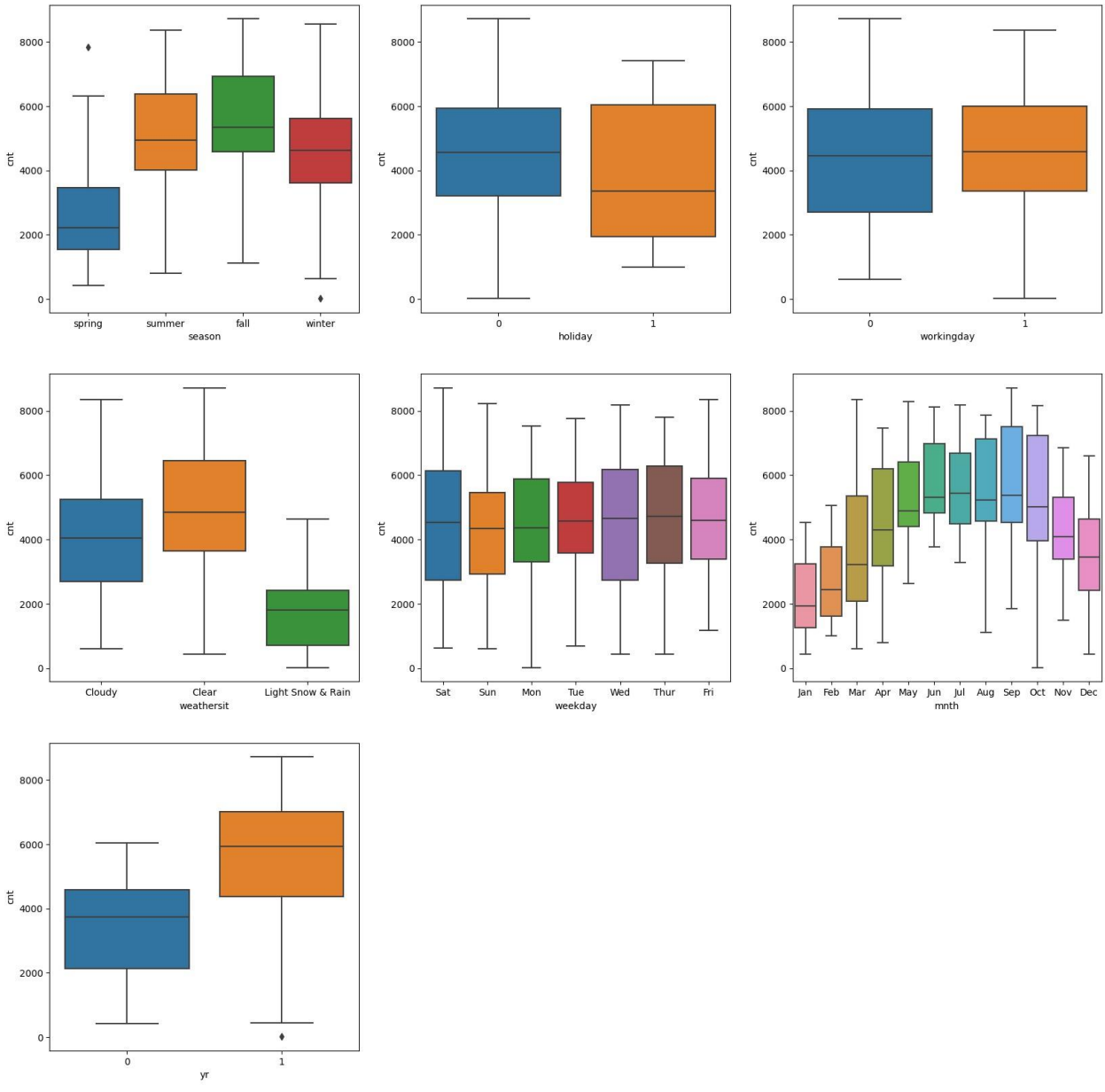
Assignment-based Subjective Questions

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



A.

The dataset includes categorical variables: season, year (yr), holiday, weekday, workingday, weathersit (weather situation), and month (mnth). These were represented visually using box plots.

These variables had the following impact on our dependent variable:

- Season: Among the seasons, Fall (category 3) exhibited the highest median demand, while Spring (category 1) showed the lowest.
- Year: There was a higher user count in 2019 compared to 2018.
- Holiday: Rentals decreased during holidays.
- Weekday: Bike demand remained relatively consistent throughout the week.
- Workingday: Bookings remained relatively constant between 4000 and 6000, indicating consistent user counts whether it was a working day or not.
- Weathersit: No users were observed during heavy rain/snow, indicating adverse weather conditions. The highest user count was seen during Clear or Partly Cloudy weather situations.

- Month: Rental numbers peaked in September and showed a similar peak in December, likely influenced by substantial snowfall during that time, leading to potential declines in rentals.

