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)**

   Answer:

   I have done analysis on categorical columns using the **boxplot** and **barplot**. Below are the few points I have infered from the visualization –

   - Fall season seems to have attracted more booking. And, the bike booking count has increased exceedingly from 2018 to 2019.
   - Most of the bike bookings has been done during the month of june, july, aug, sep and oct. It started increasing from the year till mid of the year and then it started decreasing gradually as we approached the end of the year.
   - Clear weather attracted more booking which is charismatic.
   - Thursday, Friday and Sunday has more number of bookings as compared to other day of the week.
   - When it's holiday, booking are relatively less in number which seems reasonable as on holidays, people might love spending time with their loved ones.
   - Booking seemed to be almost equal either on working day or non-working day.
   - 2019 attracted more number of booking from the previous year, which shows good progress in terms of business.

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

   Answer:

   'drop_first = True' is important to use, as it helps in reducing the extra column created during dummy variable creation.

   It also reduces the correlations created among dummy variables.

   Syntax -

   drop_first : bool, default False, which says whether to get 'k-1' dummies out of k categorical levels by removing the first level.

   The drop_first parameter specifies whether or not you want to drop the first category of the categorical variable you're encoding.

   Eg : If there are 3 levels, the drop_first will drop the first column.

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

   Answer:

   'temp' or 'temperature' variable has the highest correlation with the target variable (cnt).

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 assumption of Linear Regression Model based on below 5 assumptions -

   - ✓ **Normality of error terms :-**
     - There should be 'normal distribution' in Error Terms.

- ✓ **Multicollinearity :-**
  - There should be 'insignificant multicollinearity' among variables.
- ✓ **Linear Relationship :-**
  - There should be visible 'Linearity' among variables.
- ✓ **Homoscedasticity :-**
  - There should be 'no visible pattern' in residual values.
- ✓ **Residual Value Independent :-**
  - There should be no auto-correlation among variables.

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

Answer:

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- temperature
- season
- year

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 or target variable with given set of independent variables or features.

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 or target variable or label that we are trying to predict.

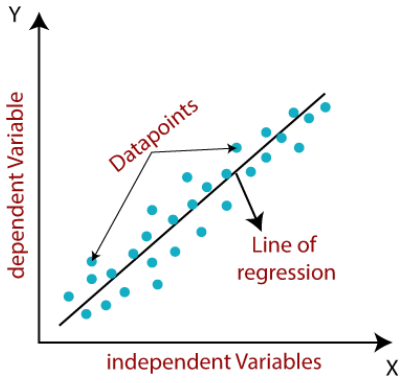
'**X**' is the independent variable or features that we are using to make predictions.

'**m**' is the slope of the regression line which represents the effect '**X**' has on '**Y**'.

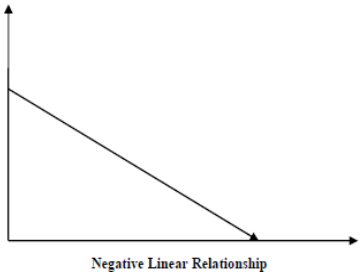
'**c**' is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

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

- **Positive Linear Relationship:**
  - A linear relationship is called positive when both independent and dependent variable increases.
  - i.e When both of them are directly proportional to each other.
  - Explaining below with the help of the graph -



- Negative Linear relationship:
  - A linear relationship is called positive when independent increases and dependent variable decreases.
  - i.e When both of them are inversely proportional to each other.
  - Explaining below with the help of a graph –



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 –

1. **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.
2. **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.
3. **Relationship between variables** –
   - Linear regression model says that the relationship between response and feature variables must be linear.
4. **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.** **(3 marks)**

Answer:

Anscombe's Quartet is composed of four datasets, That contain eleven (x, y) pairs. The important thing to note about these datasets is that they share the same descriptive statistics. But things change completely, when they are graphed.

Each graph tells a different story irrespective of their similar summary statistics.

