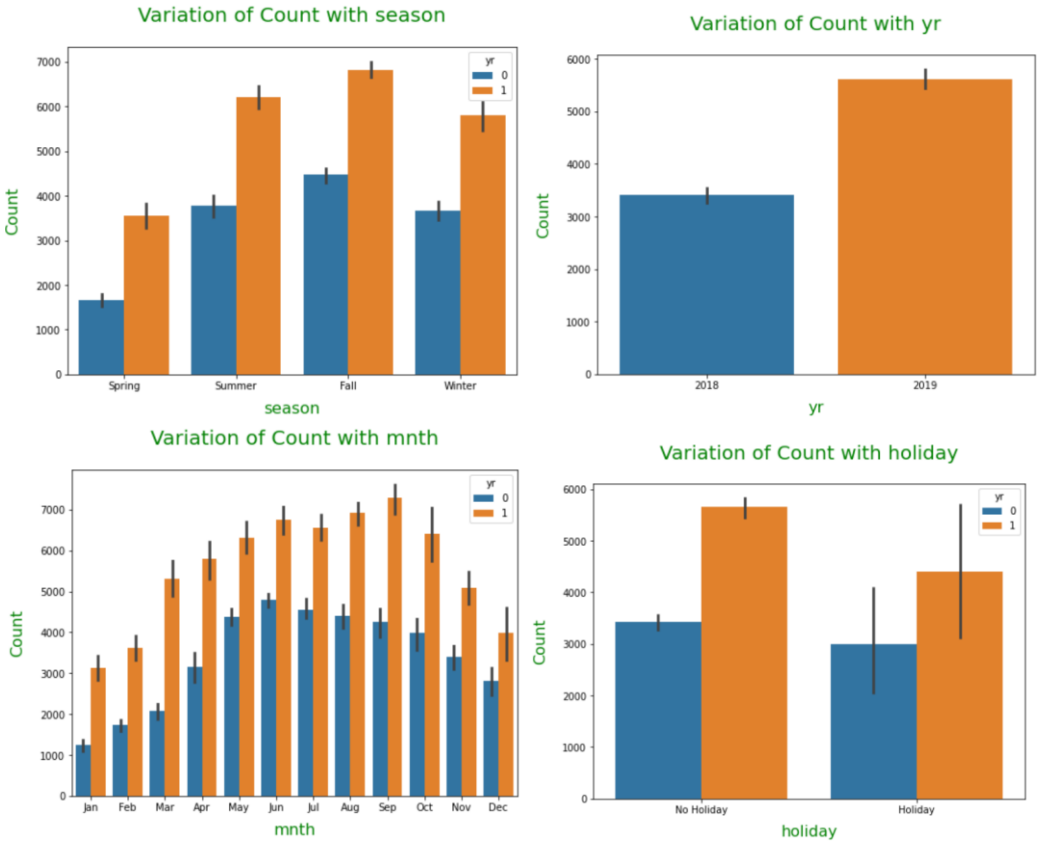
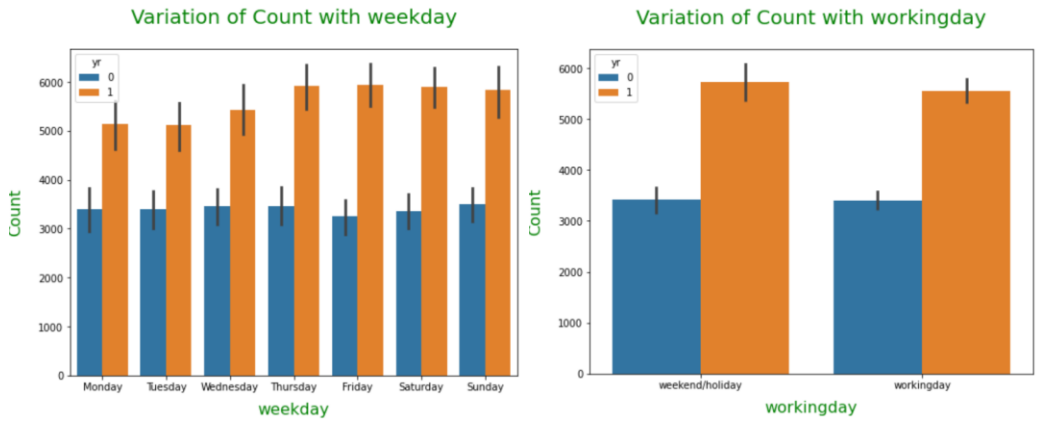
Assignment-based Subjective Questions

