Assignment-based Subjective Questions

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

   Ans:-

   - The months of June, September, and August typically experience the highest booking rates. Observing the trend on the bar plot, it's evident that bookings show an upward trend at the beginning of the year, peaking around June, September, and August, and then gradually decreasing from October onwards. Additionally, there was an overall increase in bookings from 2018 to 2019.
   - There doesn't appear to be a significant difference in the increase or decrease of bookings based on whether it's a working day or not. However, there was an overall increase in bookings from 2018 to 2019.
   - It appears that Friday ,Saturday ,Sunday and Thursday tend to have higher booking rates compared to other days of the week. Additionally, there was an overall increase in bookings from 2018 to 2019.
   - Bookings appear to be higher when it is not a holiday. Additionally, there was an overall increase in bookings from 2018 to 2019.
   - There appears to be a higher number of bookings on clear weather days compared to other weather conditions.
   - Fall season typically sees higher bookings compared to other seasons, and there was an increase in bookings from 2018 to 2019.
   - 2019 seems to have more bookings than 2018, which show good progress in terms of business.

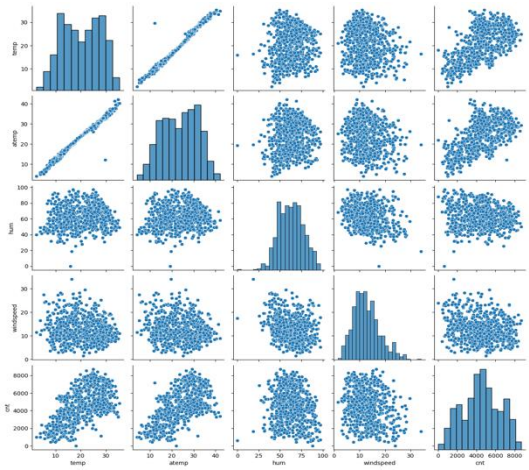
2. Why is it important to use **drop_first=True** during dummy variable creation?(2 mark)

   Ans:-

   - The purpose of creating dummy variables for a categorical variable with 'n' levels is to generate 'n-1' new columns, each indicating the presence or absence of a particular level using binary values (0 or 1). Therefore, setting drop_first=True ensures that the resulting dummy variables correspond to 'n-1' levels, reducing correlation among them. For instance, if there are 3 levels, drop_first will eliminate the first column.

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

   Ans:-



- The variables 'temp' and 'atemp' exhibit the strongest correlation with the target variable 'cnt' compared to the other variables.

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

   Ans:-

   - Residual Analysis: I examined the residuals (the differences between actual and predicted values) to ensure they were normally distributed around zero with constant variance. This was done by plotting histograms or Q-Q plots of the residuals.
   - Homoscedasticity: I checked for homoscedasticity by plotting the residuals against the predicted values. A random scatter of points around zero indicated homoscedasticity, while patterns or trends suggested heteroscedasticity.
   - Multicollinearity: I assessed multicollinearity among predictor variables using methods like variance inflation factor (VIF) or correlation matrices. VIF values below 5 or correlation coefficients below 0.7 indicated acceptable levels of multicollinearity.
   - Linearity: I verified the linearity assumption by plotting the predicted values against the actual values using ''sm.graphics.plot_ccpr''. A strong linear relationship indicated that the model adequately captures the linear relationship between predictors and the target variable.
   - Independence of Residuals: I confirmed the independence of residuals by examining autocorrelation using techniques like the Durbin-Watson test or residual plots against time or other predictors.

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

   Ans:-

Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

- temp
- winter
- sep

General Subjective Questions

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

   Ans:-

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable and one or more independent variables. It is widely employed for predictive analysis and understanding the association between variables. Here's a detailed

explanation of the linear regression algorithm:

Basic Concept:

- Linear regression assumes that there is a linear relationship between the independent variables(X)and the dependent variable (Y). The relationship can be expressed as:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn + \varepsilon$$

- Y is the dependent variable.
- $X1, X2, \ldots, Xn$ are the independent variables.
- $\beta 0, \beta 1, \beta 2, \ldots, \beta n$ are the coefficients, representing the intercept and slopes of the regression line.
- ε is the error term, representing the difference between the observed and predicted values of Y.

Objective :

The main objective of linear regression is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the sum of squared differences between the observed and predicted values of the dependent variable.

Steps Of Linear Regression :

1. **Data Collection:** Gather the dataset consisting of both the independent and dependent variables.
2. **Data Preprocessing:** This step involves handling missing values, outliers, and scaling the features if necessary.
3. **Splitting the Data:** Divide the dataset into training and testing sets. The training set is used to train the model, while the testing set is used to evaluate its performance.
4. **Model Training:** Use the training data to estimate the coefficients (β) of the regression equation. This is typically done using methods like Ordinary Least Squares (OLS), Gradient Descent, or other optimization techniques.
5. **Model Evaluation:** Once the model is trained, evaluate its performance using the testing data. Common evaluation metrics for linear regression include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared ($R2$), etc.
6. **Prediction:** After evaluating the model, it can be used to make predictions on new or unseen data.

