Assignment-Based Subjective Questions

1. **From your analysis of the categorical values from the dataset, what could you infer about their effect on the dependent variable? [3 Marks]**

Answer: After analysis on categorical columns using boxplot and bar plot. We can infer the below from the visualization.

1) More number of bikes were booked during the fall season.
2) Bike booking seemed almost same on weekday or weekend.
3) Bookings were more on the months june, july, aug, sept. Trend increases and is maximum mid-year but dips towards the end of year.
4) Bike booking increased drastically from 2018 to 2019. So it is good progress in terms of business.
5) Bookings in more on a holiday to a working day.
6) Wed, thu, fri, sat have more bookings than other days of the week.

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

Answer: We use drop_first = True as it reduces the extra column created during dummy variable creation. Hence reducing the correlations created among dummy variables.

Syntax: drop_first is of Boolean type and by default it takes value False.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi-furnished, then It is obvious unfurnished. So, we do not need 3rd variable to identify the unfurnished.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target-variable? [1 Mark]**

Answer: From the pair-plot, '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? [3 Marks]**

Answer: I have validated the 5 assumptions of Linear Regression Model as below:

1) Error terms are normally distributed.
2) There is insignificant multicollinearity between variables.
3) Validated Linear relationship among variables
4) No pattern observed in residual values.
5) No auto-correlation of residuals.

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

Answer: The top 3 features are:

1) Temp
2) winter
3) sep

General Subjective Questions

1. **Explain the Linear Regression Algorithm in detail. [4 Marks]**

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:

$Y = mX + c$

Y is the dependent variable we are trying to predict.

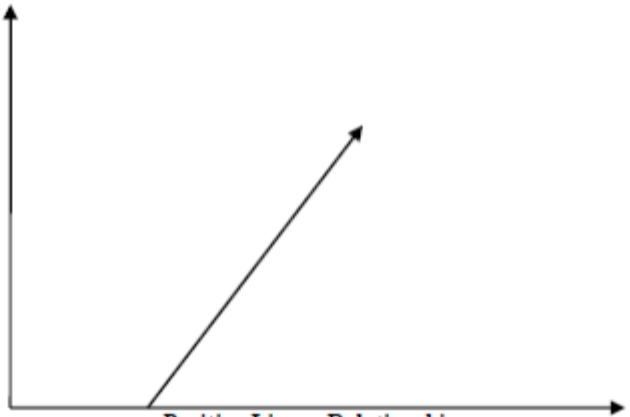
X is the independent variable we are using to make predictions.

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:

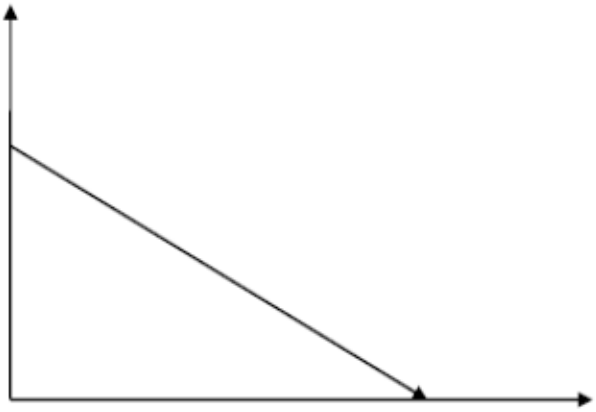
- Positive Linear Relationship:
  - A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph:



Positive Linear Relationship

- 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 following graph:



Negative Linear Relationship

Linear regression is of the following two types:

- Simple Linear Regression
- Multiple Linear Regression

Assumptions:

1) Multi-collinearity: Model assumes there is little or no multi-collinearity in the data.
2) Auto-correlation: Model assumes there is little or no Auto-correlation in the data.
3) Linear relationship validation: Relationship between variables must be linear.

4) Normality of error terms: Error terms must be normally distributed.

