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)

Ans 1.) During the examination of categorical variables through bar plots, I drew the following inferences-

- Number of bikes rented in 2019 are more than that of 2018
- Majority of the bookings happen from May-September.
- In comparison to 2018, bike bookings have seen growth for every month throughout 2019
- The bulk of reservations occurs in the latter part of the week, specifically from Thursday through Sunday. This suggests that people tend to make their bookings closer to the weekend.
- Bookings tend to be higher on non-holiday days, as it aligns with the idea that people prefer to stay at home or spend quality time with their families on holidays
- "Fall" season has the highest booking followed by "Summer" season
- The most popular weather condition for bookings has been "Clear" weather, with "Misty" weather coming in as the second most preferred choice for bike bookings

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

Ans 2.) Using “drop_first=True” during dummy variable creation is a common practice in regression modeling to improve model stability, interpretability, and efficiency, while also avoiding multicollinearity issues that can arise when including all dummy variables. It simplifies the interpretation of coefficients and helps prevent potential problems in the regression analysis.

For e.g., Suppose we have a dataset containing a categorical feature "Color" with three categories: "Red," "Green," and "Blue," and we want to use this feature in a linear regression model.

```
#Creating dummy variables for "month"
df_Color = pd.get_dummies(df.Color, drop_first = True)
```

Dumm variables will be created as below where “Color” Red is represented by Geen = 0 and Blue = 0

Green	Blue
0	0
1	0
0	1

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

Ans 3.) Looking at the pair-plot of all the numerical variables, the "temp" and "feeling_temp" (Originally "atemp") variables have high correlation with the target variable "count_of_bike".

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

Ans 4.) Once the model was constructed using the training dataset, I assessed the assumptions of Linear Regression in the following manner:

- Assumption 1: Linear relationship
  - One crucial assumption posits the existence of a linear relationship between the independent and dependent variables
  - ***Steps taken for validating the assumption***- Plotting the pair plot between a few predictor variables and the target variable "count_of_bike"

- Assumption 2: No auto-correlation i.e., Assumption of Error Terms Being Independent
  - The error terms, or residuals, exhibit independence from one another, indicating that there is no correlation between successive error terms in the time series data
  - ***Steps taken for validating the assumption***- For this we performed The Durbin-Watson statistic test

    The value typically falls between 0 and 4, where:
    - A value close to 2 indicates no significant autocorrelation
    - A value less than 2 suggests positive autocorrelation (residuals are correlated)
    - A value greater than 2 suggests negative autocorrelation (residuals are inversely correlated)

- Assumption 3: No Multicollinearity
  - It is important that the independent variables do not exhibit correlations. If multicollinearity is present among the independent variables, it becomes difficult to make accurate predictions using the model
  - ***Steps taken for validating the assumption***- For this, we calculated the VIF (Variation Inflation Factor)
    - VIF <=5 implies no multicollinearity.
    - VIF >5 implies serious multicollinearity.

- Assumption 4: Assumption of Homoscedasticity
  - Homoscedasticity refers to the condition where the residuals maintain a consistent variance across all levels of the independent variable
  - ***Steps taken for validating the assumption***- For this, we created a scatter plot to assess homoscedasticity by plotting the residuals against the fitted values
    - In a scatter plot showing homoscedasticity, the data points are spread out evenly without a discernible pattern

- Assumption 5: Assumption of Normally Distributed Error Terms
  - We need to prove that the distribution of error terms is normally distributed
  - ***Steps taken for validating the assumption***- For this, we plotted a distplot of the "residuals" (y_train_pred - y_train) and made sure that the errors are normally distributed across zero.

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

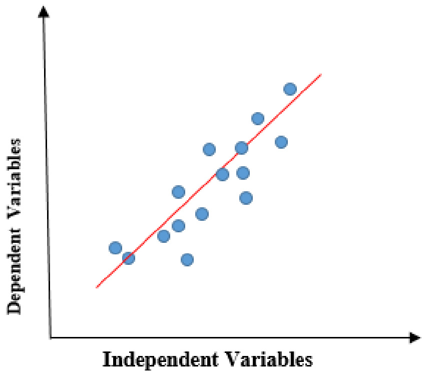
Ans 5.) According to the best-fit model, following are the top 3 features along with their parameter value that are contributing significantly towards explaining the demand of the shared bikes:

- temp (0.4910)
- year (0.2336)
- Winter (0.0817)

General Subjective Questions

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

Ans 1.) Linear regression is a widely used statistical method for modeling the relationship between a dependent variable (Y) and one or more independent variables (X). It's called "linear" because it assumes that this relationship can be approximated by a straight line.



Line of Best Fit: The goal of linear regression is to find the best-fitting straight line through the data points. This line is represented by the equation:

For Simple Linear Regression Model-

Also called simple regression, linear regression establishes the relationship between two variables. Linear regression is graphically depicted using a straight line with the slope defining how the change in one variable impacts a change in the other. The y-intercept of a linear regression relationship represents the value of one variable when the value of the other is 0.

