Assignment based subjective questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: The categorical variables in the dataset were “season”, “year”, “month”, “holiday”, “weekday”, “workingday” and “weathersit”.

These were visualized using a boxplot. These variables had the following effect on our dependent variable.

Season:

Fall: Bike demand is highest in Fall.

Summer and Winter: Summer and winter have intermediate value of count with summer having greater count among the two.

Spring: The demand for bikes is lowest in spring, possibly due to less favorable weather.

Year:

2019 vs. 2018: There is a clear increase in bike demand from the year 2018 to 2019. This trend suggests that the bike sharing program gained popularity over this period.

Month:

High-Demand Months: June, July, August and September are the months with the highest bike demand.

Out of all these months, September has seen the highest no of rentals.

This could be due to the warm summer weather, which is ideal for biking.

Low-Demand Months: December has seen the lowest no of rentals. January, February, and December see the less bike demand, likely due to colder winter weather, which discourages biking

Holiday:

Holidays vs Non-holidays: Bike demand is higher on holidays compared to non-holidays. This increase can be attributed

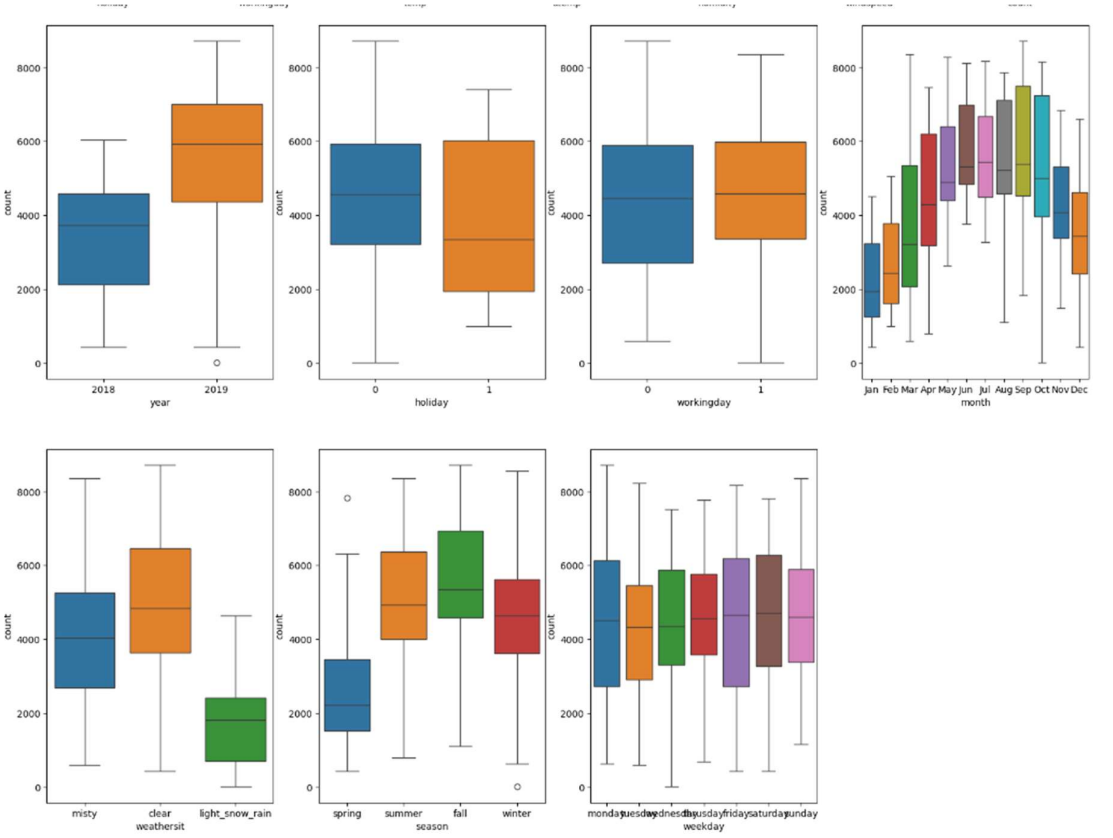
to people having more leisure time and choosing to bike for recreation or errands on holidays.

