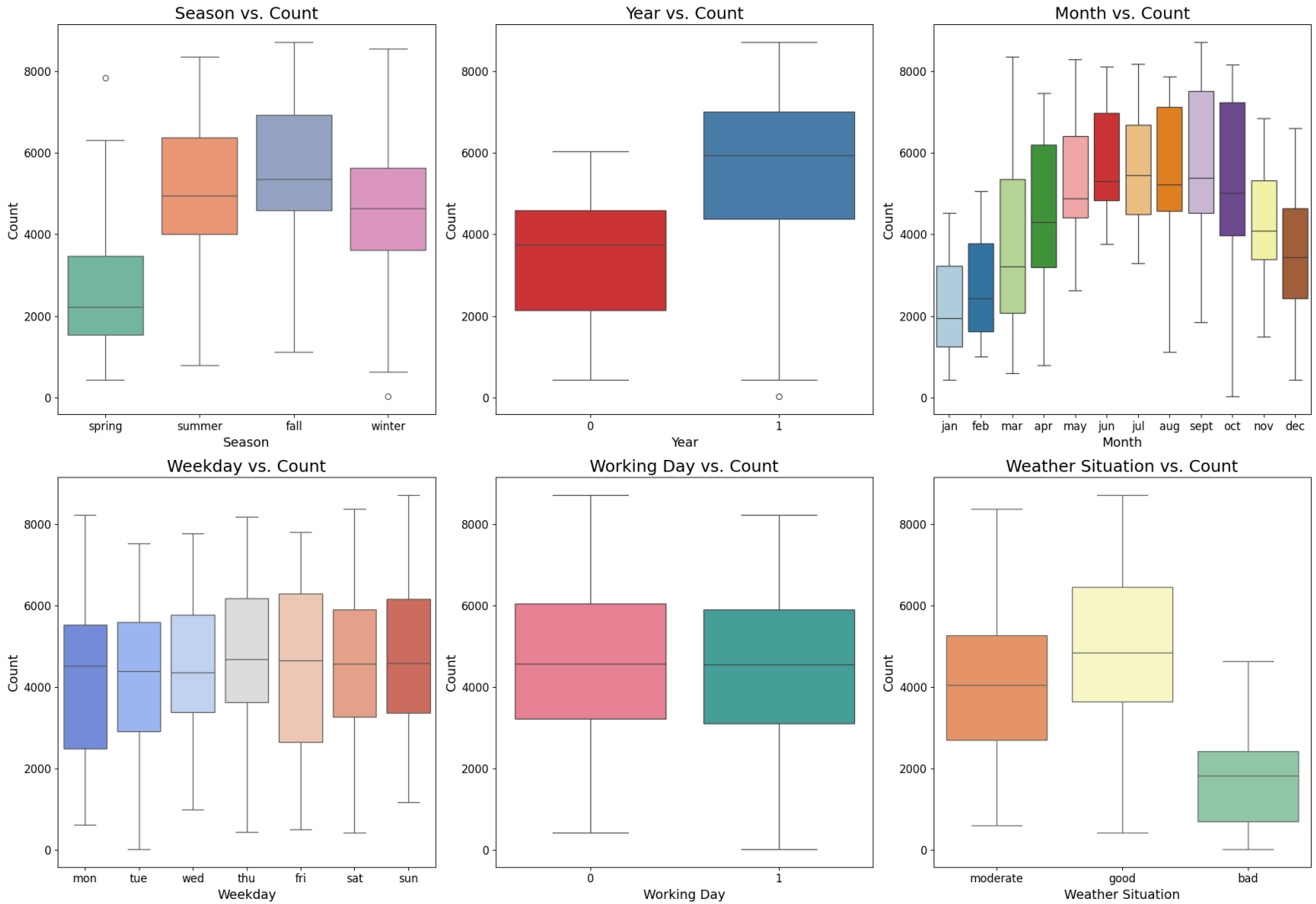
Assignment-based Subjective Questions

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

- Summer and Fall seasons have higher median counts (cnt) compared to Winter and Spring.This suggests that bike demand is higher during warmer months, likely due to favorable weather conditions for biking.
- The year 2019 (yr=1) shows a significantly higher count compared to 2018 (yr=0).This indicates that overall bike demand has increased over time, possibly due to an increase in the popularity of biking or other external factors.
- The demand varies significantly by month.September (mnth=9) and June (mnth=6) seem to have the highest median counts, whereas January (mnth=1) has the lowest.This could be due to weather variations or holiday seasons affecting biking preferences.
- Good weather (weathersit=1) has the highest median count, followed by Moderate (weathersit=2) and then Bad (weathersit=3).This indicates that poor weather conditions have a strong negative impact on bike rentals, which is expected as people are less likely to bike in unfavorable weather.
- The median count is relatively consistent across all weekdays, with a slight increase on Sundays (weekday=0).This suggests that bike rentals are relatively stable across the week, with a slight increase during weekends, possibly due to more leisure activities.
