Bike Sharing Assignment

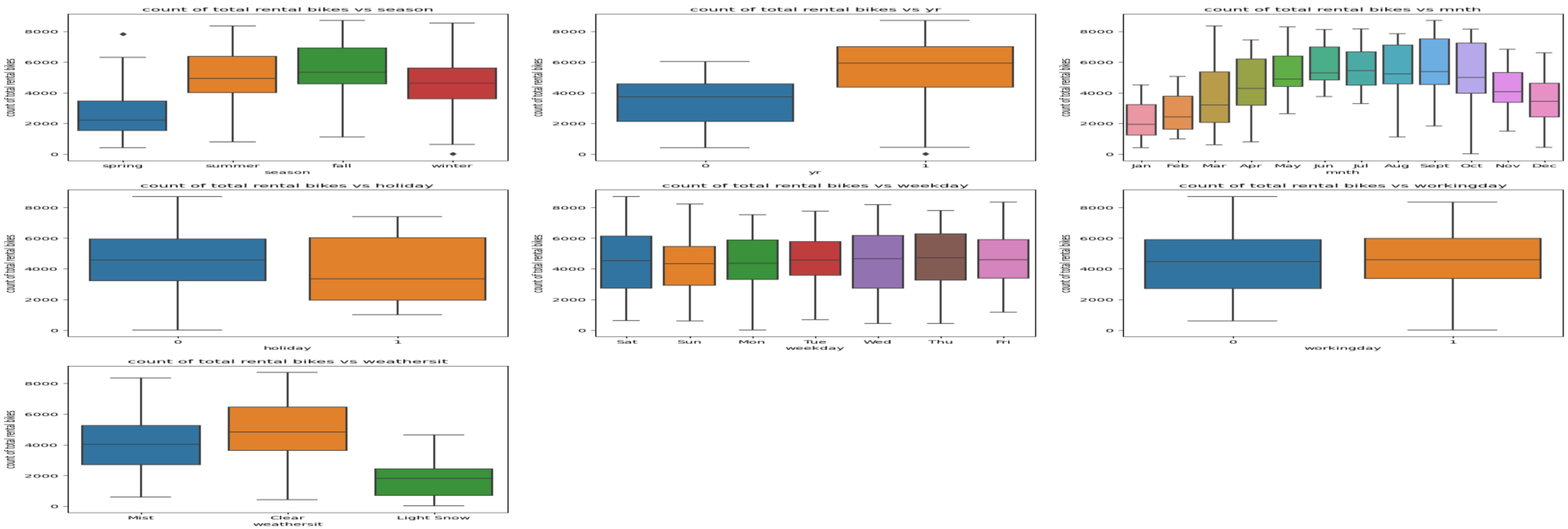
By- Kushagra Sengar

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? (3 marks)**

Ans. From our analysis of the categorical variables from the dataset**,** we can inference that :

- During **fall** season, the demand for rental bikes is highest.
- **2019** seems to be way ahead than **2018** in terms of rental bikes demand.
- There is a constant increase in the bike demand till mid-year ,**September** being the highest of them all and after that the demand decreases.
- When there is a **holiday,** there is a drop in the mean demand.
- From the plot, it looks like that the demand remains same irrespective of **weekday** or a **working day**.
- Demand is highest when the weather situation is **Clear**.



.

Assignment-based Subjective Questions

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

Ans. When we have a multilevel categorical feature in the dataset then we one-hot encode it or create dummy variables for it. If there are ‘n’ levels , then it is usually recommended to have ‘n-1’ dummy variables because ‘n-1’ levels are enough to explain features for all ‘n’ level.

When we specify **drop_first=‘True**’, it drops the first level that is available at first row of the dataset , splitting rest of the levels into ‘n-1’ columns. This also avoids a condition of ‘dummy variable trap’, which essentially means that one variable can be predicted from the others, making it difficult to interpret predicted coefficient variables in regression models.

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

Ans . Looking at the pair-plot among the numerical variables, temp and atemp have the highest correlation (0.63) with the target variable.

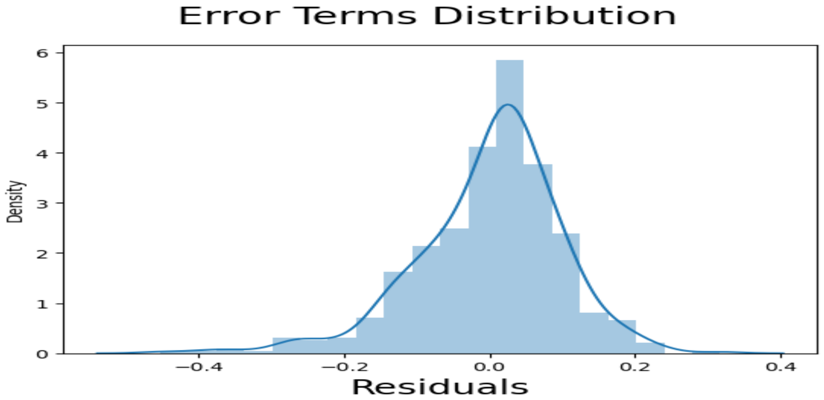
.

