Assignment-based Subjective Questions

Question: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)
Answer: I have done analysis on categorical columns season, mnth, weekday and weathersit using the boxplot. Below are the few points we can infer from the visualization –

1. In the Fall season we can see company has got a greater number of bookings and in each season the booking count has increased drastically from 2018 to 2019.
2. For each month we can see bookings were increased by very high margin from 2018 to 2019. There is a trend we can observe which is booking keeps on increasing in the first half of the year and it began to drop as we approached year end.
3. Months including may, june, july, aug, sep and oct experienced the highest number of bookings.
4. Thu, Fir, Sat have a greater number of bookings as compared to the Sunday and Monday. Tuesday and Wednesday also have decent number of bookings.
5. Clear weather attracted more booking which seems obvious.
6. I observed company has make a good business in 2019 as compared to 2018 as bookings increased drastically in 2019.

Question: Why is it important to use drop_first=True during dummy variable creation? (2 mark)
Answer: When creating dummy variables, drop_first=True is used to avoid multicollinearity. Multicollinearity occurs when two or more variables are highly correlated, which can lead to unstable estimates and inaccurate predictions.
In dummy variable creation, drop_first=True drops the first category of each feature, which helps to:

1. Avoid perfect multicollinearity: By dropping the first category, we ensure that the dummy variables are not perfectly correlated with each other.
2. Prevent the "dummy variable trap": This is a situation where the model becomes unstable due to the presence of multiple dummy variables that are highly correlated.

By using drop_first=True, we can create more stable and interpretable models. However, it's important to note that this may not always be necessary, and it depends totally on the dataset and use case.

Question: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Answer: In my observation “temp” has the highest correlation with the target variable.

Question: How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)
Answer: To validate the assumptions of Linear Regression, I’ve performed the following steps:

1. **Linearity**: Check if the relationship between the predictors and target variable is linear by visualizing the data using scatter plots or partial regression plots.
2. **Homoscedasticity**: Check if the variance of residuals is constant across all levels of predictors by visualizing residuals vs. fitted values
3. **Normality**: Verify that residuals follow a normal distribution using histograms

4. **No multicollinearity**: Check for multicollinearity using Variance Inflation Factor (VIF) or correlation matrices to ensure predictors are not highly correlated.

Question: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)
Answer: In my final model below are the top 3 features contributing significantly towards explaining the demand of the shared bikes:

1. Temp
2. Year which is (yr)
3. Winter

General Subjective Questions

Question: Explain the linear regression algorithm in detail. (4 marks)
Answer: Linear regression is a statistical method used to model the relationship between a dependent variable (target) and one or more independent variables (features). In its simplest form, linear regression models the relationship using a straight line.

Types of Linear Regression:

1. Simple Linear Regression: One independent variable.
2. Multiple Linear Regression: More than one independent variable.

1. **Simple Linear Regression**
This is the case when we try to predict a dependent variable y using a single independent variable x.
Equation**: y = mx + b**

- y: The dependent variable (what we are predicting).
- x: The independent variable (input feature).
- m: The slope of the line (how much y changes for each unit change in x).
- b: The intercept (the value of y when x is 0).

Objective: The goal is to find the best values for m (slope) and b (intercept) such that the line fits the data as closely as possible.
Example: Suppose we want to predict the price of a house based on its size. We have the following data:

Size (sq ft)	Price (in 1000s)
1500	330
2000	360
2500	400
3000	430

We want to find the line that best fits this data. The linear regression model will compute the slope and intercept, resulting in an equation such as:
Price = 0.1 * Size + 300

Here, the slope m = 0.1 tells us that for every additional square foot, the price increases by 0.1K, and the intercept b = 300 means that the starting price of a house is 300K.

Prediction: If we want to predict the price of a house with 3500 sq ft, we can use the equation:
Price = 0.1 * 3500 + 300 = 650
So, the predicted price is 650K.

2. **Multiple Linear Regression**:
When there are multiple independent variables, we extend the equation:
y = m1 * x1 + m2 * x2 + ... + mn * xn + b

Where:
- x1, x2, ..., xn are the independent variables (features).
- m1, m2, ..., mn are the coefficients (slopes) for each feature.