$y = mx + c$

$y = mx + c \quad \text{->} \; y = a0 + a1x$

y= Dependent Variable

x= Independent Variable

a0= intercept of the line

a1 = Linear regression coefficient

For Multiple Regression Model-

For complex connections between data, the relationship might be explained by more than one variable. In this case, an analyst uses multiple regression which attempts to explain a dependent variable using more than one independent variable.

$Y = \beta0 + \beta1X1 + \beta2X2 + ... + \beta n*Xn$

Y: The dependent variable you're trying to predict.
X1, X2, ... Xn: The independent variables.
β0, β1, β2, ... βn: The coefficients (parameters) of the model.

Assumptions in Linear Regression-

- Assumption 1: Linear relationship
    - One crucial assumption posits the existence of a linear relationship between the independent and dependent variables

- Assumption 2: No auto-correlation i.e., Assumption of Error Terms Being Independent
    - The error terms, or residuals, exhibit independence from one another, indicating that there is no correlation between successive error terms in the time series data The value typically falls between 0 and 4, where:
        - A value close to 2 indicates no significant autocorrelation
        - A value less than 2 suggests positive autocorrelation (residuals are correlated)
        - A value greater than 2 suggests negative autocorrelation (residuals are inversely correlated)

- Assumption 3: No Multicollinearity
    - It is important that the independent variables do not exhibit correlations. If multicollinearity is present among the independent variables, it becomes difficult to make accurate predictions using the model
        - VIF <=5 implies no multicollinearity.
        - VIF >5 implies serious multicollinearity.

- Assumption 4: Assumption of Homoscedasticity
    - Homoscedasticity refers to the condition where the residuals maintain a consistent variance across all levels of the independent variable

- Assumption 5: Assumption of Normally Distributed Error Terms
    - We need to prove that the distribution of error terms is normally distributed across zero

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

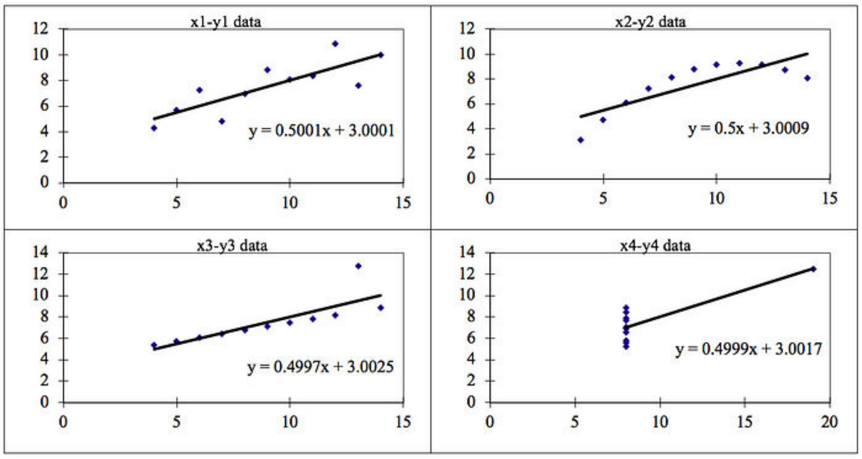
Ans 2.) Anscombe's quartet is a famous dataset in statistics that consists of four sets of small data points. What makes this dataset remarkable is that each of the four sets has nearly identical simple descriptive statistics (like mean, variance, and correlation) but looks entirely different when graphed. It was created by the statistician Francis Anscombe in 1973 to emphasize the importance of graphing data and not relying solely on summary statistics.

Summary Statistics-
When you calculate basic summary statistics for each dataset, such as mean, variance, correlation, and linear regression parameters, you'll find that they are remarkably similar across all four datasets. This is what makes Anscombe's quartet intriguing. People who rely solely on these summary statistics might think the datasets are essentially the same.

Graphical Representation-

The real lesson of Anscombe's quartet becomes apparent when you graph each dataset. When you create scatterplots for each dataset, you see that they have vastly different patterns and relationships.

- Dataset I shows a clear linear relationship
- Dataset II shows a non-linear, curved relationship
- Dataset III appears to have a linear relationship but is heavily influenced by an outlier
- Dataset IV has no clear relationship; the points are scattered randomly



Anscombe's quartet highlights the importance of visualizing data. It demonstrates that summary statistics alone can be deceiving and do not capture the full story of the data. It's a reminder that data analysis should involve not only quantitative measures but also visual exploration.

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

Ans 3.) Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. In other words, it tells you how closely the data points in a scatterplot cluster around a straight line.

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

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

Range-

Pearson's r can range from -1 to 1:

- If r = 1, it indicates a perfect positive linear relationship, meaning that as one variable increases, the other also increases linearly
- If r = -1, it indicates a perfect negative linear relationship, meaning that as one variable increases, the other decreases linearly
- If r = 0, it suggests no linear relationship; the variables are not linearly related

Interpretation-

- Positive r values (between 0 and 1) indicate a positive linear relationship. As one variable increases, the other tends to increase.
- Negative r values (between -1 and 0) indicate a negative linear relationship. As one variable increases, the other tends to decrease.
- The closer r is to 1 or -1, the stronger the linear relationship. The closer it is to 0, the weaker the linear relationship.

