Assignment-based Subjective Questions

QUES 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 -

- Fall season seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
- Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
- Clear weather attracted more booking which seems obvious.
- Thu, Fir, Sat and Sun have more number of bookings as compared to the start of
  the week.
- When it's not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy
  with family.
- 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.

QUES 2- 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. 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 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.

QUES 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.

QUES 4 - How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

I have validated the assumption of Linear Regression Model based on below 4 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. Independence of residuals- No auto-correlation

QUES 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 –

A. temp

B. winter

C. sep

General Subjective Questions

QUES 1- Explain the linear regression algorithm in detail.

Answer -

Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (predictors). It assumes a linear relationship between the variables.

Simple Linear Regression: Involves one independent variable and one dependent variable. The model can be represented as:

$y = mx + b$
where:

- y is the dependent variable
- x is the independent variable
- m is the slope of the line
- b is the y-intercept

Multiple Linear Regression: Involves multiple independent variables and one dependent variable. The model can be represented as:

y = b0 + b1*x1 + b2*x2 + ... + bn*xn + e
where:

- y is the dependent variable
- x1, x2, ..., xn are the independent variables
- b0, b1, b2, ..., bn are the coefficients
- e is the error term

How it Works

1. **Data Collection:** Gather data for the dependent and independent variables.
2. **Model Creation:** Define the linear regression model based on the number of independent variables.
3. **Parameter Estimation:** Determine the values of the coefficients (slope and intercept) that best fit the data. This is typically done using the Ordinary Least Squares (OLS) method, which minimises the sum of the squared differences between the predicted and actual values.
4. **Model Evaluation:** Assess the model's performance using metrics like R-squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Adjusted R-squared.

Ordinary Least Squares (OLS)

OLS is a common method for estimating the parameters in linear regression. It finds the values of the coefficients that minimise the sum of the squared residuals (differences between the predicted and actual values).

Assumptions of Linear Regression

Linear regression makes several assumptions about the data:

- Linearity: The relationship between the variables is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: The variance of the residuals is constant.
- Normality: The residuals are normally distributed.

Applications of Linear Regression

Linear regression is widely used in various fields, including:

- Economics: Predicting stock prices, GDP growth
- Finance: Predicting asset returns, risk assessment

- Marketing: Predicting sales, customer churn
- Science: Modelling physical phenomena, analysing experimental data
-

Challenges and Considerations

- **Multicollinearity:** When independent variables are highly correlated, it can affect the model's stability and interpretation.
- **Outliers:** Outliers can significantly impact the model's performance.
- **Heteroscedasticity:** If the variance of the residuals is not constant, it can lead to biased estimates.
- **Model Selection:** Choosing the right independent variables is crucial for model performance.

Beyond Linear Regression

While linear regression is a powerful tool, it has limitations. For non-linear relationships, consider polynomial regression or other non-linear models. For categorical predictors, consider techniques like logistic regression or decision trees.

QUES 2- Explain the Anscombe's quartet in detail.

Answer-

Anscombe's Quartet: Understanding Through Simple Words

Anscombe's Quartet is a set of four different datasets that all have nearly identical simple descriptive statistics, yet appear very different when graphed. The idea behind Anscombe's Quartet is to highlight the importance of graphing data before analysing it, as different datasets can tell very different stories even if their summary statistics (like mean, variance, etc.) are the same.

Components of Anscombe's Quartet

1. Mean of X values:The average value of the X variable is the same for all four datasets.

2. Mean of Y values: The average value of the Y variable is the same for all four datasets.

3. Variance of X values: The measure of how spread out the X values are from the mean is the same.

4. Variance of Y values: The measure of how spread out the Y values are from the mean is the same.

5. Correlation between X and Y: The strength and direction of the relationship between X and Y are the same.

6. Linear Regression Line: The best-fit line (least-squares line) for predicting Y from X is the same in all datasets.

The Four Datasets

Each dataset in Anscombe's Quartet is made up of 11 (X, Y) points. Here is a brief description of each:

1. Dataset 1:

- This dataset appears to follow a linear relationship. When plotted, it shows a scatter plot where the points fall closely around a straight line.

