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: The demand bike cnt increased in the year 2019 when compared with year 2018

bike demand cnt is high when weather is clear and Few clouds.

Bike demand cnt is less in holidays in comparison to not being holiday

demand cnt of bike is almost similar throughout the weekdays

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

Ans: If we do not use drop_first = True, then n dummy variables will be created, and these predictors(n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to Dummy Variable Trap.

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

Ans: temp(temperature) and atemp(actual feeling temperature) has highest correlation

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

Ans: 30% of test was separated from the training set. Using model that was finalized sent an x_pred to get y_pred and have drawn a scatter plot on y_test(actual) vs y_pred(predicted). Found very residues. Calculated mean squared error was coming as 0.0101 which is less.

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: Positive coefficients like temp,yr indicate that an increase in these values will lead to an increase in the value of cnt. We can infer that year by year popularity of bike booking is getting increased similary for if

temperature is increased then bike booking are getting increased. If if weather is Light Snow & Rain it is negatively effected 0.488987 * temp + 0.234799 * yr -0.253042 * weathersit_Light Snow & Rain

General Subjective Questions:

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

Ans: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors

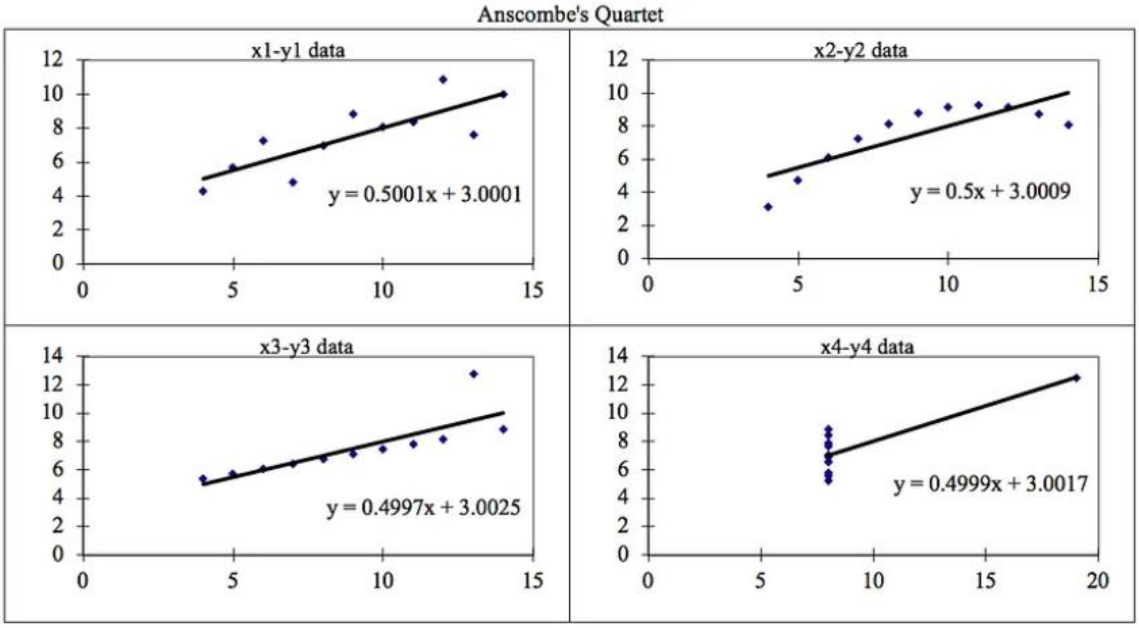
Assumptions of simple linear regression:

1. **Linear relationship between X and y.**
2. **Normal distribution of error terms.**
3. **Independence of error terms.**
4. **Constant variance of error terms.**

MLR: interpretation of coefficients, multicollinearity, model complexity and feature selection (VIF calc) which get adds from SLR.

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.



3. What is Pearson’s R? (3 marks)

Ans: Pearson correlation coefficient (PCC) — also known as Pearson's r. The Pearson product-moment correlation coefficient (PPMCC), the bivariate correlation, or colloquially simply as the correlation coefficient — is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1

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

Ans: scaling is step o_**f 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 must 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.

Normalization/Min-Max Scaling: It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

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.

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

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).

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 Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q-Q plots is to find out

if two sets of data come from the same distribution. A 45-degree angle is plotted on the Q-Q plot. If the two data sets come from a common distribution, the points will fall on that reference line.