- Non-working days (workingday=0) show a slightly higher demand compared to working days (workingday=1).This implies that people are more likely to rent bikes during weekends or holidays, possibly for recreational purposes.

Boxplots of Categorical Variables vs. Demand Count

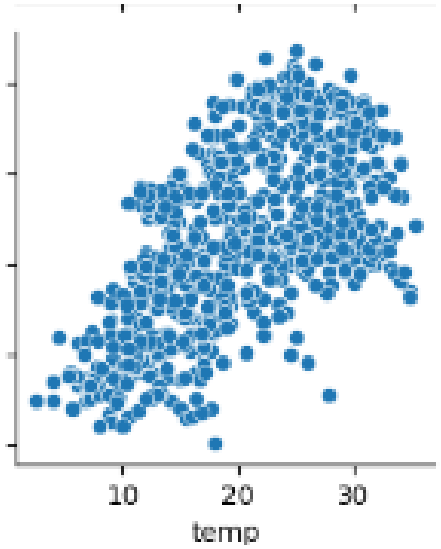


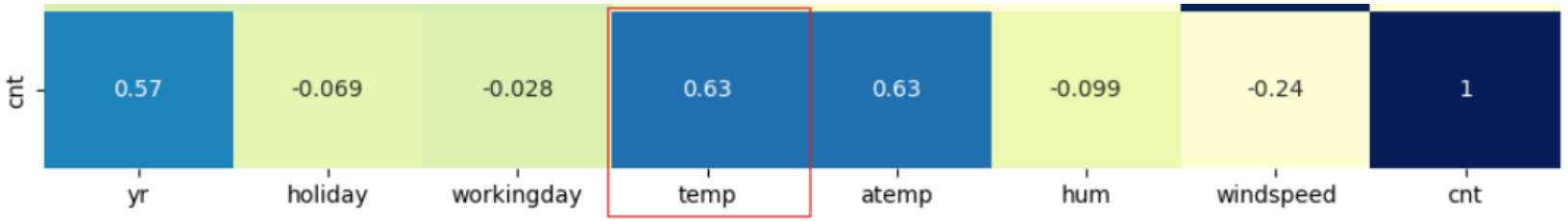
2. Why is it important to use drop_first=True during dummy variable creation?

- When you create dummy variables for a categorical feature with k categories, you typically generate k dummy variables. If you include all k variables in your regression model, it introduces redundancy because the kth variable can be perfectly predicted from the other $k-1$ dummy variables (i.e., if the first $k-1$ dummies are 0, the kth must be 1). This results in a situation known as the **"dummy variable trap."**
- By using drop_first=True, you only **create $k-1$ dummy** variables, **thus preventing multicollinearity** and allowing for a more stable model.

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

- "Temp" has highest correlation with target variable . (Can be seen in picture.1 for reference)





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

I have validated the assumption of Linear Regression Model based on below 5 assumptions -

Normality of error terms

- Error terms should be normally distributed

Multicollinearity check

- There should be insignificant multicollinearity among variables.

