Assignment-based Subjective Questions

Ques 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.

Observations from above boxplots for categorical variables:

- The season bar plots indicates that more bikes are rented during fall season followed by summer and winters
- The weather bar plots indicates that more bikes are rented during the most followed by cloudy and rainy weather
- The month bar plots indicates that more bikes are rented during June and September month.

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

Ans.

drop_first=True is important to use, as it helps in reducing the extra columns that gets created during dummy variable creation. Hence it reduces the correlations created among dummy variables. If we do not remove these extra columns, it may affect some models adversely

---

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

Ans.

By looking at the pair plot temp variable has the highest (0.63) correlation with target variable 'cnt'.

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

Ans.

The Dependent variable and Independent variable must have **a linear relationship**.

It can be seen from pair plots, there is some linear relation between temp, atemp with Count

It is also seen from heatmap that variables have high collinearity with the target variable.

Perfect Linear regression equation:

cnt = 0.5841 - 0.2368 * season_spring - 0.1105 * mnth_dec - 0.1181 * mnth_jan

- 0.1095 * mnth_nov - 0.0868 * weathersit_Cloudy - 0.3054 * weathersit_Rain + 0.2475 * yr - 0.2178 * windspeed

Little or no Multicollinearity between the features:

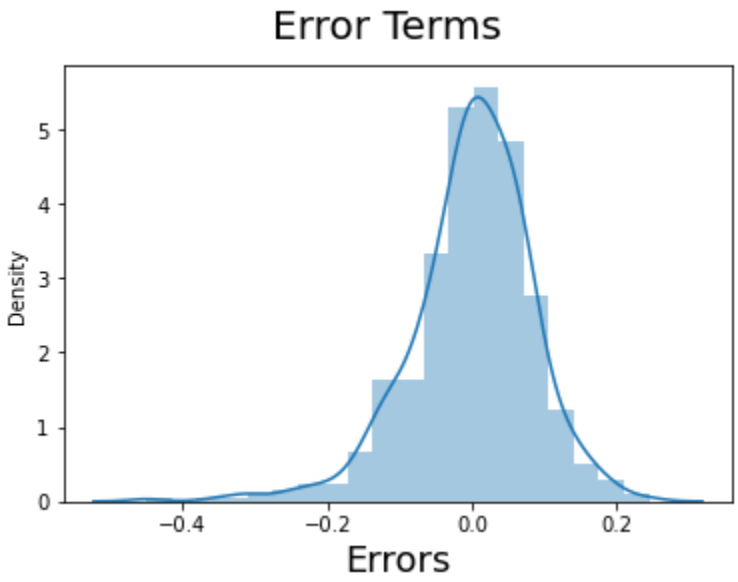
It was seen temp and atemp were highly co-related (correlation more than .9) with each other.

These two were independent variables and were highly correlated. Also atemp seems to be derived from temp so atemp field was dropped in our analysis.

It was also noted that If VIF < 3 which means Less Multicollinearity

Residual Analysis of the train data:

Residual distribution are normal distribution and centered around 0(mean =0) as shown in the image below



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

The top 3 features are:

1. 0.3054 * weathersit_Rain
2. 0.2475 * yr
3. 0.2178 * windspeed

General Subjective Questions

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

Linear regression is a statistical regression method used for predictive analysis and 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 gives a sloped straight line describing the relationship within the variables.

The linear regression model can be represented by the following equation:

$$y= a0+a1x+ \varepsilon$$

The linear regression model provides a sloped straight line representing the relationship between the variables.

y= Dependent Variable (Target Variable)
x= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each input value).
ε = random error

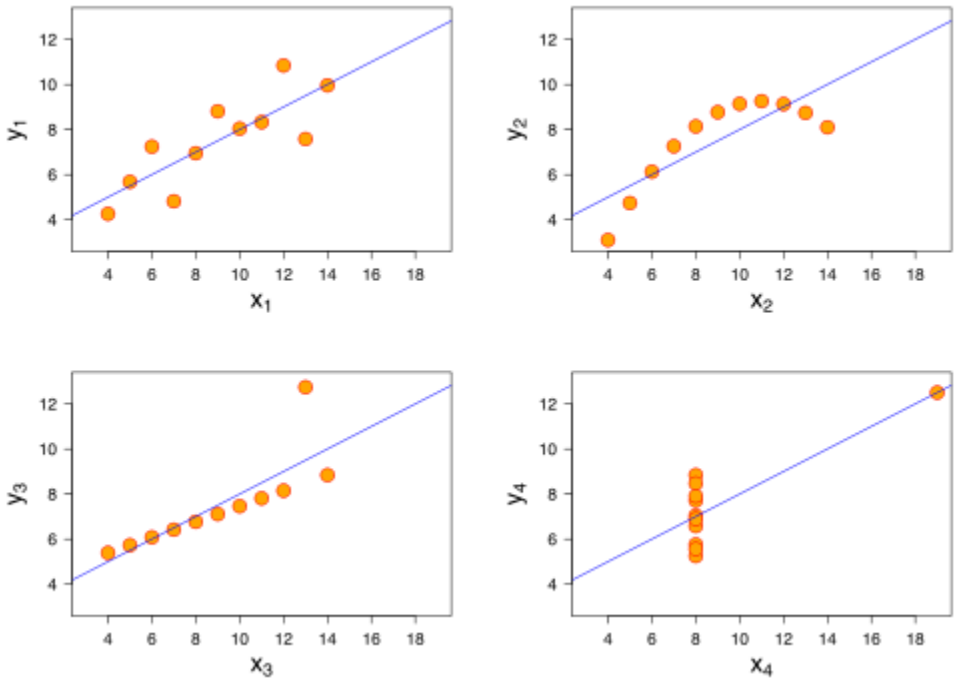
The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.

