**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?

Ans : From the analysis of categorical variable over dependent variable is that

1. In the first, second and third season we can gradually see high in the graph where it increases with dependent variable .
2. With respect to year first year we have gradually less bike rented when compared to next year
3. In months we can see up ligfting till 8 month after that there is a small deviation towards down
4. Non Holiday there is slight higher median when compared to holiday
5. WeekDays and Working days were not much effect on the dependent variable where they approximately a constant value

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

Ans: We will use drop_first while creating dummy variables from categorical variables in the context of linear regression or other modeling techniques, using drop_first=True is a common practice. The primary reason for using drop_first=True is to avoid multicollinearity, a situation where one predictor variable in a multiple regression model can be predicted from the others with a high degree of accuracy.

We also use it to decrease the number of variables created while creating dummy variables. In the given data set we will remove the season column where we have derived a dummy column and there we will be dropping the first column.

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

Ans: atemp and temp has the highest correlation with the target variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans :

1. Error terms should be normally distributed.

Residual Analysis:

Check if the error terms are also normally distributed. Plot the histogram of the error terms and see what it looks like.

Scatterplots:

The simple way to determine if this assumption is met or not is by creating a scatter plot ytest vs pred y test.

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

Ans: Based on my final model, the top three features contributing significantly towards explaining the demand of the shared bikes are

1. atemp
2. Year
3. Light Snow/Rain

**General Subjective Questions**

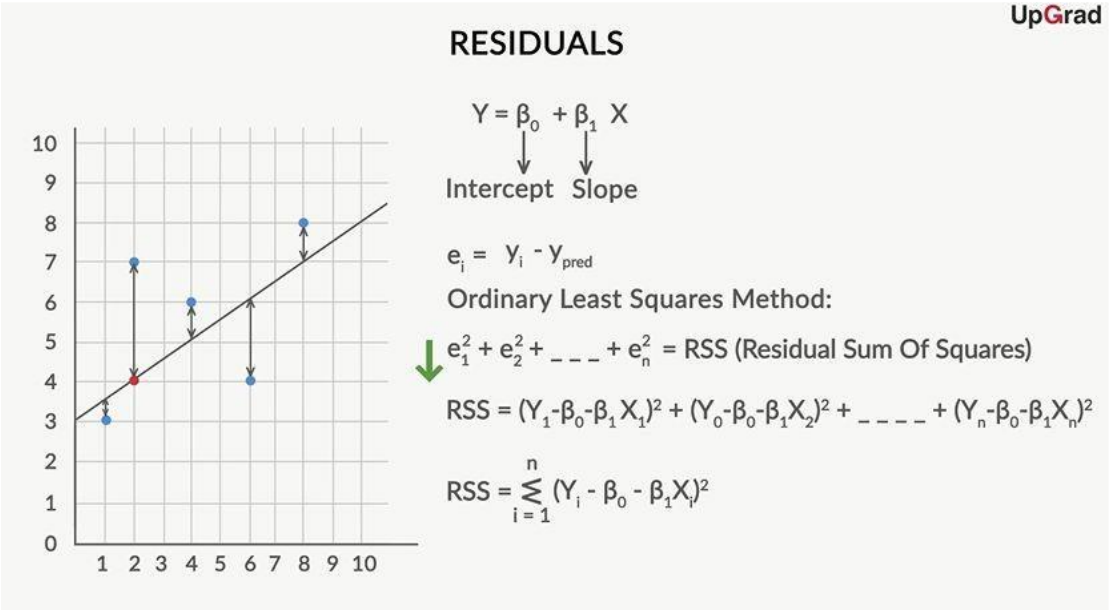
1. Explain the linear regression algorithm in detail.

Ans: Linear regression is a used for **predicting numerical values based on input features**. It assumes a linear relationship between the independent variables (features) and the dependent variable (target).

1. linear equation y = mx + c

where: y: dependent variable x: independent variable m: slope of the line c: y-intercept

**2.** The best-fit line is found by minimizing the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot.




**3. Algorithm Steps:**

- **Data Preparation**:
    1. Collect data and differentiate the target and featured variables .
    2. Do EDA's and Scaling the data and adding the dummy variables for categorical variable and split data to train and test
- **Model Building**:
    1. Build the model with linearRegression
    2. Use optimization algorithm (e.g., gradient descent) to minimize the RSS or go for libraries like sklearn and build the model and get the model to more significant and having lease VIF's
- **Model Evaluation**:

1. Evaluate the performance of the model using metrics like Rsquare
2. Plot a scatter plot and evaluate based on predicted and test data

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a set of four data sets designed to highlight the limitations of relying solely on summary statistics to analyze data. Each data set has the same mean, variance, correlation coefficient, and linear regression line, yet they appear very different when plotted visually.

