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:

Based on my analysis of the dataset, I've identified two types of data: categorical and numeric. For analyzing the categorical data, I used box plots to visually explore relationships and distributions. Here are the key insights from my categorical data analysis:

Initial Exploration:

I started by checking the list of columns in the dataset using df.columns, which returned all the column names from the pandas DataFrame created from the CSV file. Then, I examined the first few rows of the data using head() to identify which columns were numeric and which were categorical.

Categorical Data Analysis Using Box Plots:

For the categorical variables, I created box plots to investigate their relationships with the dependent variable. The key categorical variables I analyzed are:

- **Weathersit**: This variable represents the weather conditions and has four categories: "Clear", "Misty", "Light Snow Rain", and "Heavy Snow Rain". From the box plot, I observed that the majority of the bike rentals occurred in "Clear" weather, which was the most frequent category.
- **Workingday**: This variable indicates whether the day is a working day or a holiday. The analysis revealed that bike rentals were more frequent on working days compared to holidays, suggesting a higher demand for bikes during weekdays.
- **Yr (Year)**: This variable compares bike rentals in 2018 and 2019. The data showed that 2019 had a higher count of bike rentals compared to 2018, indicating an overall increase in bike usage in the more recent year.
- **Month**: The analysis of monthly bike rentals revealed that the highest number of bike rides occurred in **September**, **August**, **October**, and **May**, with these months showing more activity compared to others.
- **Season**: The season variable showed that bike rentals were most frequent during the **Summer** and **Autumn** seasons, likely due to favorable weather conditions in those months.

Implications for Dummy Variables:

Based on these insights, I determined that creating dummy variables for these categorical columns would be essential for preparing the data for linear regression modeling. Dummy variables will allow us to convert these categorical features into a format that can be used in machine learning models, ensuring compatibility with both training and testing datasets.
By using these box plots and analyzing the trends in the categorical data, I gained a deeper understanding of the relationships between weather, workdays, seasons, and bike rental patterns, which will inform the creation of the necessary dummy variables for further modeling.

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:

When converting categorical variables with pd.get_dummies(), the drop_first parameter controls whether the first category is included in the dummy variables.

- **drop_first=False**: All categories are included as separate binary columns.
- **drop_first=True**: The first category is dropped, avoiding multicollinearity and simplifying model interpretation by using it as the reference category.

In summary, drop_first=True reduces redundancy and makes the model easier to interpret.
When using pd.get_dummies(), setting drop_first=True drops the first category and avoids multicollinearity by using it as the reference category. This results in fewer dummy variables and simplifies model interpretation.

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:

“temp” column has the highest correlation in the data.

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:

I validated the assumptions of the Linear Regression model based on the following key criteria:

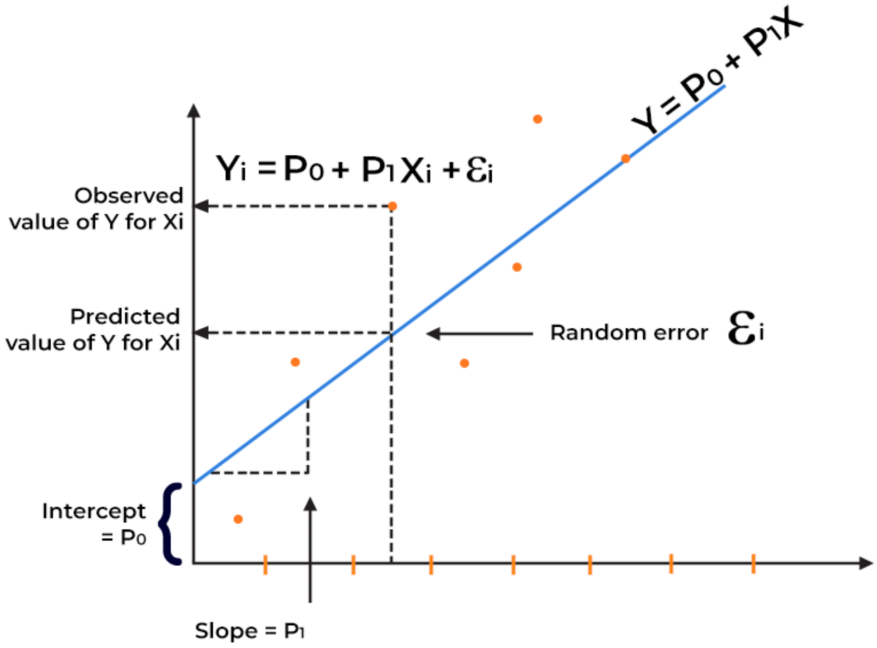
- **Normality of Residuals**: The residuals (error terms) should follow a normal distribution for accurate statistical inference. I tested this assumption to ensure the error terms are normally distributed.
- **Multicollinearity**: Multicollinearity occurs when independent variables are highly correlated, which can affect model stability. I checked for multicollinearity using variance inflation factors (VIFs) to ensure no significant multicollinearity among predictors.
- **Linearity**: Linear regression assumes a linear relationship between independent and dependent variables. I validated this assumption by examining scatter plots and residual plots to ensure a linear trend.
- **Homoscedasticity**: Homoscedasticity means the variance of residuals is constant across all levels of the independent variables. I checked residual plots for any patterns and confirmed that the variance remains consistent.
- **Independence of Residuals**: Residuals should be independent, with no autocorrelation between them. I tested this assumption using the Durbin-Watson statistic to confirm there is no significant autocorrelation.
- **Outliers & Influential Points:** Outliers can distort model results. I identified and addressed any influential points using leverage and Cook’s distance to ensure they did not unduly affect the model’s performance.

