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:
Boxplots and bar plots have been used for the analysis of categorical columns.

Observation on Categorical Variable: -

1. **Season**: Fall season seems to have attracted more bookings. And, in each season the booking count has increased drastically from 2018 to 2019.

2. **Month:** Higher bookings have been done during the month of May, June, July, Aug, Sep, and Oct. The trend increased starting of the year till mid of year and then it started decreasing as we approached the end of the year.

3. **Weather:** Clear weather attracted more bookings.

4. **Weekday:** Thu, Fri, Sat, and Sun have more bookings as compared to the start of the week.

5. **Holiday:** People rent more on non-holiday, the reason might be people prefer to spend time at home and enjoy themselves with family.

6. **Working day:** Bookings are almost equal either on the working day or non-working day.

7. **Year:** 2019 has higher bookings than 2018, which shows a good rise in terms of business.

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

Answer:
A variable with n levels can be represented by an n-1 dummy variable. So if we remove the first column then also we can represent the data.

drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

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

Let's say we have 3 types of values in the Categorical column and we want to create a 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.

A	B
0	1
1	0
0	0

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

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? (3 marks)**

Answer:
I have validated the assumption of the Linear Regression Model based on below assumptions –

1. **Normality of error terms:** Error terms should be normally distributed
2. **Multicollinearity check:** There should be insignificant multicollinearity among variables.
3. **Linear relationship validation:** Linearity should be visible among variables
4. **Homoscedasticity:** There should be no visible pattern in residual values.

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 –

1. temp
2. winter
3. sep

General Subjective Questions

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

Answer:
Linear regression is defined as the statistical model that analyses the linear relationship between a dependent variable with a 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 the dependent variable will also change accordingly (increase or decrease).

Linear Equation : Y = m X + c

Y is the dependent variable.
X is the independent variable.
m is the slope of the regression line.
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Linear relationships can be positive or negative in nature –
Positive Linear Relationship: A linear relationship will be called positive if both the independent and dependent variable increases.

Negative Linear relationship: A linear relationship will be called negative if the independent increases and the dependent variable decreases.

Linear regression is of the following **two types** −

- **Simple Linear Regression**
- **Multiple Linear Regression**

Assumptions -
Multi-collinearity –There is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency among them.
Auto-correlation –There is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is a dependency between residual errors.
Relationship between variables -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.

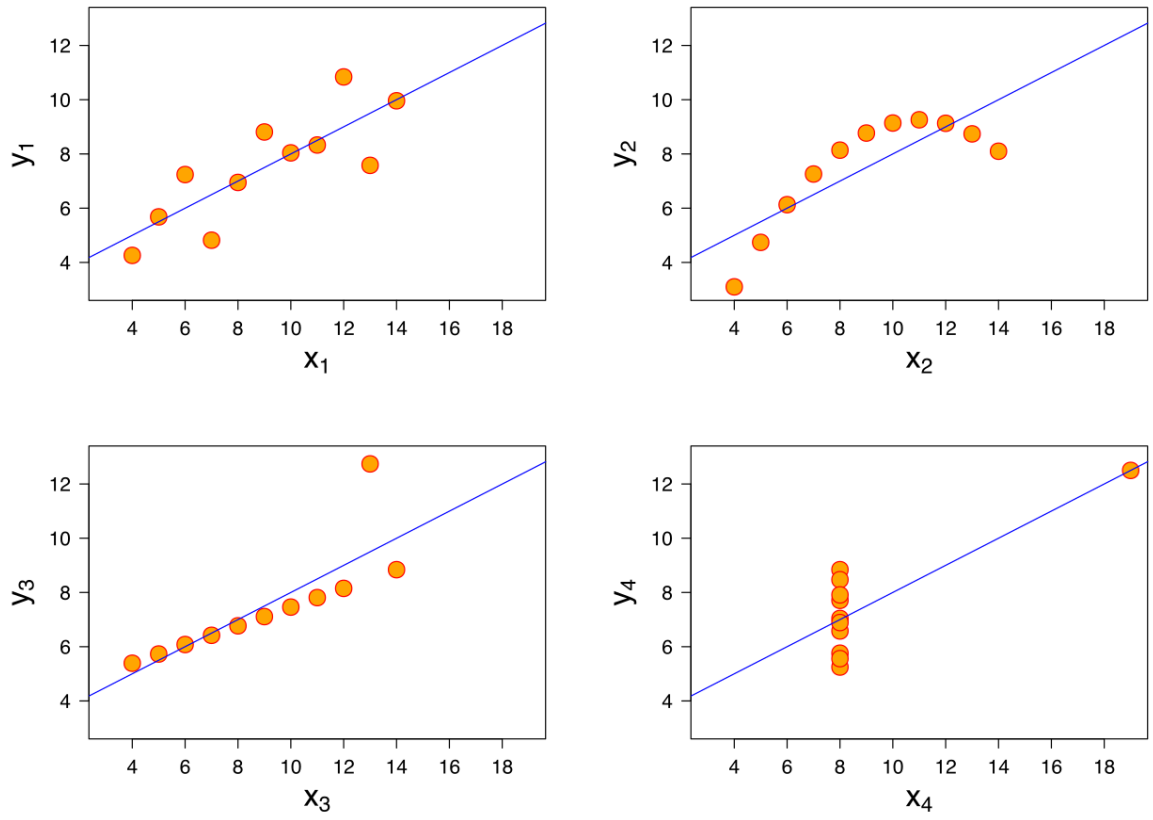
2. **Explain the Anscombe's quartet in detail. (3 marks)**