Weekday:

Even Distribution: Bike demand is relatively evenly distributed across all weekdays, indicating consistent usage throughout the week.

Weathersit:

1. Clear Weather: The highest bike demand occurs during clear weather conditions, due to favorable weather.
2. Adverse Weather: Bike demand decreases significantly during misty conditions, light snow/rain, and heavy snow/rain. There are no users when there is heavy snow/rain. The least demand is observed during light snow/rain, as adverse weather conditions make biking less appealing and potentially hazardous.



Q2. Why is it important to use drop_first=True during dummy variable creation?

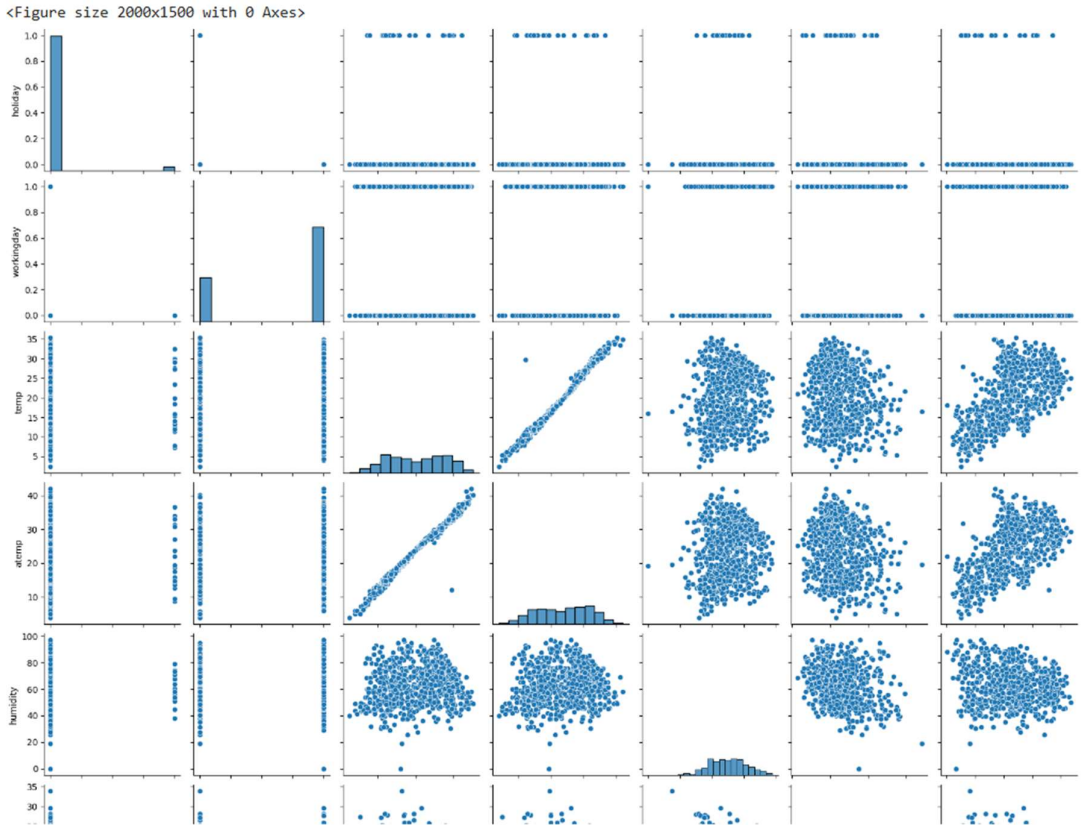
Ans: Using 'drop_first=True' during dummy variable creation is important for the following reasons:

- Preventing Multicollinearity: Including all dummy variables for a categorical feature can lead to multicollinearity, which occurs when predictor variables are highly correlated. This can make it difficult to determine the individual effect of each variable on the target variable. By dropping the first dummy variable, we avoid this issue and ensure that the remaining dummies can provide the necessary information without redundancy.