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:
The top three features significantly influencing the demand for shared bikes are:
Temperature (temp): Higher temperatures lead to increased bike demand, as people are more likely to ride bikes in warmer weather.
Winter (winter): The winter season typically results in a decrease in bike rentals due to colder weather conditions.
September (sep): Bike demand is notably higher in September, likely due to favorable weather and the end of summer vacations.

General Subjective Questions
Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer:

Linear Regression Algorithm

Linear Regression is a supervised machine learning algorithm used to model the relationship between a dependent variable (target) and one or more independent variables (predictors).


$Y = P_0 + P_1X$
$Y_i = P_0 + P_1X_i + \mathcal{E}_i$
Observed value of Y for Xi
Predicted value of Y for Xi
Random error $\mathcal{E}_i$
Intercept = P_0
Slope = P_1

1. Basic Concept:

In **simple linear regression**, the relationship between a single independent variable xx and the dependent variable yy is modeled as a straight line:

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

For **multiple linear regression**, the equation extends to include multiple predictors:

$$y=\beta0+\beta1 \cdot x1+\beta 2 \cdot x2+\cdots+\beta p \cdot xp+\epsilon$$

2. Objective:

The goal is to minimize the **Mean Squared Error (MSE)** between the actual and predicted values:

MSE=1/n(∑ n(yi−y^i)^2)

Where:

- yi is the actual value
- y^i is the predicted value
- i =0
- n = 0 ….n

3. Model Fitting:

Linear regression is fitted using:

- **Ordinary Least Squares (OLS)**: Direct computation of coefficients.
- **Gradient Descent**: Iterative optimization to minimize MSE.

4. Assumptions:

Linear regression assumes:

- **Linearity**: The relationship between variables is linear.
- **Independence of errors**: Residuals are independent.
- **Homoscedasticity**: Residuals have constant variance.
- **Normality of errors**: Residuals are normally distributed.
- **No multicollinearity**: Independent variables are not highly correlated.

5. Evaluation Metrics:

- **R-squared (R^2)**: Proportion of variance explained by the model.
- **Mean Absolute Error (MAE)**: Average absolute difference between actual and predicted values.
- **Root Mean Squared Error (RMSE)**: Square root of the average squared errors.
- **F-statistic**: Tests model significance in multiple regression.

6. Positive and Negative Linear Regression:

Linear regression models can have either a **positive** or **negative** relationship between the dependent and independent variables based on the sign of the slope coefficient β1\beta_1.

- **Positive Linear Regression** ($\beta1>0$): This indicates a **positive relationship**. As the independent variable xx increases, the dependent variable yy also increases. Mathematically, the equation becomes:

  $y=\beta0+\beta1 \cdot x$ where $\beta1>0$

  Example: Higher advertising spending may lead to higher sales.