Linear relationship validation

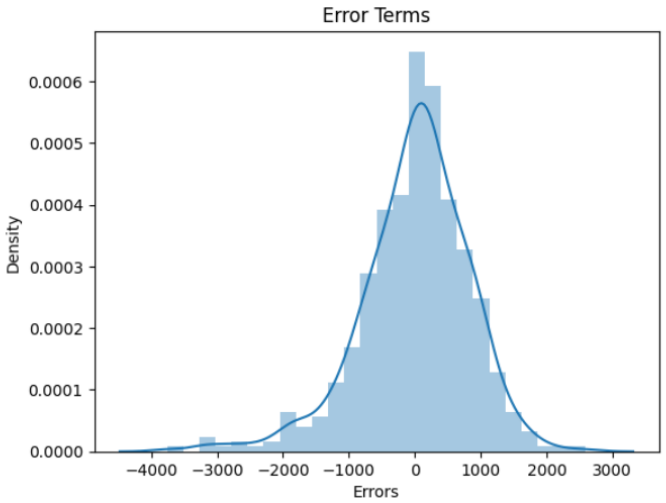
- Linearity should be visible among variables

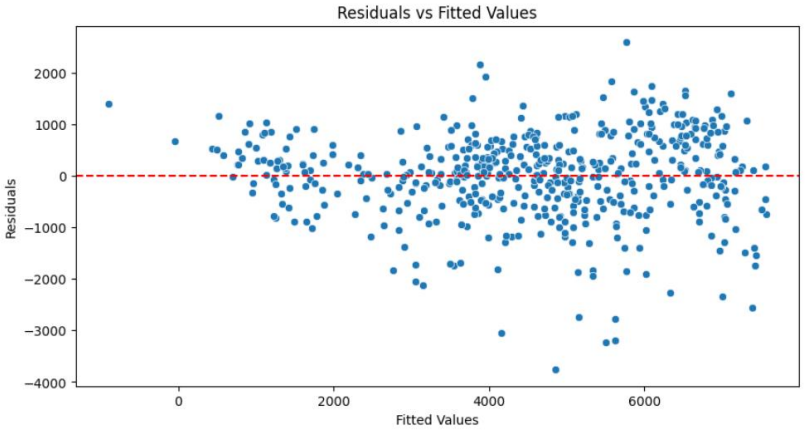
Homoscedasticity

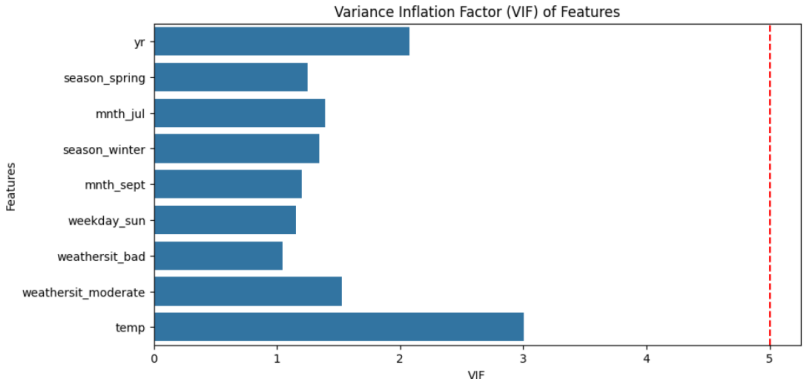
- There should be no visible pattern in residual values.

Independence of residuals

- No auto-correlation







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

Temp (TEMPERATURE):

- **Coefficient:** 3984.26
- **P-value:** 0.000
- **Interpretation:** This variable has a significant positive effect on bike demand. For each unit increase in temperature, the demand increases by approximately 3984 bikes, holding other factors constant.The year 2019 (yr=1) shows a significantly higher count compared to 2018 (yr=0).This indicates that overall bike demand has increased over time, possibly due to an increase in the popularity of biking or other external factors.

