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?**
   - *Month variable* – Analysis of the "Month" variable reveals a seasonal trend in bike bookings. Bookings peak between April and October, with a median exceeding 4,000 per month. This suggests a strong correlation between month and booking volume, potentially indicating seasonality as a significant factor influencing demand.
   - *Weathersit variable* – Analysis of the "weathersit" variable reveals a positive correlation with bike bookings. The median booking volume is highest when the weather is "Clear," reaching nearly 5,000 bookings per month. This suggests weather conditions, particularly clear skies, significantly influence demand.
   - *Workingday variable* – The "workingday" variable shows a potential influence on bike bookings. The median booking volume is close to 5,000 on weekdays, suggesting weekdays might see higher demand compared to weekends.
   - *Season variable* – Analysis of the "season" variable suggests a seasonal pattern in bike bookings. Bookings are highest during "Season fall" with a median exceeding 5,000 per month. Both "Season Summer" and "Season Winter" also show significant booking activity. This indicates a strong correlation between season and booking volume, suggesting seasonality as a key factor influencing demand.

2. **Why is it important to use drop_first=True during dummy variable creation?**
   - It's important to use drop_first=True during dummy variable creation to avoid a statistical issue called multicollinearity
     i. Multicollinearity – Multicollinearity occurs when there's a high degree of correlation between two or more independent variables in a regression model. This creates redundancy and makes it difficult to isolate the true effect of each variable on the dependent variable.
     ii. Dummy Variables and Redundancy – When you create dummy variables for a categorical variable with n categories, you end up with n-1 dummy variables. This is because one dummy variable can be mathematically recreated from the others.
     iii. Drop_First Solves Redundancy – By setting drop_first=True, you essentially remove one of the dummy variables. This eliminates the perfect linear relationship between the remaining dummy variables and prevents multicollinearity.
     iv. Impact on Model – Multicollinearity can lead to several problems in regression models:
        1. Unreliable Coefficients: The coefficients (measures of the effect of each variable on the dependent variable) become unreliable and difficult to interpret.
        2. Increased Variance: The variance of the coefficients increases, making it harder to determine if a relationship between a variable and the dependent variable is statistically significant.
     v. In summary, using drop_first=True when creating dummy variables helps to avoid multicollinearity, leading to a more reliable and interpretable regression model.

- For Example – take Weekday, in our data set weekday has 0 to 6 values correspondingly 0: 'Sunday', 1: 'Monday', 2: 'Tuesday', 3: 'Wednesday', 4: 'Thursday', 5: 'Friday', 6: 'Saturday' Now if we don't use drop_first = True, it will create all the 7 columns in the data set for weekday. If we use drop_first = True, we can get only 6 columns except WeekDay_Sunday, as we don't need this in data set, if it is not there also, we can analyse it by looking all the other columns i.e., if all the columns have a zero value, then obviously it is Sunday

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

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   - Here's how you can validate the assumptions of linear regression after building the model on the training set:
     i. Linearity:
        Scatter Plots: Plot the independent variable(s) vs. the dependent variable. Look for a roughly linear pattern. Non-linear relationships might require transformations or alternative models.
     ii. Independence of Errors:
        Residual Plots: Plot the residuals (errors) vs. fitted values. Look for random scatter with no discernible patterns. Patterns like curves or funnels indicate potential issues with model fit or hidden variables.
     iii. 3. Homoscedasticity (Constant Variance):
        Scale-Location Plots: Plot the squared residuals vs. fitted values. Look for a horizontal band, indicating constant variance across the range of fitted values. Uneven spread suggests heteroscedasticity, which can be addressed by transformations or robust regression techniques.
     iv. Normality of Errors:
        Q-Q Plots: Plot the quantiles of the residuals vs. the quantiles of a normal distribution. A straight line indicates normality. Deviations from the line suggest non-normality, which might not be a major concern for large datasets but could affect hypothesis testing in smaller ones.
     v. No Multicollinearity:
        Correlation Matrix: Calculate the correlation coefficients between independent variables. High correlations (> 0.8 or < -0.8) suggest multicollinearity. Consider dropping redundant variables or using dimensionality reduction techniques.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   - Temperature (0.4515) – Positive coefficient indicates that as temperature increases, the demand for bikes increases and Higher temperatures likely encourage more bike usage due to more comfortable riding conditions.
   - Year 2019 (0.2341) – Positive coefficient suggests that demand was higher in 2019 compared to the baseline year (2018) and reflects growing popularity and usage of bike-sharing services over time.
   - Month September (0.0577) – Positive coefficient indicates higher demand in September and September might have favourable weather conditions for biking or could coincide with specific events or back-to-school periods.