- Reducing Redundancy: By dropping the first category, the total number of dummy variables is reduced, which simplifies the model and improves efficiency.
- Model Interpretability: This practice ensures that the model remains interpretable and free from redundant variables, making it easier to understand and analyze the effects of other predictors.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: In the pair-plot analysis, the two temperature variables, "temp" and "atemp" , show the highest correlation with the target variable "count" or "cnt" . This strong positive correlation indicates that higher temperatures are associated with an increase in bike bookings.



Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans:- We have validated the assumption of Linear Regression Model based on below 5 assumptions –

- **Normality of Error Terms**: If the residuals follow a normal distribution, the assumption is met.

  Histogram: Plotted a histogram of the residuals. If the residuals are normally distributed, the histogram should resemble a bell curve.

  Q-Q Plot: Plotted a Q-Q plot of the residuals. If the residuals are normally distributed, the points should lie along the 45-degree line.

- **Multicollinearity Check:**

  Variance Inflation Factor (VIF): Calculated the VIF for each predictor variable. VIF values less than 10 indicate that multicollinearity is not a concern.

- **Linear Relationship Validation:**

  Residual Plot: Plotted residuals against the predicted values. If the residuals are randomly scattered around zero, it suggests that there is a linear relationship between the predictors and the response variable.

- **Homoscedasticity:**

  Residuals vs. Predicted Plot: Plotted residuals against the predicted values to check for constant variance. The absence of a clear pattern indicates homoscedasticity.

- **Independence of residuals:**

  Tested with residuals and curves.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand for the shared bikes?

Ans: -

Top features for the model

- temp coef:0.57

- year_2019 coef:0.23

- light snow coef:-0.2425

Final Equation

count (cnt) = 0.1907 + workingday * 0.0526+ temp * 0.5684 - humidity * 0.1643 - windspeed * 0.1943 + year_2019 * 0.2296 - month_Jan * 0.0401 - month_Jul * 0.0429 + month_Sep * 0.0909 + weekday_monday * 0.0629 + season_summer * 0.0765 + season_winter * 0.1251 - weathersit_light_snow_rain * 0.2425 - weathersit_misty * -0.0538

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.846
Model:	OLS	**Adj. R-squared:**	0.842
Method:	Least Squares	**F-statistic:**	194.7
Date:	Wed, 04 Sep 2024	**Prob (F-statistic):**	6.46e-191
Time:	06:46:23	**Log-Likelihood:**	516.45
No. Observations:	510	**AIC:**	-1003.
Df Residuals:	495	**BIC:**	-939.4
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.1953	0.030	6.610	0.000	0.137	0.253
workingday	0.0518	0.011	4.769	0.000	0.030	0.073
temp	0.5319	0.030	17.569	0.000	0.472	0.591
humidity	-0.1619	0.037	-4.337	0.000	-0.235	-0.089
windspeed	-0.1902	0.026	-7.457	0.000	-0.240	-0.140
year_2019	0.2298	0.008	28.598	0.000	0.214	0.246
month_Aug	0.0418	0.019	2.170	0.030	0.004	0.080
month_Jan	-0.0382	0.017	-2.200	0.028	-0.072	-0.004
month_Jul	-0.0176	0.021	-0.829	0.407	-0.059	0.024
month_Sep	0.1085	0.018	6.096	0.000	0.074	0.144
weekday_monday	0.0612	0.014	4.369	0.000	0.034	0.089
season_summer	0.0917	0.013	7.077	0.000	0.066	0.117
season_winter	0.1340	0.012	11.119	0.000	0.110	0.158
weathersit_light_snow_rain	-0.2429	0.026	-9.302	0.000	-0.294	-0.192
weathersit_misty	-0.0557	0.010	-5.355	0.000	-0.076	-0.035

