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)
Answer.

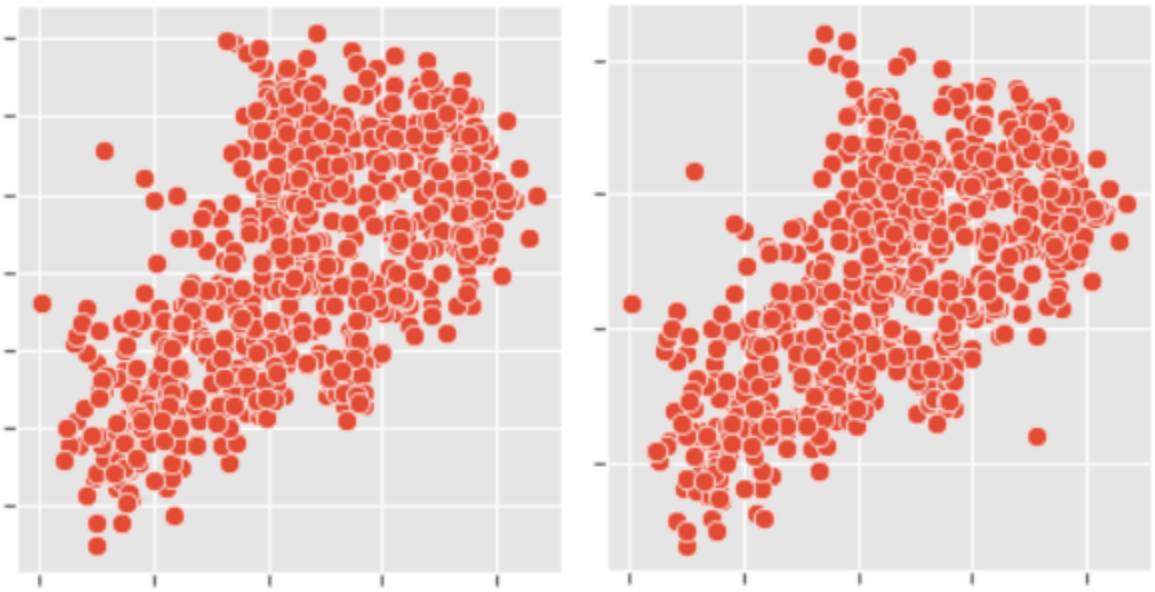
- **Year**: The dependent variable shows a positive trend over time, with a notable increase in 2019 compared to 2018. This suggests a time-based growth in the dependent variable, possibly indicating rising demand or usage over the years.
- **Season, Month, and Weather**: Warmer seasons (Summer and Fall) and months (June, July, August) show higher counts, indicating that the dependent variable is strongly influenced by favorable weather conditions. Clear and mild weather conditions further boost activity, while colder months and adverse weather (like light snow and rain) reduce the dependent variable's counts.
- **Holiday and Working Day**: Non-holidays and working days are associated with higher activity, implying that the dependent variable is closely tied to daily work routines, possibly due to commuting. This shows that work-related factors have a stronger effect compared to leisure-related days like holidays.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)
A- It reduces the extra column that can be ignored which was created during dummy variable creation.
From the case study for the column season which were having 4 seasons: (spring,summer,fall, winter)
It created three only, which means if spring, summer and winter all are 0 then it is fall.

```
[608] season_dummy.head()
```

dteday	spring	summer	winter
2018-01-01	1	0	0
2018-01-02	1	0	0
2018-01-03	1	0	0
2018-01-04	1	0	0
2018-01-05	1	0	0

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
Answer- temp and atemp is having the highest correlation with the target variable with 0.64



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

- To determine if the assumption is met or not we create a scatter plot for X vs Y graph. The data points should have a straight line in the graph to confirm there is a linear relationship between the dependent and independent variables.
- There's no multicollinearity in the data. It can be checked via VIF and correlation matrix.
- There Should be Homoscedasticity Among the Data. The data is said to be homoscedastic when the residuals are equal across the line of regression. In other words, the variance is equal. It can be checked by plotting the residuals vs. fitted (predicted) values. If the spread of residuals is consistent across all fitted values, the homoscedasticity assumption is valid.
- Ensure that the residuals are normally distributed. This can be validated by plotting graphs of residuals like histogram or Q-Q plot.

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

A- The top 3 features contributing significantly towards explaining the demand of the shared bikes are:
yr (Year): Coefficient = 0.2343

- This indicates that each additional year (likely indicating a time series from year to year) is associated with a significant positive change in the target variable.

temp (Temperature): Coefficient = 0.4799

- This shows that temperature has a strong positive impact on the prediction. As temperature increases, the target variable increases significantly.