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

A. Reducing the number of columns is crucial for simplifying the model, facilitating easier interpretation of variables that significantly affect the dependent variable. Creating dummy variables for each category in a categorical variable may introduce multicollinearity issues, which can be mitigated by using 'drop_first=True'.

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

A. The numerical variable 'registered' initially displayed the highest correlation with the target variable 'cnt' among all features. However, post data preparation, when 'registered' was dropped due to multicollinearity concerns, the numerical variables 'temp' and 'atemp' emerged with the highest correlation to the target variable 'cnt'.

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

A. Validating Linear Regression Assumptions:

1. Normality of Errors: Ensuring error terms follow a normal distribution.
2. Multicollinearity Check: Identifying insignificant multicollinearity among variables.
3. Linear Relationship: Verifying visible linearity among variables.
4. Homoscedasticity: No discernible pattern in residual values.
5. Independence of Residuals: Absence of autocorrelation among residuals.

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

A. The top 3 features contributing significantly towards explaining the demand of the shared bikes based on the final model are:

1. ‘yr’
2. ‘holiday’
3. ‘workingday’

General Subjective Questions

1. Explain the linear regression algorithm in detail.

A. Linear regression is a statistical model examining the linear connection between a dependent variable and independent variables. It expresses how changes in independent variables relate to changes in the dependent variable through the equation Y = mX + c, where Y is the dependent variable, X is the independent variable, m is the slope, and c is the Y-intercept.

Y = m1X1 + m2X2 + …… + c (for multiple variables)

Positive linear relationships occur when both variables increase, while negative relationships see one variable increase as the other decreases.

There are two types: Simple Linear Regression (one predictor) and Multiple Linear Regression (multiple predictors).

Cost Function: Measures the model's performance by evaluating the difference between predicted and actual values (MSE or MAE).

Evaluation: Assesses model performance using metrics like R-squared, RMSE, or MAE to gauge how well the model fits the data.

Assumptions: Linear regression assumes linearity, independence of errors, constant variance, normal distribution of errors, and minimal multicollinearity.

2. Explain the Anscombe's quartet in detail.

A. Anscombe's Quartet, created by Francis Anscombe, comprises four datasets with nearly identical statistical properties but vastly different distributions and appearances when plotted. It underscores the importance of graphing data prior to analysis and highlights the impact of outliers and influential observations on statistical characteristics.

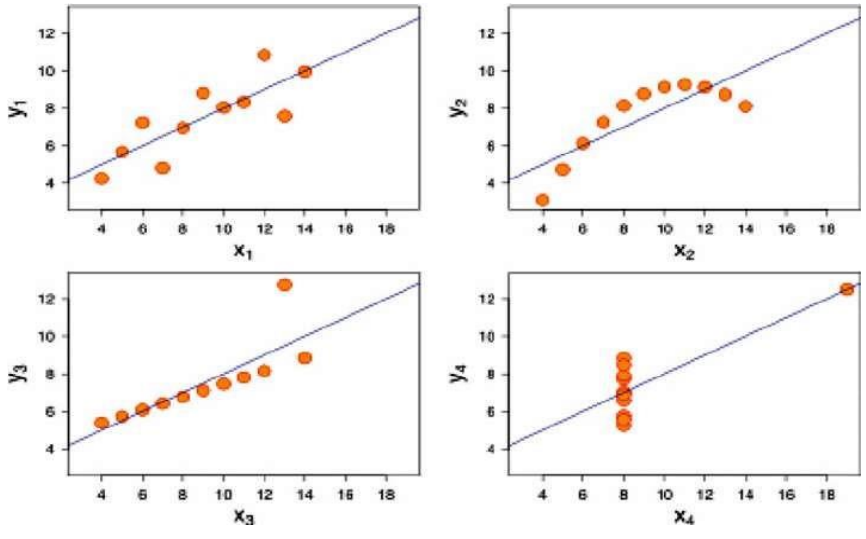
First Graph (Top Left): Displays a seemingly simple linear relationship between variables.

Second Graph (Top Right): Doesn't exhibit a normal distribution; although a relationship exists, it's not linear.

Third Graph (Bottom Left): Shows a linear distribution but is heavily influenced by one outlier, altering the regression line and reducing the correlation coefficient from 1 to 0.816.

Fourth Graph (Bottom Right): Illustrates how a single high-leverage point can inflate the correlation coefficient significantly, despite other data points indicating no relationship between the variables.

Each dataset within Anscombe's Quartet demonstrates the significance of visualization in understanding data and highlights how outliers or influential points can distort statistical analysis and interpretations.