Omnibus:	67.878	**Durbin-Watson:**	2.057
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	166.574
Skew:	-0.688	**Prob(JB):**	6.74e-37
Kurtosis:	5.438	**Cond. No.**	20.6

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans) Linear regression is one of the most fundamental and widely used algorithms in statistical modeling and machine learning. It models the relationship between a dependent variable (target) and one or more independent variables (features) by fitting a linear equation to the observed data.

a) Linear regression aims to predict the value of a dependent variable Y based on the value(s) of one or more independent variables X1, X2, …., Xn. The relationship between the dependent and independent variables is modeled as a linear equation:

Simple Linear Regression (one independent variable):

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Y is the dependent variable.

- X is the independent variable.

- β_0 is the intercept (the value of Y) when (X = 0)

- β_1 is the slope of the line (how much Y changes for a unit change in X)

- ϵ is the error term (the difference between the observed and predicted values).

- Multiple Linear Regression (more than one independent variable):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

- Y is the dependent variable.

- $X_1, X_2, \ldots, X_n$ are the independent variables.

- β_0 is the intercept.

- $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients corresponding to each independent variable.

- ϵ is the error term.

b) **Objective of Linear Regression:-**

The main objective of linear regression is to find the best-fitting line through the data points. This line is called the **regression line**. The best-fitting line is the one that

minimizes the sum of the squared differences (errors) between the observed values and the values predicted by the line. This method is known as **Ordinary Least Squares (OLS)**

c) **Ordinary Least Squares (OLS) Method:-**

- **Error Calculation:** The difference between the observed value Y_i and the predicted value $\hat{Y}_i$ is called the residual or error. For each data point, the error is:

$$\text{Error} = Y_i - \hat{Y}_i$$

- **Sum of Squared Errors (SSE):** The OLS method seeks to minimize the sum of the squared errors:

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

- **Minimizing SSE:** The coefficients $\beta_1, \beta_2, \ldots, \beta_n$ are chosen to minimize the SSE. The solution involves calculus, specifically taking the derivative of SSE with respect to each coefficient and setting it to zero to find the minimum.

d) Assumptions of Linear Regression:-

Linear regression relies on several key assumptions:

1. Linearity: The relationship between the independent and dependent variables is linear.

2. Independence: The observations are independent of each other.

3. Homoscedasticity: The variance of the residuals (errors) is constant across all levels of the independent variables.

4. Normality of Errors: The residuals (errors) are normally distributed.

Violating these assumptions can lead to biased or misleading results.

e) Evaluating the Model:-

After fitting the linear regression model, it's important to evaluate its performance. Key metrics include:

(i) R-squared (Coefficient of Determination):

- Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

- Ranges from 0 to 1, where 0 indicates that the model explains none of the variance and 1 indicates that it explains all the variance.

- **Adjusted R-squared** is used in multiple regression to account for the number of predictors in the model.

(ii) p-values:

- Test the hypothesis that a given coefficient is different from zero (i.e., the independent variable has an effect on the dependent variable).

- A low p-value (typically < 0.05) indicates that the corresponding variable is statistically significant.

(iii) F-statistic:

- Tests the overall significance of the model (whether the model with predictors is better than a model with no predictors).

- A high F-statistic with a low p-value suggests that the model is statistically significant.

(iv) Mean Squared Error (MSE) and Root Mean Squared Error (RMSE):

- MSE is the average of the squared differences between the observed and predicted values.

- RMSE is the square root of MSE and provides a measure of the average error in the same units as the dependent variable.

f) Advantages of Linear Regression:-

- Simplicity: Easy to understand and implement.

- Interpretability: The coefficients provide clear insight into the relationship between the dependent and independent variables.

- Efficiency: Computationally efficient for small to medium-sized datasets.

g) Limitations of Linear Regression:-

- Linearity Assumption: Assumes a linear relationship, which may not always be appropriate.

- Sensitivity to Outliers: Outliers can disproportionately affect the regression line.