- **Negative Linear Regression** ($\beta1<0$): This indicates a **negative relationship**. As the independent variable xx increases, the dependent variable y decreases. Mathematically, the equation becomes:

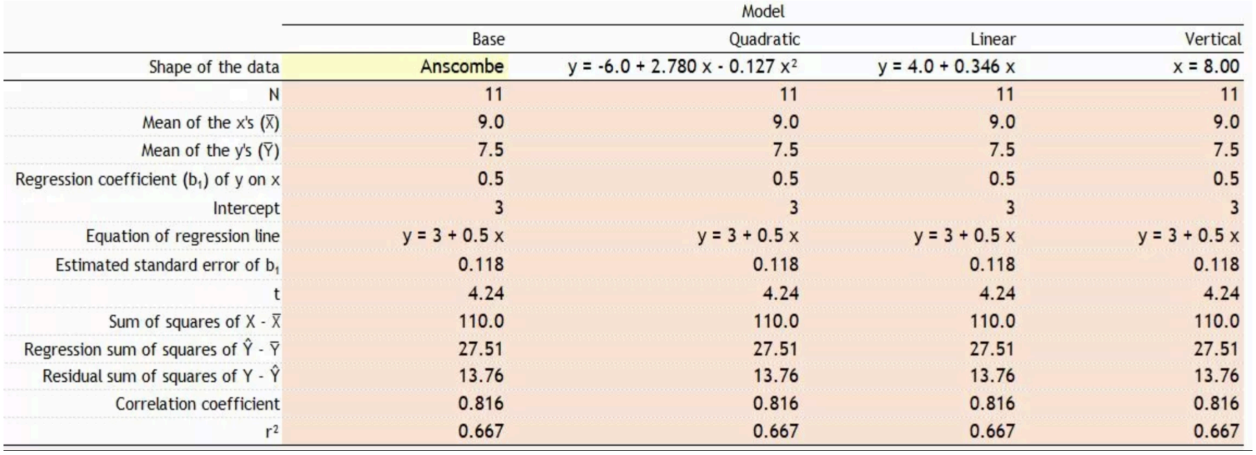
  $y=\beta0+\beta1 \cdot x$ where $\beta1<0$

Example: Increased hours of work might decrease free time, resulting in lower satisfaction.

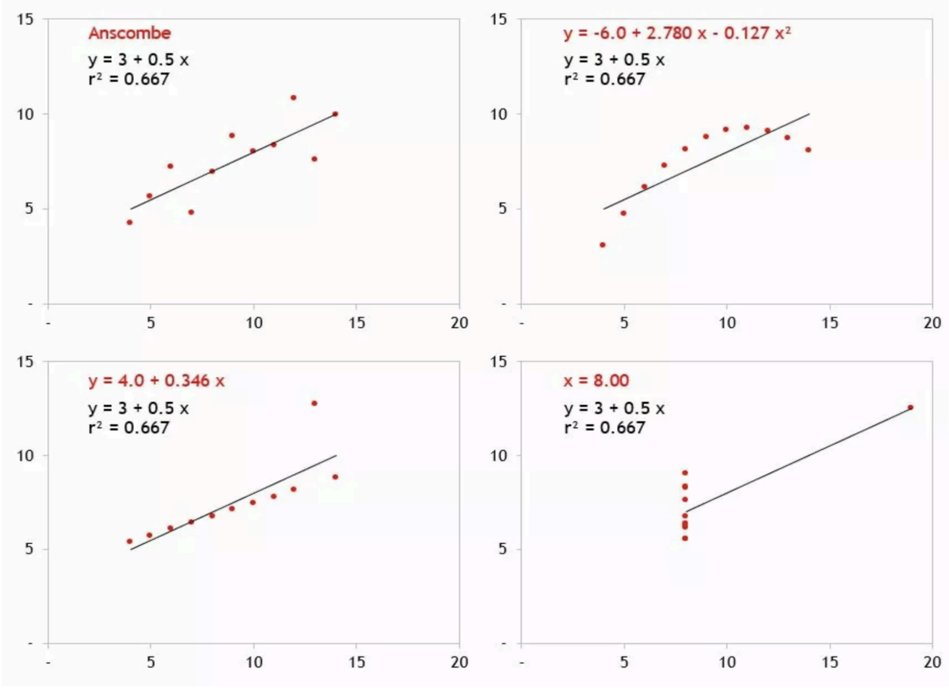
In **multiple linear regression**, a similar concept applies where the effect of each independent variable is reflected by the corresponding coefficient βi\beta_i. A **positive coefficient** means the predictor is positively correlated with the target variable, and a **negative coefficient** means an inverse relationship.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer:

Anscombe's Quartet is a set of four datasets, each with 11 data points, that were constructed by the statistician **Francis Anscombe** in 1973. The purpose of the quartet is to demonstrate how summary statistics (such as mean, variance, and correlation) can be misleading if the underlying data is not properly visualized. Despite having nearly identical statistical properties, the datasets exhibit very different characteristics when plotted.



