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)

By analyzing categorical variables and their interaction with other features (e.g., temperature), we can gain a deeper understanding of the factors influencing bike rental demand. The model will capture these relationships and predict demand based on specific combinations of these categorical features.

Day of the Week: We can expect to see a cyclical pattern in bike rentals across the days of the week. Weekends (Saturday and Sunday) likely experience higher demand compared to weekdays (Monday-Friday) due to increased leisure activities.

Working Day and Holiday: Working Day may have lower demand, unless the rental place in a tourist location that receives tourists all week in specific seasons. Whereas Holiday especially during Holiday seasons may drive higher demand.

Month: A seasonal pattern might be present, with warmer months (e.g., summer) potentially leading to higher demand compared to colder months (e.g., winter) when people are less likely to cycle outdoors.

Season: Similar to month, the "Season" variable can further solidify the seasonal trends. Spring and Fall might see higher demand, while Summer could have moderate, and Winter the lowest.

Weather Condition: This variable can have a complex impact on demand. Sunny and clear weather likely leads to higher rentals, while rainy or snowy conditions might deter potential users. However, some riders might prefer using bikes during mild rain or pleasant breezes.

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

In Pandas' get_dummies function, the drop_first=True argument is important for two main reasons:

Reduces Multicollinearity: When creating dummy variables for categorical features, each category gets its own binary column. By default, get_dummies creates one column for each category level (k categories -> k dummy columns). However, this introduces multicollinearity, a statistical issue where features are highly correlated with each other. This can negatively impact model performance and coefficient interpretation.

With drop_first=True, one category level is arbitrarily excluded from the dummy variable creation. This removes one column, leaving k-1 dummy variables. Since the information about the missing category can be inferred from the remaining ones (e.g., if someone isn't Male or Female, they must be the other), multicollinearity is reduced.

Improves Model Interpretation: With drop_first=True, the coefficients of the remaining dummy variables represent the difference in the target variable compared to the dropped category. This simplifies interpretation as coefficients directly reflect the impact of each category relative to the baseline (dropped one).

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

Temperature and Feeling Temperature have the highest correlation with the target variable – Cnt. This makes sense as the bicycle renters prefer moderate temperatures for a good cycling day.

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

Here are some of the ways to validate assumptions after building the model on the training set:

1. **Linearity - Visualization**: Plotted the independent variables against the dependent variable to see if the relationship appears linear using Scatter Plots. Pair plots to visualize relationships between all features and the target variable was also done.

2. **Independence - Residual Analysis**: Plotted the residuals (the difference between predicted and actual values) versus the fitted values (predicted values). The residuals were randomly scattered around zero with no apparent pattern, suggesting independence.

3. **Homoscedasticity (Constant Variance): Visualization**: Plot the residuals against the fitted values. In a homoscedastic scenario, the spread of the residuals should be constant across the range of fitted values.

4. **Normality of Residuals**: Plot the residuals in a histogram - a normal distribution suggests the residuals are normally distributed.

5. **No Multicollinearity - Correlation Matrix:** Calculated the correlation matrix between all independent variables. Highly correlated features (correlation coefficient close to 1 or -1) were addressed by removing redundant features.

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

The following are the top 3 features contributing significantly:

Feature	Coefficient	P-Value	Remarks
atemp	0.4534	0.000	Positive Correlation with Feeling Temparature – which makes sense since people prefer moderate temparatures for cycling
weathersit_bad	-0.2608	0.000	Negative Correlation since bad weather is bad day for cycling
season_spring	-0.1174	0.000	Negative Correlation – it appears people prefer Summers and Fall over Spring

General Subjective Questions

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

Linear regression (LR) is a supervised machine learning algorithm that models the relationship between a dependent variable and one or more independent variables using a linear equation. The goal for LR is to find the line of best fit that minimizes the error between the predicted values and the actual values in the dataset.