- Multicollinearity: High correlation between independent variables can make the model unstable.

- Overfitting: In the case of multiple linear regression, adding too many variables can lead to overfitting, where the model performs well on training data but poorly on unseen data.

h) Extensions of Linear Regression:-

- **Polynomial Regression:** Extends linear regression by adding polynomial terms of the independent variables, allowing for modeling of non-linear relationships.

- **Ridge and Lasso Regression:** Regularized versions of linear regression that include a penalty term to prevent overfitting and handle multicollinearity.

Q2. Explain the Anscombe’s quartet in detail.

Ans) Anscombe's quartet is a group of four datasets that have nearly identical simple descriptive statistics, such as mean, variance, correlation, and linear regression line. However, when you visualize these datasets, they reveal significantly different patterns. This quartet was created by statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before performing statistical analysis.

Despite these identical statistics, the four datasets display very different behaviours when plotted.

Dataset 1: A Simple Linear Relationship

Characteristics:

The points in this dataset follow a nearly perfect linear relationship.

The data fits well with the linear regression line, and there are no significant outliers.

Dataset 2: Non-linear Relationship

Characteristics:

This dataset features a non-linear (curved) relationship between x and y.

Although the summary statistics are identical to Dataset 1, the actual relationship is quadratic rather than linear.

Dataset 3: Linear Relationship with an Outlier

Characteristics:

In this dataset, most of the data points lie along a linear trend, but one point is a significant outlier.

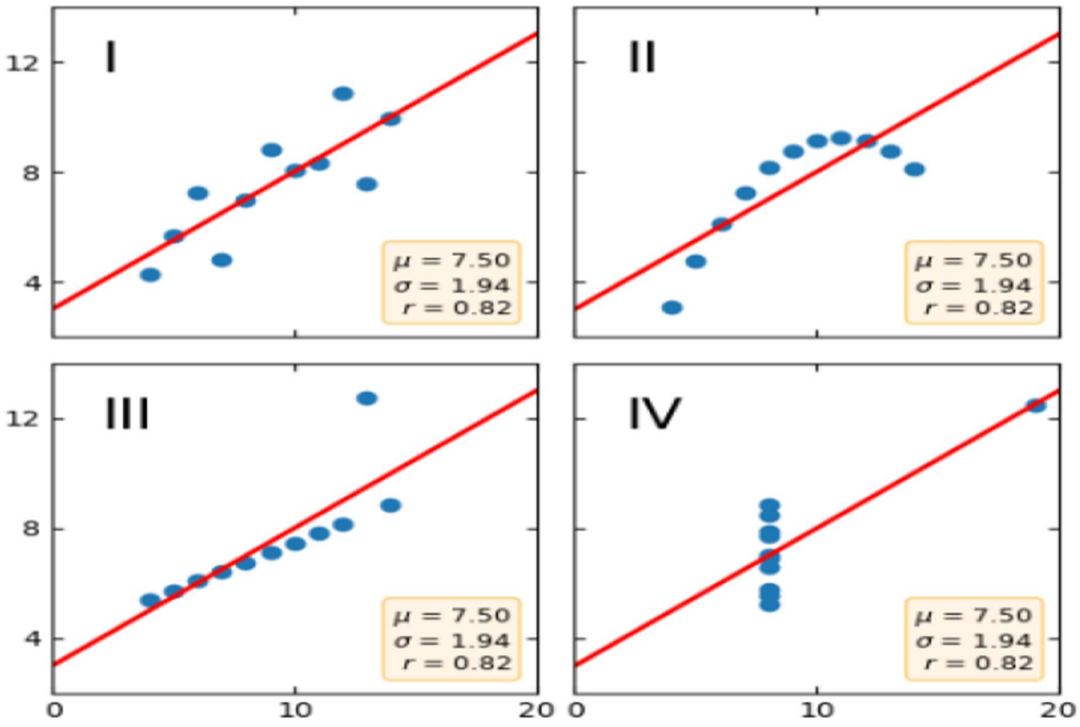
The presence of the outlier greatly affects the linear regression line, even though the summary statistics are the same as those in Dataset 1.

