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?

Answer: I have analysed the data based on categorical variables using box plot and pair plot. Below are my observations.

1. Season Fall seems to have more bookings demand in year 2019 compared to 2018.
2. Bookings are increased starting from April to October in the year 2019.
3. Clear weather attracted more booking which seems obvious.
4. When holidays are plotted, bikes demand is high.
5. There is no major difference in bookings either on weekdays or weekends.
6. In all aspects Year 2019 has made good business.

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

Answer: We use drop_first = True while creating dummy variables for the following reasons:

a. helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.
b. It will remove first level of dummies to get k-1 dummies out of k categorical levels. For example: in case of dummies generating for gender variable, it will create (Male and Female) instead of (Male, Female and Other).

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

Answer: 'temp' variable has the highest correlation with the target variable.

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

Answer: Below are the validations that are done on the model:

a. Normality of error.
    a. Error terms are following normal distribution
b. Multi Collinearity
    a. There is no multicollinearity among the variables
c. Linear relationship validation
    a. Linearity should be visible among variables.

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

Answer: Below are the three features that explains the demand for shared bikes

a. Temp
b. Winter
c. Sep

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation

$$Y = mX + c$$

Here, Y is the dependent variable we are trying to predict. X is the independent variable we are using to make predictions. m is the slope of the regression line which represents the effect X has on Y c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

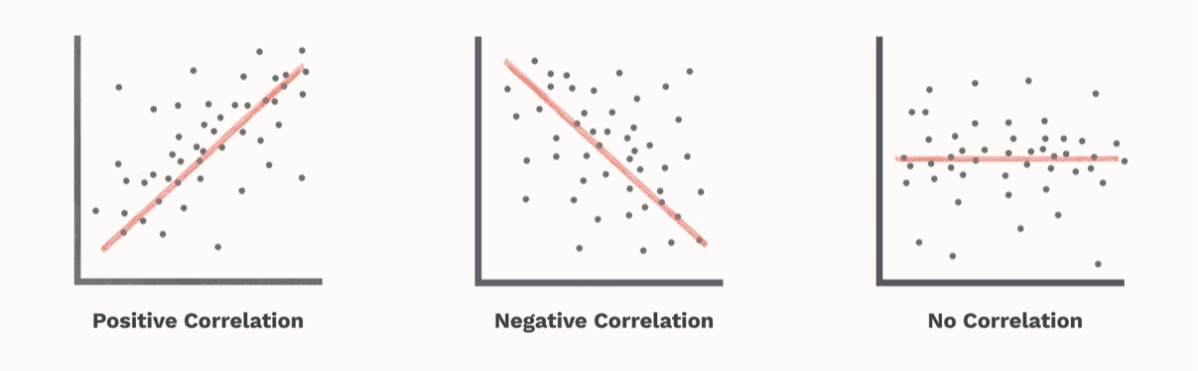
Furthermore, the linear relationship can be positive or negative in nature as explained below:

1. Positive Linear Relationship:

A linear relationship will be called positive if both independent and dependent ariable increases. It can be understood with the help of below image.

2. Negative Linear relationship:

A linear relationship will be called positive if independent increases and dependent Variable decreases. It can be understood with the help of below image.



Linear regression is of the following two types:

- Simple Linear Regression
- Multiple Linear Regression

Assumptions:

The following are some assumptions about dataset that is made by Linear Regression model

- Multi-collinearity
    - Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.

- Auto-correlation
    - Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- Relationship between variables
    - Linear regression model assumes that the relationship between response and feature variables must be linear.
- Normality of error terms
    - Error terms should be normally distributed
- Homoscedasticity
    - There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail.

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.

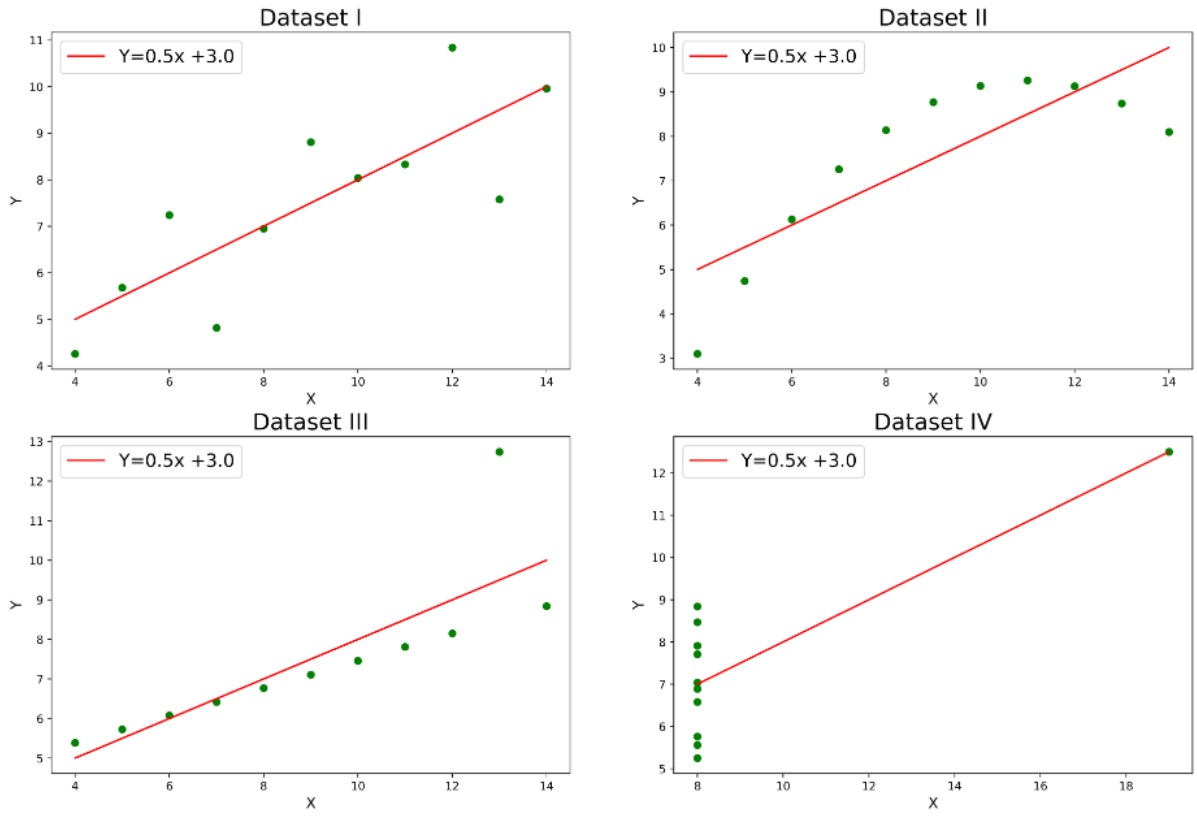
The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