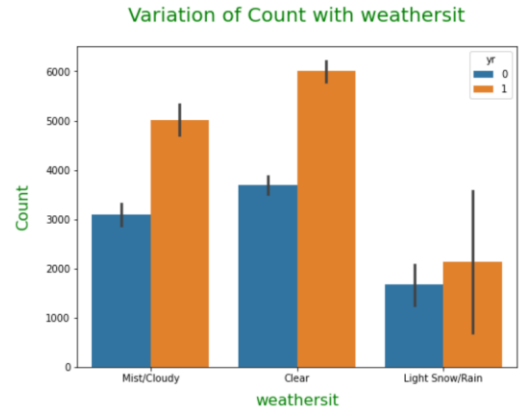
1. **From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable? (3 marks)**

Answer:

- The number of bikes rented in the fall season is high in both 2018 and 2019.
- When comparing the two years, 2019 has a higher rental count of over 6500 (approx.) compared to 2018, which had a count of over 4000 (approx.) in the fall.
- In both 2018 and 2019, the number of bike rentals is low in the spring.
- In 2019, there are more bike rentals than in 2018.
- The rental count in 2018 was around 3500 (approximately), but it increased to more than 5500 in 2019. (approx.).
- In September of this year, the rental count is at an all-time high.
- In the month of June this year, the rental count is at an all-time high.







- In contrast, for the months of January in both 2018 and 2019, demand for bike rentals is low.
- When there are no holidays in both 2018 and 2019, the number of bike rentals is high.
- In the year 2019, the demand for bike rentals is higher on Thursday, Friday, and Saturday than on other weekdays.
- In the year 2018, the demand for bike rentals is particularly high on Mondays, Tuesdays, Wednesdays, and Thursdays.
- In both 2018 and 2019, the demand for bike rentals is high when the weather is clear.
- In both the winter and summer, demand decreases due to light snowfall.
- In the aforementioned visualizations of numerical and categorical columns, there are no outliers.
- For the fourth circumstance (Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog), the Weathersit boxplot contains 0 cnt values.
- The Season and Months box plot shows that cnt values increase during the summer months of 4,5,6 months and gradually drop during the winter months of 10,11,12 months.
- It can be seen from the working day box plot that cnt values are lower on weekends/holidays than on weekdays.

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

Answer:

- If we don't drop the first column, the dummy variables will be correlated (redundant). This may have a negative impact on some models, and the effect is amplified when the cardinality is low.

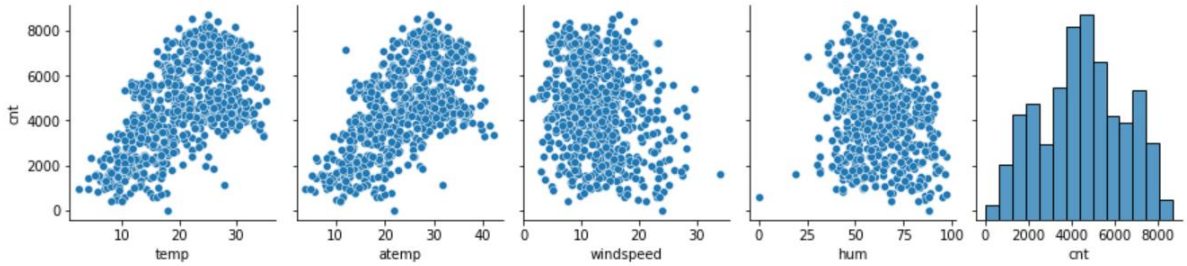
n-1 dummies out of n categorical levels

Example: Iterative models may have difficulty converging, resulting in skewed lists of variable relevance. Another argument is that having all dummy variables results in multicollinearity between them. We lose one column to keep everything under control.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlationwith the target variable? (1 mark)

Answer:

- The numerical variables "temp" and "atemp" are closely associated with the target variable "cnt."



