Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer:

From our analysis using box plots and bar plots, we've gathered some key insights:

1. Fall season appears to be the most popular time for bookings, with a significant increase in bookings observed from 2018 to 2019 across all seasons.
2. The months of May through October show the highest booking activity, peaking from early to mid-year and tapering off towards the end of the year.Number of booking for each month seems to have increased from 2018 to 2019.
3. Unsurprisingly, clear weather conditions correlate with higher booking numbers. And in comparison, to previous year, i.e. 2018, booking increased for each weather situation in 2019.
4. Wednesday through Saturday consistently see higher booking rates compared to the early days of the week.
5. Considering 0 as not holiday, booking were made on non-holiday days are more.
6. Slightly More bookings done on working day. But it seems nearly an equal distribution of bookings between working days and non-working days.
7. Overall, 2019 saw a notable increase in bookings compared to the previous year, indicating positive business growth.

Q2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

Answer :

Using **drop_first=True** is important because it reduces the number of dummy variables created from categorical data. By leaving out one dummy variable (usually the first one), we avoid potential problems of multicollinearity that can occur when variables are perfectly correlated. This approach makes our models easier to understand and more efficient, ensuring that each category is accurately represented without unnecessary duplication.

Syntax:

- **drop_first:** This is a boolean setting (default is False) that decides whether to create k-1 dummy variables out of k categorical levels by excluding the first level.

Example Explanation: For example, if we have a categorical column with three distinct values (let's say A, B, and C) and we create dummy variables without using drop_first=True, we would end up with three dummy variables. However, since the absence of A and B automatically indicates C, we only need two variables (one for A and one for B) to effectively cover all three categories. This simplification not only saves memory but also helps our models generalize better to new data.

In summary, using drop_first=True during dummy variable creation is crucial for enhancing model performance by reducing unnecessary complexity and correlations among dummy variables.

Q3. 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

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

I checked the assumptions of the Linear Regression Model based on these 5 criteria:

- **Normality of error terms**: The error terms should follow a normal distribution.
- **Multicollinearity check**: There shouldn't be significant multicollinearity among the variables.
- **Validation of linear relationship**: There should be a clear linear relationship among the variables.
- **Homoscedasticity**: The residuals should not show any noticeable pattern.
- **Independence of residuals**: There should be no autocorrelation among the residuals.

These checks help ensure that the Linear Regression Model is appropriately validated and can provide reliable predictions based on the data provided. Each of these criteria contributes to the overall robustness and accuracy of the model's predictions.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer :

To determine the top 3 features contributing significantly towards explaining the demand of shared bikes, we need to look at the absolute values of the coefficients in the linear regression equation. The larger the absolute value of the coefficient, the more impact that feature has on the prediction.

Let's list out the absolute values of the coefficients:

1. year: |0.2341|
2. holiday: |0.0963|
3. temp: |0.4777|
4. windspeed: |0.1481|
5. sep: |0.0910|
6. Lightsnowrain: |0.2850|
7. Misty: |0.0787|
8. spring: |0.0554|
9. summer: |0.0621|
10. winter: |0.0945|

Ordering these from highest to lowest, the top 3 features are:

1. temp: 0.4777
2. Lightsnowrain: 0.2850

3. year: 0.2341

Therefore, based on this linear regression model, the top 3 features contributing significantly towards explaining the demand of shared bikes are:

1. Temperature (temp)
2. Light snow or rain conditions (Lightsnowrain)
3. Year

Temperature has the largest impact, with a positive coefficient indicating that higher temperatures are associated with increased bike demand. Light snow or rain has a negative impact, suggesting fewer bike rentals in these weather conditions. The year also has a positive impact, which might indicate a general trend of increasing bike usage over time.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail. (4 marks)

Answer : Linear regression is a fundamental algorithm in machine learning and statistics used for predicting a continuous target variable based on one or more input features. Let me break down the linear regression algorithm in detail:

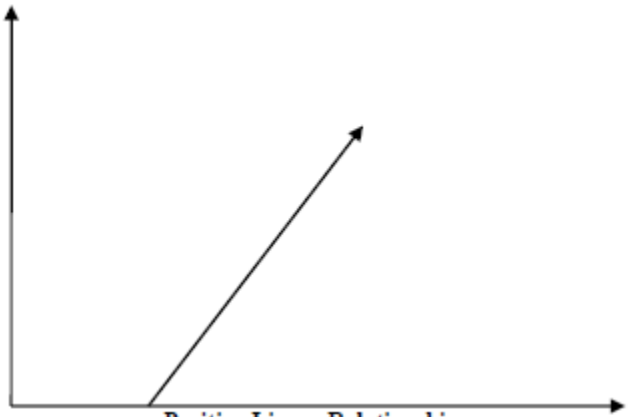
a) Basic Concept: Linear regression attempts to model the relationship between variables by fitting a linear equation to observed data. The simplest form is simple linear regression, with one independent variable: y = mx + b Where y is the dependent variable, x is the independent variable, m is the slope, and b is the y-intercept.