Assumptions of Linear regression :

- Linearity: The relationship between the independent and dependent variables should be linear.
- Independence: The residuals (errors) should be independent of each other.
- Homoscedasticity: The variance of the residuals should be constant across all levels of the independent variables.

- Normality: The residuals should follow a normal distribution.
- No Multicollinearity: The independent variables should not be highly correlated with each other.

Types of Linear Regression :

- Simple Linear Regression: When there is only one independent variable.
- Multiple Linear Regression: When there are multiple independent variables.
- Polynomial Regression: When the relationship between the variables is not linear but can be approximated by a polynomial function.
- Ridge Regression and Lasso Regression: Variants of linear regression that include regularization to prevent overfitting.

Applications :

Linear regression finds applications in various fields including economics, finance, healthcare, social sciences, and more. It is used for prediction, forecasting, trend analysis, and understanding the relationship between variables.

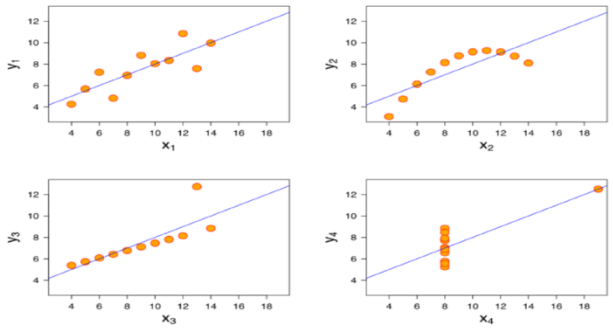
2. Explain the Anscombe's quartet in detail. (3 marks)

Ans:-

- Anscombe's quartet is a fascinating statistical phenomenon that highlights the importance of data visualization in understanding and interpreting data. It consists of four datasets that have nearly identical statistical properties, yet appear very different when graphed. This quartet was introduced by the statistician Francis Anscombe in 1973 to emphasize the dangers of relying solely on summary statistics (such as mean, variance, correlation, etc.) without visually inspecting the data.

Overview:

- Dataset I: This dataset consists of linearly related variables. When plotted, the relationship between the variables appears clearly linear, with a correlation coefficient close to 1.
- Dataset II: This dataset also shows a linear relationship between variables, but with an outlier that significantly affects the regression line. Despite the outlier, the correlation coefficient remains similar to Dataset I.
- Dataset III: Unlike the first two datasets, Dataset III exhibits a non-linear relationship between variables. When plotted, it forms a clear quadratic curve, but the correlation coefficient is still the same as in the previous datasets.
- Dataset IV: This dataset is designed to demonstrate the impact of a single data point on correlation. It consists of three data points with identical x-coordinates and varying y-coordinates, except for one outlier. The correlation coefficient is 0, yet the data points appear somewhat linear when plotted.



- The first scatter plot (top left) appears to be a simple linear relationship.
- The second graph (top right) is not distributed normally; while there is a relation between them is not linear.
- In the third graph (bottom left), the distribution is linear, but should have a different regression line the calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

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

Ans:-

- Pearson's r, often referred to as Pearson's correlation coefficient or simply Pearson's r, is a measure of the linear relationship between two variables in a dataset. It quantifies the strength and direction of the relationship between two continuous variables.

Pearson's r ranges from -1 to 1:

- A r value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.
- A r value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.
- A r value of 0 indicates no linear relationship between the variables.

- In other words, r measures how much one variable changes when the other variable changes. If r is close to 1 or -1, it indicates a strong linear relationship, while values closer to 0 indicate a weaker linear relationship or no relationship at all.

- It's important to note that Pearson's r only measures linear relationships. It may not accurately capture non-linear relationships between variables. Additionally, Pearson's r is sensitive to outliers and may not be appropriate for datasets with non-normally distributed variables or heteroscedasticity.

The formula for Pearson's r is:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where:

- xi and yi are the individual data points
- $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively.

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

Ans:-

Scaling in the context of technology or business typically refers to the process of adjusting the capacity, size, or scope of a system or operation to accommodate increased demand or growth. This could involve increasing the resources (such as hardware, software, or personnel) to handle greater workload, expanding infrastructure to serve more users, or optimizing processes to improve efficiency and effectiveness.