Example:
Suppose we want to predict house prices based on both size and number of bedrooms. We might have the following data:

Size (sq ft)	Bedrooms	Price (in 1000s)
1500	3	330
2000	4	360
2500	3	400
3000	5	430

The linear regression equation might be:
Price = 0.08 * Size + 20 * Bedrooms + 250
Here:
- For each additional square foot, the price increases by 0.08K.
- For each additional bedroom, the price increases by 20K.
- The intercept of 250K is the base price when the size and number of bedrooms are 0.

Prediction:
If a house has 2700 sq ft and 4 bedrooms, the predicted price would be:
Price = 0.08 * 2700 + 20 * 4 + 250 = 466
So, the predicted price is 466K.

3. **Steps to Perform Linear Regression:**

1. Gather Data: Collect the dependent variable y and the independent variable(s) x.
2. Hypothesis Function: Define the linear equation y = mx + b.
3. Loss Function (Cost Function): The difference between the predicted value and actual value is calculated using a cost function. A common cost function is Mean Squared Error (MSE):

   MSE = (1/n) * Σ (y_i - ŷ_i)^2

   Where:
   - y_i is the actual value.
   - ŷ_i is the predicted value.
   - n is the number of observations.
4. Optimize Parameters (Gradient Descent):

1. Adjust the values of m and b to minimize the cost function.
2. Gradient descent is an iterative optimization algorithm used to find the minimum of a function. It adjusts the parameters m and b until the cost function reaches its minimum.

5. Model Evaluation: After finding the optimal parameters, evaluate the model using metrics like R-squared, which shows how well the model fits the data.

4. **Assumptions of Linear Regression:**

1. Linearity: The relationship between the independent and dependent variable is linear.
2. Homoscedasticity: The variance of errors is consistent across all levels of the independent variables.
3. No Multicollinearity: In multiple linear regression, the independent variables should not be highly correlated with each other.
4. Normality of Errors: The residuals (errors) of the model should be normally distributed.

Conclusion:

Linear regression is a fundamental machine learning algorithm that assumes a linear relationship between the input variables and the target variable. It's easy to interpret and widely used in many fields, from economics to engineering.

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

Answer: Anscombe's quartet is a collection of four datasets, each with 11 points, that illustrate the importance of visualizing data and the limitations of statistical summaries. The datasets were created by Francis Anscombe in 1973.

Here's a detailed explanation:

Datasets:

1. Dataset 1: (x, y) = (10, 8.04), (8, 6.95), ..., (5, 5.68)
2. Dataset 2: (x, y) = (10, 9.14), (8, 8.14), ..., (5, 4.77)
3. Dataset 3: (x, y) = (10, 7.46), (8, 6.77), ..., (5, 4.44)
4. Dataset 4: (x, y) = (8, 6.58), (8, 5.76), ..., (8, 5.42)

Properties:

1. Means: All datasets have nearly identical means for x and y.
2. Variances: All datasets have nearly identical variances for x and y.
3. Correlations: All datasets have nearly identical correlations between x and y (around 0.816).
4. Linear Regression: All datasets yield nearly identical linear regression lines (y = 3 + 0.5x).

Visualization:

When plotted, each dataset reveals a different relationship between x and y:

1. Dataset 1: A straightforward linear relationship.
2. Dataset 2: A curved relationship, with a quadratic shape.
3. Dataset 3: A linear relationship with an outlier.
4. Dataset 4: A horizontal line with constant y-values.

Conclusion:

Anscombe's quartet demonstrates that:

1. Statistical summaries (means, variances, correlations) can be identical for different datasets.
2. Visualizing data is crucial to understanding the relationships and patterns.
3. Linear regression can be misleading if the data is not appropriately examined.