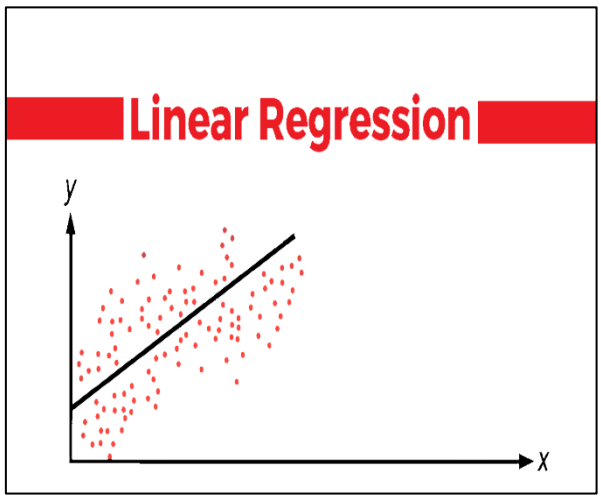
General Subjective Questions

1. Explain the linear regression algorithm in detail.

a. Definition - Linear regression is a fundamental supervised machine learning algorithm used for predicting a continuous dependent variable based on one or more independent variables. It assumes a linear relationship between the features (independent variables) and the target variable (dependent variable).

b. Model Representation –

i. Image -



Equation

$y = b_o + b_1x_1 + b_2x_2 + ... + bnxn + \varepsilon$

Where

- y is the predicted dependent variable (what we're trying to predict)
- b_o is the intercept (the y-axis value where the line crosses)
- b_1 to bn are the coefficients (slopes) for each independent variable x_1 to xn (what factors influence y)
- ε is the error term (the difference between the actual and predicted value)

c. The Learning Process: Minimizing the Mistake Meter – The goal is to find the coefficients (b_o, b_1, ..., bn) that minimize the discrepancy between the predicted (y) and actual (y_actual) values of the dependent variable. Here's how we achieve this:

i. Cost Function: We use the sum of squared errors (SSE) as a measure of discrepancy. The lower the SSE, the better the model fits the data. SSE = Σ(y_actual - y)^2

ii. Optimization: An iterative technique like gradient descent is employed. It adjusts the coefficients slightly in the direction that minimizes the SSE. Imagine rolling a ball down a hill; the ball eventually settles at the lowest point (minimum SSE).

d. Unveiling the Mystery: Interpreting Coefficients – Once trained, the coefficients become storytellers. Their values and signs reveal the following:

i. Strength of Influence: A larger coefficient magnitude indicates a stronger influence of the corresponding variable on the dependent variable.

ii. Direction of Relationship: A positive coefficient suggests that as the independent variable increases, the predicted dependent variable also increases (and vice versa for a negative coefficient).

e. Assumptions: The Foundation of Trust – Linear regression thrives under certain assumptions about the data:
   i. Linear Relationship: The relationship between the features (independent variables) and the target variable (dependent variable) should be linear.
   ii. Independence of Errors: The errors for each data point should be independent of each other.
   iii. Homoscedasticity (Constant Variance): The variance of the errors should be constant across all levels of the independent variables.
   iv. Normality of Errors: The errors should be normally distributed.
   v. No Multicollinearity: The independent variables should not be highly correlated with each other.
   vi. Violating these assumptions can lead to inaccurate or misleading results. Techniques like data transformations or alternative models can be used to address these issues.

f. Unveiling the Future: Prediction Powerhouse – With a trained model, you can predict the dependent variable for new data points. This makes linear regression a valuable tool for various tasks:
   i. Real Estate: Predicting house prices based on factors like size, location, and number of bedrooms.
   ii. Customer Behaviour Analysis: Understanding customer churn (when a customer stop using your service) based on purchase history and demographics.
   iii. Sales Forecasting: Predicting future sales based on historical data and marketing campaigns.

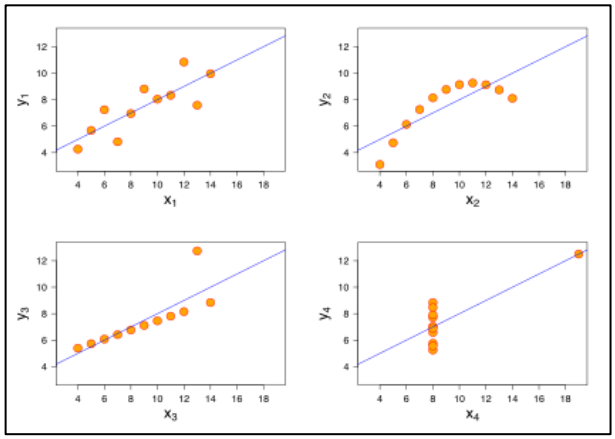
2. Explain the Anscombe's quartet in detail.