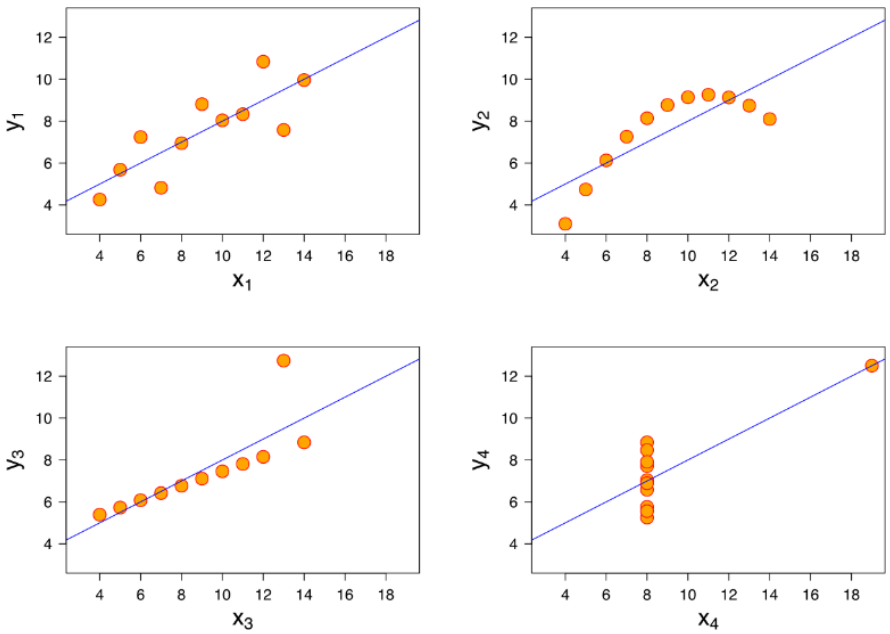
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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 between x and y is 0.816 for all dataset.

When we plot these four datasets on an x or y coordinate plane, we can observe that they show the same regression lines as well but each dataset tells a different story:



- Dataset 1 shows clean and well-fitting linear models.
- Dataset 2 shows no normal distribution.
- Dataset 3 shows linear distribution, but in the calculated regression,an outlier can also be seen.
- Dataset 4 shows that one outlier is enough to produce a high correlation coefficient.

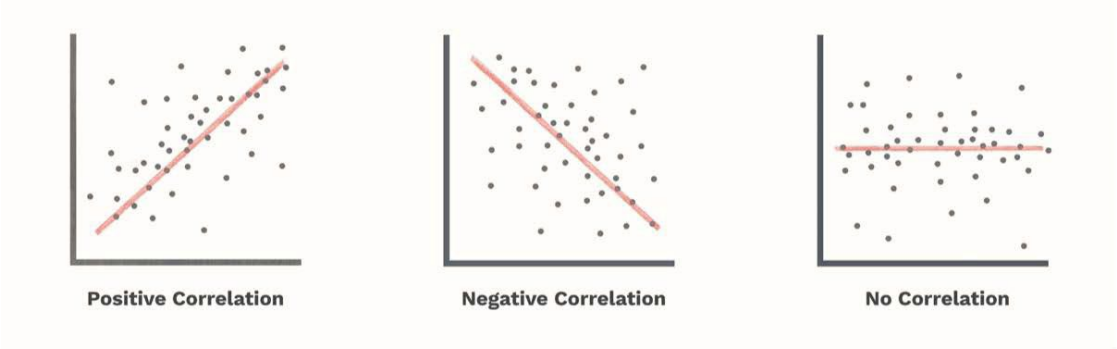
This emphasizes the importance of visualization in Data Analysis.
Looking at the data, it reveals a lot of the structure and a clear picture of the dataset.

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

Answer:

Pearson's R is a numerical summary of the strength of the linear association among 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,i.e. as the value of one variable increases, the value of the other variable also increases. A value less than 0 indicates a negative association,i.e. as the value of one variable increases, the value of the other variable decreases. This can be seen in the diagram below :-



In Statistics, the Pearson's Correlation Coefficient is also referred to as Pearson's R, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between variables.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\bar{y}$ =mean of values in y variable

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 2000 meter to be greater than 10 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.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
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.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
6.	Normalized scaling is used when we don't know about the distribution.	Standardized scaling is used when distribution is normal.
7.	Normalized scaling is called as scaling normalization.	Standardized scaling is called as Z-Score Normalization.

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

Answer:

VIF (Variance Inflation Factor) basically helps in explaining the relationship between one independent variable with all the other independent variables.

A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately. A very high VIF value shows a perfect correlation between two independent variables. In the case of perfect correlation, we get $R^2 = 1$, which lead to $1/(1-R^2)$ infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

VIF under 5 is acceptable and is perfectly considered.

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

Answer:

Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform 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.

Advantages:

- It can be used with sample size also.
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot Q-Q plot use on two datasets to check
- If both datasets came from population with common distribution
- If both datasets have common location and common scale
- If both datasets have similar type of distribution shape
- If both datasets have tail behaviour.