# Assignment-based Subjective Questions

# Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

# Answer:

# There are some independent variables that have effect on count of sales or the dependent variables like seasons ,year, , month and Weathersit.

# 

# 

**Question 2.** Why is it important to use **drop\_first=True** during dummy variable creation? (Do not edit)

**Total Marks:**

# Answer:)

# Drop\_first is important as it will drop the extra columns that are not required as we make dummy columns and if these are not dropped it will create a big data which is not good for large set of data frames and will create a confusion

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

# Answer:

# Temp and atemp has highest correlation with target variable.

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

**Total Marks:**

# Linear Regression are validated on the basis of Normality of errors , Multicollinearity, Homoscedasticity , independence of residuals.

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

**Total Marks:**

# Answer:

# *The top three features contributing significantly are* *Temp, Holiday and weathersit.*

# General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail.

**Total Marks:**

**Answer:**

# *Linear regression is defined as an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events.*

# *The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.*

# *A sloped straight line represents the linear regression model.*

# *Linear Regression*

**Question 7.** Explain the Anscombe’s quartet in detail.

**Total Marks:**

**Answer:**

# *Anscombe’s quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.*

# *The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.*

# *There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.*

# *This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc.*

# *Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.*

# *These four plots can be defined as follows:*

# *image*

# *The statistical information for all these four datasets are approximately similar and can be computed as follows:*

# 

**Question 8.** What is Pearson’s R? (Do not edit)

**Total Marks:**

**Answer:**

The **Pearson correlation coefficient (*r*)** 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.

The Pearson correlation coefficient (*r*) is the most widely used correlation coefficient and is known by many names:

* Pearson’s *r*
* Bivariate correlation
* Pearson product-moment correlation coefficient (PPMCC)
* The correlation coefficient

The Pearson correlation coefficient is a [descriptive statistic](https://www.scribbr.com/statistics/descriptive-statistics/), meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables.

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

**Total Marks:**

**Answer:**

# Scaling, or feature scaling, refers to the process of transforming the values of variables to a specific range. This is often done to ensure that all variables have a comparable impact on the regression model. Scaling can help prevent certain variables from dominating the model due to their larger magnitude.

# Scaling can help prevent certain variables from dominating the model due to their larger magnitude.

# In both cases, you're transforming the values of numeric variables so that the transformed data points have specific helpful properties. The difference is that: in scaling, you're changing the range of your data, while. in normalization, you're changing the shape of the distribution of your data.

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

**Total Marks:**

**Answer:**

# A Variance Inflation Factor (VIF) can be infinite when there is a perfect correlation between variables, which is called multicollinearity.

# When two variables are perfectly collinear, the columns in X ˜ are linearly dependent, resulting in an infinite CN. In other words, all data points in the plane x1 × x2 fall on a straight line.

# *A large VIF indicates a high degree of multicollinearity. Multicollinearity is a problem because it can undermine the statistical significance of an independent variable.*

# *We can consider adjusting the model structure or selecting different independent variables.*

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(Do not edit)

**Total Marks:**

**Answer:**

# *A quantile-quantile (QQ) plot is a graphical tool that compares two data sets to determine if they come from the same distribution. It's used in linear regression to check if the residuals are normally distributed, as well as to ensure that a machine learning model is based on the correct distribution.*

# *A QQ plot is a scatterplot that plots the quantiles of one data set against the quantiles of another. A quantile is the point at which a certain percentage of the data falls below.*

# *Interpret*

# *If the two data sets come from the same distribution, the points on the plot will form a roughly straight line. The greater the deviation from the line, the more likely the data sets come from different distributions.*

# *Use*

# *QQ plots can be used to:*

# *Check if a data set is normally distributed*

# *Compare different data sets*

# *Ensure that a machine learning model is based on the correct distribution*

# *Some limitations of QQ plots include:*

# *They can be sensitive to the sample size and the choice of reference distribution*

# *They can be subjective and hard to quantify*