analysis of the target variable cnt- No of Bike rides, per Season , we can see that the highest number of rides were in Fall season followed by Summer and then winter with the least number of rides.



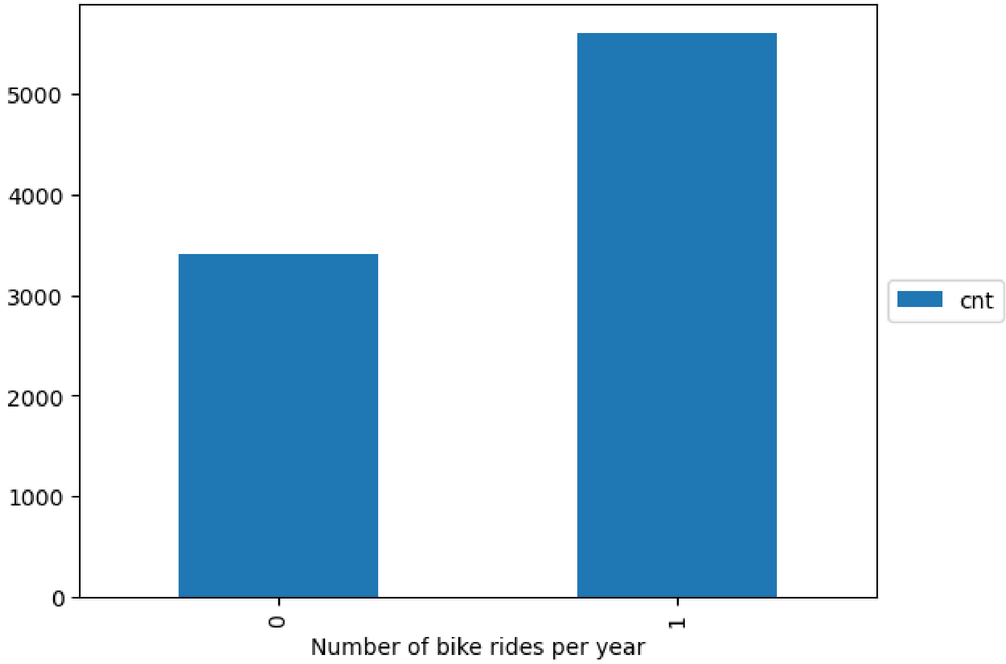
- Weather Situation:
  On analysis of Weather Situation and target variable we can see that most rides have been  taken on Clear weather, followed by Misty weather and then the least rides on days with light rain. There is no data on heavy rain which might be because no bike rides were booked on rainy days



- Day of the week
  On analysis of Day of the week and bike rides with the number of bike rides we can see very minimal increase in demand on Fridays and Saturdays.



- Year:
  We can see that the demand of Bike rides have increased in 2018 compared to 2017



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

**Answer**:

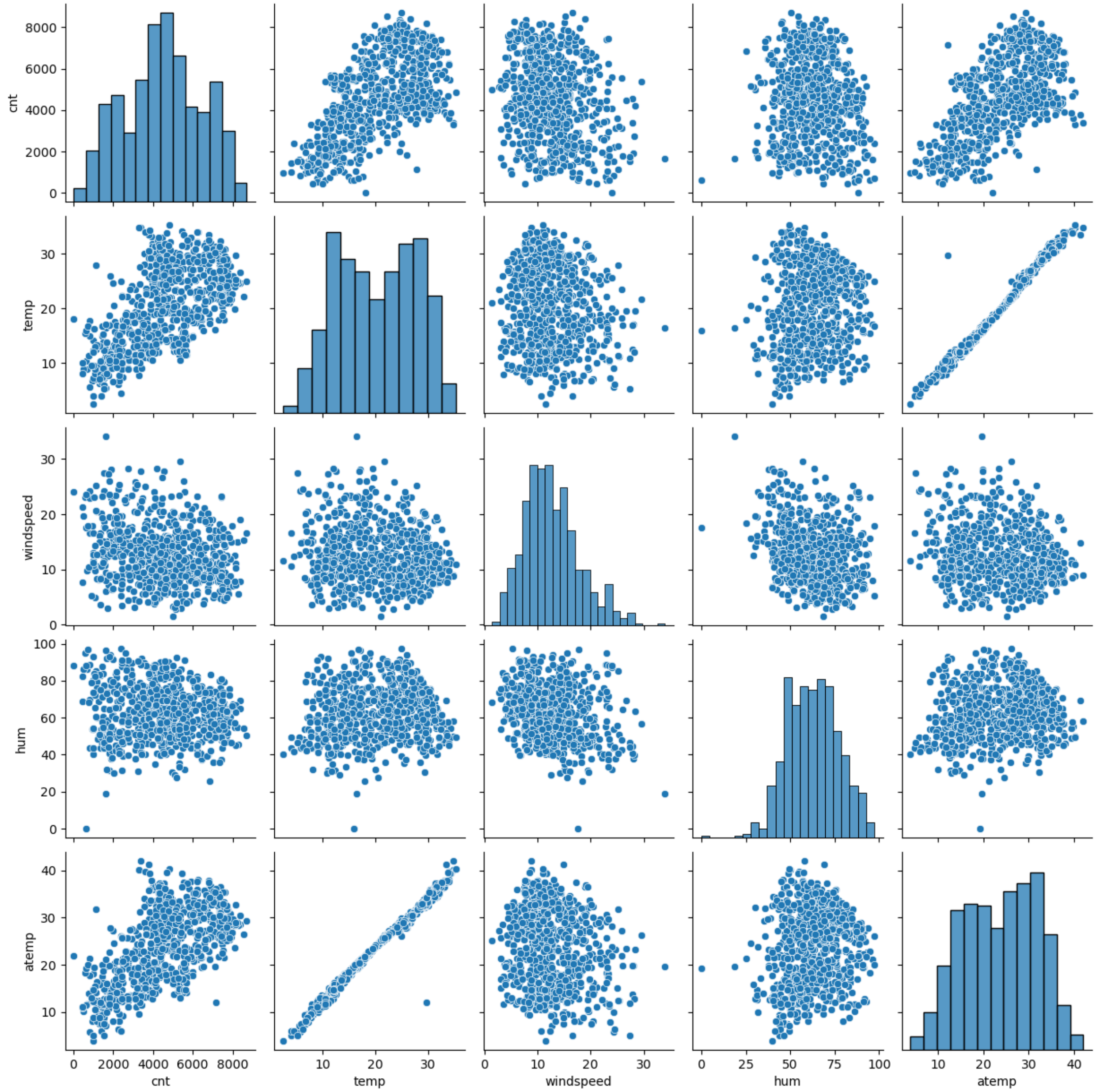
The **drop_first=True** argument of the create dummies function drops the 1st dummy variable it creates. This is done as the 1st variable can be anyway referenced using the other dummies created. For example, let's take the **'weathersit'** variable in the data set. It is a categorical variable with categories **'Clear'**, **'Mist'**, **'Drizzle'** and **'Rain'**. **drop_first=True** will remove the dummy for 'Clear' because it can be represented by the other dummies created for the remaining categories.We can see in the below table, that **'Clear'** can be represented by 0 in the **other 3 categories**. Therefore, we can drop the 1st Dummy and make interpretation easier as well as avoid overfitting.