The linear regression equation takes the form:

$y = b0 + b1*x1 + b2*x2 + ... + bn*xn$

Where:
- y is the dependent variable
- x1, x2, ..., xn are the independent variables
- b0 is the y-intercept (value of y when all x's are 0)
- b1, b2, ..., bn are the coefficients that represent the change in y per unit change in the respective x variable

The coefficients are determined using the method of least squares, which minimizes the sum of the squared residuals (difference between predicted and actual values). This is an iterative process to find the line that best fits the data.

There are two main types of linear regression:
1. **Simple linear regression**: Predicts y from a single x variable using a straight line
2. **Multiple linear regression**: Predicts y from multiple x variables using a hyperplane

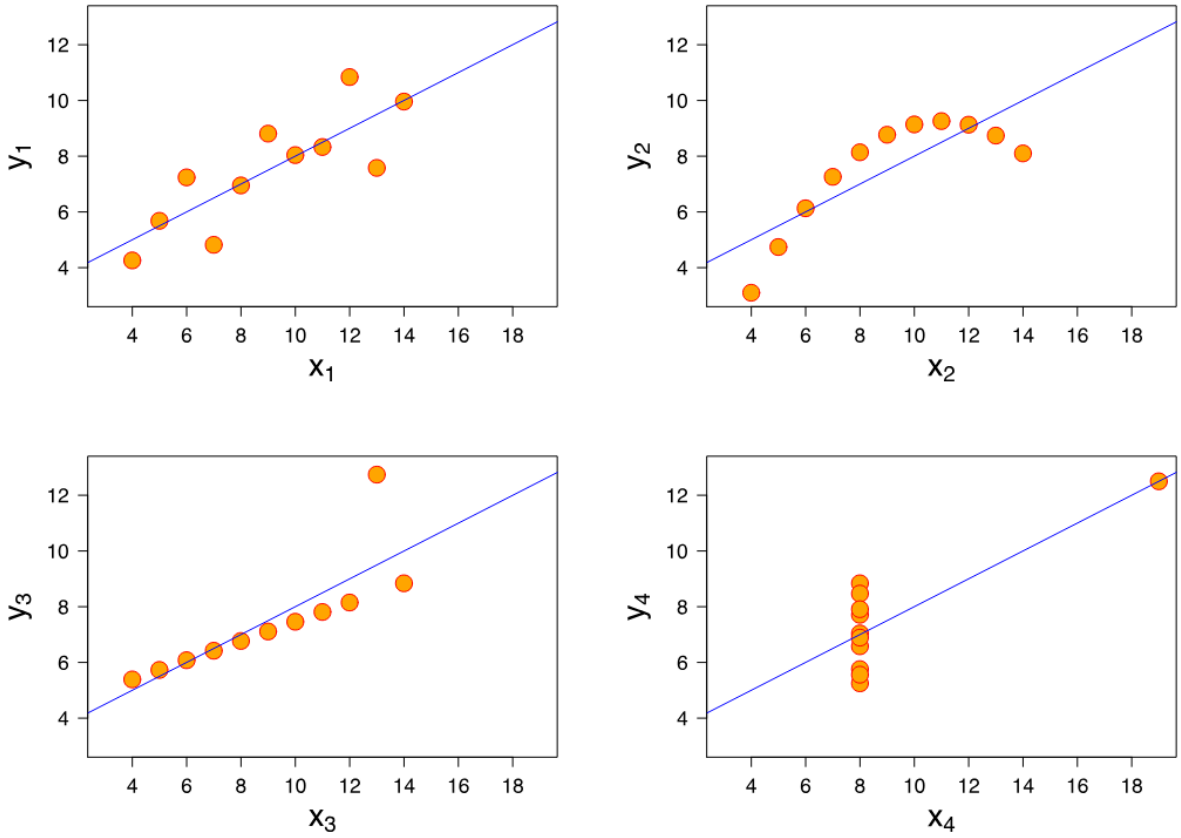
Some key assumptions of linear regression include:

Assumption	Description
Linearity	The relationship between x and y is linear
Homoscedasticity	The variance of the residuals is constant
Multicollinearity	The independent variables are not highly correlated with each other
Independence	The residuals are independent

The performance of a linear regression model is typically evaluated using metrics like R-squared (proportion of variance explained) and p-values (statistical significance of coefficients). Linear regression is a powerful tool for prediction and understanding relationships between variables but may not capture complex nonlinear patterns.

