Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 1 goes below this line> (Do not edit)

- **Year**: Bike demand increased in 2018 compared to the previous year, showing a positive trend in usage.
- **Weekday**: Demand remained steady across most weekdays, with a small rise observed on Sundays.
- **Holiday**: While demand patterns on holidays and non-holidays were similar, the average demand was slightly higher on non-holiday days.
- **Month**: Demand was highest in August, September, and October, while it was lower during the extreme months of January and July.
- **Weather**: Clear weather was associated with higher demand for bikes, whereas misty or cloudy conditions tended to reduce demand.
- **Season**: Demand for bikes was higher in the summer and winter seasons, while spring and fall saw relatively lower levels of demand.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 2 goes below this line> (Do not edit)

By removing perfect correlation among dummy variables, this approach effectively prevents multi collinearity and ensures linear independence by reducing the number of dummy variables from n to $n-1$. This reduction simplifies the regression model, minimizing complexity from redundant variables and enhancing the model's interpretability by concentrating on the influence of the retained categories. It preserves all essential information related to the categorical variable while eliminating unnecessary redundancy.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

The variables **temp** and **atemp** exhibit a strong correlation, with the heat map indicating a correlation coefficient of 0.63.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

a. Checked for Multi collinearity by calculating **VIF** (Variance Inflation Factor). In the features

considered for building the model the **VIF** is less than 5 thereby ensuring no redundant variables are considered for building the model

b. Validated whether Residuals have a Normal Distribution or not and centered around 0 (i.e., Mean = 0)
c. Validated the Linearity by plotting the residuals & predicted values & noticed the points are randomly dispersed without skewness
d. Homoscedasticity - Looking for constant variance across all levels of fitted values

---

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

- Variations in weather plays a key role in bike demand. Days with warmer temperature has more demand.
- Rainy/Cloudy weather tend to decrease the bike demand
- Increase in bike rentals compared in 2019 to 2018

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear Regression Overview

Linear regression is a statistical technique used to examine the relationship between variables, helping us estimate the value of one variable based on another.

Independent Variable (X): This variable is used for making predictions.
Dependent Variable (Y): This is the outcome variable we aim to predict.

The objective of linear regression is to determine the best-fitting line, known as the regression line, which represents the relationship between the independent and dependent variables.

Types of Linear Regression

Simple Linear Regression: Involves a single predictor variable, represented by the model:

$$y=\beta 0+\beta 1x+\epsilon$$

Multiple Linear Regression: Incorporates several predictor variables, represented by the model as:

$$y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta nxn+\epsilon$$

---

where:

β represents the coefficients of each predictor, while
x stands for the independent variables.
Y is the dependent variable.
ϵ is the residual, or error term.

How Linear Regression Works

Linearity: A linear relationship exists between the predictors and the target variable.
Independence: Observations are independent of one another.
Homoscedasticity: Errors show constant variance.
Normality: Errors are normally distributed.

Fitting the Model: The goal is to draw the line that best fits the data points by minimizing the squared differences between actual and predicted values (least squares method).

Coefficient Estimation: Using the least squares method, the coefficients β are calculated to minimize the sum of squared residuals:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Where:
yi = observed value (actual value)
yi^ = predicted value (from the model)
n = number of observations

Evaluating the Model
R-squared (R^2): Indicates the proportion of variance in the dependent variable that is explained by the independent variables, ranging from 0 to 1.
Adjusted R-squared: Provides a more accurate measure of R^2, adjusting for the number of predictors when multiple variables are involved.
Mean Squared Error (MSE): Calculates the average of the squared differences between actual and predicted values.

Limitations of Linear Regression
Linear regression is limited to linear relationships, which may not always reflect real-world data. Issues such as overfitting, multi collinearity, and the presence of outliers can also adversely impact the model's accuracy and reliability.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

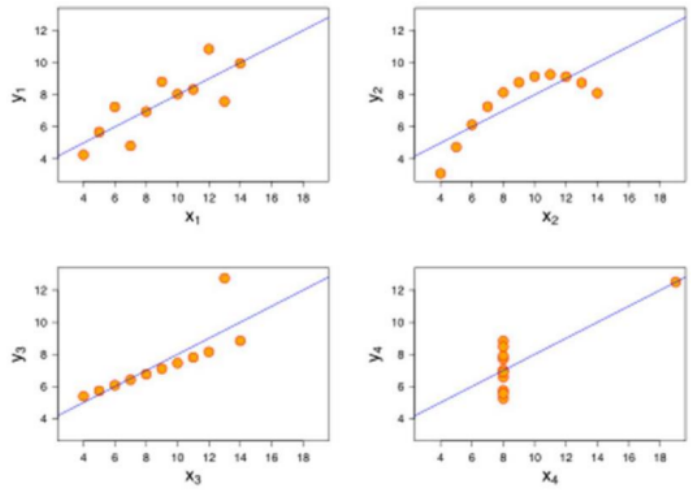
Anscombe's Quartet is a collection of four sets of data that look very different when you plot them, but all have the same basic statistics (like average and correlation).

