Assignment-based Subjective Questions

Ques1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer1: I have used boxplot and bar plot for analysing categorical data. Key results of the analysis:
1. 2019 has shown more growth in attracting people for shared bike as compare to 2018, this represents the growth in business.

2. On Holiday, we can see a drop on shared bike services, however the max counts on holiday and non holidays seems to be same.

3. On workdays and non workdays, seems to have same trend for bike sharing.

4. The trend on going for shared bike services on different seasons is as follows: Most in fall, than summer followed by winter and least in spring.

5. Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend in Bike sharing increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.

6. Clear weather attracted more booking, followed by mist and lowest in Light_snowrain.

7. Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.

Ques2: Why is it important to use drop_first=True during dummy variable creation?

Answer2: drop_first = True is important to use, as it helps in removing redundant column created during dummy variable creation.

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

Example

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 A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

Ques3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Answer3: 'temp' variable has the highest correlation with the target variable.

Ques4: How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer: Assumption of Linear Regression Model is validated by the following ways

1. Normality of error terms:
   Error term showed the normally distribution

2. Multicollinearity Minimised:
   This is done using VIFs and feature selection.
3. Linear relationship validation:
   Linearity should be visible among variables
4. Homoscedasticity is cross checked:
   Almost equal variance is seen in error terms.

Ques5. 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 top 3 features contributing significantly towards explaining the demand of the shared bikes :

temp, Light_snowrain, windspeed

General Subjective Questions

Ques1. Explain the linear regression algorithm in detail?

Answer: 1. 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).

2. 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

3. Linear regression is of the following two types: Simple Linear Regression and Multiple Linear Regression

4. Assumptions for linear regression:

4.1 Multi-collinearity : No

4.2 Auto-correlation : No

4.3 Relationship between variables : Linear

4.4. Normality of error terms: Normally distributed

4.5 Homoscedasticity: There should be no visible pattern in residual values.

Ques2: Explain the Anscombe's quartet in detail.

Answer: 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

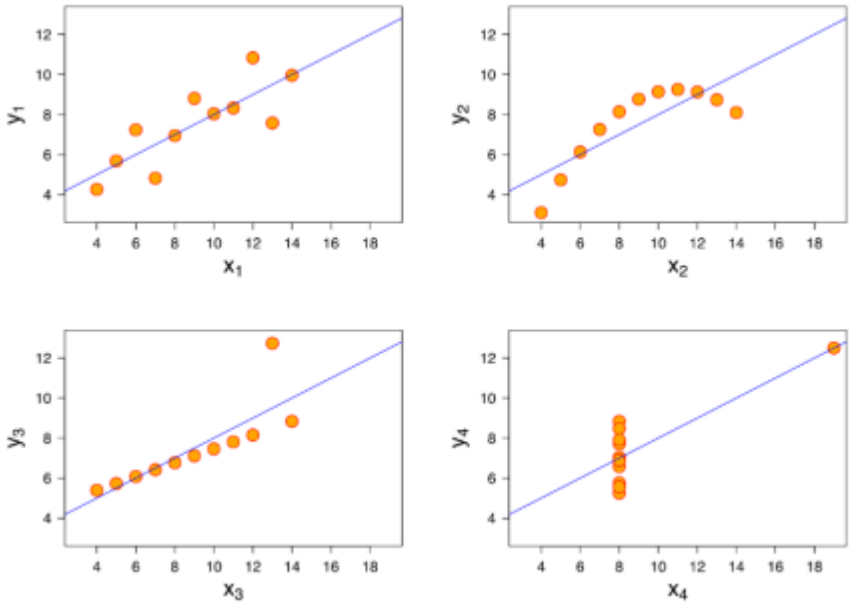
summary statistics. The summary statistics show that the means and the variances were identical for x and y across the groups:

	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

• 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.

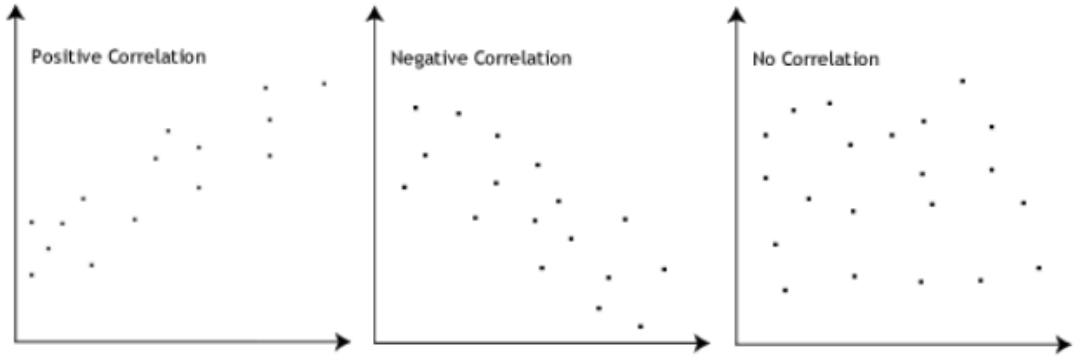
• In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.

• 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.

Ques3. What is Pearson's R?

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:



Ques4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer: Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.

Its importance:

1.It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

2. 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.

3. It makes interpretation easy and model run time less.

Normalized scaling	Standardized scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
It is really affected by outliers.	It is much less affected by outliers

Ques5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

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.

Ques6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer:

1. **Q-Q:** The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.
2. **Use:** 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.
3. **Importance:** 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