| Clear | Mist | Drizzle | Rain |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 1 |

**3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

<u>**Answer**</u>:

The Highest correlation between the numerical variables **'cnt'**, **'temp'**, **'windspeed'**, **'hum'**, **'atemp'** is observed between **‘temp’** and **‘atemp’** which is natural as both depict the temperature of the day. While ‘**temp’ depicts the actual temperature** of the day, **‘atemp’ depicts the feels-like temperature** which changes according to wind speed, humidity etc. However, the next useful correlation is seen **between ‘temp’ and ‘cnt’** where we can see a positive correlation between the two, thereby indicating the **Bike-rides tend to increase with increase in Temperature**.

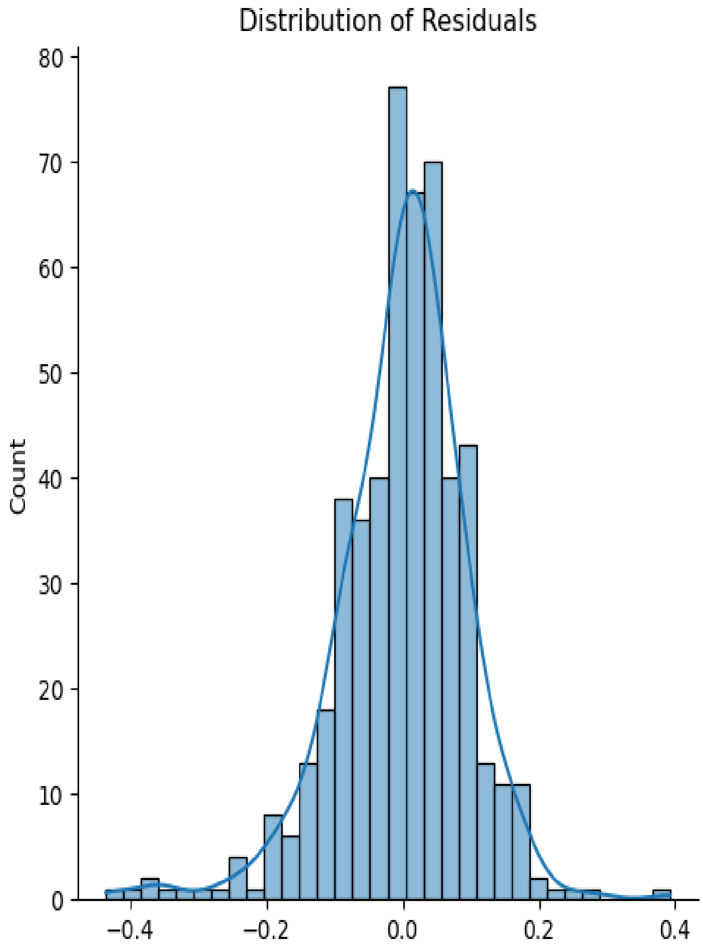


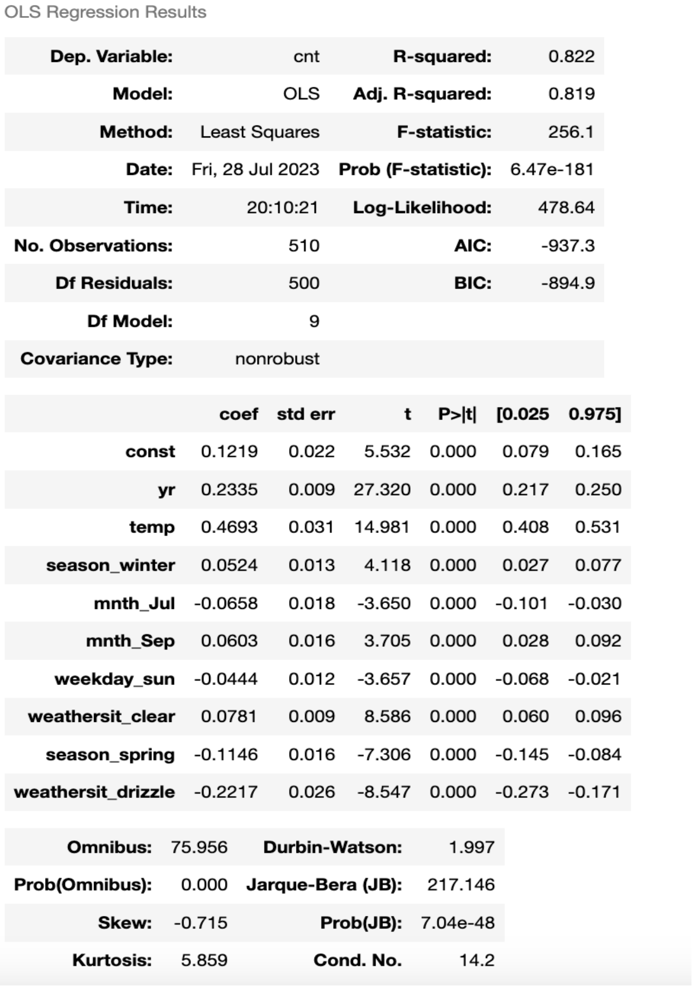
**4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

**Answer:**

After the model is built we check the **R2-Score** and **Adjusted-R2-Score** to see the accuracy of our model. We also check the **P-Value** of the features to determine the Significance level of the features used to build the model. The **Variance Inflation Factor (VIF)** is checked to verify that the features are not collinear and are within the acceptable range. The residuals calculated should be normally distributed. Furthermore, the model is used to predict the Test Dataset and the difference between **y_pred and y_test** is checked to verify whether the **R2 Score** obtained in the Train Dataset is **within 2%-5%.**

