QUESTION AND ANSWER’S

ASSIGNMENT-BASED SUBJECTIVE QUESTION: -

Q1. Why is it important to use (drop\_first=True) during dummy variable creation?

Ans. Dummy variables are useful because they enable us to use a single regression equation to represent multiple groups. This means that we don't need to write out separate equation models for each subgroup.

drop\_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

Q2. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans. Bike rental are more correlated to Temperature Column i.e. (temp column)

Q3. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans. In the Final Model the Top 3 Features are: -

1. Monday Column (Mon)
2. Tuesday Column (Tues)
3. Holiday Column (holiday)

Q4. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans. Analysis of the categorical variables from the dataset are: -

1. Bike rentals more at high humidity.
2. Bike Rentals are observed at higher temperatures.
3. Bike Rentals are more in the year 2019 compared to 2018.
4. Bike Rentals are more in partly cloudy weather.
5. Bike Rentals are more during the Fall season and then in summer.

Q5. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans.  **1) The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis.**

**2) A rule of thumb commonly used in practice is if a VIF is > 10, you have high multicollinearity. In our case, with values less than 5, we are in good shape, and can proceed with our regression.**

**3) R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale. And we have the R-square value of 0.796 or 79%.**

**4)The adjusted R-squared adjusts for the number of terms in the model. And we got it around 0.791 or 79%.**

##### General Subjective Questions: -

Q1. Explain the linear regression algorithm in detail.

Ans “Modelling means machine learning algorithm, in which the machine learns from the data.”

1) Machine Learning model can be classified into three types based on the task performed and nature of the output.

1. Regression: The output variable to be predicted is a continuous variable.
2. Classification: The output variable to be predicted is a categorical variable.
3. Clustering: No predefined notion of a label is allocated to the group formed.

2) Machine Learning model can be classified into two broad categories:

1) Supervised Learning Method

-> Past data with labels is used for building the model.

-> Regression and Classification fall under this category.

2) Unsupervised Learning Method

-> No predefined label is assigned to the past data.

-> Clustering algorithm falls under this category.

**“Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task. A simple Linear Regression model attempts to explain the relationship between a dependent variable and an independent variable using one straight line.”**

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. The regression line is the best fit line for our model.

# **Hypothesis function for Linear Regression: -**

# Y = θ1 + X\*θ2

While training the model we are given:  
**x:** input training data (univariate – one input variable(parameter))  
**y:** labels to data (supervised learning)

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.  
**θ1:** intercept  
**θ2:** coefficient of x

Once we find the best θ1 and θ2 values, we get the best fit line. So, when we are finally using our model for prediction, it will predict the value of y for the input value of x.

##### **🡪How to update θ1 and θ2 values to get the best fit line?**

[Gradient Descent](https://www.geeksforgeeks.org/gradient-descent-in-linear-regression/)**:**  
To update θ1 and θ2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ1 and θ2 values and then iteratively updating the values, reaching minimum cost.

Cost Function:

Typically, Machine Learning model define cost function for a particular problem and then try to minimize the cost function. In this regression, our cost function is RSS (Residual Sum of Square) which is the cost function which we want to minimize and we will pick that **θ1 and θ2** .

Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).

Q2. Explain the Anscombe’s quartet in detail.

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

 They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers on statistical properties.

Q3. What is Pearson’s R?

Ans. The [Pearson correlation coefficient (r)](https://www.scribbr.com/statistics/pearson-correlation-coefficient/) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables. It is the ratio between the [covariance](https://en.wikipedia.org/wiki/Covariance) of two variables and the product of their [standard deviations](https://en.wikipedia.org/wiki/Standard_deviation); thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

### For a population;

Pearson's correlation coefficient, when applied to a [population](https://en.wikipedia.org/wiki/Statistical_population), is commonly represented by the Greek letter (rho) and may be referred to as the *population correlation coefficient* or the *population Pearson correlation coefficient*. Given a pair of random variables {\displaystyle (X,Y)}(X, Y).

correlation coefficient = (Covariance of X and Y (cov (X, Y)) / (standard deviation of X) \* (standard deviation of Y)

Q4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans It 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.

Most of the times, 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.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

1. Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1.

sklearn.preprocessing.MinMaxScalerhelps to implement normalization in python.

MinMaxScaling = (x-min(x))/(max(x)-min(x))

# Standardization Scaling:

Standardization 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**).

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

Standardisation = (x-mean(x))/sd(x)

Q5. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans. Quantile-Quantile (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.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

 Q6 You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans. If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).