2. Dataset 2:

- This dataset forms a curve. Even though the summary statistics are similar, the scatter plot reveals a quadratic relationship.

3. Dataset 3:

- In this dataset, most points lie on a straight line, but there is one outlier that can heavily influence the correlation and regression line.

4. Dataset 4:

- Here, most of the X values are the same except one, making a vertical line. The Y values are quite spread out, with one extreme outlier that affects the overall statistics.

Why It’s Important

Anscombe's Quartet teaches a crucial lesson: relying solely on summary statistics can be misleading. It's essential to visualize data because:

- Outliers: Can dramatically affect the results.

- Patterns: Different patterns (linear, curved, etc.) can emerge only when plotted.

- True Relationship: The true nature of the data, whether it's linear, non-linear, or scattered, is visible only through graphs.

Key Takeaways

1. Always Graph Your Data: Before drawing conclusions, plot the data to see its shape and trends.

2. Summary Statistics Aren't Enough: Mean, variance, and correlation can't capture the entire story.

3. Look for Patterns and Outliers:** Visualisation helps in identifying anomalies and true relationships.

Anscombe's Quartet is a powerful reminder of the importance of data visualisation in understanding the true nature of data.

QUES 3 - What is Pearson's R?

Answer-

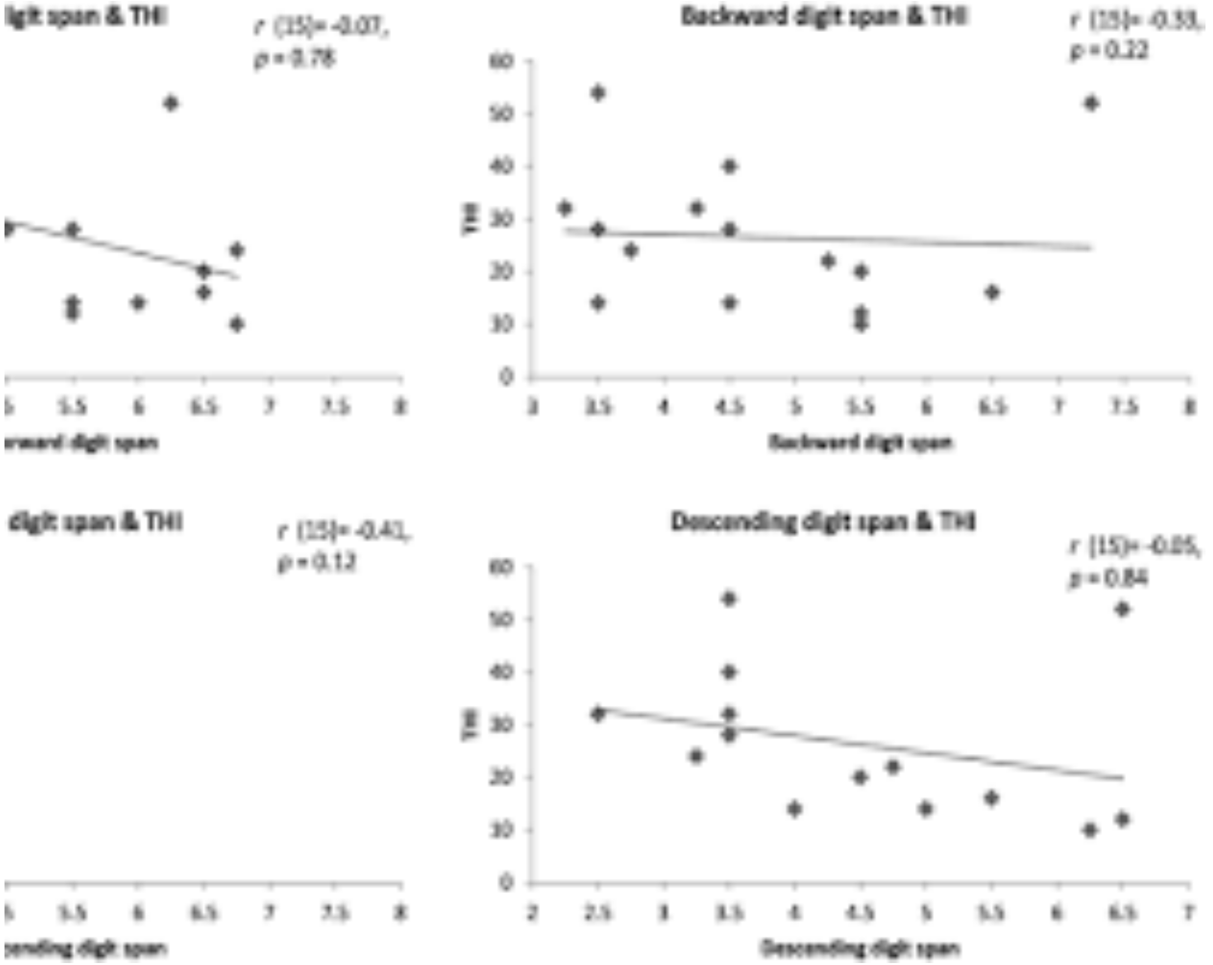
Pearson's r is a statistical measure that quantifies the linear relationship between two continuous variables. It provides information about both the strength and direction of the relationship.

Key Characteristics:

- **Range:** Values range from -1 to 1.
- **Interpretation:**
  - **-1:** Perfect negative correlation (as one variable increases, the other decreases)
  - **0:** No correlation between the variables

- **1:** Perfect positive correlation (as one variable increases, the other increases)

Visual Representation:



Assumptions:

- **Linearity:** The relationship between the variables is linear.
- **Normality:** Both variables are normally distributed.
- **Outliers:** There are no significant outliers in the data.

Calculation:

While the formula can be complex, statistical software packages typically handle the calculation.

When to Use Pearson's r:

- When you want to measure the strength and direction of a linear relationship between two continuous variables.
- When the data meets the assumptions of linearity, normality, and no outliers.

Limitations:

- Only measures linear relationships.
- Sensitive to outliers.
- Does not imply causation, only correlation.

In summary, Pearson's r is a valuable tool for understanding the relationship between two continuous variables, but it's essential to consider its assumptions and limitations when interpreting results.

QUES 4- What is scaling? Why is scaling performed? What is the difference between normalised scaling and standardised scaling?

Answer-

Scaling is a data preprocessing technique used to standardise the range of independent variables or features of data. In simple terms, scaling transforms the features into a common scale without distorting differences in the ranges of values.

Why is Scaling Performed?

Scaling is performed for several reasons:

1. **Improved Model Performance:** Many machine learning algorithms (e.g., gradient descent-based algorithms, k-nearest neighbours, and support vector machines) perform better and converge faster when the data is scaled.
2. **Fair Comparisons:** Features with different scales can disproportionately affect the model’s performance. Scaling ensures that all features contribute equally to the result.
3. **Reduced Bias:** Features with larger scales can dominate distance calculations and bias the results, especially in distance-based algorithms like k-means clustering and principal component analysis (PCA).

Difference Between Normalised Scaling and Standardised Scaling
Normalised Scaling:

- **Definition:** Normalisation (also known as Min-Max scaling) rescales the data to fit within a specific range, usually 0 to 1.
- **Formula:** [X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}]
- **Use Case:** Suitable for algorithms that assume bounded input values or where a uniform distribution is expected.
- **Example:** For a feature with values ranging from 10 to 100, normalisation would scale the values to range between 0 and 1.

Standardised Scaling:

- **Definition:** Standardisation (also known as Z-score normalisation) transforms the data to have a mean of 0 and a standard deviation of 1.
- **Formula:** [X_{std} = \frac{X - \mu}{\sigma}] Where (\mu) is the mean of the feature, and (\sigma) is the standard deviation.
- **Use Case:** Suitable for algorithms that assume Gaussian distribution of the data or when features have different units and scales.
- **Example:** For a feature with a mean of 50 and a standard deviation of 10, standardisation would transform the data so that it has a mean of 0 and a standard deviation of 1.

Summary:

- **Scaling** adjusts the range of features to a common scale.
- **Normalisation** rescales features to a fixed range (0 to 1), suitable for bounded data.
- **Standardisation** transforms features to have a mean of 0 and a standard deviation of 1, suitable for normally distributed data or data with different units and scales.

QUES 5- 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.

QUES 6-What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

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.