The F-Statistic and P-value of F-Statistic also gives us the significance level of the model.





OLS Regression Results

| | | | |
|---|---|---|---|
| **Dep. Variable:** | cnt | **R-squared:** | 0.822 |
| **Model:** | OLS | **Adj. R-squared:** | 0.819 |
| **Method:** | Least Squares | **F-statistic:** | 256.1 |
| **Date:** | Fri, 28 Jul 2023 | **Prob (F-statistic):** | 6.47e-181 |
| **Time:** | 20:10:21 | **Log-Likelihood:** | 478.64 |
| **No. Observations:** | 510 | **AIC:** | -937.3 |
| **Df Residuals:** | 500 | **BIC:** | -894.9 |
| **Df Model:** | 9 | | |
| **Covariance Type:** | nonrobust | | |

| | **coef** | **std err** | **t** | **P>\|t\|** | **[0.025** | **0.975]** |
|---|---|---|---|---|---|---|
| **const** | 0.1219 | 0.022 | 5.532 | 0.000 | 0.079 | 0.165 |
| **yr** | 0.2335 | 0.009 | 27.320 | 0.000 | 0.217 | 0.250 |
| **temp** | 0.4693 | 0.031 | 14.981 | 0.000 | 0.408 | 0.531 |
| **season_winter** | 0.0524 | 0.013 | 4.118 | 0.000 | 0.027 | 0.077 |
| **mnth_Jul** | -0.0658 | 0.018 | -3.650 | 0.000 | -0.101 | -0.030 |
| **mnth_Sep** | 0.0603 | 0.016 | 3.705 | 0.000 | 0.028 | 0.092 |
| **weekday_sun** | -0.0444 | 0.012 | -3.657 | 0.000 | -0.068 | -0.021 |
| **weathersit_clear** | 0.0781 | 0.009 | 8.586 | 0.000 | 0.060 | 0.096 |
| **season_spring** | -0.1146 | 0.016 | -7.306 | 0.000 | -0.145 | -0.084 |
| **weathersit_drizzle** | -0.2217 | 0.026 | -8.547 | 0.000 | -0.273 | -0.171 |