	Model			
	Base	Quadratic	Linear	Vertical
Shape of the data	Anscombe	$y = -6.0 + 2.780\,x - 0.127\,x^2$	$y = 4.0 + 0.346\,x$	$x = 8.00$
N	11	11	11	11
Mean of the x's ($\bar{X}$)	9.0	9.0	9.0	9.0
Mean of the y's ($\bar{Y}$)	7.5	7.5	7.5	7.5
Regression coefficient (b_1) of y on x	0.5	0.5	0.5	0.5
Intercept	3	3	3	3
Equation of regression line	$y = 3 + 0.5\,x$	$y = 3 + 0.5\,x$	$y = 3 + 0.5\,x$	$y = 3 + 0.5\,x$
Estimated standard error of b_1	0.118	0.118	0.118	0.118
t	4.24	4.24	4.24	4.24
Sum of squares of $X - \bar{X}$	110.0	110.0	110.0	110.0
Regression sum of squares of $\hat{Y} - \bar{Y}$	27.51	27.51	27.51	27.51
Residual sum of squares of $Y - \hat{Y}$	13.76	13.76	13.76	13.76
Correlation coefficient	0.816	0.816	0.816	0.816
r^2	0.667	0.667	0.667	0.667



Summary of the Datasets:

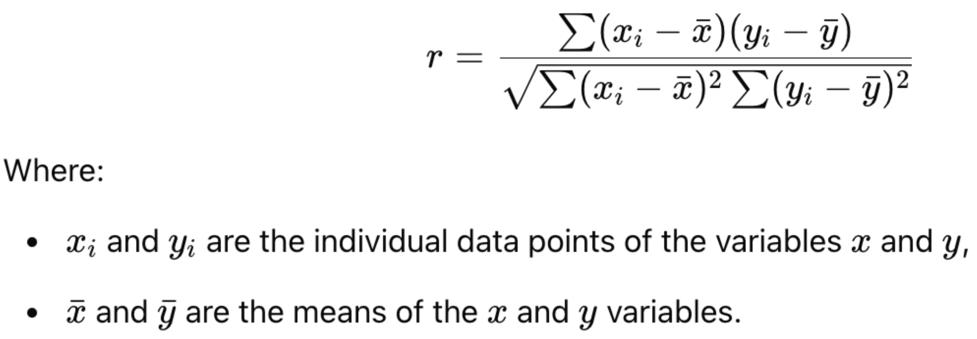
1. **Dataset 1 (Linear Relationship)**: This dataset shows a **linear relationship** between the variables xxx and yyy. The data points align closely with the regression line, confirming the correlation and the positive linear trend.
2. **Dataset 2 (Nonlinear Relationship)**: In this dataset, the relationship between xxx and yyy is **nonlinear**. While the correlation remains high, the points deviate from the regression line, indicating a **curved** pattern rather than a linear one.
3. **Dataset 3 (Outliers)**: This dataset is influenced by **outliers**, which skew the relationship between xxx and yyy. Although most of the points follow the linear trend, one extreme outlier causes the regression line to be affected, distorting the true relationship.
4. **Dataset 4 (Vertical Line of Points)**: This dataset shows a scenario where all but one data point lie on a **vertical line** (indicating that xxx is constant), except for one outlier. The correlation value is artificially high due to the presence of the outlier, but the dataset fails to provide a meaningful linear relationship.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer:

Pearson's R, also known as the **Pearson correlation coefficient**, is a measure of the **linear relationship** between two continuous variables. It quantifies the strength and direction of the association between the variables, with a value ranging from -1 to 1.

Key Points:

1. **Formula**:
   Pearson's R is calculated using the following formula:

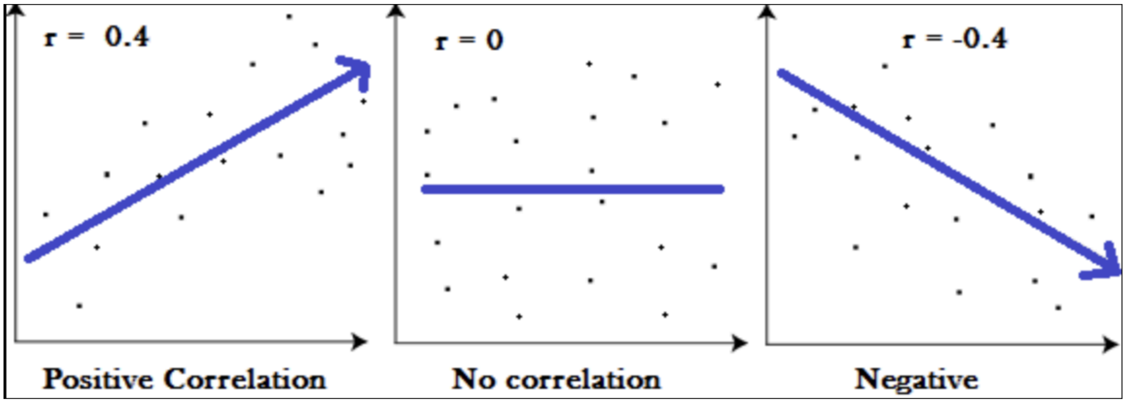


$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Where:

- x_i and y_i are the individual data points of the variables x and y,
- $\bar{x}$ and $\bar{y}$ are the means of the x and y variables.

2. **Interpretation**:
   a. **r=1**: Perfect positive linear relationship (as one variable increases, the other also increases).
   b. **r=−1**: Perfect negative linear relationship (as one variable increases, the other decreases).
   c. **r=0:** No linear relationship between the variables.
   d. **0 < r < 1**: A positive correlation, with higher values indicating a stronger positive relationship.
   e. **-1 < r < 0**: A negative correlation, with lower values indicating a stronger negative relationship.



3. **Assumptions**:
   a. The variables should have a **linear relationship**.
   b. Both variables should be **continuous**.
   c. The data should be **normally distributed** for meaningful statistical inference.

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:

Scaling refers to the process of adjusting the range or distribution of features (independent variables) in a dataset so that they are on a similar scale. This is particularly important for algorithms like **linear regression** where the magnitude of features can impact model performance, especially when features vary widely in scale.

Why is Scaling Performed?

1. **Faster Convergence**: Scaling prevents features with larger values from dominating the optimization process, enabling faster convergence in algorithms like gradient descent.
2. **Equal Contribution**: Ensures each feature contributes equally to the model, improving model accuracy.
3. **Improved Coefficient Interpretation**: Scaled data helps in better interpreting how each feature affects the target variable.

Normalized vs Standardized Scaling

1. **Normalized Scaling (Min-Max)**:
   ○ Rescales the data to a fixed range, typically between 0 and 1.
   ○ Sensitive to outliers: Extreme values can skew the scaling.
   ○ Useful when the features have a known, bounded range (e.g., pixel values or rates between 0 and 1).
   ○ Maintains the relative distribution of the data but may distort if the range is wide or outliers are present.

2. **Standardized Scaling (Z-score)**:
    - Centers the data around 0 with a standard deviation of 1.
    - Less sensitive to outliers compared to normalization, but extreme values still influence the scaling.
    - Preferred when the data has varying units or a distribution that is not bounded, like age, income, or other continuous variables.
    - Does not constrain the values to a specific range, making it suitable for models that assume normally distributed data.

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:

The **Variance Inflation Factor (VIF)** is a measure used to assess **multicollinearity** in a dataset, indicating how much the variance of a regression coefficient is inflated due to correlations with other independent variables.

- **VIF and Perfect Correlation**:
  If there is **perfect correlation** between two independent variables, the **VIF becomes infinite (∞)**. This happens because the R^2 value equals 1, leading to a division by zero in the VIF formula:

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

- **Large VIF**:
  A **large VIF** indicates **multicollinearity**, meaning the variance of the regression coefficient is inflated due to correlations between predictors. For example, a VIF of 4 means the coefficient variance is four times larger than it would be without multicollinearity.

- **Solution for Perfect Multicollinearity**:
  If VIF is infinite, it shows **perfect multicollinearity**. To resolve this, **drop one of the correlated variables** from the dataset.

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:

A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, such as the **normal distribution**. It plots the quantiles of the observed data against the quantiles of the expected distribution.

Use and Importance of a Q-Q Plot in Linear Regression:

- **Check for Normality of Residuals**:
  In linear regression, one of the assumptions is that the residuals (the differences between the observed and predicted values) should be normally distributed. A Q-Q plot helps assess this assumption. If the residuals fall roughly along a straight line in the plot, it suggests that they are normally distributed, which is crucial for valid hypothesis testing and confidence intervals.
- **Identify Outliers**:
  Points that deviate significantly from the line in a Q-Q plot represent **outliers** in the data. Outliers can influence regression results, making it important to identify and address them for more robust models.
- **Model Diagnostics**:
  By examining the Q-Q plot of residuals, you can identify potential problems with the model, such as non-normality of residuals, which could affect the reliability of regression estimates. Non-normal residuals may suggest the need for a data transformation or a different modeling approach.