Assignment-based Subjective Questions

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

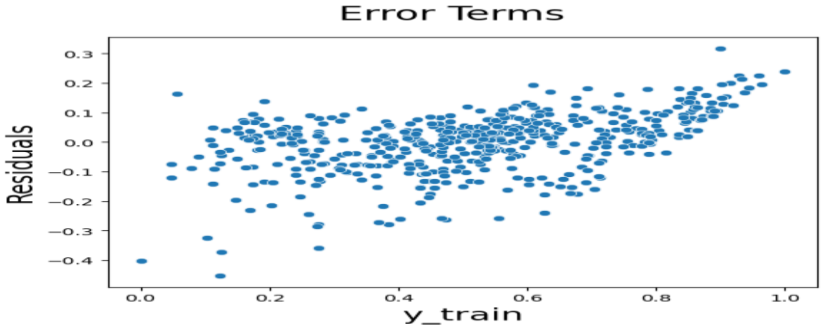
Ans. Ways to validate the assumptions of Linear Regression is to :

- Plot a histogram of the error terms to check whether they are normally distributed.



As we can see from our model that the residuals are normally distributed with mean 0. Hence, it validates the first assumption of a linear regression

- We need to plot the error terms, this time with either of X or y to check for any patterns to validate the another assumption that the error terms should be independent of each other



From the plot above (from our model) we can confirm that the Error terms are independent of each other as they exhibit randomness

Assignment-based Subjective Questions

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

Ans. The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

- **atemp** - Coefficient of atemp indicates that the unit increase in the atemp , increases the bike hiring by 0.475352
- **Light Snow (weathersit)** - Coefficient of Light Snow weather situation indicates that the unit increase in the Light Snow weather situation , decreases the bike hiring by -0.230065.
- **yr-** Coefficient of yr indicates that the unit increase in the yr , increases the bike hiring by 0.226318

General Subjective Questions

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

Ans.

- Linear regression is a supervised machine learning algorithm that shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis. If there is a single input variable (x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called multiple linear regression.

- The linear regression model can be represented by the following equation: $Y = \beta_0 + \beta_1 X$.

  where β_0 – constant and β_1 – slope of the line.

- The equation of the best fit regression line $Y = \beta_0 + \beta_1 X$ can be found by minimising the cost function RSS i.e Residual sum of squares. It is done using the following two methods:
  1. Differentiation
  2. Gradient descent method

- The strength of a linear regression model is mainly explained by R^2, where $R^2 = 1 - (RSS / TSS)$

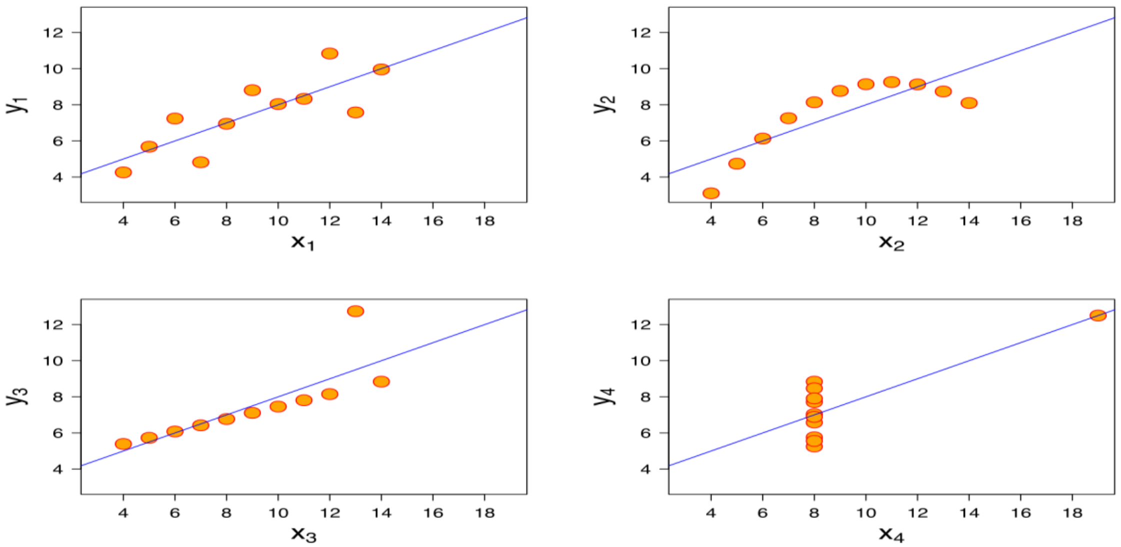
- The various assumptions related to linear regression algorithms are :
  1. Linear relationship between X and y.
  2. Normal distribution of error terms.
  3. Independence of error terms.
  4. Constant variance of error terms

General Subjective Questions

- Hypothesis testing in linear regression is done to determine the significance of beta coefficients. The null hypothesis is that the beta coefficients are 0 and if p- value for the coefficients given by t-score = βi/SE(βi) is less than 5% then we can reject the null hypothesis.