Light Snow & Rain: Coefficient = −0.2865

- This feature has a negative impact, meaning that the presence of light snow and rain is associated with a decrease in the target variable.

General Subjective Questions

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

Answer: Linear regression is a supervised learning algorithm used for predicting a continuous dependent variable based on one or more independent variables. It assumes a linear relationship between the dependent variable (output) and the independent variables (inputs). The goal of linear regression is to find the best-fitting line that minimizes the error between the predicted and actual values.

a. Types of Linear Regression

- Simple Linear Regression: This is when there is only one independent variable. The relationship is modeled as:
  $y=\beta0+\beta1x+\epsilon$
  Where:
  - y is the dependent variable (the variable being predicted)
  - x is the independent variable
  - β_0 is the intercept (value of y when x is 0)
  - β_1 is the slope (rate of change of y with respect to x)
  - ϵ is the error term (residuals)
- Multiple Linear Regression: This is when there are multiple independent variables. The relationship is modeled as:
  $y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$

  Where:

  - x1,x2,…,xn are the independent variables (features)
  - β1,β2,…,βn are the coefficients (weights)

b. **Assumptions of Linear Regression**

- Linearity: The relationship between the independent and dependent variables is linear.
- Independence: Observations are independent of each other.
- Homoscedasticity: Constant variance of errors across the data.
- Normality: The residuals (differences between observed and predicted values) should follow a normal distribution.
- No Multicollinearity: In multiple linear regression, the independent variables should not be highly correlated with each other.

c. **Cost Function**

The cost function for linear regression is the Mean Squared Error (MSE), which measures the average of the squared differences between the predicted and actual values:

$J(\beta 0,\beta 1,\ldots,\beta n)=m1i=1\sum m(yi-yi)2$

Where:

- y_i is the actual value
- y_i^ is the predicted value
- mmm is the number of data points

The goal is to minimize the cost function to find the optimal values of $\beta 0,\beta 1,\ldots,\beta n$

d. **Optimization using Gradient Descent**

- Gradient Descent is an optimization algorithm used to minimize the cost function.
- It starts with random initial values for the coefficients and iteratively updates them in the direction of the negative gradient (downhill) to minimize the cost.
- The update rule for the coefficients is:

  $\beta j:=\beta j-\alpha\frac{\partial J}{\partial \beta j}$

  Where:

  - α is the learning rate (step size)
  - $\frac{\partial J}{\partial \beta j}$ is the partial derivative of the cost function with respect to βj

This process is repeated until the cost function converges to a minimum.

e. **Evaluating Linear Regression Model**

- R-squared (Coefficient of Determination): Measures how well the independent variables explain the variability of the dependent variable. Ranges from 0 to 1, where 1 indicates a perfect fit.
- Adjusted R-squared: Adjusts the R-squared value for the number of predictors in the model. It penalizes the inclusion of unnecessary variables.

- Root Mean Squared Error (RMSE): Measures the square root of the average squared differences between actual and predicted values. Lower values indicate better performance.

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, such as mean, variance, and correlation, yet differ significantly when graphed. It was created by statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data rather than relying solely on statistical summaries.

a. The Four Datasets

Each dataset in Anscombe's Quartet has the following statistical properties:

- Mean of x: 9 for all datasets
- Mean of y: 7.5 for all datasets
- Variance of x: 11 for all datasets
- Variance of y: 4.12 for all datasets
- Correlation between x and y: ~0.816 for all datasets
- Linear regression line (y = 3 + 0.5x): Nearly the same for all datasets

Despite these identical statistical properties, the four datasets tell very different stories when plotted.

b. Significance of the Four Graphs

Each of the four datasets highlights a unique characteristic that is missed when only analyzing the data numerically:

- Dataset 1: This dataset resembles what you would expect from a typical linear regression analysis. It shows a clear linear relationship between x and y.
- Dataset 2: Despite the same statistical measures, this dataset has a curved relationship. Linear regression would be inappropriate here.
- Dataset 3: This dataset contains an outlier that drastically affects the correlation and regression line, although most of the data shows little variation in y.
- Dataset 4: This dataset has nearly identical x-values, except for one outlier, which skews the regression line. Without the outlier, the data would show no relationship.

c. Lessons from Anscombe's Quartet

- Importance of Data Visualization: Anscombe's Quartet emphasizes that relying solely on statistical measures like mean, variance, or correlation can be misleading. Data visualization helps uncover patterns, trends, and anomalies that are not visible through numbers alone.

- Outliers and Non-linearity: The datasets show that outliers and non-linear relationships can strongly affect summary statistics. Visual inspection is necessary to identify these issues.
- Caution with Statistical Models: It demonstrates that applying the same statistical model, like linear regression, across datasets with similar statistics can lead to vastly different conclusions if the data patterns are not understood.