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

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics yet appear very different when graphed. The quartet was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and not relying solely on numerical summaries.



The key points about Anscombe's quartet are:

1. **Identical Statistics**: Despite having very different distributions and patterns, the four datasets in the quartet have almost the same mean, variance, correlation coefficient, and linear regression line.

2. **Diverse Visualizations**: When the datasets are plotted on scatter plots, each one exhibits a distinct pattern - some show linear relationships, others have non-linear trends, and some contain outliers that significantly impact the regression line.

3. **Importance of Visualization**: Anscombe's quartet highlights that numerical summaries alone can be misleading, and data visualization is crucial to uncover the true nature of the relationships in the data. Plotting the data reveals insights that the statistics alone cannot capture.

4. **Limitations of Linear Regression**: The quartet demonstrates that linear regression may not be appropriate for certain types of data, such as those with non-linear patterns or influential outliers. Relying solely on regression analysis can lead to incorrect conclusions in such cases.

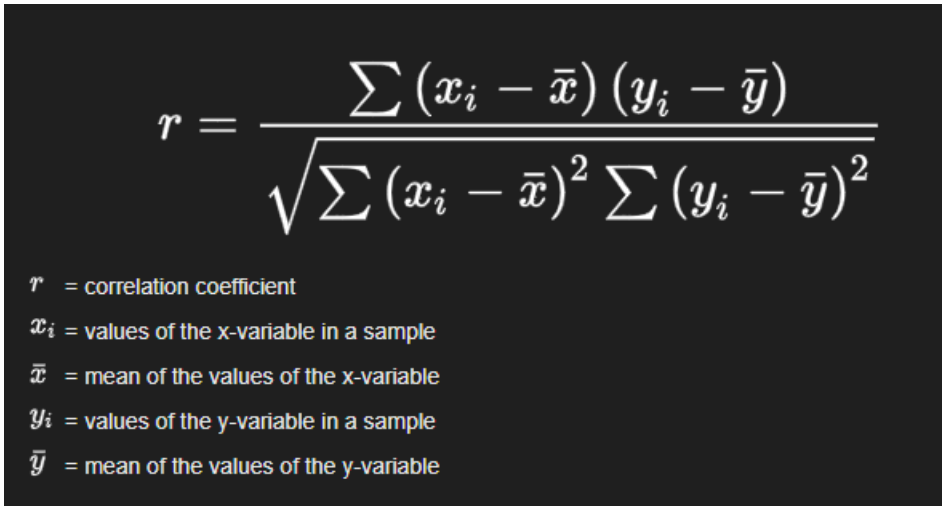
5. **Need for Exploratory Data Analysis**: The quartet emphasizes the importance of thoroughly examining data through exploratory data analysis (EDA) techniques, such as creating visualizations, before applying statistical models. EDA can help identify anomalies, patterns, and relationships that may not be evident from the summary statistics.

3. What is Pearson’s R? (3 marks)

Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. Key points about Pearson's R:

Definition: Pearson's R is a statistic that measures the strength and direction of the linear association between two variables. It ranges from -1 to 1, with -1 indicating a perfect negative linear relationship, 0 indicating no linear relationship, and 1 indicating a perfect positive linear relationship.

Formula: For a sample, Pearson's R is calculated as:



Interpretation: The closer Pearson's R is to 1 or -1, the stronger the linear relationship between the two variables. The sign indicates the direction of the relationship (positive or negative). Values close to 0 indicate a weak or no linear relationship.

Assumptions: Pearson's R assumes that the relationship between the variables is linear, the variables are approximately normally distributed, and the variance of one variable is constant for all values of the other variable.