b) Multiple Linear Regression: When there are multiple independent variables, it becomes multiple linear regression: y = b0 + b1x1 + b2x2 + ... + bnxn Where b0 is the y-intercept and b1, b2, ..., bn are the coefficients for the independent variables x1, x2, ..., xn.

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

Positive Linear Relationship:

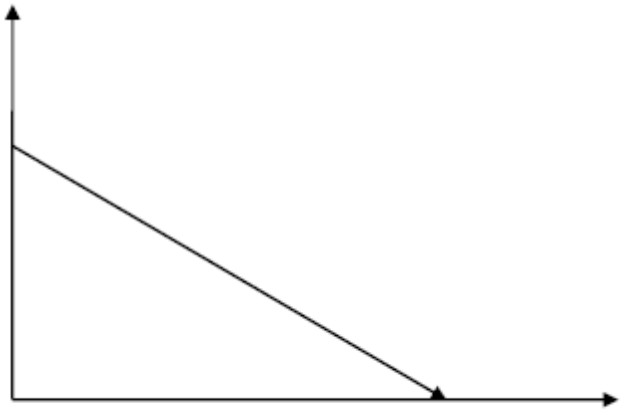
- A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph −



Positive Linear Relationship

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

c) Objective: The goal is to find the best-fitting line (or hyperplane in multiple dimensions) that minimizes the difference between predicted and actual values.
d) Ordinary Least Squares (OLS): This is the most common method for estimating the parameters of a linear regression model. It minimizes the sum of squared residuals (the differences between observed and predicted values).
e) Cost Function: The Mean Squared Error (MSE) is typically used as the cost function: MSE = (1/n) * Σ(y_actual - y_predicted)^2 Where n is the number of data points.
f) Gradient Descent: This is an optimization algorithm used to minimize the cost function by iteratively moving towards the minimum.
g) Assumptions:

- 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.
- Linearity: The relationship between X and Y is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: The variance of residual is the same for any value of X. There should be no visible pattern in residual values.
- Normality: For any fixed value of X, Y is normally distributed. Error terms should be normally distributed

h) Model Evaluation:

- R-squared (coefficient of determination): Indicates the proportion of the variance in the dependent variable that is predictable from the independent variable(s).
- Adjusted R-squared: Adjusts for the number of predictors in the model.
- F-statistic: Tests the overall significance of the model.
- t-statistics: Test the significance of individual coefficients.

i) Limitations:

- Assumes a linear relationship, which may not always be the case.
- Sensitive to outliers.
- Can struggle with multicollinearity (high correlation between independent variables).

j) Implementations: Linear regression can be implemented using various libraries in different programming languages, such as scikit-learn in Python, statsmodels in Python for more detailed statistical output, or built-in functions in R.

Linear regression serves as a foundation for understanding more complex machine learning algorithms and is widely used in various fields for prediction and understanding relationships between variables.

