Assignment-Based Subjective

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

→ Based on the analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer from the visualization.

Here are the rephrased points based on the analysis of categorical columns using boxplots and bar plots:

- Fall season appears to have garnered higher booking numbers, with a notable increase in bookings observed across all seasons from 2018 to 2019.
- The months of May, June, July, August, September, and October witnessed the highest booking volumes, exhibiting an increasing trend from the beginning of the year until mid-year, followed by a decline towards the year's end.
- Clear weather conditions unsurprisingly attracted more bookings.
- Thursdays, Fridays, Saturdays, and Sundays recorded higher booking counts compared to the early days of the week.
- Non-holiday periods corresponded to fewer bookings, which aligns with expectations as people may prefer to spend holidays at home with family.
- Booking volumes appeared relatively consistent between working and non-working days.
- There was a noticeable increase in bookings in 2019 compared to the previous year, indicating positive growth in business.

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

→ Setting `drop_first = True` is crucial as it helps to mitigate multicollinearity issues by reducing the number of dummy variables created. This parameter instructs the function to generate k-1 dummies out of k categorical levels by excluding the first level. This approach is based on the assumption that if a data point is not represented by the first k-1 dummy variables, it must belong to the kth category. This helps to prevent redundancy and improve the interpretability of the model.

For instance, consider a categorical column with 3 distinct values: A, B, and C. By setting `drop_first = True`, the function will create dummy variables for B and C only, as the absence of B and C implies the presence of A. This reduces the number of dummy variables required and avoids the multicollinearity issue associated with including all three dummy variables.

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

→ '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)

→ I have validated the assumptions of the Linear Regression Model based on the following five criteria:

1. Normality of error terms:
    - The error terms should exhibit a normal distribution.
2. Multicollinearity check:
    - There should be no significant multicollinearity among the independent variables.
3. Linear relationship validation:
    - The relationships among the variables should be linear.
4. Homoscedasticity:
    - The residuals should display no discernible pattern.
5. Independence of residuals:
    - There should be no autocorrelation present in the residuals.

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

→ The top three features significantly contributing to explaining the demand for shared bikes are:

- temp
- winter
- sep

General Subjective Questions

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

→ Linear regression can be defined as a statistical model that analyses the linear relationship between a dependent variable and a set of independent variables.

This relationship implies that changes in the independent variables lead to corresponding changes in the dependent variable.

Mathematically, this relationship is represented by the equation

Y = mX + c,

where Y is the dependent variable,

X is the independent variable,

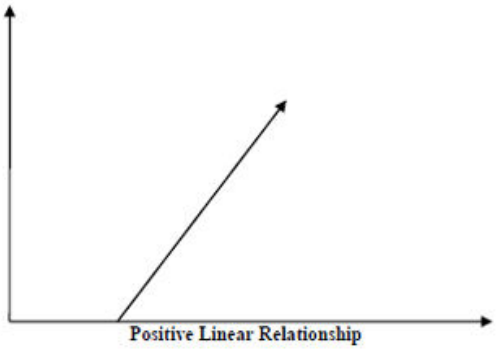
m is the slope representing the effect of X on Y, and

c is the Y-intercept.

A positive linear relationship occurs when both variables increase, while a negative linear relationship occurs when the independent variable increases and the dependent variable decreases.

- Positive Linear Relationship:

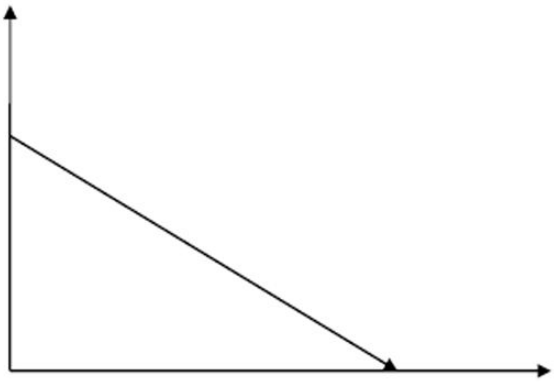
  A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph –



- 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

There are two types of linear regression:

1. Simple Linear Regression.
2. Multiple Linear Regression.

Several assumptions are made by the linear regression model about the dataset:

1. Multi-collinearity:
- The model assumes minimal or no multi-collinearity among the independent variables, which occurs when the features have dependencies.

2. Auto-correlation:
- Linear regression assumes minimal or no auto-correlation in the data, where auto-correlation refers to dependencies between residual errors.

3. Relationship between variables:
- The model assumes a linear relationship between the response and feature variables.

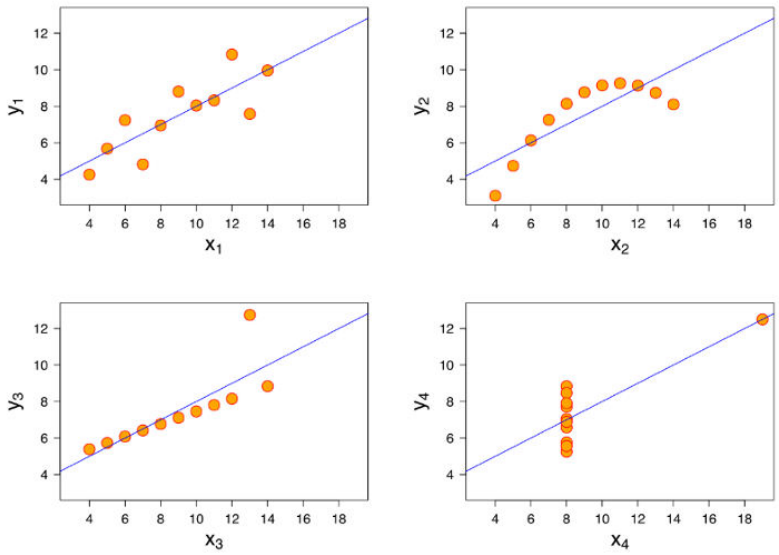
4. Normality of error terms:
- Error terms are expected to follow a normal distribution.

