1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

A. When the weather is Clear, then the overall number of rides are high compared with disturbing weather conditions such as high “Mist” and “Light snow / Light Rain”. During the “Light snow / Light Rain” conditions, the number of rides are very less. However, impact on the number of rides when the weather is Clear vs Mist is less. Spring season has lesser rides compared to all other three seasons. While the number of rides are increasing year on year, Fall season has highest rides compared to other seasons in this category across a single year.

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

A. If there are n dummy variables of one category, n-1 columns would also give the same information as n columns. So, in-order to be more efficient drop_first=True removes the column that has the first True value which essentially has no impact on the overall information we get from the data set.

   If we consider an example, furnished column has three values 1. Furnished, 2. Semi-Furnished 3.Not Furnished. This needs three dummy variables to be appended to the data set with three column names as above. If a house is not furnished or not semi-furnished, it will have false in the first two columns and true in the third column ‘Not Furnished’. The true in the last column is redundant as we get the information from the other two columns.

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

A. While “temp” and “atemp” columns have a great correlation, next comes “registered” and “cnt” variables

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

A. Here are the main assumptions of Linear Regression:

No Multicollinearity: the correlation matrix between all the independent variables is calculated. Correlations of close to 1 or -1 indicate a strong collinearity.

Linear Relationship: The initial scatter plot of the independent and dependant variables is almost in line with the relationship between the predicted and the actual values at the end. Confirming the assumption to be true

Equal Variance of Residuals: The scatter plot on Residuals vs. Fitted Values, the residuals on the y-axis and the predicted values on the x-axis and if the spread of the residuals is constant across the range of fitted values, it suggests homoscedasticity. This is confirmed with scatter plot after on the predicted values vs residuals.

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

A. The variables with Higher magnitude Coefficients, Lowest P-values will contribute significantly to explain the demand of the shared bokes. From the model parameters, we could see that year, Sep, Working Day (indicates business day) are contributing positively towards the counts whereas the other variables like Weather situation Light Snow or Mist+Cloudy, Spring season and winter are in the order of negatively impacting the demand of the shared bikes

Coefficient

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4935	0.019	25.728	0.000	0.456	0.531
yr	0.2468	0.010	25.058	0.000	0.227	0.266
holiday	-0.0606	0.034	-1.809	0.071	-0.127	0.005
workingday	0.0486	0.014	3.388	0.001	0.020	0.077
Jan	-0.0883	0.021	-4.259	0.000	-0.129	-0.048
July	-0.0086	0.022	-0.389	0.698	-0.052	0.035
Sep	0.0765	0.020	3.742	0.000	0.036	0.117
Sat	0.0527	0.018	2.927	0.004	0.017	0.088
Light Snow	-0.3220	0.030	-10.867	0.000	-0.380	-0.264
Mist + Cloudy	-0.0868	0.011	-8.226	0.000	-0.108	-0.066
spring	-0.2844	0.018	-15.570	0.000	-0.320	-0.249
summer	-0.0580	0.017	-3.501	0.001	-0.091	-0.025
winter	-0.0795	0.016	-4.926	0.000	-0.111	-0.048

General Subjective Questions

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

Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation.

This is achieved by fitting a line to the data using least squares. The line tries to minimize the sum of the squares of the residuals. It assumes a linear relationship between these features and the target. Linear regression

aims to find a straight line that best fits this data. This line represents the equation of your linear model, which can be used to predict the value of Y for new unseen X values.

There are two types of linear regression depending on the number of independent variables:

Simple Linear Regression: This has only one independent variable (X) to predict the dependent

variable (Y).

Multiple Linear Regression: This involves multiple independent variables (X1, X2, X3, etc.) to predict the

dependent variable (Y).

This can be represented using an equation $y = b0+b1x1 + b2x2 +…..$ Based on the number of independent variables represented as xn

The key aspect of linear regression is finding the equation for the best fit line. This is achieved by minimizing the residuals. Residuals represent the vertical distances between each data point and the corresponding point on the fitted line.