Q2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

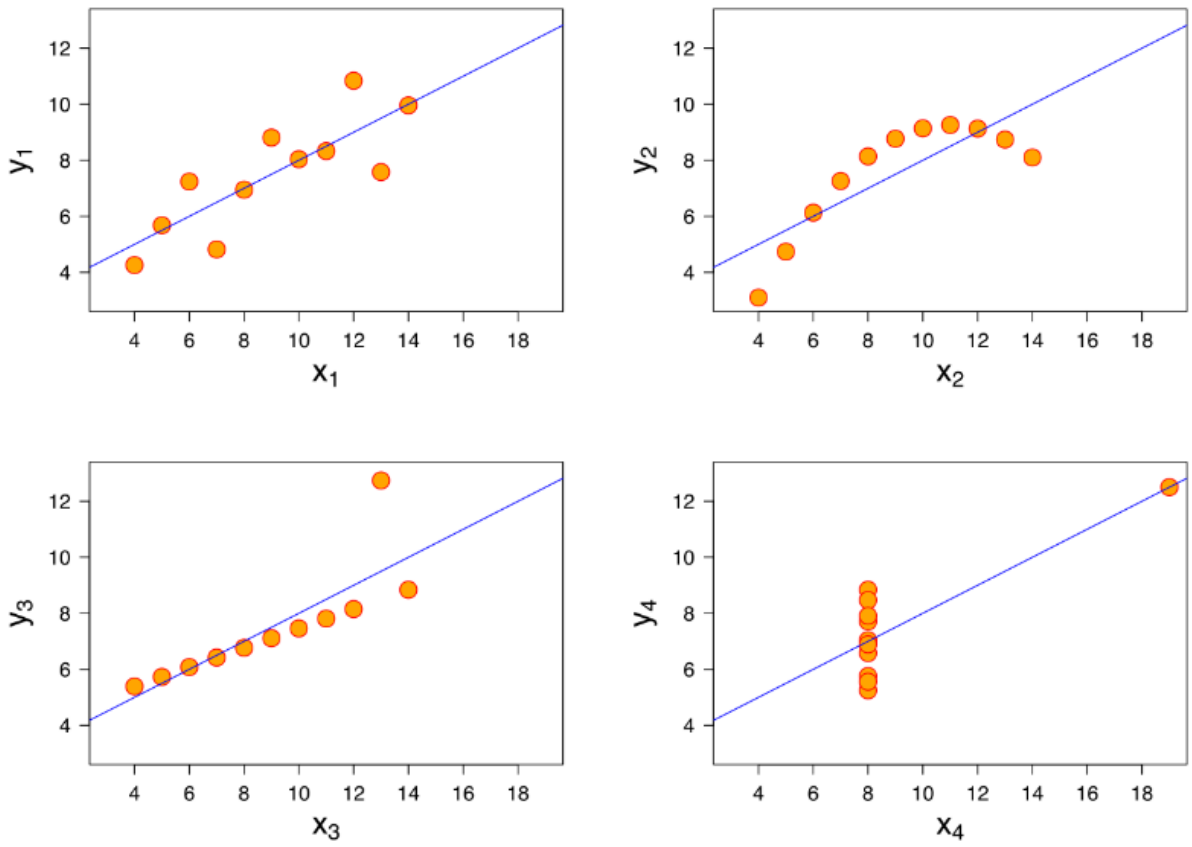
Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	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 (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.

Q3. What is Pearson's R? (3 marks)

Answer:

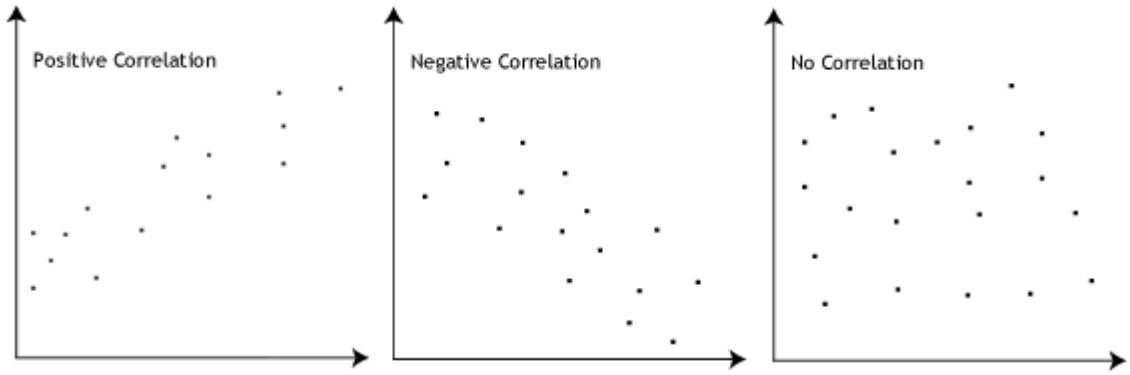
Pearson's R, also known as Pearson's correlation coefficient or simply the correlation coefficient, is a measure of the linear correlation between two variables. It's named after Karl Pearson, who developed it in the 1890s. Here's a detailed explanation of Pearson's R:

a) Definition: Pearson's R measures the strength and direction of the linear relationship between two continuous variables.

b) Range: It ranges from -1 to +1, where:

- +1 indicates a perfect positive linear correlation

- 0 indicates no linear correlation
- -1 indicates a perfect negative linear correlation



c) Formula: $R = \Sigma((X - \bar{X})(Y - \bar{Y})) / \sqrt{(\Sigma(X - \bar{X})^2 * \Sigma(Y - \bar{Y})^2)}$ Where X and Y are the variables, $\bar{X}$ and $\bar{Y}$ are their means.
d) Interpretation:
e) to 0.19: Very weak correlation