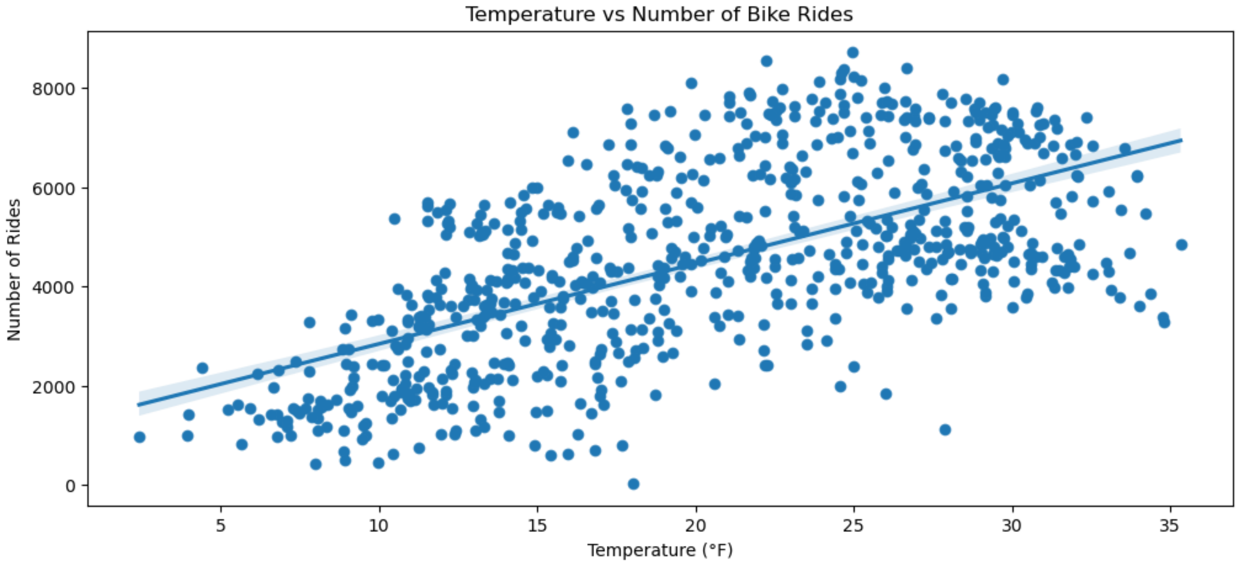
| | | | |
|---|---|---|---|
| **Omnibus:** | 75.956 | **Durbin-Watson:** | 1.997 |
| **Prob(Omnibus):** | 0.000 | **Jarque-Bera (JB):** | 217.146 |
| **Skew:** | -0.715 | **Prob(JB):** | 7.04e-48 |
| **Kurtosis:** | 5.859 | **Cond. No.** | 14.2 |

**5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

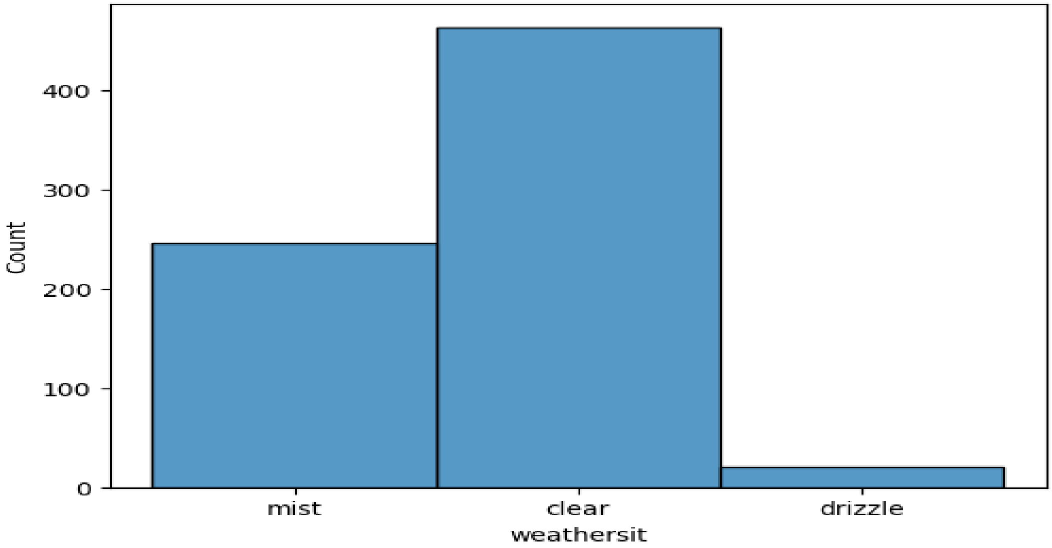
**Answer:**

The top 3 features contributing to the demand of bikes are as follows:

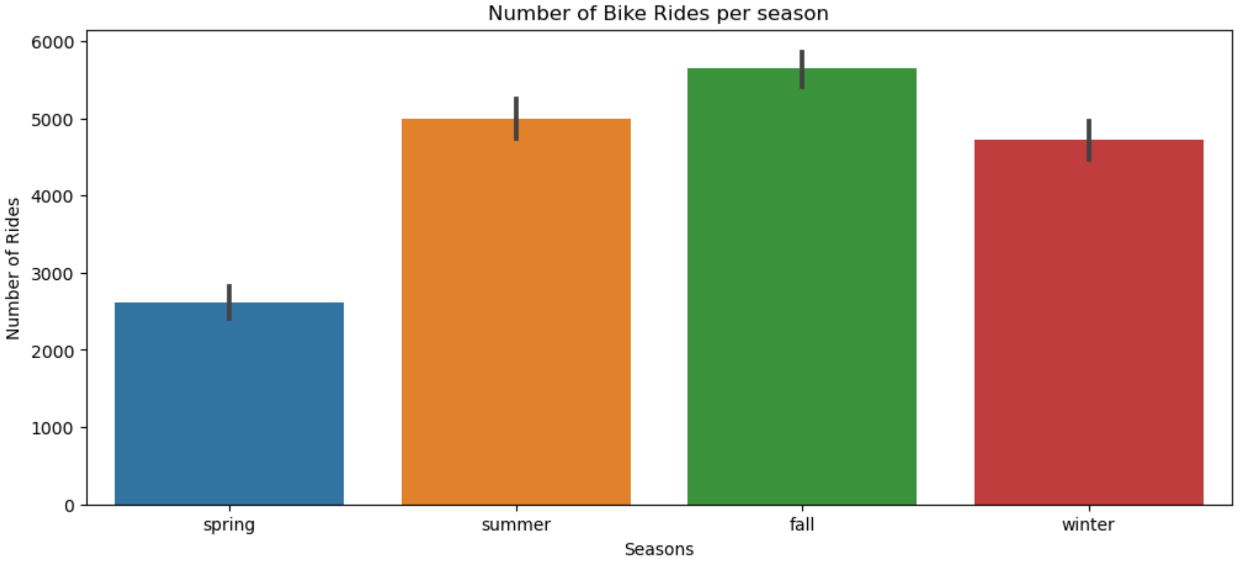
- Temperature:
  We can see a clear positive relation between the temperature and the demand for bikes. We can see as the temperature goes up, so does the demand for bikes.



- Weather Conditions:
  We also see a positive correlation between weather conditions, i.e. the bike demand is highest during Clear weather, followed by Mist and Drizzle. We do not have any data on Heavy Rain weather conditions, which may be because there was no demand for bikes on Rainy days, or the location did not see any heavy rain days.



- Seasons:
  We do see a correlation between the seasons and the Bike ride demands. As per the data available, the highest demand is during Fall, followed by Summers and then Winter. However we notice a decrease in demand during Spring.

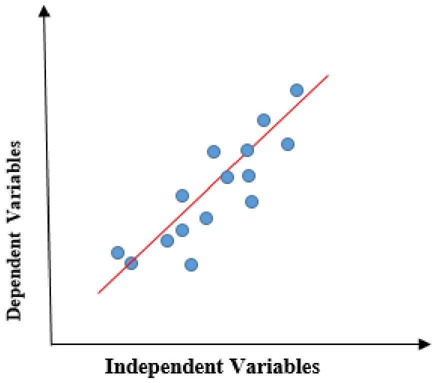


## General Subjective Questions

**1. Explain the linear regression algorithm in detail. (4 marks)**

**Answer:**

Linear Regression Algorithm in Machine Learning is a Supervised Learning Algorithm where we try to predict the relation between a dependent continuous variable (Numerical) and a set of independent features, that are Categorical and/or Continuous.



Using the Linear Regression algorithm we try to predict a best fit line that passes through the data points such that the difference between the individual Target data points (y) and the respective Predicted Data point (y_predicted) is minimum.

The underlying calculations can be explained as the best fit line that minimizes the sum of squared residuals (also known as the sum of squared errors) between the predicted values and the actual target values.

This line is represented by the equation:

$$\mathbf{y=\beta0 + \beta1X1 + \beta2X2 +..... + \beta nXn}$$

Where:

| | |
|---|---|
| Y | : Target Variable |
| B0 | : is the y intercept- value of y when X=0 |
| β0, β1, ..., βn | : are the coefficients of the respective independent variables |
| X1, X2, ..Xn | : are the values of the independent variables (features) |

The goal of linear regression is to find the values of the coefficients (β0, β1, ..., βn) that minimize the sum of squared residuals between the predicted values and the actual target values in the training data. This is achieved using the Ordinary Least Squares (OLS) method. Using the OLS method we derive the coefficients such that we have a best fit line.

However, before we can derive the coefficients and build the model, we need to pre-process the data such that the features used are independent and even if they are collinear, they should be within the acceptable range. We do that by creating dummy variables for the Categorical variables and then calculating the Variance Inflation Factor (VIF). There are libraries that are utilized for calculating these eg: variance_inflation_factor from statsmodels.stats.outliers_influence.

It is important to do this for the fact that the coefficients are calculated by calculating the target variable by changing the coefficients of each feature by keeping all other features constant. This cannot happen if the features are collinear, as increasing the coefficient of 1 might increase/decrease the coefficients of one or more features.