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. It quantifies how closely the data points in a scatterplot follow a straight line. The value of Pearson's R ranges from -1 to 1.

a. Formula for Pearson's R

The formula for calculating Pearson's R is:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where:

- r is Pearson's correlation coefficient
- n is the number of data points
- x and y are the two variables

Alternatively, it can be written as:

$$r = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Where:

- **Cov(X, Y)** is the covariance between X and Y
- **σX** and **σY** are the standard deviations of X and Y

b. Interpretation of Pearson's R

- r = 1: Perfect positive linear relationship (as one variable increases, the other increases proportionally).
- r = -1: Perfect negative linear relationship (as one variable increases, the other decreases proportionally).

- r = 0: No linear relationship (the variables are uncorrelated).

Values between -1 and 1 indicate the strength and direction of the relationship:

- 0.1 to 0.3 (or -0.1 to -0.3): Weak linear relationship
- 0.3 to 0.7 (or -0.3 to -0.7): Moderate linear relationship
- 0.7 to 1.0 (or -0.7 to -1.0): Strong linear relationship

c. Limitations

- Pearson's R only measures linear relationships. It may not be accurate for non-linear data.
- It is sensitive to outliers, which can distort the correlation.
- It assumes the variables are continuous and follow a normal distribution.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
Answer- Scaling is a data pre-processing procedure that normalizes data within a specific range. It's used to ensure that all values in a dataset are within a certain range. It accelerate algorithmic calculations.

The primary purpose of scaling is to ensure that all variables have equal influence on the analysis, as many statistical and machine learning algorithms are sensitive to the scale of the input data. Scaling helps in making comparisons and calculations more meaningful and can improve the performance of certain algorithms.

Normalization and standardization are two scaling methods. Normalization typically rescales values into a range in which the data doesn't have Gaussian distribution and is highly affected by outliers

Standardization typically rescales data to have a mean of 0 and a standard deviation of 1 and is being used on data having Gaussian distribution and this is not bounded by range.

Key differences between normalization and standardization:

- Range: Normalization scales values to a specific range (usually 0 to 1), while standardization rescales values to have a mean of 0 and a standard deviation of 1.
- Impact on Distribution: Normalization maintains the distribution's shape and only changes the scale, while standardization centers the data around 0 and changes the spread.
- Outliers: Standardization is less affected by outliers compared to normalization. Outliers can disproportionately impact the range-based scaling used in normalization.
- Interpretability: Normalization retains the original units of measurement, while standardization transforms data into z-scores, making it unit less.

The choice between normalization and standardization depends on the specific requirements of your analysis and the characteristics of your data. For instance, if you're working with data where the variable values are on different scales, normalization may be more appropriate. If your data is expected to follow a normal distribution and you want to mitigate the influence of outliers, standardization may be preferred. In practice, it's common to experiment with both methods and choose the one that results in better model performance or more meaningful insights.

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

Answer: Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression models. It quantifies how much the variance of a regression coefficient is inflated due to collinearity with other independent variables. A higher VIF indicates a higher degree of multicollinearity.

The formula for VIF is:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}$$

Where R_i^2 is the coefficient of determination (R-squared) when the iii-th independent variable is regressed on the other independent variables.

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 plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically a normal distribution. It plots the quantiles of the observed data against the quantiles of the theoretical distribution. If the points in the Q-Q plot fall roughly along a straight line, it indicates that the data follows the theoretical distribution (e.g., normal distribution).

Use and Importance of Q-Q Plot in Linear Regression

In linear regression, a Q-Q plot is crucial for evaluating the assumption of normality of the residuals (errors). Linear regression models assume that the residuals are normally distributed. If this assumption holds, the model's predictions are more reliable, and hypothesis tests (like t-tests and F-tests) are valid.

a. Checking Normality of Residuals

- In linear regression, the residuals (difference between actual and predicted values) should follow a normal distribution.

- A Q-Q plot helps visualize this by plotting the quantiles of the residuals against the quantiles of a normal distribution. If the residuals are normally distributed, the points will lie along a 45-degree straight line.

b. Importance of Normality in Linear Regression

- Validating statistical tests: The assumption of normality in residuals ensures that the p-values and confidence intervals calculated from t-tests and F-tests are valid.
- Predictive accuracy: A normal distribution of residuals ensures that errors are randomly distributed, leading to more reliable predictions from the model.

c. Interpreting a Q-Q Plot

- Straight Line: If the points in the Q-Q plot fall along a straight line, it indicates that the residuals are normally distributed, and the linear regression assumptions are likely valid.
- Deviations: If the points deviate from the straight line, it suggests that the residuals are not normally distributed, indicating potential issues such as skewness, heavy tails, or outliers. This could signal that the model might not fit the data well, and the normality assumption is violated.