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

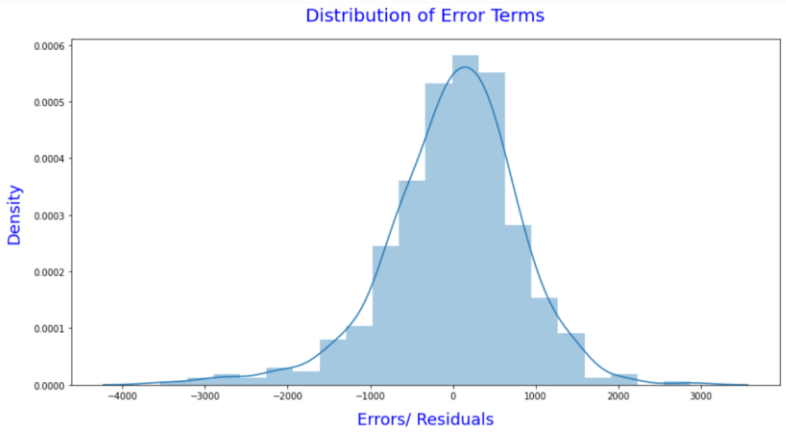
Answer:

The following Linear Regression assumptions have been verified.

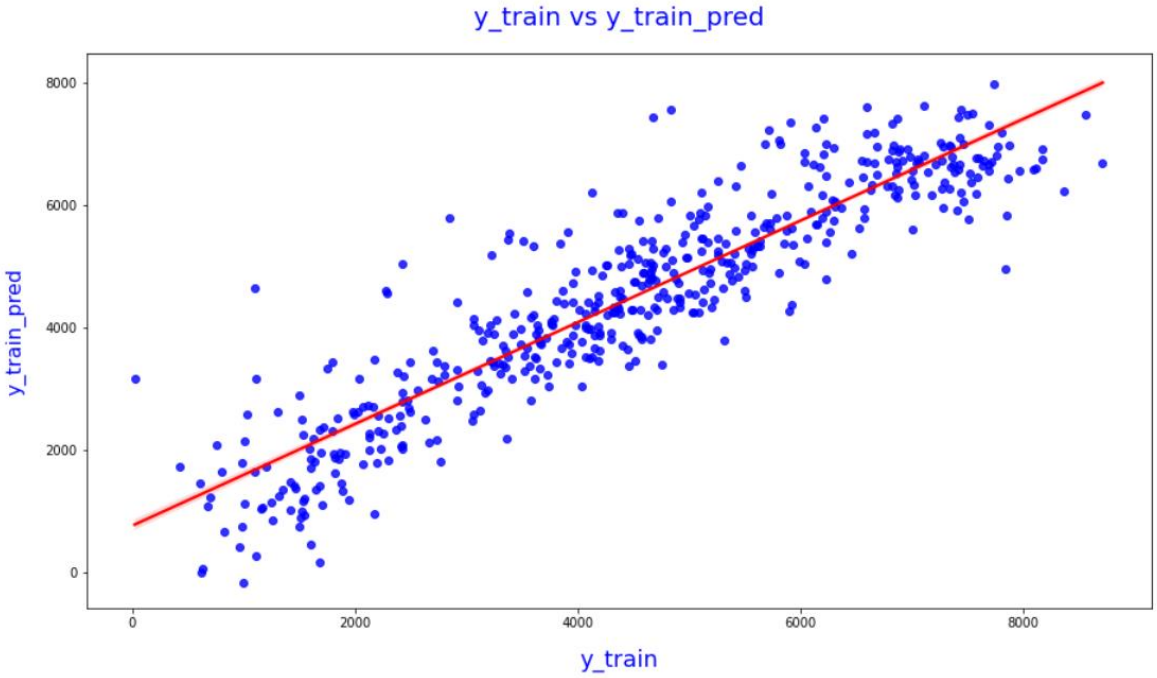
• **Normality:** Error terms are normally distributed with zero mean.

• **Homoscedasticity:** For any value of X, the variance of the residual is the same.

Normality: The Error terms are normally distributed with a 0 mean, as seen in the distribution plot. As a result, our assumption of Normality was proved.



Homoscedasticity: The variance is constant and the regression plot between y_train and y_train_pred is evenly distributed along the regression line. As a result, the assumption of homoscedasticity was proven.



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

Answer:

The top three contributors to the Boom Bike sharing business are listed below.

Temperature: As we can see from the model equation above, the dependent variable, namely the count of bike rentals (cnt), increases as the temperature rises. As a result, Boom Bikes, a US bike-sharing company, can focus more on temperature, as rising temperatures will increase demand for bikes.

Light snow: The dependent variable, i.e., the count of bike rentals (cnt), drops in Light Snow weather conditions, according to the model equation. Because there is less demand in light snow weather, businesses might provide special offers/discounts or set up a bike shield to increase demand during this time.

Year: Bikes were in more demand in 2019 than in 2018. As more people become aware of Boom bike sharing company through various marketing methods, the demand for bike rentals is expected to rise in the future years.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**

Answer:

The most basic and widely used strategy for predicting a continuous variable is linear regression.