It is also noted that there are few key assumptions made about Linear Regression

- Linearity: The relationship between the independent and dependent variables is linear. This means that the best-fitting line should be a straight line.

- Homoscedasticity: The variance of the residuals is constant across all values of the independent variable. This means that the residuals should be evenly spread around the line.

- Normality: The residuals are normally distributed. This means that the residuals should follow a normal distribution.

- Independence: The residuals are independent of each other. This means that the residuals should not be correlated with each other.

- Zero mean: The mean of the residuals is zero. This means that the residuals should be centered around zero.

There is also the matter of a Cost Function, which helps us find the best possible values for the coefficients and the constant (Y-intercept) for a best fit line.

Here the cost function is Mean Squared Error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

Gradient Descent is another method for optimizing the model where it selects a coefficient and then keeps reducing it iteratively based on the Learning Rate to reach the minima. Here the learning rate decides how quickly the function converges to minima. A gradual decrease is considered to avoid the oscillation caused when Learning rate is high.

In conclusion, linear regression is a powerful and versatile statistical method that can be used to predict a wide variety of continuous variables. However, it is important to be aware of its limitations before using it.

| Advantages of Linear Regression | Disadvantages of Linear Regression |
|---|---|
| It is a simple and easy-to-understand algorithm. | It assumes that the relationship between the independent and dependent variables is linear. |
| It is very versatile and can be used to predict a wide variety of variables. | It is sensitive to outliers. |
| It is relatively efficient and can be used to analyze large datasets. | It can be difficult to interpret the results of the model. |

**2. Explain the Anscombe's quartet in detail. (3 marks)**

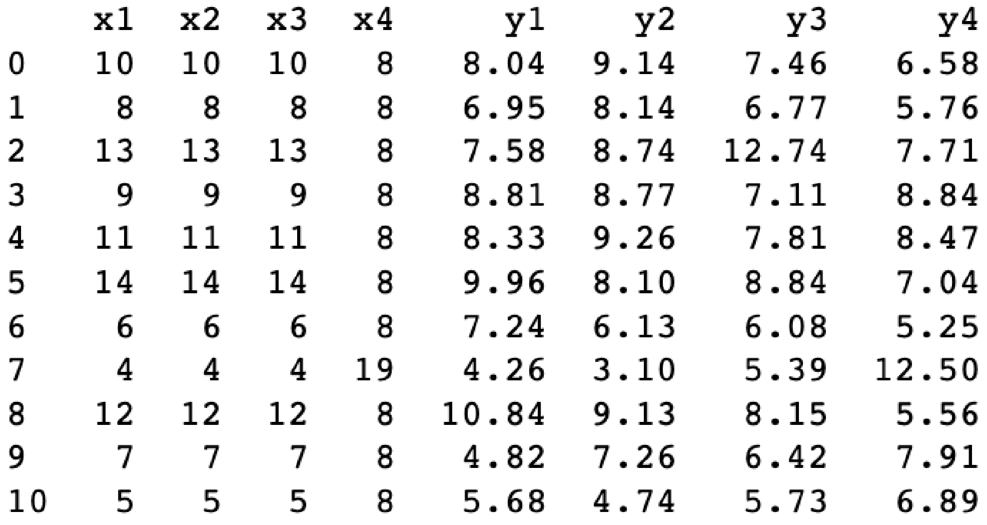
**Answer:**

Anscombe's quartet is a set of four small datasets that have nearly identical statistical properties but exhibit drastically different relationships when plotted. It was introduced by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not solely relying on summary statistics.

The four datasets of Anscombe's quartet share the following statistical properties:

Mean of x and y: Each dataset has nearly the same mean for both the x and y variables.
Variance of x and y: The variance of x and y is almost identical across all datasets.
Correlation coefficient: Each dataset has a correlation coefficient close to 0.816, indicating a strong linear relationship between x and y.
However, when we plot these datasets, we observe significant differences:

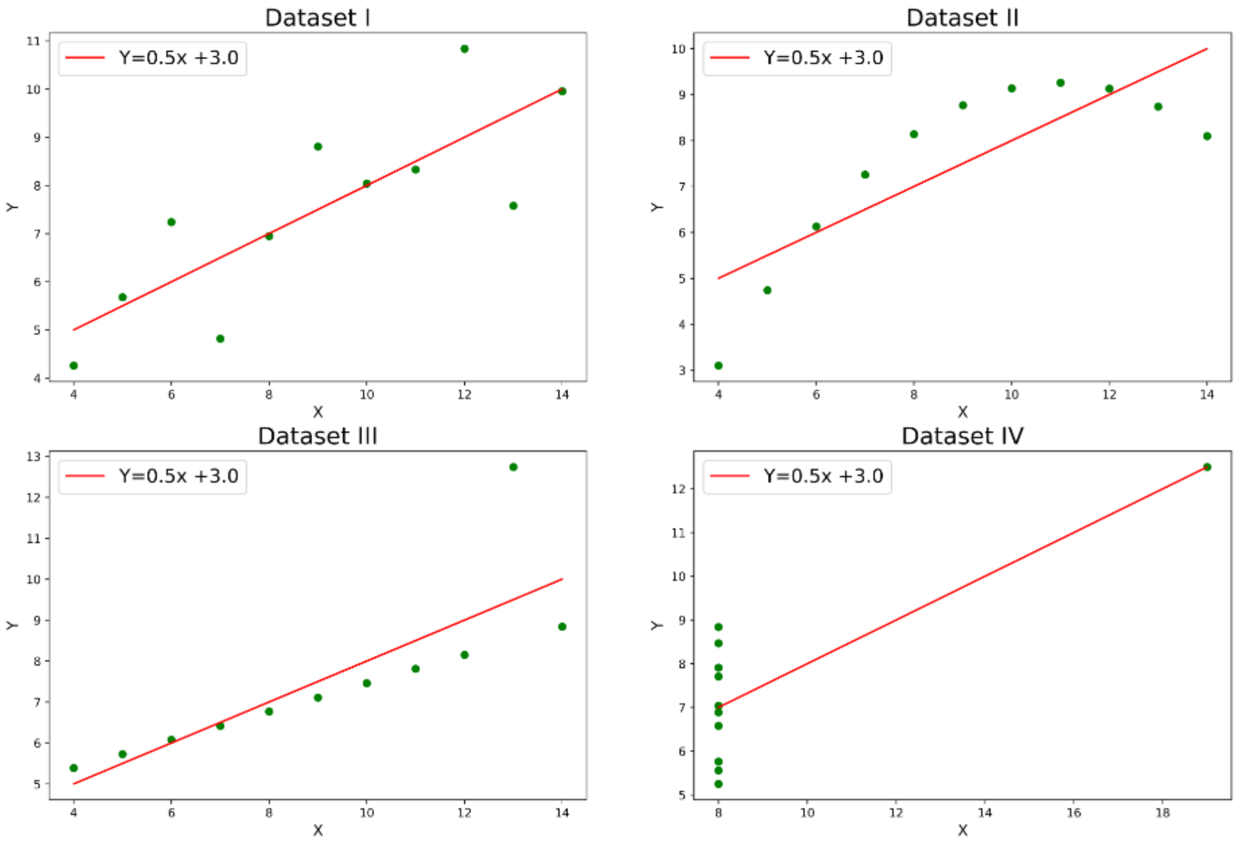
Dataset 1: A simple linear relationship
Dataset 2: A non-linear relationship following a quadratic curve.
Dataset 3: An outlier significantly influences the regression line.
Dataset 4: The linear relationship remains, but one outlier has a substantial effect on the regression line.



```
    x1  x2  x3  x4     y1    y2     y3     y4
0   10  10  10   8   8.04  9.14   7.46   6.58
1    8   8   8   8   6.95  8.14   6.77   5.76
2   13  13  13   8   7.58  8.74  12.74   7.71
3    9   9   9   8   8.81  8.77   7.11   8.84
4   11  11  11   8   8.33  9.26   7.81   8.47
5   14  14  14   8   9.96  8.10   8.84   7.04
6    6   6   6   8   7.24  6.13   6.08   5.25
7    4   4   4  19   4.26  3.10   5.39  12.50
8   12  12  12   8  10.84  9.13   8.15   5.56
9    7   7   7   8   4.82  7.26   6.42   7.91
10   5   5   5   8   5.68  4.74   5.73   6.89
```

| | I | II | III | IV |
|---|---|---|---|---|
| Mean_X | 9.000000 | 9.000000 | 9.000000 | 9.000000 |
| Variance_X | 11.000000 | 11.000000 | 11.000000 | 11.000000 |
| Mean_y | 7.500909 | 7.500909 | 7.500909 | 7.500909 |

| | | | | |
|---|---|---|---|---|
| Variance_y | 4.127269 | 4.127269 | 4.127269 | 4.127269 |
| Correlation | 0.816421 | 0.816421 | 0.816421 | 0.816421 |
| Linear Regression Slope | 0.500091 | 0.500000 | 0.499727 | 0.499909 |
| Linear Regression Intercept | 3.000091 | 3.000909 | 3.002455 | 3.001727 |



Here are some of the key takeaways from Anscombe's quartet:

Linear regression statistics can be misleading. It is important to visualize the data.

**3. What is Pearson’s R? (3 marks)**

**Answer:**

Pearson correlation coefficient or Pearson’s correlation coefficient or Pearson’s r is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other.

Pearson’s R is used to calculate the effect of change in one variable when the other variable changes. It seeks to draw a line through the data of two variables to show their relationship. The relationship of the variables is measured with the help of Pearson correlation coefficient calculator.

The linear relationship can be positive or negative. It ranges between -1 and 1.
Anything equal to or greater than 0.80 is considered to be a strong correlation between the concerned variables. A value of 0 indicates no correlation.

The formula for calculating Pearson’s R is

$r = \sum(x - \bar{x})(y - \bar{y}) / \sum(x - \bar{x})^2$
where:
$r$ is the Pearson correlation coefficient
$x$ and $y$ are the values of the two variables
$\bar{x}$ and $\bar{y}$ are the mean values of the two variables

Advantages of Pearson's r:
It is a simple and easy-to-calculate measure of correlation.
It is a widely used measure of correlation, so it is easy to compare results with other studies.