Dataset 4: No Relationship with an Outlier

Characteristics:

This dataset appears random, with no clear relationship between x and y.

There is one outlier that significantly influences the correlation and regression line, but without it, there would be no discernible pattern in the data.



Q3. What is Pearson's R?

Ans) Pearson's correlation coefficient (r) is a statistical measure that quantifies the linear relationship between two variables. It ranges from -1 to 1, where:

- **-1:** Indicates a perfect negative correlation. As one variable increases, the other decreases.
- **0:** Indicates no correlation between the variables.
- **1:** Indicates a perfect positive correlation. As one variable increases, the other increases.

Here are some key points to remember about Pearson's r:

- **Linearity:** It measures the strength of a linear relationship. It doesn't capture non-linear relationships.
- **Strength and Direction:** The magnitude of r indicates the strength of the relationship, while the sign indicates the direction (positive or negative).
- **Assumptions:** Pearson's r assumes that the data are normally distributed and have no outliers.

When to Use Pearson's r:

- When you want to measure the strength and direction of a linear relationship between two numerical variables.
- When the data meet the assumptions of normality and no outliers.

Example: If r = 0.8 between height and weight, it means there's a strong positive correlation. Taller people tend to weigh more.

In summary, Pearson's r is a valuable tool for understanding the relationship between two variables, but it's important to ensure the data meet its assumptions and consider the limitations of linear relationships.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans) Scaling in Data Analysis:

Scaling is a process of transforming data to a common scale, usually between 0 and 1 or -1 and 1. This is essential in many machine learning algorithms, especially those that rely on distance calculations or gradient descent.

Why is Scaling Performed?

a. **Normalization:**

1. Equalizes the range: Ensures that all features contribute equally to the model, preventing features with larger magnitudes from dominating.
2. Improves convergence: Helps algorithms converge faster, especially for gradient-based methods.
3. Prevents numerical instability: Avoids issues like vanishing gradients.

b. **Standardization:**

1. Centers the data: Brings the mean of each feature to 0.
2. Scales to unit variance: Makes the standard deviation of each feature1.
3. Handles outliers: Less sensitive to outliers compared to normalization.

Difference Between Normalized Scaling and Standardized Scaling:- Feature	Normalized Scaling	Standardized Scaling
Range	**0 to 1**	**-∞ to ∞**
Mean	**Preserved**	**0**
Standard Deviation	**Not preserved**	**1**
Outlier Sensitivity	**Sensitive**	**Less sensitive**
Use Cases	**When data range is important (e.g., image pixels)**	**When data distribution is skewed or outliers are present**

Choosing the Right Method:

1. Normalization: Use when you want to ensure all features are on a comparable scale and the range is important.
2. Standardization: Use when you want to center the data and scale it to unit variance, especially for algorithms that assume normally distributed data.

Example:

1. Normalization: For image data, where pixel values range from 0 to 255, normalization ensures all pixels contribute equally to the model.
2. Standardization: For a dataset with features that have different units or scales (e.g., height in inches and weight in pounds), standardization can make the features comparable.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans) VIF (Variance Inflation Factor) can become infinite when there is perfect multicollinearity, meaning that one predictor variable is an exact linear combination of one or more other predictor variables. This situation causes the denominator in the VIF formula to be zero, leading to an infinite VIF value. Perfect multicollinearity implies that the regression coefficients are not uniquely determined, making the model unstable.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans) A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to assess if a dataset follows a particular distribution, usually the normal distribution. In the context of linear regression, a Q-Q plot

is used to check the normality of the residuals (errors). If the residuals are normally distributed, the points on the Q-Q plot will fall along a 45 degree straight line. This is important because one of the key assumptions of linear regression is that the residuals are normally distributed. Non-normality can indicate issues with the model, such as the presence of outliers, skewness, or other violations of model assumptions.