5) Homoscedasticity: There should be no visible pattern in residual values

2. Explain the Anscombe's quartet in detail. [3 Marks]

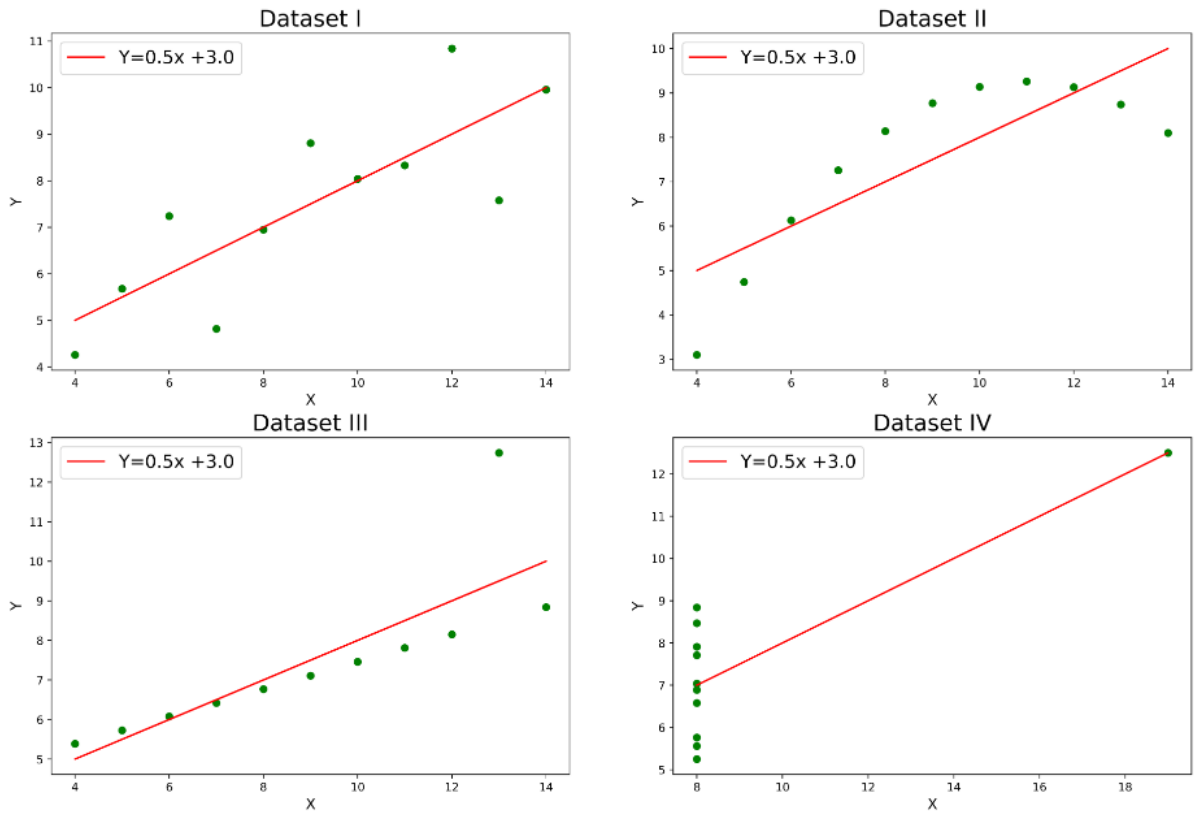
Answer: Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The 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.

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

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



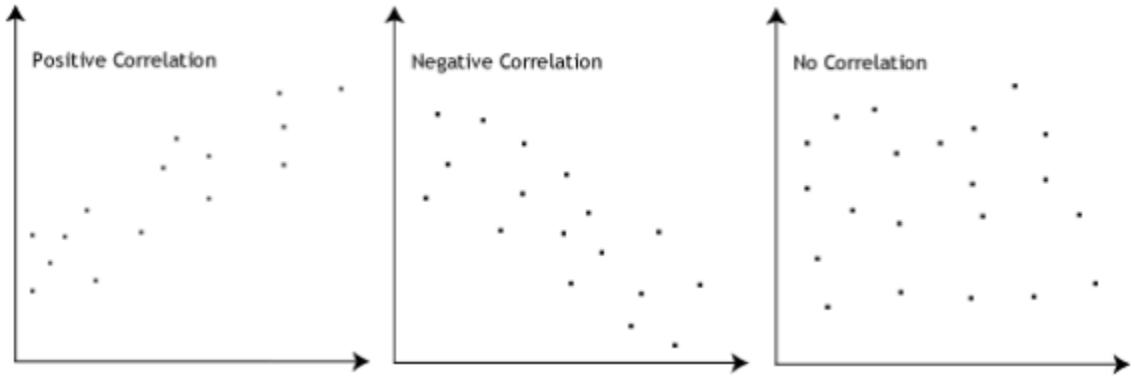
- **Dataset I** appears to have clean and well-fitting linear models.
- **Dataset II** is not distributed normally.
- In **Dataset III** the distribution is linear, but the calculated regression is thrown off by an outlier.
- **Dataset IV** shows that one outlier is enough to produce a high correlation coefficient.

Conclusion:
While the descriptive statistics of Anscombe's Quartet may appear uniform, the accompanying visualizations reveal distinct patterns, showcasing the necessity of combining statistical analysis with graphical exploration for robust data interpretation.

3. What is Pearson's R? [3 Marks]

Answer: 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 of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Answer: 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 actually 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.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scales values between [0,1] and [-1,1]	It is not bounded to a certain range.

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

Answer: 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.
[3 marks]

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:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.