- It teaches us that looking at the actual data through graphs is really important. Just because numbers look the same doesn't mean the data behaves the same way.
- If we only look at numbers (like averages and correlations), we might think the data is telling us one thing, but the graphs can show a completely different story.
- So, basically reminds us to always visualize our data because it can reveal important details that numbers alone might hide!

Here are the key points: Each dataset consists of 11 pairs of (x,y) values. The four datasets are:
Dataset I: A linear relationship with a positive correlation.
Dataset II: A nonlinear relationship with a parabolic shape.
Dataset III: A linear relationship with an outlier that heavily influences the slope.
Dataset IV: A vertical line indicating a constant xxx value with varying y values, showing no correlation.

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



Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R is also known as Pearson correlation coefficient. It is a statistical measure that evaluates the strength and direction of a linear relationship between two continuous variables.

Pearson's R ranges from -1 to 1.

- -1 indicates a perfect negative linear relationship.
- 0 indicates no linear relationship.
- 1 indicates a perfect positive linear relationship.

Values close to 1 or -1 signify a strong relationship, while values close to 0 suggest a weak relationship.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

If one feature has significantly larger values (like annual revenue in millions) compared to another feature with smaller values (such as the number of employees), it can dominate the model's predictions, potentially skewing the results.

- Scaling is a technique to adjust feature values so they are on a similar scale.
- This adjustment ensures no single feature overshadows others due to its magnitude.

In linear regression, there are two primary scaling techniques:

1. **Normalized Scaling / Min-Max Scaling**

- Normalization scales feature values to lie between 0 and 1.
- This method is beneficial when we do not assume any particular distribution and simply want all values within a defined, consistent range.

Formula for Normalization: $x' = (x - \min(x)) / (\max(x) - \min(x))$

Where:

- x is the original value.
- min(x) is the minimum value of the feature.
- max(x) is the maximum value of the feature.
- x' is the normalized value.

When to Use: Normalization is appropriate when the data does not follow a normal (Gaussian) distribution, and when we want values constrained within a specific range.

2. **Standardized Scaling**
   - Standardization centers feature values around 0, with most values ranging between -1 and 1.
   - This approach is most effective when data roughly follows a bell curve, as it enhances the

accuracy of various predictive models.

Formula for Standardization: $x' = x - \mu / \sigma$

- x is the original value.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.
- x' is the standardized value.

When to Use: Standardization is useful for comparing features with different units or scales, especially if the data approximates a Gaussian distribution.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

Here is the rephrased version with formal and simple language, along with a new example:

From my analysis, I have identified three main reasons why Variance Inflation Factor (VIF) can sometimes be infinite. These are:

1. **Perfect Multi collinearity** occurs when one independent variable is an exact linear combination of other variables in the model.

Example: Consider variables such as "Years of Experience" and "Months of Experience." These variables represent the same information but in different units, leading to redundancy and perfect multi collinearity.

2. **Absence of the Constant (Intercept) Term**
When building a model, especially using the `statsmodels` package, it is essential to explicitly add a constant term to the model. This term does not appear by default. If the constant term is omitted, some variables may become linearly dependent on others, resulting in high or infinite VIF values.

Example: Omitting the intercept term during regression analysis can distort the relationship between variables, often leading to artificially high VIF values.

3. **Errors During Dummy Variable Creation**
When creating dummy variables for categorical data, it is important to drop one category to avoid multi collinearity. The number of dummy variables should always be one less than the number of unique categories in the original variable.

Example: When encoding a "City" variable with categories such as "New York," "Los Angeles," and "Chicago," we must drop one of these cities (e.g., "New York") to prevent perfect collinearity and

avoid high VIF values.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

- In the term Q-Q Plot, Q-Q Stands for Quantile-Quantile.
- It is a graph that helps us see if data follows a specific distribution, usually a normal distribution.
- In Linear Regression, Q-Q Plot is used in Residual Analysis. It helps to check if the Residuals Are Normally distributed.
- We will plot the quantiles of the residuals against the quantiles of a normal distribution.
- If the points mostly line up in a straight line, it indicates the data is normally distributed.
- Deviations from a straight line suggest the data might have issues like skewness, which could affect the accuracy of the regression model.

How a Q-Q Plot Works:

1. Quantiles: Quantiles are points taken at regular intervals from the cumulative distribution function (CDF) of a random variable. For example, the median is the 0.5 quantile.
2. Plotting: In a Q-Q plot, the quantiles of the sample data are plotted against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points should lie approximately on a straight line

Intercepting a Q-Q Plot:

Straight Line: If the points lie on or near the straight line, the residuals are approximately normally distributed.
S-shaped Curve: Indicates heavy tails
Inverted S-shaped Curve: Indicates light tails
Deviations at Ends: Suggest skewness in the data.