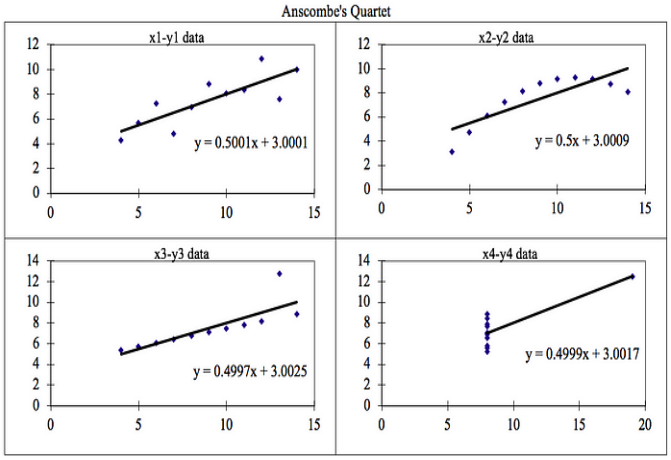
a. Definition – Anscombe's quartet is a collection of four datasets created by statistician Francis Anscombe in 1973 to illustrate the importance of data visualization alongside statistical analysis. Here's why it's so interesting:
   i. The Identical Statistics:
      a) Imagine four datasets, each with 11 data points. Surprisingly, all four datasets share the following statistical properties:
      Mean (average) for x-values: Around 9
      Mean (average) for y-values: Around 7.5
      Variance (spread) for x-values: Around 3.16
      Correlation coefficient between x and y: Around 0.8 (indicating a positive linear relationship)
      Linear regression line: Similar slope and intercept for all four datasets
      b) This can be defined by four data sets having 11 data points each, which are nearly identical. Even though these data points look identical but it appear differently when plot scatter plots. The Four Datasets looks like:

Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

c) Visualizing the data (Graphs)





d) The Visual Deception – Now comes the surprising part. When you plot these datasets as scatter graphs, you'll see a completely different story. Each dataset reveals a distinct visual pattern, shattering the illusion of identical data:

Dataset 1: This dataset exhibits a relatively linear relationship between x and y, resembling what you might expect based on the statistical summaries.

Dataset 2: This dataset throws a curveball. It contains a single clear outlier that skews the data upwards, distorting the linear relationship.

Dataset 3: This dataset takes an unexpected turn. The data points follow a non-linear, curved pattern, defying the expectation of a straight line.

Dataset 4: This dataset presents a puzzling scenario. Most data points cluster tightly on the left side, with a single outlier far away, creating a misleading picture.

e) The Importance of Data Exploration – Anscombe's quartet emphasizes the significance of data visualization in exploratory data analysis (EDA). By visually inspecting your data, you can uncover:

Patterns and trends: You might identify hidden patterns or relationships not readily apparent from statistical summaries.

Outliers: These can be data points that deviate significantly from the overall trend and might require further investigation.

Non-linearities: Linear regression might not always be the best fit. Visualization can help identify non-linear relationships that statistical summaries might miss.

3. What is Pearson's R?

a. Definition – Pearson's R coefficient, denoted by r, is a statistical measure that quantifies the linear relationship between two continuous variables. It essentially tells you the strength and direction of the association between those variables.

b. Key aspects
   i. Range: r ranges from -1 to +1
   ii. Interpretation:
      a) Positive correlation ($0 < r < 1$): As the value of one variable increases, the value of the other variable also tends to increase (think: height and weight).
      b) Negative correlation ($-1 < r < 0$): As the value of one variable increases, the value of the other variable tends to decrease (think: study time and test scores - ideally!).
      c) Zero correlation ($r = 0$): There's no linear relationship between the two variables (think: shoe size and vocabulary).

c. Formula and Calculation – $r = \sum(xi - \bar{x})(yi - \bar{y}) / \sqrt{\sum(xi - \bar{x})^2 * \sum(yi - \bar{y})^2}$
   i. r is the Pearson correlation coefficient
   ii. xi and yi are individual data points for variables x and y, respectively.
   iii. $\bar{x}$ and $\bar{y}$ are the means of variables x and y, respectively.

d. Pearson's correlation coefficient formula, when applied to a population is:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \quad \text{(Eq.1)}$$

cov is the covariance,

σ_X Is the standard deviation of X σ_Y and is the standard deviation of Y

The formula for p can be expressed in terms of mean and expectation.