There are different techniques to achieve this, but a common approach is the method of least squares. It involves minimizing the sum of squared residuals, essentially finding the line that makes the sum of these squared distances as small as possible.

The equation of a straight line is generally represented in mathematics as: $Y = b0 + b1 * X$

where, b0 is the y-intercept (the point where the line crosses the Y-axis).

b1 is the slope of the line (represents the change in Y for a unit change in X).

The linear regression algorithm determines the optimal values for b0 and b1 that minimize the residuals.

The main assumption is that the relationship between X and Y is linear. If the data exhibits a non-linear the prediction may not be so accurate.

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

Anscombe's quartet is developed by Francis Anscombe in 1973. This model illustrates the importance of plotting the data before starting the in-depth analysis. This model consists of four sets of data, each containing 11 (x, y) data points. These datasets have identical basic summary statistics like mean, variance, standard deviation, correlation coefficient, etc. But when we plot these data sets, they reveal themselves to be very different visually. This explains the crucial role of data visualization in understanding the underlying patterns in the data. Eventually, it explains the necessity of performing Exploratory Data Analysis (EDA) on the data sets before we build any ML modal.

Anscombe's quartet demonstrates that any numerical statistics can be misleading and fail to capture the true essence of the data if we do not visualize the data. By visualizing the data, we can uncover important aspects like:

Non-linear relationships: Even with similar means and correlations, the data points might not follow a straight line, indicating a non-linear relationship between the variables.

Outliers: Summary statistics might not reveal the presence of outliers that can significantly influence the fitted line.

Underlying distributions: The visual representation can show us the spread and distribution of the data points, which can be crucial for choosing appropriate statistical methods.

The Anscombe’s quartlet explains that if there are four different Data Sets and let us say the respective scatter plots drawn on these Datasets may represent as follows:

Data Set 1: This could be a tight scatter plot with a linear relationship.

Data Set 2: This might show a non-linear pattern, but the relation between the variables is pretty obvious. Coefficient determination would be more appropriate.

Data Set 3: This modelled relationship is linear, but should have a different regression line. This could be a random scatter with one extreme outlier heavily influencing the fitted line

Data Set 4: This might depict a straight line, but with all the data points clustered on one side, making the line irrelevant for most predictions.

However, the mean, variance, standard deviation of the above all four Datasets might be same / nearly matching though the data points represent different scatter plots.

This model makes an imperative suggestions to do the following:

perform EDA / visualize the data before building any complex models based on statistics.

Visualize your data: Create scatter plots, histograms, and other visualizations to understand the

distribution of your variables and identify potential issues.

Look for outliers: Investigate outliers and assess their impact on the analysis.

Consider non-linear relationships: Not to always assume a linear relationship. Explore if transformations or different models might be more suitable.

3. What is Pearson's R? (3 marks)

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure used to quantify the linear relationship between two continuous variables. It represents the strength and direction of that association.

Strength: The value of R ranges from -1 to +1.

Positive R (between 0 and +1): Indicates a positive correlation. As the value of one variable increases, the other variable tends to increase as well. (Stronger positive correlation as the value gets closer to 1).

Negative R (between -1 and 0): Indicates a negative correlation. As the value of one variable increases, the other variable tends to decrease. (Stronger negative correlation as the value gets closer to -1).

Direction: The positive or negative sign indicates the direction of the relationship.

No Correlation: A value of 0 indicates no linear correlation. The changes in one variable are unrelated to the changes in the other.

Pearson's R only measures linear relationships. It won't capture non-linear patterns like curves or exponential trends.

It assumes normally distributed data for both variables. If the data is not normally distributed, the R value might be misleading.

Pearson's R is widely used in various fields to understand relationships between variables. Examples like Analyzing the correlation between stock prices and market movements, Investigating the correlation between environmental factors and plant growth, etc

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

Scaling is a data preprocessing technique where we adjust the values of features within a dataset to a specific range proportionally matching with original data values. This is done to ensure all features contribute equally to the model's training process and improve the overall performance.