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

Ans. Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, yet have very different in their distribution when plotted. That is why its always recommended to visualizing the data before deriving to any solutions via algorithms that can help in identifying the various anomalies present in the data like outliers, patters,etc.

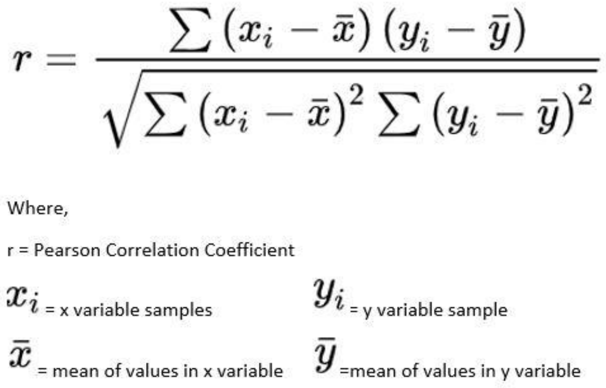


(Image Source: Wikipedia)

General Subjective Questions

3. What is Pearson’s R? (3 marks)

Ans. Pearson's r, is the correlation between the two variables . In simple linear regression, if we calculate correlation between x and y variable , then the R squared value of the model is square of the correlation and hence the term Pearson’s R is given to the coefficient. It measure strength of linear relationship between two variables.



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

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\bar{y}$ =mean of values in y variable

(Image Source: Wikipedia)

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

General Subjective Questions

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

Ans. Scaling is the process to normalize the range of features**.** It aims to create features that have similar ranges to each other. Scaling is done so that we don't over weigh a particular feature just because its values are much larger than other features. Also, it makes the cost optimization algorithm i.e. gradient descent very fast. Scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

There are two major methods to scale the variables –

1. **Standardization :**

- Standardization brings all of the data into a standard normal distribution with mean zero and standard deviation one. The formulae used for this methods is given by:

Standardisation: $x=x-\text{mean}(x)/\text{sd}(x)$

- Standardisation lowers the range of a feature but does not limit it between 0 and 1. Therefore, outliers can be detected even after scaling.

2. Minmax scaling :

- Minmax scaling, on the other hand, brings all of the data in the range of 0 and 1

The formulae used for this methods is given by :

MinMax Scaling: $x=x-\min(x)/\max(x)-\min(x)$

- Since outliers are also limited between 0 and 1 , it takes them also into consideration while model building and they are hard to find out after scaling.

General Subjective Questions

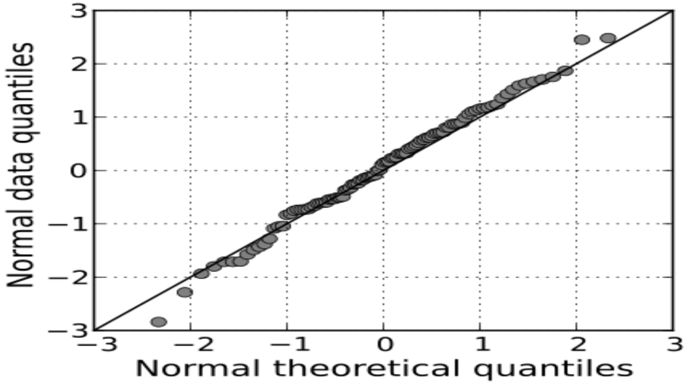
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans. When VIF of a feature is infinite ,it means that feature is perfectly correlated with one or more features. As per the VIF formulae, when VIF is infinite , then 1-R2 is equal to 0, which means R2 score for that variable is 1 i.e. perfectly correlated and that feature can be explained by the other. In such cases we need to drop the feature with infinite VIF.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans. As per the medium source**,** Q-Q(quantile-quantile) plots compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

- Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution, etc. Since we are most interested in normal distribution, so way to verify it using Q=Q plot is If all the points plotted on the graph perfectly lies on a straight line then we can clearly say that this distribution is Normally distribution because it is evenly aligned with the standard normal variate which is the simple concept of Q-Q plot



(Image Source: Wikipedia)

- Its importance in linear regression lies when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.