- Linear regression is used to predict the value of a dependent variable (y) based on the value of an independent variable (x). As a result of this regression technique, a linear relationship between x (input) and y (output) is discovered (output).
- When the dependent variable is a continuous data type, regression is used, and the predictors or independent variables can be of any data type, such as continuous, nominal, or categorical.
- The regression approach aims to identify the best fit line that accurately depicts the connection between the dependent variable and the predictors.
- The output/dependent variable is a function of the independent variable, the coefficient, and the error term in regression.

There are two types of regression: basic linear regression and multiple linear regression.

Simple Linear Regression: SLR is utilised when only one independent variable is used to predict the dependent variable.

- It's termed "Simple Linear Regression" because there's only one predictor variable involved, and the relationship between two variables x and y is explained using a straight line called the Regression Line.

Equation of regression line for simple linear Regression: $y = \beta_0 + \beta_1 x + \epsilon$

- y = Dependent variable.
- x = explanatory / Predictor variables
- β0 = y-intercept (constant term)
- β1 = Slope coefficients for each explanatory variable
- ϵ = The model's error term (also known as the residuals)

Multiple Linear Regression: When many independent factors are employed to predict the dependent variable, MLR is used. Multiple regression is a variant of simple linear regression in which more than one explanatory variable is used.

Equation of regression line for Multiple linear Regression: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_p x_p + \epsilon$

- for i = p observations:
- y = Dependent variable
- xi = Explanatory / Predictor variables
- β0 = y-intercept (constant term)
- βp = Slope coefficients for each explanatory variable xi
- ϵ = The model's error term (also known as the residuals)

Assumptions of Linear Regression:

1. Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.

2. Independence: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.

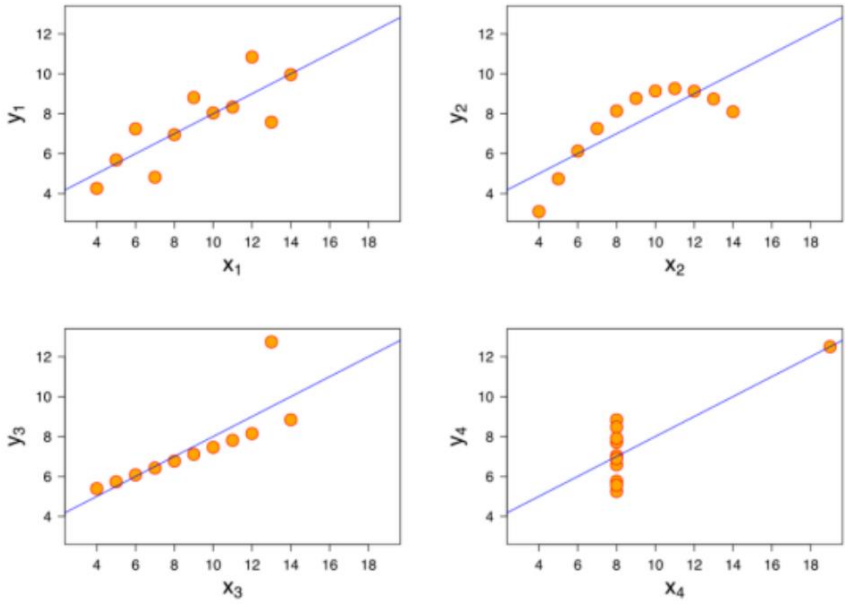
3. Homoscedasticity: The residuals have constant variance at every level of x.

4. Normality: The residuals of the model are normally distributed.

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's quartet consists of four datasets with essentially equal simple statistical features but distinct visual appearances when graphed. There are eleven (x, y) points in each dataset.



- They were created by statistician Francis Anscombe in 1973 to show the significance of charting data before analyzing it, as well as the impact of outliers on statistical features.
- The scatter plot (top left) looks to show a straightforward linear relationship.
- The second graph (top right) is not normally distributed; there is a relationship between the two, but it is not linear.
- The distribution in the third graph (bottom left) is linear, but the regression line should be different. The estimated regression is thrown off by one outlier, which has a large enough impact to reduce the correlation coefficient from 1 to 0.816.
- Finally, the fourth graph (bottom right) illustrates how one high-leverage point can yield a high correlation coefficient even when the other data points show no association between the variables.

3. What is Pearson's R? (3 marks)

Answer:

Pearson's r is a numerical representation of the strength of the linear relationship between two variables. Its value varies from -1 to +1.