Question: What is Pearson's R? (3 marks)
Answer: Pearson's R, also known as the Pearson product-moment correlation coefficient, is a statistical measure that evaluates the strength and direction of a linear relationship between two continuous variables, X and Y. It's denoted by the symbol "R" or "r".
Here's a breakdown:
Values:

1. R ranges from -1 to 1
2. R = 1: Perfect positive linear correlation
3. R = -1: Perfect negative linear correlation
4. R = 0: No linear correlation

Interpretation:

1) Strength: The absolute value of R (|R|) indicates the strength of the correlation
   a) $0 < |R| < 0.3$: Weak correlation
   b) $0.3 \leq |R| < 0.7$: Moderate correlation
   c) $0.7 \leq |R| < 1$: Strong correlation
2) Direction: The sign of R indicates the direction of the correlation
   a) Positive R: As X increases, Y tends to increase.
   b) Negative R: As X increases, Y tends to decrease.

Assumptions:

1) Both variables should be continuous and normally distributed
2) The relationship between X and Y should be linear.

Use cases:

1. Pearson's R is widely used in various fields, such as social sciences, biology, and finance, to:
   a. Analyse relationships between variables.
   b. Identify correlations.
   c. Make predictions.

Question: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Answer: Scaling is the process of transforming data into a common range, usually between 0 and 1, to:

a) Prevent features with large ranges from dominating the model.
b) Improve model convergence and performance.
c) Enhance interpretability.

There are two types of scaling:

a) Normalized scaling (Min-Max Scaling):
   - Rescales data to a common range [0, 1] using the minimum and maximum values.
   - Formula: (x - min) / (max - min)
   - Preserves the original distribution and shape.
b) Standardized scaling (Z-Score Scaling):
   - Rescales data to have a mean of 0 and a standard deviation of 1.
   - Formula: (x - mean) / std_dev

- Transforms data to a normal distribution (Gaussian distribution)

Key differences:

a) Normalized scaling preserves the original distribution, while standardized scaling transforms data to a normal distribution
b) Normalized scaling is sensitive to outliers, while standardized scaling is more robust.

When to use each:

a) Normalized scaling:
   - When the data has a fixed range (e.g., ratings, scores)
   - When preserving the original distribution is important
b) Standardized scaling:
   - When the data has a varying range (e.g., age, income)
   - When normality is assumed or required (e.g., statistical modelling)

Scaling is a crucial preprocessing step in machine learning and data analysis.

Question: You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)
Answer: VIF (Variance Inflation Factor) is a measure of multicollinearity between predictors in a regression model. An infinite VIF value indicates perfect multicollinearity, which occurs when:

- Two or more predictors are identical (e.g., duplicate variables).
- One predictor is a linear combination of others (e.g., x1 = 2_x2 + 3_x3).
- There is a perfect correlation between predictors (e.g., x1 = x2).

In these cases, the matrix operation to calculate VIF becomes singular (non-invertible), leading to an infinite value. Infinite VIF indicates that the model is over specified, and at least one predictor is redundant. To address this:

1. Remove redundant predictors.
2. Use dimensionality reduction techniques (e.g., PCA, feature selection).
3. Regularization techniques (e.g., LASSO, Ridge regression) can help mitigate multicollinearity.

Infinite VIF is a sign of perfect multicollinearity, which can lead to unstable estimates and inaccurate predictions. Addressing multicollinearity is crucial for building reliable regression models.

Question: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)
Answer: A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of two datasets or a dataset and a theoretical distribution. In linear regression, Q-Q plots are used to:

- Check normality of residuals: Verify if the residuals follow a normal distribution, which is a key assumption in linear regression.
- Identify departures from normality: Detect skewness, heavy tails, or outliers in the residuals.
- Compare distributions: Examine the distribution of residuals against a theoretical normal distribution or another reference distribution.

Importance:

- Diagnostic tool: Q-Q plots help diagnose departures from normality, which can lead to inaccurate inference and unreliable predictions.
- Model validation: Q-Q plots validate the assumption of normality, ensuring the linear regression model is appropriate.
- Improvement opportunities: Q-Q plots can guide transformations or corrections to achieve normality and improve model performance.

Interpretation:

- If the points lie close to a straight line, the residuals are approximately normally distributed.
- Deviations from the line indicate departures from normality.

By using Q-Q plots, we can ensure the validity of our linear regression model and identify areas for improvement.