The four datasets of **Anscombe's quartet.**

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.



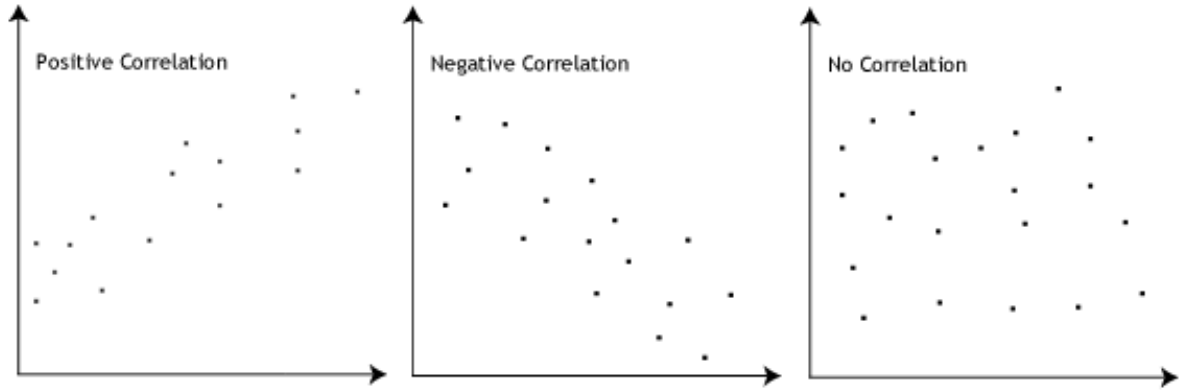
It is mentioned in the definition that Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed.

3. What is Pearson's R?

Answer: Pearson's R, often referred to as the Pearson correlation coefficient, is a measure of the linear correlation between two variables. It quantifies the strength and direction of the linear relationship between two continuous variables.

The Pearson correlation coefficient R ranges from -1 to 1:

- R=1: Perfect positive correlation. As one variable increases, the other variable increases proportionally.
- R=−1: Perfect negative correlation. As one variable increases, the other variable decreases proportionally.
- R=0: No linear correlation. There is no systematic linear relationship between the variables.



The formula to calculate Pearson's correlation coefficient between variables X and Y with n observations is:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Where:

- Xi and Yi are the individual observations of variables X and Y respectively.
- $\bar{x}$ and $\bar{y}$ are the means of variables X and Y respectively.

Pearson's correlation coefficient is a parametric measure and assumes that the relationship between variables is linear and that the variables are normally distributed. It is sensitive to outliers and can be influenced by extreme values.

Pearson's R is widely used in various fields such as statistics, social sciences, engineering, and finance to assess the strength and direction of the linear relationship between variables. It provides valuable insights into the association between variables, which can inform decision-making and further analysis.

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

Answer: Scaling also known as feature scaling is a technique to standardize the independent features present in the data in a fixed range.
It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
Example: If an algorithm is not using feature scaling method, then it can consider the value 3000 meter to be greater than 5 km but that's not true and, in this case, the algorithm will

give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a functionality called MinMaxScaler for Normalization.	Scikit-Learn provides a functionality as called StandardScaler for standardization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer: This happens when there is a perfect correlation between two independent variables. Whenever there is a perfect correlation then we get R-squared values as 1. When we use R-squared as 1. Then VIF = 1/(1-RSquare) = 1/ (1-1) = infinite.
To solve this, we need to drop one of the variables from the dataset which is causing perfect multicollinearity.

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

Answer: The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

1. Assess distributional assumptions, such as normality.
2. Detect outliers and skewness in the data.
3. Evaluate the normality of model residuals in regression analysis.
4. Compare distributions between datasets or with theoretical distributions.
5. Serve as diagnostic tools for model validation and improvement.

Top of Form

Q-Q plots are used in linear regression to assess the normality assumption of residuals. They help in Calculate the residuals by subtracting the observed values from the predicted values obtained from the linear regression model.
Used to construct a Q-Q plot of the residuals against the quantiles of a normal distribution. The plot compares the distribution of the residuals with that of a theoretical normal distribution.

If the residuals follow a normal distribution, the points on the Q-Q plot will fall approximately along the diagonal line. Deviations from the diagonal line indicate departures from normality.

Detecting departures from normality in the Q-Q plot suggests potential issues with the linear regression model. Non-normal residuals may indicate violations of assumptions, such as outliers, influential observations, or mis specified functional forms.

If the Q-Q plot reveals non-normality, adjustments may be needed in the regression model. This could involve transformation of variables, addressing outliers, or considering alternative regression techniques.