**CREATED BY : EZAZAEHMAD VAHIDBHAI PALIZA**

**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)

**Ans:** 1. From ‘Year’ variable, we can say that count of bike rentals increased in year 2019 than 2018.

2. From ‘Weathersit’ variable, we can say that count of rentals of the bikes are more when weather is clear.

3. In Holiday Avg Ride count is increases.

4. Fall and Summer are more favorable than spring and winter.

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

**Ans:** **drop\_first=True** is used to avoid multicollinearity.

- It is also used to avoid Redundant features.



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

**Ans:** Target variable has a highest correlation with ‘temp’ variable.

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

**Ans:** 1. Residual error follows the normal distribution.

2. Scatter Plot Shows Linear Relationship between

Actual Test Data Points & Predicted Test Data Points

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

**Ans:**  Based on the final model we can say that “Temperature”, “Year” and “Season” are most contributing features.

**General Subjective Questions**

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

**Ans:** Linear regression is a supervised Machine Learning algorithm and the most used regression algorithm. In simple words, linear regression means fitting the best fit line between independent and target variables with the least mean square error method.

Linear regression is basic forms of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

we can write a linear regression equation as: y = mx + c

where, y is dependent variable, x is independent variable, m is a slope of the line and c is a intercept of the line.

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

**Ans:** Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

3. What is Pearson’s R? (3 marks)

**Ans:**  Pearson’s R is also known as Pearson’s correlation coefficient, Pearson correlation coefficient is a measure of the strength of a linear association between two variables or a measure of linear correlation between two sets of data.

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

**Ans:**  Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

* Sometimes, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.
* Normalized scaling brings all of the data in the range of 0 and 1.
* Standardized scaling replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

**Ans:**  VIF(i) = 1/(1-(R(i))^2) is a formula for finding VIF, where I refers the ith variable.

If R-squared value is equal to 1 then the denominator of the above formula become 0 and the

overall value become infinite. It denotes perfect correlation in variables.

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

(3 marks)

**Ans:**  Q-Q plot are also known as Quantile-Quantile (Q-Q) plot, Q-Q plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

**Benefits**

* It can be used with sample sizes also
* Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.