Yr (YEAR):

- **Coefficient:** 1994.68
- **P-value**: 0.000
- **Interpretation:** This indicates that bike demand increases by approximately 1995 bikes each year, showing a trend of increasing demand over the years.

weathersit_bad (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds):

- **Coefficient:** -2256.10
- **P-value:** 0.000
- **Interpretation:** This feature has a significant negative effect on bike demand. When the weather is categorized as "bad," demand decreases by approximately 2256 bikes, holding other factors constant.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical technique that models the relationship between a dependent variable and one or more independent variables using a straight line.

For simple linear regression: $Y=b0+b1X$

For multiple linear regression: $Y=b0+b1X1+b2X2+...+bnXn.$

Goal: Minimize the sum of squared errors using the Ordinary Least Squares (OLS) method.

Interpretation: Coefficients represent the impact of independent variables on the dependent variable.

Assumptions:

Linearity: Relationship between X and Y is linear.

Independence: Observations are independent of each other.

Homoscedasticity: Constant variance of error terms.

No Multicollinearity: Independent variables are not highly correlated.

Normality of Residuals: Errors are normally distributed.

2. Explain the Anscombe’s Quartet in Detail

A set of four different datasets created by statistician Francis Anscombe.

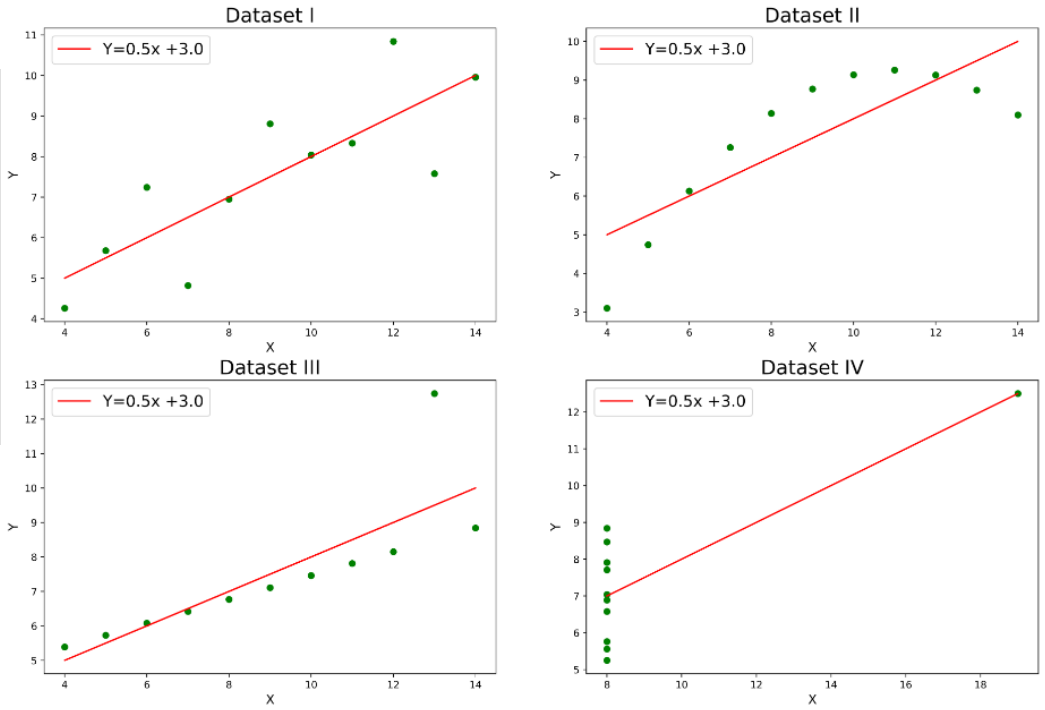
Purpose: To show that data visualization is crucial for understanding data distributions, as summary statistics can be misleading.