Scaling is performed for several reasons:

1. **Meeting Increased Demand**: As a product or service gains popularity or usage, the demand for it often grows. Scaling ensures that the system can handle the increased load without experiencing performance degradation or downtime.
2. **Improving Performance**: Scaling allows for better performance under heavy usage. By distributing the workload across multiple resources or optimizing processes, the system can maintain responsiveness and speed even as the demand increases.
3. **Enhancing Reliability and Resilience**: Scaling can involve redundancy and failover mechanisms, which improve the system's reliability and resilience to failures. Redundant resources can take over if one component fails, ensuring continuous operation.
4. **Supporting Business Growth**: Scaling is essential for businesses aiming to grow. It enables them to expand their customer base, enter new markets, and handle larger volumes of transactions or interactions.
5. **Cost Optimization**: Scaling can also help optimize costs by ensuring that resources are allocated efficiently. Scaling up or down based on demand can prevent overprovisioning of resources, thereby reducing unnecessary expenses.

Difference between normalized scalingand standardized scaling

Aspect	**Normalized Scaling**	**Standardized Scaling**
Definition	Adjusting values to a common scale (e.g., 0 to 1 range).	Adjusting values to have a common mean and standard deviation.
Purpose	Allows for comparisons between variables with different units or scales.	Centers the data around a mean of 0 and a standard deviation of 1.
Formula	xnormalized=max(X) / min(X)x−min(X)	*xstandardized*= $\frac{\chi-\bar{X}}{\sigma}$
Range of Values	0 to 1, where 0 represents the minimum value and 1 represents the maximum value.	Can vary, but typically centered around 0 with values ranging above and below.
Effect on Distribution	Preserves the shape of the original distribution.	Centers the distribution around 0 and adjusts the spread.
Outliers	Sensitive to outliers, as they can disproportionately affect the range.	Less sensitive to outliers compared to normalized scaling.
Interpretation	Useful when the absolute values are not as important as their relative positions.	Useful when data needs to be centered and spread evenly for certain statistical methods.

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

(3 marks)

Ans:-

- The Variance Inflation Factor (VIF) is a measure used in regression analysis to detect multicollinearity, which occurs when predictor variables in a regression model are highly correlated with each other. A high VIF value indicates that a predictor variable is highly correlated with other predictor variables in the model, making it difficult to isolate the effect of that particular variable on the response variable.

- VIF is calculated as $\frac{1}{1-R^2}$, where R^2 is the coefficient of determination obtained by regressing a predictor variable on all other predictor variables in the model.

- In some cases, the VIF value can become infinite. This happens when the coefficient of determination (R^2) for a predictor variable regressed on other predictor variables is equal to 1. When R^2=1, it means that the predictor variable is a perfect linear combination of the other predictor variables in the model, indicating perfect multicollinearity. In such situations, the VIF value becomes infinite because the denominator in the VIF formula becomes zero (1 - 1 = 0).

- Perfect multicollinearity can arise due to various reasons, such as including redundant variables in the model, linear dependencies among predictor variables, or data errors.

- To address this issue, it's essential to identify and remove redundant variables or linear dependencies from the model before performing regression analysis. This can involve techniques like dropping one of the correlated variables, transforming variables, or using regularization methods like ridge regression. Additionally, careful variable selection and preprocessing of the data can help mitigate multicollinearity issues and prevent infinite VIF values.

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

(3 marks)

Ans:-

A Q-Q (quantile-quantile) plot is a graphical tool used to assess whether a dataset follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the dataset's empirical distribution (observed data) to the quantiles of a theoretical distribution (expected data). The Q-Q plot is particularly useful for visually inspecting the similarity between the empirical distribution and the theoretical distribution.

Here's how a Q-Q plot is constructed:

Sort the values of the dataset in ascending order.

Calculate the quantiles of the dataset (the percentiles of the sorted data).

Calculate the expected quantiles based on the chosen theoretical distribution.

Plot the observed quantiles against the expected quantiles.

In a Q-Q plot, if the dataset follows the theoretical distribution closely, the points on the plot will fall approximately along a straight line. Any deviation from a straight line indicates a departure from the assumed distribution.

In linear regression, Q-Q plots are often used to assess the assumption of normality of residuals. Residuals are the differences between observed and predicted values in a regression analysis. The assumption of normality of residuals is crucial for making valid inferences and predictions from the regression model.

Here's how a Q-Q plot is used in linear regression:

Calculate the residuals of the regression model by subtracting the predicted values from the observed values.

Construct a Q-Q plot of the residuals.

If the residuals follow a normal distribution, the points on the Q-Q plot will approximately fall along a straight line.

Deviations from a straight line suggest non-normality of residuals, indicating potential issues with the regression model.

The importance of a Q-Q plot in linear regression lies in its ability to visually identify departures from the assumption of normality. If the Q-Q plot reveals significant deviations from a straight line, further investigation may be needed to understand the underlying reasons and potentially improve the regression model. Addressing non-normality of residuals can lead to more accurate and reliable regression analyses and interpretations.