3. What is Pearson's R?

A. Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

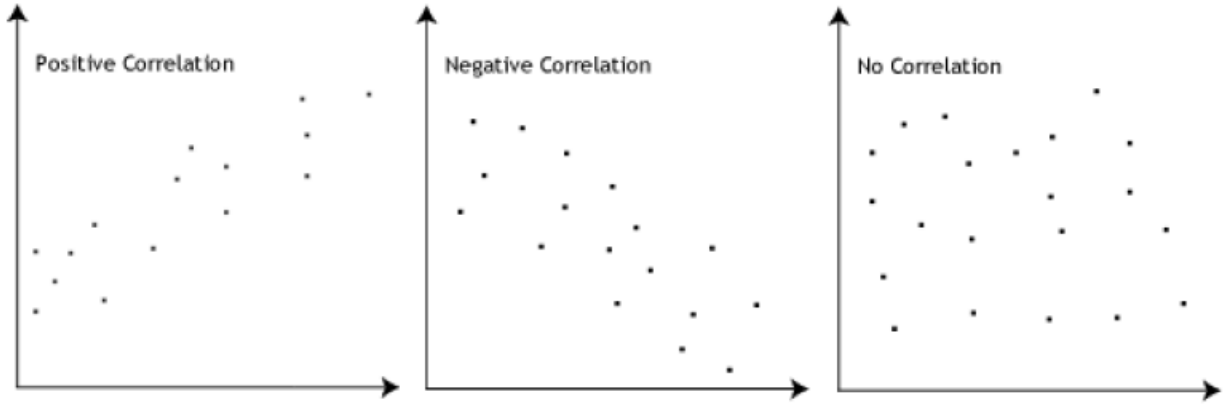
Strength of Association :-
R values from +1 to -
R = 1 Perfect positive linear relationship
R = -1 Perfect negative linear relationship.
R = 0 No linear relationship.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1.

1. A value of 0 indicates that there is no association between the two variables.

2. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.

3. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.



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

A.

Scaling, an integral part of data preprocessing, harmonizes independent variables to a consistent range. This aids in computational efficiency and accuracy within algorithms.

Why is it done? Well, datasets often encompass features spanning diverse magnitudes, units, and scales. Without scaling, algorithms prioritize magnitude over unit, leading to flawed modeling. To counteract this, scaling aligns variables to a unified magnitude level, ensuring fair consideration for all factors.

It's crucial to clarify that scaling exclusively impacts coefficients, leaving other parameters such as t-statistic, F-statistic, p-values, R-squared, etc., untouched.

Normalization, also known as Min-Max Scaling, confines data within a 0 to 1 range. Implementation in Python can be executed using sklearn.preprocessing.MinMaxScaler.

The formula for MinMax Scaling is: x = x - min(x) / max(x) – min(x)

Standardization, on the other hand, transforms values into their Z scores, aligning data to a standard normal distribution with a mean (μ) of zero and a standard deviation (σ) of one.

The formula for Standardization is x = x – mean(x) / sd(x)

For implementation in Python, sklearn.preprocessing.scale is employed. It's worth noting that normalization may sacrifice some data information, particularly concerning outliers, compared to standardization

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

A.

The occurrence of an infinite value for the Variance Inflation Factor (VIF) is tied to perfect correlation between features, known as multicollinearity. VIF measures how much the variance of a regression coefficient inflates due to multicollinearity. When there's perfect correlation among features, it leads to an infinite value for VIF.

The VIF value is determined by the formula:

$VIF = 1/1-R^2$

Where R^2 represents the coefficient of determination for the relationship between one variable and the remaining set of independent variables. A high VIF signals significant correlation between variables, indicating potential multicollinearity issues.

When the value of R^2 approaches 1 (perfect correlation), the denominator in the VIF formula approaches zero, leading to an infinitely high VIF value. This scenario points towards an exact or near-exact linear relationship among the variables, causing the VIF to become infinite.

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

A.

A Q-Q (quantile-quantile) plot serves as a visual tool to assess whether two datasets derive from populations sharing a common distribution.

Use of Q-Q Plot:
This plot compares the quantiles of one dataset against the quantiles of another. Quantiles represent the fraction or percentage of data points below a given value. For instance, the 0.3 quantile denotes the point where 30% of the data lies below and 70% above that value. The plot includes a 45-degree reference line. Ideally, if both datasets originate from populations with the same distribution, the plotted points should align closely along this reference line. Any deviation from this line indicates differences between the distributions of the datasets.

Importance of Q-Q Plot:
The Q-Q plot is crucial in assessing whether the assumption of a shared distribution between two datasets is valid. If the assumption holds, estimators for location and scale can effectively combine both datasets to derive common estimates. On the contrary, if differences exist between the datasets, the Q-Q plot provides insights into the nature of these differences. It offers a more intuitive understanding compared to analytical methods like the chi-square and Kolmogorov-Smirnov 2-sample tests, enabling a deeper grasp of the disparities between datasets. This visual representation aids in assessing the nature and extent of differences in distributions between the compared datasets.