Answer:
Anscombe's Quartet can be defined as a group of four data sets that are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

This is used to demonstrate both the importance of graphing data when analyzing it, and the effect of outliers and other influential observations on statistical properties

When we plot the 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:

All four sets are identical when examined using simple summary statistics, but vary considerably when graphed



• 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)**

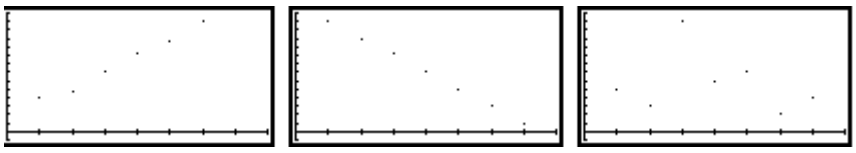
Answer:
Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated 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.

The figure below shows some data sets and their correlation coefficients. The first data set has an r=0.996, the second has an r = -0.999 and the third has an r= -0.233



r=0.996 r = -0.999 r= -0.233

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

Scaling:- It is a step of data Pre-Processing that is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Scaling is performed because most of the time, the collected data set contains features highly varying in magnitudes, units, and range. If scaling is not done then the algorithm only takes magnitude into account and not units hence incorrect modeling.
Ex: If an algorithm is not using the feature scaling method then it can consider the value 3000 meters to be greater than 5 km which is incorrect.

To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalized scaling vs Standardized scaling

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 transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
5.	$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$	$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$

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

Answer:

If there is a perfect correlation, then VIF = infinity.

The value of VIF is calculated by the below formula:

$$VIF_i = \frac{1}{1-R_i^2}$$

Where 'i' refers to the ith variable.

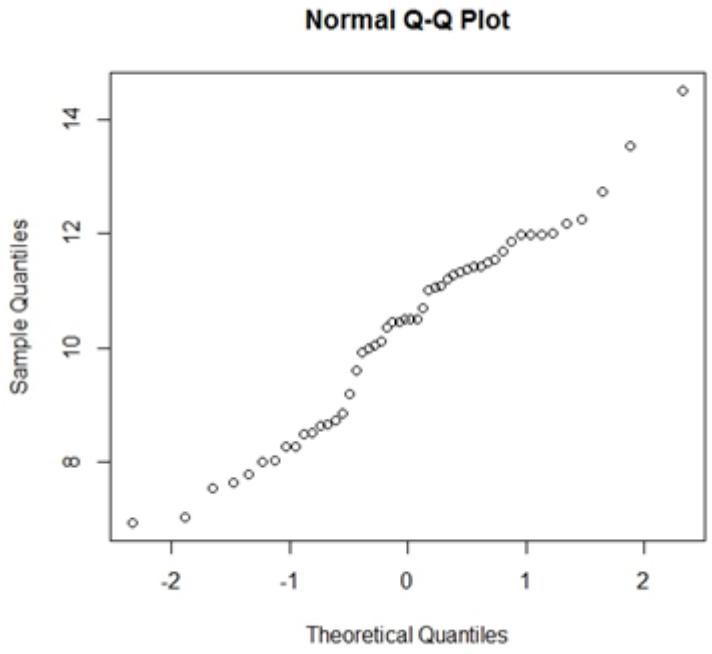
A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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 leads 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.

It is a scatterplot, created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we see the points forming a line that's roughly straight.



Use of Q-Q plot:

It is used to check the below scenarios, If two data sets —

- come from populations with a common distribution
- have a common location and scale
- have similar distributional shapes
- have similar tail behavior

A 45-degree reference line is 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:

- The sample sizes do not need to be equal.
- Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers.
- The q-q plot can provide more insight into the nature of the difference than analytical methods.