**Purpose:**

- Demonstrates the importance of data visualization in statistical analysis.
- Shows that summary statistics can be misleading if the underlying data distribution is not considered.
- Highlights the potential for outliers and other influential observations to distort statistical analysis.

**Each data set consists of 11 points:**

- **Data set 1:** Linear relationship with no outliers.
- **Data set 2:** Non-linear relationship with an outlier.
- **Data set 3:** Linear relationship with a point far from the regression line.
- **Data set 4:** Non-linear relationship with several outliers

3. What is Pearson's R?

Ans: Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It ranges from -1 to 1, where:

✧ 1 indicates a perfect positive linear relationship,
✧ 0 indicates no linear relationship,
✧ -1 indicates a perfect negative linear relationship.

Pearson's correlation is calculated using the following formula:

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \cdot \Sigma(y_i - \bar{y})^2}}$$

where:

- $x_i$ and $y_i$ are the individual data points,
- $\bar{x}$ and $\bar{y}$ are the means of the variables $x$ and $y$.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is a preprocessing technique used in data analysis and machine learning to transform numerical features of different scales into a common scale. The goal of scaling is to standardize the range of the independent variables or features of the dataset. This is important because many machine learning algorithms are sensitive to the scale of the input features. Features with different scales can lead to biased or incorrect predictions, as some features may dominate others simply due to their scale.

Why Scaling is Performed:

Algorithm Sensitivity: Some machine learning algorithms, like k-nearest neighbors or support vector machines, are distance-based and sensitive to the scale of the features. Scaling helps these algorithms perform better.

Convergence in Gradient Descent: For optimization algorithms like gradient descent, scaling can help in faster convergence by ensuring that the steps taken during optimization are more consistent across features.

Regularization: Regularization techniques, such as L1 or L2 regularization, can be sensitive to the scale of features. Scaling helps in ensuring that regularization penalties are applied uniformly.

Min Max Scaling:

1. Transforms data to a range of 0 to 1 or -1 to 1 by subtracting the minimum value and dividing by the range (maximum - minimum).
2. Preserves the relative order of the original data points.
3. Sensitive to outliers, which can significantly affect the scaling range.
4. Useful when the data is not normally distributed.

Standardized Scaling:

1. Also known as Z-score normalization.
2. Transforms data to a mean of 0 and a standard deviation of 1.
3. Subtracts the mean and divides by the standard deviation.
4. Does not preserve the relative order of the original data points.
5. Less sensitive to outliers than normalized scaling.
6. Useful when the data is normally distributed or needs to be normalized for statistical analysis.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: The Variance Inflation Factor (VIF) is a measure used to assess multicollinearity in a regression analysis. High VIF values indicate that there may be a problematic degree of correlation between predictor variables, which can lead to issues such as unstable coefficient estimates and inflated standard errors.

1. Perfect Multicollinearity:

This occurs when one variable is perfectly linear combination of other variables in the model. This means one variable can be predicted exactly from the other variables, leading to an undefined variance and an infinite VIF.

2. Near Perfect Multicollinearity:

Even when perfect multicollinearity doesn't exist, high correlations between independent variables can significantly inflate the VIF value. If two or more variables are highly correlated, they share a lot of information, leading to inflated variances and potentially infinite VIF values.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: A Q-Q (quantile-quantile) plot is a graphical tool used to assess whether a given dataset follows a specific theoretical distribution, such as a normal distribution. It compares the quantiles of the observed data against the quantiles expected from the theoretical distribution. In the context of linear regression, Q-Q plots are particularly useful for checking the normality assumption of residuals.

Use and Importance of Q-Q Plot in Linear Regression:

1. Normality Assumption:

- Purpose: One of the key assumptions in linear regression is that the residuals (the differences between observed and predicted values) are normally distributed. Deviations from normality can impact the validity of statistical inferences.
- Use of Q-Q Plot: By comparing the observed residuals against the quantiles of a standard normal distribution in a Q-Q plot, you can visually assess whether the residuals follow a normal distribution. If the points on the Q-Q plot roughly align with a straight line, it suggests that the residuals are approximately normally distributed.

2. Identification of Outliers:

- Purpose: Q-Q plots can help identify outliers or heavy-tailed distributions in the residuals.
- Use of Q-Q Plot: If the tails of the Q-Q plot deviate from a straight line, it may indicate the presence of outliers or non-normality in the residuals. This can guide further investigation and potentially lead to model refinement.

3. Model Adequacy Checking:

- Purpose: Q-Q plots are part of a set of diagnostic tools used to check the overall adequacy of the regression model.
- Use of Q-Q Plot: A well-behaved Q-Q plot supports the assumption that the model is appropriate for the data. Deviations from the expected pattern may suggest issues with the model or violations of underlying assumptions.