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

e. Important Considerations
   i. Pearson's R is best suited for linear relationships. If the relationship between your variables is not linear, it may not be a good choice.
   ii. The variables you're analysing should be continuous (like hours studied or exam scores).
   iii. Be mindful of outliers that can skew the results.

f. Example
   i. Imagine you're researching the relationship between study hours (x) and exam scores (y) for a group of students. You collect data and calculate a Pearson's R of 0.7. This indicates a positive correlation, meaning students who dedicate more study hours tend to score higher on exams. The strength of the correlation (0.7) suggests a moderately strong positive relationship.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   a. Scaling – It refers to the process of transforming your features (independent variables) to a new range. This is often done to bring all features onto a similar scale, which can improve the efficiency and accuracy of the model, especially when using optimization algorithms like gradient descent.
   b. Why is scaling performed –
      i. Improves convergence – When features have vastly different scales, the optimization algorithm might prioritize updates for features with larger magnitudes, making convergence slower. Scaling levels the playing field for all features, allowing the algorithm to find the optimal solution more efficiently.
      ii. Reduces bias towards features with larger scales – Without scaling, features with larger values will have a greater influence on the model's coefficients (slopes), potentially biasing the model towards those features. Scaling ensures all features contribute proportionally.
      iii. Numerical stability – Some optimization algorithms can become unstable with features of very different scales. Scaling promotes numerical stability and helps the algorithm perform calculations more accurately.
   c. What is the difference between normalized scaling and standardized scaling – Both normalized scaling and standardized scaling are techniques used to transform features (independent variables) in machine learning, particularly in linear regression, but they differ in their approach and output:
      i. Normalized Scaling – (Min - Max)
         a) Focus: Rescales features to a specific user-defined range, typically between 0 and 1 (or -1 and 1).
         b) Formula: x_scaled = (x - min(x)) / (max(x) - min(x))
         c) Impact on Distribution: Doesn't make any assumptions about the underlying data distribution. It simply stretches or shrinks the data to fit the chosen range. Outliers can have a significant impact on the scaling.
         d) Interpretation of Coefficients: Coefficients become less interpretable in terms of their original units since the scaling is based on arbitrary bounds.
      ii. Standardized Scaling –
         a) Focus: Transforms features to have a standard normal distribution (mean of 0 and standard deviation of 1).
         b) Formula: x_scaled = (x - mean(x)) / standard_deviation(x)
         c) Impact on Distribution: Assumes the underlying data is (or can be approximated to be) normally distributed. This transformation forces the data to follow a specific bell-shaped curve.
         d) Interpretation of Coefficients: Coefficients become more interpretable. They represent the change in the target variable (in units of standard deviations) for a one-unit change in the feature.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   a. In linear regression, VIF (Variance Inflation Factor) can skyrocket to infinity due to severe multicollinearity**.** This happens when two or more features are extremely correlated, meaning one can be almost perfectly predicted from the others.
   b. Why infinity – VIF is calculated as 1 / (1 - R^2), where R^2 is the fit of a model where one feature is regressed against all others. In perfect multicollinearity, R^2 becomes 1 (perfect fit). Dividing by 1 minus 1 (0) result in infinity.
   c. Impact –
      i. Unstable coefficients: Small changes in data can drastically alter the coefficients.
      ii. Difficulty in interpretation: Coefficients become hard to understand because features are intertwined.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?**
   a. A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used in linear regression to assess whether the errors (residuals) of your model follow a normal distribution.
   b. How it works
      i. Quantiles: Imagine dividing your data (errors in this case) into equal portions. The quantiles represent the values that separate these portions.
      ii. Plotting: The Q-Q plot takes the quantiles of your errors and plots them against the quantiles of a theoretical normal distribution (usually a straight line at a 45-degree angle).
      iii. Interpretation: If the points in the plot fall roughly along the diagonal line, it suggests that the errors are normally distributed. Deviations from the line indicate departures from normality.
   c. Importance in Linear Regression:
      i. Validity of Tests: Many statistical tests used to assess the significance of the model and its coefficients rely on the normality of errors.
      ii. Confidence Intervals: Confidence intervals, which provide a range for the true value of a coefficient, are also based on the normality assumption.
      iii. Model Robustness: While not a strict requirement, normality can make the model more robust to outliers and improve its generalizability.
   d. Example: