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)

Boxplot and bar plot were used to for Categorical column analysis and findings are

- Seems that Fall season has best bookings, then summer and winter follows and spring is the least
- Compared to 2018 all season bookings are higher in 2019
- Most sales are happening in June, July, Aug and Sep months
- 2019 has more increase in every month

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

When we are creating dummy variables for a categorical feature, the number of columns in the resulting dataframe will be equal to the number of unique values in the original feature minus 1. This is because one of the columns is redundant, the redundant column is also referred to as a "dummy trap".

We can avoid the dummy trap by setting drop_first=True

If we have 3 types of Categorical column and we want to create dummy variable for that column. We can manage with two and can reduce 1 as if a variable is neither A not B means it is always C

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

   Numerical variable 'temp' has highest correlation with target variable 'cnt'

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

   Some of the assumptions

   - Residual plots : Distribution of residuals are randomly distributed and have a mean close to zero
   - Normality of error terms : Error terms should be normally distributed
   - Multicollinearity check : There should be insignificant multicollinearity among variables. This refers to the assumption that the independent variables are not highly correlated with each other. High correlation between the independent variables can lead to unstable estimates of the regression
   - Independence of residuals: This refers to the assumption that the residuals are independent of each other, meaning that the residuals from one observation are not related to the residuals from another observation.
   - Homoscedasticity: There should be no visible pattern in residual values.
   - Independence of residuals : No auto-correlation

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
    - temp
    - winter
    - Sep

General Subjective Questions

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

Linear regression is a supervised machine learning algorithm used for predicting continuous values based on a set of independent variables.
The goal of linear regression is to find the best linear relationship between the independent variables (also known as predictors or features) and the dependent variable (also known as the target).

The basic idea behind linear regression is to fit a line to the data that minimizes the sum of the squared differences between the observed target values and the predicted target values. This line is represented by the equation of a line:

$y = b0 + b1 * x1 + b2 * x2 + ... + bn * xn$

or

$Y = mX + c$

where y is the target, x1, x2, ..., xn are the independent variables, b0 is the intercept or constant, and b1, b2, ..., bn are the coefficients. The coefficients represent the change in the target for a one unit change in the corresponding independent variable, while holding all other independent variables constant.

To find the best coefficients, the linear regression algorithm uses an optimization algorithm, such as gradient descent, to minimize the sum of the squared differences between the observed target values and the predicted target values.

Once the coefficients are found, the linear regression model can be used to make predictions for new data by plugging in the values for the independent variables and using the equation of the line to calculate the predicted target value.

Linear regression is of the following two types
- Simple Linear Regression
- Multiple Linear Regression

It is important to keep in mind that linear regression makes several assumptions about the data and the relationship between the independent variables and the target.

- Residual plots : Distribution of residuals are randomly distributed and have a mean close to zero
- Normality of error terms : Error terms should be normally distributed
- Multicollinearity check : There should be insignificant multicollinearity among variables. This refers to the assumption that the independent variables are not highly correlated with each other. High correlation between the independent variables can lead to unstable estimates of the regression
- Independence of residuals: This refers to the assumption that the residuals are independent of each other, meaning that the residuals from one observation are not related to the residuals from another observation.
- Homoscedasticity: There should be no visible pattern in residual values.
- Independence of residuals : No auto-correlation

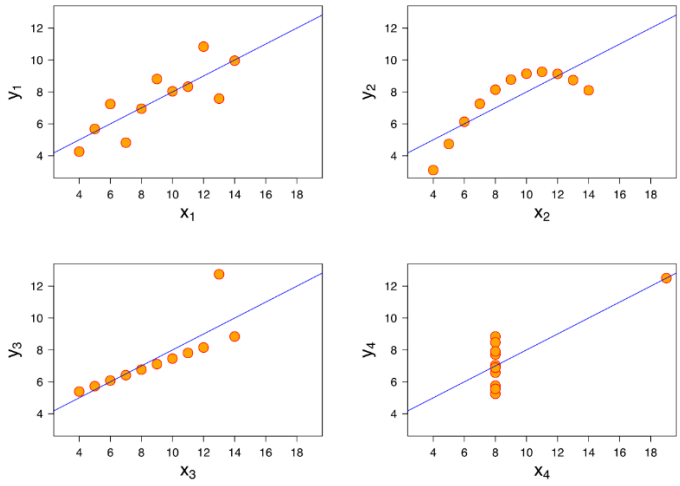
2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

The summary statistics show that the means and the variances were identical for x and y across the groups:

- Mean of x is 9 and mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset
- The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset
  When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

• Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

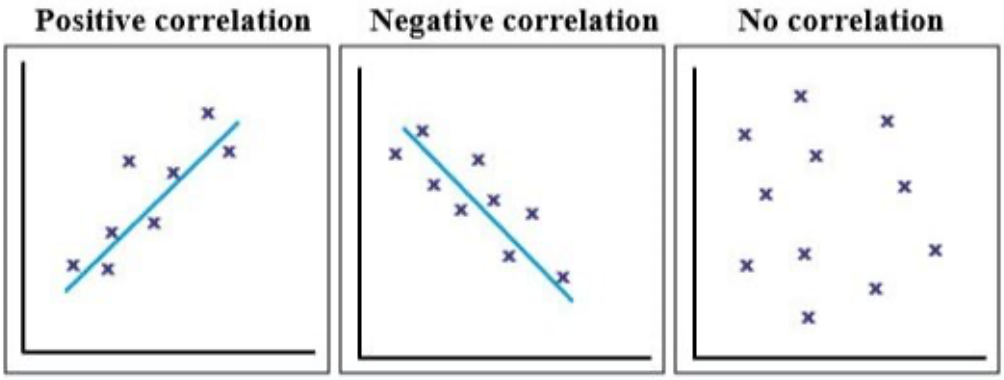
Pearson's r is a numerical summary of the strength of the linear association between the variables.

- If the variables tend to go up and down together, the correlation coefficient will be positive.
- If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1.

- A value = 0 indicates that there is no association between the two variables.
- A value > 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.
- A value < 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

Same is depicted in below graphs



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

   Scaling is the process of transforming the values of a feature or set of features so that they have a common scale. The goal of scaling is to ensure that the features are on a similar scale, which is important for many machine learning algorithms that require features to be in a specific range, or to have a common scale, in order to work effectively.

   There are several reasons why scaling is performed:

- Improving the performance of machine learning algorithms
- Handling heterogeneous features: In many datasets, features can have different units of measurement or scales, making it difficult to compare them. Scaling can help to handle this problem by transforming the features so that they have a common scale.
- Reducing the influence of outliers: Outliers can have a large influence on the mean and standard deviation of a feature, which can be problematic for some machine learning algorithms. Scaling can help
- Improving interpretability: by making it easier to compare the features and understand their relative importance.

   There are several methods for scaling features, including normalization, standardization.

Normalized scaling:

   Also known as min-max scaling, transforms the data so that the values for each feature fall within a specified range, usually between 0 and 1. This is done by subtracting the minimum value for each feature and dividing by the range (the difference between the minimum and maximum values). Normalized scaling is useful when the data contains large outliers or extreme values, as it can help to prevent these values from dominating the analysis.

Standardized scaling:

   Also known as z-score normalization, transforms the data so that each feature has a mean of 0 and a standard deviation of 1. This is done by subtracting the mean and dividing by the standard deviation. Standardized scaling is useful when the data contains features with different units or scales, as it allows for a comparison of the features on a common scale.

   The choice of method depends on the characteristics of the data and the goals of the analysi

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

   If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

   When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

   A Q-Q plot, also known as a quantile-quantile plot, is a graphical tool used to assess if a set of data follows a specific theoretical distribution. It is a type of scatter plot that plots the quantiles of the data against the quantiles of a theoretical distribution. The plot is used to visually assess the goodness-of-fit of the data to the theoretical distribution

   The Q-Q plot is created by plotting the quantiles of the data against the quantiles of the theoretical distribution. The quantiles of the data are calculated by dividing the data into equal-sized groups, and the quantiles of the theoretical distribution are calculated from the cumulative distribution function of the distribution. If the data follows the theoretical distribution, the points on the Q-Q plot should form a roughly straight line. If the data does not follow the theoretical distribution, the points on the Q-Q plot will deviate from the straight line, indicating that the data does not fit the theoretical distribution well.

   In linear regression, a Q-Q plot can be used to assess the normality of the residuals. The residuals are the differences between the observed values and the values predicted by the linear regression model. It is assumed that the residuals are normally distributed, and the Q-Q plot can be used to visually assess this assumption.

   If the residuals are normally distributed, the points on the Q-Q plot of the residuals against a normal distribution should form a roughly straight line. This indicates that the residuals follow a normal distribution and that the linear regression model is a good fit for the data. If the residuals do not follow a normal distribution, the Q-Q plot will show deviations from a straight line, indicating that the linear regression model is not a good fit for the data.

   In summary, the Q-Q plot is an important tool for assessing the normality of the residuals in linear regression. It helps to ensure that the assumptions and statistical inferences of linear regression are valid, and it can also be used to identify outliers and leverage points in the data.