Limitations-

- Pearson's r only measures linear relationships. It may not capture nonlinear associations.
- It is sensitive to outliers, which can distort the correlation value.
- A correlation does not imply causation; just because two variables are correlated does not mean one causes the other

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

Ans 4.) Feature scaling is a preprocessing technique in data analysis and machine learning where you transform the values of your dataset's features (variables) to fit within a specific scale or range. The primary goal of feature scaling is to make sure that the features are on a similar scale, as this can be beneficial for various algorithms and analytical methods.

Why is Feature Scaling Performed?

Feature scaling is performed for following reasons:

- Equalize Variable Impact: Many machine learning algorithms are sensitive to the scale of the input features. Features with larger scales may have a disproportionate impact on the model's predictions. Scaling helps ensure that all features contribute equally to the model
- Convergence Speed: Some optimization algorithms, like gradient descent, converge faster when features are scaled. This can significantly reduce training time for machine learning models

Difference between normalized scaling and standardized scaling-

Aspect	Normalized Scaling (Min-Max Scaling)	Standardized Scaling (Z-score Scaling)
Formula	X_new = (X - X_min) / (X_max - X_min)	X_new = (X - mean)/Std
Range	Typically scales data to [0, 1]	Transforms data to have mean = 0, standard deviation = 1
Impact on Distribution	Preserves the shape of the original distribution	Transforms the distribution to have a mean of 0 and a standard deviation of 1
Centering and spread	Does not center data around 0 or adjust spread	Centers data around 0 and adjusts spread to have a standard deviation of 1
Use Cases	Useful when you have a specific range constraint in mind and want to preserve the original distribution shape	Preferred when you don't have a specific range requirement, and you want to maintain the shape of the original distribution
Sensitivity to Outliers	Sensitive to outliers as they can disproportionately affect the range	Less sensitive to outliers due to the use of the mean and standard deviation
Common Application	Data visualization, neural networks with specific input range requirements	Algorithms assuming normal distribution, clustering (k-means), and distance-based models

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

Ans 5.) The Variance Inflation Factor (VIF) is a statistical measure used to assess multicollinearity in regression analysis. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other. VIF helps identify the extent to which multicollinearity is present in a regression model.

Value of VIF is infinite-

A VIF of infinity (or very large values) typically occurs when there is perfect multicollinearity in the model. Perfect multicollinearity means that one or more independent variables in the regression model are perfectly linearly related to each other, making it impossible for the regression algorithm to estimate separate coefficients for them.

Here are some common scenarios that can lead to infinite VIF values:

- Perfect Linear Relationships: If two or more independent variables in your regression model are perfectly correlated, meaning that one can be expressed as an exact linear combination of the others, this leads to multicollinearity. For example, if you have two variables X1 and X2, and X2 is always equal to 2*X1, it creates perfect multicollinearity.

- Dummies and Interactions: In some cases, when dummy variables (binary variables) and their interactions are used in a regression model, perfect multicollinearity can occur if certain combinations of dummies are always equal to each other.

- Overparameterization: Having too many predictors relative to the number of observations in your dataset can also lead to perfect multicollinearity. If the number of predictors approaches or exceeds the number of data points, the model may be overparameterized.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans 6.) A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It compares the quantiles of the observed data against the quantiles of the expected theoretical distribution. The Q-Q plot helps you visually evaluate how well the data fits the assumed distribution and identifies deviations from the expected pattern.

Use of Q-Q Plot in Linear Regression-

- Normality Assumption: In linear regression and many other statistical analyses, one common assumption is that the residuals (the differences between observed and predicted values) follow a normal distribution. This assumption is important because it impacts the validity of hypothesis tests, confidence intervals, and the overall reliability of regression results.

- Checking Residual Normality: A Q-Q plot is often used to check whether the residuals of a linear regression model are approximately normally distributed. By plotting the residuals on the y-axis against the quantiles of a normal distribution on the x-axis, you can visually assess whether the points follow a straight-line pattern.

- Interpretation: In a Q-Q plot, if the points closely follow a straight line, it indicates that the residuals are normally distributed. Deviations from the straight line suggest departures from normality.

Importance of Q-Q Plot in Linear Regression-

- Assumption Validation: Linear regression models assume that the residuals are normally distributed. The Q-Q plot provides a graphical means to validate this assumption. If the Q-Q plot shows a reasonably straight line, it suggests that the assumption holds.

- Detecting Non-Normality: If the Q-Q plot deviates from a straight line, it indicates that the residuals do not follow a normal distribution. This is crucial information because non-normality can affect the accuracy and validity of statistical inference, including p-values and confidence intervals.

- Model Improvement: Identifying departures from normality can guide you in improving the linear regression model. You may consider transforming the response variable or adding predictor variables to account for the non-normality in the residuals.

- Outlier Detection: Q-Q plots can help in detecting outliers in the data. Outliers may appear as points that deviate significantly from the expected straight-line pattern.