Disadvantages of Pearson's r:
It only measures linear correlation.
It is sensitive to outliers.

**4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

**Answer:**

Scaling is a data preprocessing step before building a Machine learning model, which is used to transform all the coefficients to a single data range. Usually, the datasets have data of different ranges and units, which helps in speeding up algorithm calculations and also reading the data.

Scaling is of two types:

1. Standardized Scaling:
   In this process the data is transformed so that the mean is 0 and standard Deviation is 1.
   It is calculated using the formula (X - mean(X))/Std X
   Here the values are centered around 0 and spread on both directions
   It is helpful where the data has lot of outliers

2. Normalized Scaling:
   This process transforms the entire data to the range 0-1
   It is calculated using the formula (X - min(X))/(max(X)-Min(X))
   Here the values range between 0 and 1 with center at 0.5
   Its disadvantage is such that it may lead to data loss when there are lot of outliers

**5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

**Answer:**

Firstly, VIF(Variance Inflation Factor) is an index using which we can calculate the estimated variance between the coefficients of the Predictors to identify the magnitude of collinearity between 2 or more supposed independent variables (Predictors) of a dataset. One of the basic requirements for a linear regression model to be accurate is that the Predictors are Not-Collinear in the ideal case, or the collinearity is very less i.e within a select VIF range.

VIF is calculated using the below formula:

VIF = 1 / (1 - R2)

where:

- VIF is the Variance Inflation Factor
- R2 the coefficient of determination for the regression model

The result of VIF formula varies from 1 to infinity, where,

| | |
|---|---|
| 1 | : No Collinearity |
| 2-5 | : Minimal collinearity (Considered the acceptable range) |
| 5 & above | : High Collinearity between the Predictors |
| Infinity | : Perfect Correlation between the Predictors |

Hence, a VIF of Infinity would mean that one or more Predictors/Features/Independent Variables have a perfect Correlation amongst each other and are perfectly Collinear.

**6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

**Answer:**

Quantile-Quantile Plot (QQ Plot) is essentially a Scatter Plot of the Target Variable across the Predictors / Features/ Independent variables. In linear regression, a Q-Q plot can be used to check the assumption of normality. The normality assumption states that the residuals of the linear regression model are normally distributed.

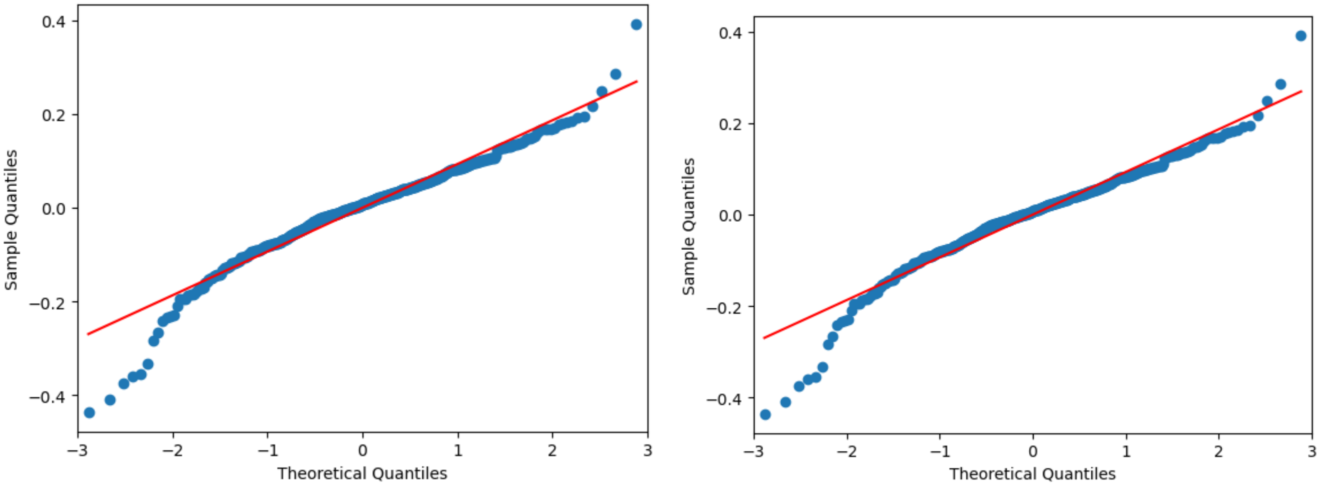
Q-Q plots for a linear regression model are created by sorting the Residuals in the ascending order. Then, the quantiles of the residuals are plotted against the quantiles of the standard normal distribution.

If the residuals are normally distributed, then the points on the Q-Q plot will fall along a straight line.

QQ plot can be plotted using the qqplot function from the StatsModel Library as below:

```
from statsmodels.api import qqplot
sm.qqplot(res, line='r')
plt.show()
```

QQ Plot for BoomBike Train Dataset (Left) and Test Data Set (Right)


Sample Quantiles
Theoretical Quantiles
Sample Quantiles
Theoretical Quantiles