Key Properties: All four datasets have the same mean, variance, correlation, and regression line, but look very different when visualized.

Interpretation: Encourages performing visual inspection along with calculating summary statistics.

Example Data Summary and Visualisation: (refer below images) [Roman Letters denotes different datasets]

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727



3. What is Pearson's R?

A measure of the strength and direction of the linear relationship between two variables.

Value Range:

+1: Perfect positive linear relationship.

0: No linear relationship.

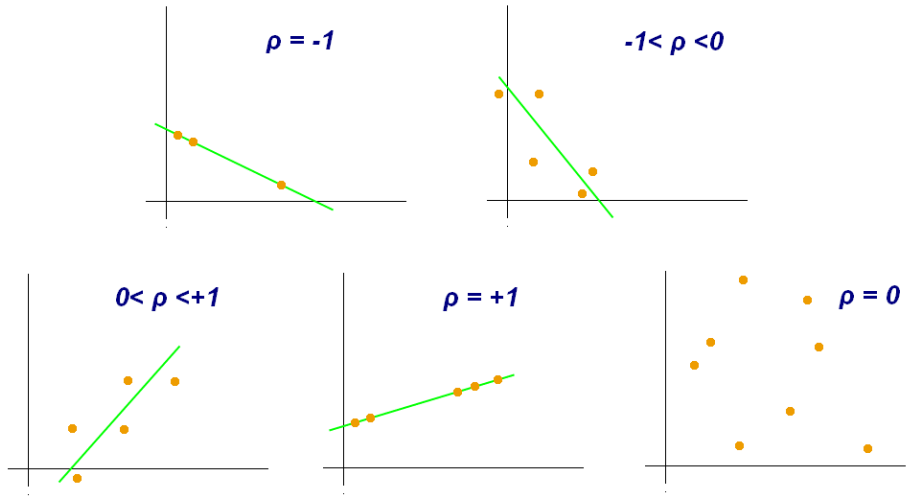
-1: Perfect negative linear relationship.

Use Case: Quantifies how well a linear model fits the data.

$$\rho_{X,Y} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}\,\sqrt{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}}.$$

Formula for Pearson coefficient

Examples of scatter diagrams with different values of correlation coefficient (ρ)



4. What is Scaling? Why is Scaling Performed? What is the Difference Between Normalized Scaling and Standardized Scaling?

Process of transforming data to a fixed range or distribution.

Why Performed:

- **Avoids Dominance:** Prevents features with larger scales from dominating model training.
- **Improves Performance:** Ensures faster convergence and prevents instability.

Normalization: Scales features to a range [0, 1]

Standardization: Scales features to have a mean of 0 and standard deviation of 1

Use Cases:

Normalization: When features have different ranges.

Standardization: When assumptions of normal distribution are important.

Normalisation Formula

```
X_new = (X - X_min)/(X_max - X_min)
```

Standardisation Formula

```
X_new = (X - mean)/Std
```

5. You Might Have Observed That Sometimes the Value of VIF is Infinite. Why Does This Happen?

VIF (Variance Inflation Factor) measures how much the variance of a regression coefficient is inflated due to multicollinearity.

When Infinite: VIF becomes infinite when there is perfect multicollinearity.

Reason: Occurs when one independent variable is a perfect linear combination of others.

This makes **1−R2=0**, leading to a division by zero in VIF calculation.

6. What is a Q-Q Plot? Explain the Use and Importance of a Q-Q Plot in Linear Regression.

A Quantile-Quantile (Q-Q) Plot compares the quantiles of a dataset with the quantiles of a specified theoretical distribution (usually normal).

Purpose: Helps check if data follows a particular distribution. Useful in validating the normality of residuals in linear regression.

Interpretation:

- Points on the line: Data is normally distributed.
- Points off the line: Indicates skewness or kurtosis.

Importance in Linear Regression:

Ensures that residuals are normally distributed, a key assumption for OLS.