- It depicts a linear relationship between two pieces of information. Simply put, it tells us whether or not we can build a line graph to describe the data.
- The Pearson correlation coefficient (PCC), sometimes known as Pearson's R, is a measure of how well two variables are related.
    - r = 1 denotes a fully linear data set with a positive slope.
    - r = -1 denotes a fully linear data set with a negative slope.
    - r = 0 indicates that there is no linear relationship.

4. What is scaling? Why is scaling performed? What is the difference between normalized scalingand standardized scaling? (3 marks)

Answer:

Scaling is a technique for normalising or standardising a set of independent variables or data elements. It is used to deal with fluctuating values in the dataset during the data pre-processing step. It should be done after the data has been separated into the train and test sets.

Why is scaling performed?

- The magnitudes, units, and ranges of features in the real-world dataset are extremely varied. When the scale of a characteristic is irrelevant or deceptive, normalisation should be used instead. When the scale is meaningful, normalise it.
- If a feature in the dataset has a large scale in comparison to other features, it becomes dominant and must be scaled. Feature scaling is used to ensure that all features are equally weighted.

Normalization:

- This technique, often called Min-Max scaling, rescales a feature or observation value with a distribution value between 0 and 1.

$$X_{new} = (X - \min(X)) / (\max(X) - \min(X))$$

- Here, max(X) and min(X) represent the feature's maximum and minimum values, respectively.

Standardization:

- It's a powerful technique for rescaling a feature value so that it has a distribution with a 0 mean value and variance equal to 1 (Xmean is the mean of the feature values).

$$X_{new} = (X - X_{mean}) / \text{Standard Deviation}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

The variance inflation factor (VIF) is used to detect multicollinearity by quantifying the level of correlation between one predictor and the other predictors in a model.

- The VIF calculates how much the variance of a regression coefficient is inflated as a result of the model's multicollinearity.

In general, VIF can be calculated based on R-square ,

$$VIF = 1 / (1 - R^2)$$

- A high VIF value implies that the variables are related; the VIF formula states that as the value of R2 climbs (R2 1), indicating that one independent variable can be explained by all other independent or predictive variables, the denominator decreases, making the entire VIF infinite.

If $R^2 \rightarrow 1$, then $(1 - R^2)$ becomes *zero*, and VIF will become *infinity*.

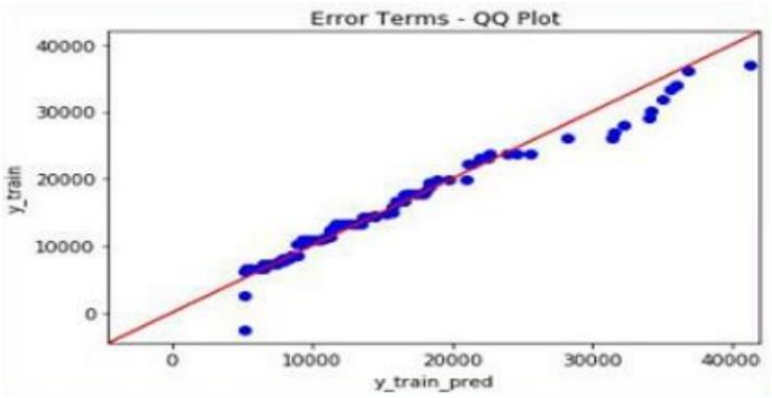
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

Answer:

The Q-Q plot is a scatterplot that is created by graphing theoretical quantiles or by charting the quantiles of the first data set against the quantiles of the second data set. The points should form a straight line if both sets of quantiles came from the same distribution**.**

- On the x-axis, it's called a normal variate, and on the y-axis, it's called ordered values for random variables.
- A q-q plot for one of the interpretations (if the y-quantiles are lower than the x-quantiles) is shown below.
- The q-q plot and q-q plot 2 samples from statsmodels.api are used to plot Q-Q graphs for single and multiple data sets, respectively.



Use of Q-Q plot in Linear regression:

- We can use this Q-Q plot in the linear regression model to determine whether our model error terms (y train – y pred) or residuals are normally distributed or not.
- The data is approximately normally distributed if the Q-Q plot shows a nearly straight line.

Importance of Q-Q plot in Linear regression:

- The Q-Q plot can be used to detect shifts in position, scale, symmetry, and the existence of outliers, among other distributional features.
- It can also be used to sample sizes.

If two data sets are present, it is utilised to examine the following scenarios:

  - are derived from groups with a similar distribution
  - having a similar scale and location
  - have distributional shapes that are similar
  - have tail behaviour that is similar