Significance Testing: The statistical significance of Pearson's R can be tested using a t-test. This tests the null hypothesis that the true correlation coefficient is zero in the population.

Applications: Pearson's R is widely used in various fields, such as psychology, economics, and social sciences, to measure the strength and direction of linear relationships between variables.

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

Scaling refers to the process of transforming or adjusting the range of values of a variable or dataset. There are a few key points about scaling:

Purpose of Scaling:
- Scaling is often performed to standardize the range of independent variables or features in a dataset, which is important for many machine learning algorithms.
- It helps ensure that variables are on a common scale, so that no single variable dominates the objective function.
- Scaling can also be used to improve the numerical stability and convergence of optimization algorithms.

Types of Scaling:
- Normalized Scaling (Min-Max Scaling): This linearly transforms the data to a common range, typically between 0 and 1. It preserves the relationships between the original data points.
- Standardized Scaling (Z-Score Scaling): This transforms the data to have a mean of 0 and a standard deviation of 1. This is useful when the original scales of the variables vary widely.

Differences between Normalized and Standardized Scaling:
- Normalized scaling preserves the original distribution shape, while standardized scaling transforms the distribution to a standard normal distribution.
- Normalized scaling is more sensitive to outliers, while standardized scaling is more robust to outliers.
- Standardized scaling is more appropriate when the variables have different units or vastly different ranges.

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

Few key reasons why the VIF (Variance Inflation Factor) can return an infinite value:

1. **Perfect Multicollinearity**: The VIF formula involves dividing by (1 - R^2), where R^2 is the coefficient of determination from regressing one independent variable on the others. If there is perfect multicollinearity, meaning one independent variable can be perfectly predicted by a linear combination of the other independent variables, then the R^2 will be 1. This causes the denominator to become 0, resulting in an infinite VIF value.

2. **More Variables than Observations**: If the number of independent variables (predictors) in the regression model exceeds the number of observations, it can also lead to perfect multicollinearity and infinite VIF values. This is because with more variables than observations, the regression model can perfectly fit the data, resulting in an R^2 of 1.

3. **Dummy Variables with Small Categories**: When using dummy variables to represent categorical predictors, if a category has only a small number of observations, the VIF for that dummy variable can become infinite, even if there is no perfect multicollinearity between the predictors.

To address this issue, the common solutions are:
1. Remove one or more of the highly correlated independent variables.
2. Use dimension reduction techniques like Principal Component Analysis (PCA) or Partial Least Squares (PLS) regression.
3. For dummy variables with small categories, consider combining categories or removing the problematic variables.

The presence of infinite VIF values indicates a serious multicollinearity problem that needs to be resolved before proceeding with the regression analysis, as it can lead to unstable and unreliable coefficient estimates.

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

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset to a theoretical probability distribution, such as the normal distribution. The key points about Q-Q plots are:

1. **Purpose**: Q-Q plots are used to assess whether a dataset follows a particular probability distribution, most commonly the normal distribution. This is important for many statistical analyses, including linear regression, which often assume normality of the data or residuals.

2. **Construction**: A Q-Q plot is constructed by plotting the quantiles of the dataset against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points will fall approximately on a straight 45-degree line.

3. **Interpretation**:
   - If the points on the Q-Q plot fall close to a straight line, it suggests the data follows the theoretical distribution.
   - Deviations from the straight line, especially at the tails, indicate the data does not follow the theoretical distribution.
   - Curvature in the Q-Q plot can reveal issues like skewness or kurtosis in the data distribution.

4. **Importance in Linear Regression**:
   - In linear regression, the normality assumption applies to the residuals (the differences between the observed and predicted values).
   - A Q-Q plot of the residuals can be used to assess whether the normality assumption is met.
   - If the residuals do not follow a normal distribution, the validity of statistical inferences from the regression model may be compromised.
   - Identifying non-normality in the residuals can help guide the selection of appropriate transformations or alternative modelling approaches.