- 0.20 to 0.39: Weak correlation
- 0.40 to 0.59: Moderate correlation
- 0.60 to 0.79: Strong correlation
- 0.80 to 1.00: Very strong correlation (The same applies for negative values)

f) Key Properties:

- It's symmetric: the correlation of X with Y is the same as Y with X.
- It's dimensionless and scale-invariant.
- It only measures linear relationships; it may miss non-linear relationships.

g) Assumptions:

- Variables should be continuous.
- There should be a linear relationship between variables.
- Variables should be normally distributed.
- There should be no significant outliers.

h) Limitations:

- Sensitive to outliers.
- Only measures linear relationships.
- Doesn't imply causation.
- Can be misleading if the relationship is non-linear.

i) Use Cases:

- In statistics for hypothesis testing.
- In machine learning for feature selection and dimensionality reduction.
- In finance for portfolio management and risk assessment.
- In social sciences for analyzing relationships between variables.

j) Relationship to R-squared: In simple linear regression, the square of Pearson's R equals the coefficient of determination (R-squared).
k) Variations:

- Spearman's rank correlation: Used for ordinal data or non-linear relationships.
- Point-biserial correlation: Used when one variable is dichotomous.

Pearson's R is a fundamental tool in statistics and data analysis, providing a simple way to quantify the strength and direction of linear relationships between variables. However, it's important to use it in conjunction with other analyses and to be aware of its limitations.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer:

Scaling is an important preprocessing technique in data analysis and machine learning. Scaling is the process of transforming numerical features in a dataset to a common scale. It involves changing the range of values for different features so they are comparable and on a similar scale.

Scaling is performed for several reasons:

a) To ensure all features contribute equally to the analysis or model training.
b) To improve the performance and convergence of many machine learning algorithms.
c) To prevent features with larger magnitudes from dominating those with smaller magnitudes.
d) To make the interpretation of certain models easier. e) To reduce the impact of outliers in some cases.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 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.

Difference between normalized scaling and standardized scaling:

Normalized Scaling (also known as Min-Max scaling):

- Scales features to a fixed range, typically between 0 and 1.
- Formula: X_scaled = (X - X_min) / (X_max - X_min)
- Preserves zero values and does not center the data.
- Useful when you need values in a bounded interval.

Standardized Scaling (also known as Z-score normalization):

- Transforms data to have a mean of 0 and a standard deviation of 1.
- Formula: X_scaled = (X - μ) / σ, where μ is the mean and σ is the standard deviation.
- Centers the data around zero and scales it to unit variance.
- Useful when you need to compare features that have different units or scales.

Key differences:

- **Range:** Normalized scaling produces values in a fixed range (usually [0,1]), while standardized scaling doesn't have a bounded range.
- **Central tendency:** Standardized scaling centers the data around zero, while normalized scaling doesn't necessarily do so.
- **Outlier sensitivity:** Normalized scaling can be more sensitive to outliers as it's based on the minimum and maximum values.
- **Interpretation:** Standardized scaling allows for easier interpretation in terms of standard deviations from the mean.

The choice between these methods depends on the specific requirements of your analysis or machine learning algorithm, the nature of your data, and the goals of your project.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

The Value Inflation Factor (VIF) becoming infinite is an important issue in multicollinearity analysis. Here's why this happens:

a. Perfect Multicollinearity: VIF becomes infinite when there is perfect multicollinearity between two or more independent variables in a regression model. Perfect multicollinearity occurs when one independent variable can be expressed as an exact linear combination of other independent variables.

b. Mathematical Explanation: VIF for a variable is calculated as: $VIF = 1 / (1 - R^2)$ Where R^2 is the coefficient of determination when this variable is regressed against all other independent variables. When perfect multicollinearity exists, R^2 becomes 1, making the denominator $(1 - R^2)$ equal to zero. Division by zero results in an infinite VIF.

c. Causes:

- Duplicate variables in the dataset
- Variables that are linear combinations of others
- Overparameterization of the model
- Inclusion of both a variable and its reciprocal
- Using dummy variables incorrectly (e.g., including all categories)

d. Implications:

- The regression model becomes unstable
- Coefficient estimates become unreliable
- Standard errors of the coefficients increase dramatically
- The model's predictive power is compromised

a) Solutions:

- Remove one of the perfectly correlated variables
- Combine the correlated variables into a single feature

- Use regularization techniques like Ridge regression
- Collect more data if possible
- Respecify the model to avoid the linear dependency

Q6. 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.

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.