5. Homoscedasticity:

- There should be no discernible pattern in the residual values.

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

→ Anscombe's Quartet refers to a group of four datasets that exhibit nearly identical simple descriptive statistics but possess unique characteristics that can deceive a regression model if built without careful analysis. Despite having similar statistical summaries, the datasets differ significantly in their distributions and scatter plot representations. This quartet serves to emphasize the importance of visualizing data before analysis and model building, highlighting the impact of outliers and other observations on statistical properties.



The four datasets are described as follows:

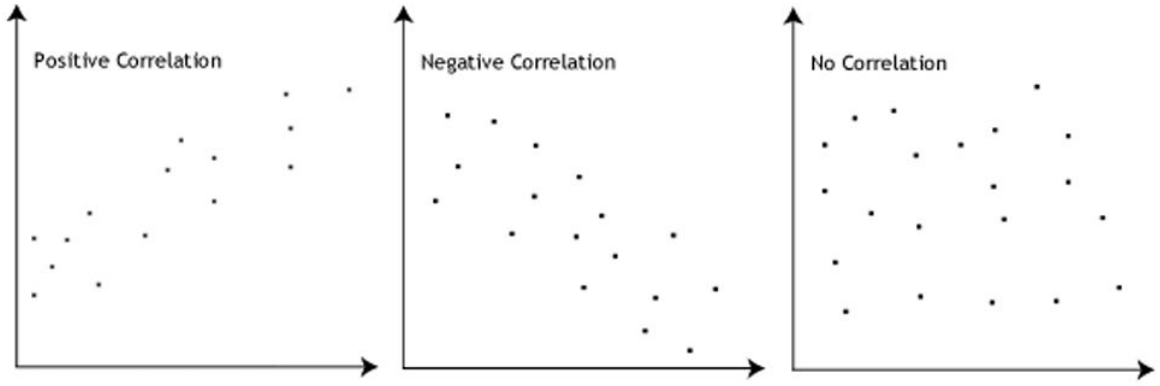
1. The first dataset displays a linear relationship between X and Y, making it suitable for a linear regression model.
2. The second dataset does not exhibit a linear relationship between X and Y, rendering it unsuitable for a linear regression model.
3. The third dataset contains outliers that cannot be effectively handled by a linear regression model.
4. The fourth dataset includes a high leverage point, resulting in a high correlation coefficient.

The key takeaway from Anscombe's Quartet is that regression algorithms can be misled, underscoring the importance of data visualization before constructing machine learning models.

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

→ Pearson's correlation coefficient (r) provides a quantitative measure of the strength of the linear association between two variables. When the variables tend to increase or decrease together, the correlation coefficient is positive. Conversely, when one variable tends to increase as the other decreases, the correlation coefficient is negative.

The Pearson correlation coefficient (r) ranges from +1 to -1. A value of 0 indicates no association between the variables. A positive value indicates a positive association, meaning that as one variable increases, the other variable also tends to increase. Conversely, a negative value indicates a negative association, where an increase in one variable corresponds to a decrease in the other variable. 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)

→ Feature Scaling is a preprocessing technique used to standardize independent features within a fixed range. Its purpose is to handle variations in magnitudes, values, or units present in the data. When feature scaling is not applied, machine learning algorithms may incorrectly weigh greater values higher and smaller values lower, regardless of their actual significance.

For example, without feature scaling, an algorithm might erroneously interpret a value of 3000 meters as greater than 5 kilometres. This can lead to inaccurate predictions. Feature scaling resolves this issue by bringing all values to the same magnitude, ensuring that the algorithm treats them appropriately.

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.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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

→ VIF (Variance Inflation Factor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF is given below:

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 R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, 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)

→ The quantile-quantile (q-q) plot is a graphical method used to assess whether two data sets originate from populations with a common distribution.

In a q-q plot, the quantiles of one data set are plotted against the quantiles of the other data set. A quantile represents the fraction (or percent) of data points below a given value. For example, the 0.3 quantile corresponds to the point at which 30% of the data fall below, and 70% fall above, that value. A reference line at a 45-degree angle is also plotted. If the two data sets share the same distribution, the points should approximately align along this reference line. Deviations from this line suggest differences in distributions.

The q-q plot is valuable because it helps determine if the assumption of a common distribution is valid when analysing two data samples. If the assumption holds true, estimators for location and scale parameters can combine both data sets to derive common estimates. Conversely, if the two samples exhibit differences, the q-q plot can provide insights into the nature of these differences, surpassing analytical methods like the chi-square and Kolmogorov-Smirnov 2-sample tests.