Algorithms like gradient descent rely on calculating the difference between predicted and actual values. Features with vastly different scales can cause the updates to become very small for features with lower scales, hindering the convergence process. Scaling helps achieve a more balanced optimization process. Features with a much larger range of values can dominate features with smaller ranges during model training. Scaling ensures each feature has a similar influence on the model. Computational power and usage of resources on scaling data is more optimum and efficient.

There are two most common scaling techniques:

Normalization: Normalization typically scales features to a range between 0 and 1 (or -1 and 1). Here's a common approach for min-max normalization:

New_Value = (Old_Value - Min_Value) / (Max_Value - Min_Value)

where,

Min_Value is the minimum value of the feature in the dataset.

MAX Value is the maximum value of the feature in the dataset.

This ensures all features are restricted to a specific range. It's useful when you want to bound the data within a known range of 0 and 1.

Standardization: Standardization transforms features to have a zero mean and unit standard deviation.

Here's the formula:

New Value = (Old Value - Mean) / Standard Deviation

This process centres the data around zero and scales it based on the spread of the data points. It's

particularly useful when the features have different units of measurement, and you want the model to

focus on the relative differences between data points rather than the absolute values.

Differences:

Normalization creates a uniform distribution of values between a specific range (often 0-1 or -1 to 1).

Standardization creates a normal distribution with a mean of 0 and a standard deviation of 1.

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

Infinite VIF is an indicator of a perfect correlation between the variables. This means one independent variable can be expressed entirely as a linear combination of the other independent variables.

Reasons for infinite VIF is explained as follows:

VIF Calculation: VIF is calculated using this formula: 1 / (1 - R^2), where R^2 is the coefficient of determination between a specific independent variable and all the other independent variables combined.

Perfect Multicollinearity: In this case, R^2 between the variable and the others becomes exactly 1. This signifies a perfect linear relationship, where one variable can be perfectly predicted by the others.

Division by Zero: Plugging R^2 of 1 into the VIF formula results in a denominator of zero (1 - 1 = 0). Dividing by zero is mathematically undefined, hence the infinite VIF. Essentially, an infinite VIF indicates that the variable you're looking at is redundant because it carries no unique information. It can be completely recreated using the other independent variables.

Unreliable Coefficients: The regression coefficients for the variables with high VIF (including the one with infinite VIF) become unreliable and difficult to interpret.

Inaccurate Predictions: The model might not be able to make accurate predictions due to the inflated variances of the coefficients.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

A. The quantile-quantile plot or shortly called as Q-Q plot, is a tool used to assess the normality of residuals in linear regression in a graphical manner. It helps visualize how closely the distribution of the residuals aligns with a theoretical normal distribution.

The steps involved are

1) Calculating Quantiles: Both the observed residuals from a linear regression model and the theoretical quantiles from a normal distribution are calculated. Quantiles represent specific percentiles of the data distribution.

2) Plotting the Quantiles: The quantiles of the residuals are plotted against the corresponding quantiles of the normal distribution. Ideally, the points should fall roughly along a straight diagonal line.

Interpreting the Q-Q Plot:

Straight Diagonal Line: If the points form a reasonably straight diagonal line, it suggests that the residuals are normally distributed. Deviations from the straight line indicate departures from normality. Points above the line indicates that the tails of the distribution are heavier than a normal distribution (fatter tails), possibly indicating outliers or skewed data.

Points below the line indicates that the lighter tails in the distribution compared to normal, potentially implying a presence of more data points near the centre.

A Q-Q plot provides a visual way to assess normality of residuals and identify potential issues. And Q-Q plot is helpful in identifying Outliers. Deviations from the line in the Q-Q plot can highlight the presence of outliers that might be affecting the model. The Q-Q plot is a tool that mostly helps for diagnosing problems with the linear regression model. With Q-Q plots into the linear regression analysis, the underlying residual distribution could be better understood.