The cost function helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points. The cost function is used to find the accuracy of the **mapping function** that maps the input variable to the output variable. This mapping function is also known as **the Hypothesis function**.

In Linear Regression, **Mean Squared Error (MSE)** cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

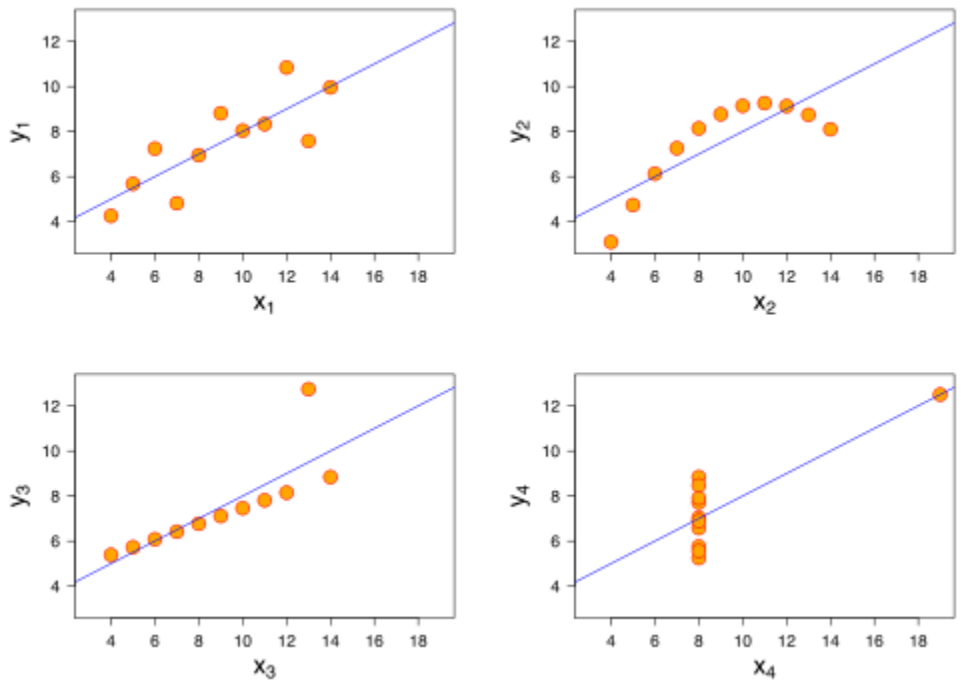
Ques 2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's quartet is a group of four data sets that are nearly identical in simple descriptive statistics. There are peculiarities that fool the regression model once you plot each data set. The data sets have very different distributions, so they look completely different from one another when you visualize the data on scatter plots.



The four data sets that make up the quartet have the same mean and standard deviation (actually very close to the same). If you look at only the values of the mean and std dev you might incorrectly conclude the fours distributions are very similar if not identical as shown in the image below.

This tells us about the importance of visualizing data before applying various algorithms to build models. The data features must be plotted to see the distribution of the samples that can help us identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.).

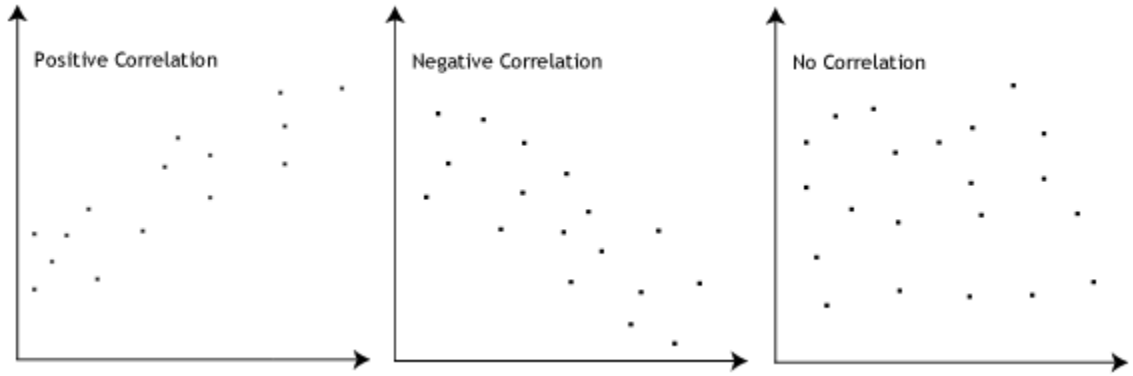


Ques 3. What is Pearson's R? (3 marks)

Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

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



Pearson r Formula

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

Here,

- =correlation coefficient
- =values of the x-variable in a sample
- =mean of the values of the x-variable
- =values of the y-variable in a sample
- =mean of the values of the y-variable

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

Scaling is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

When we collect data set it contains features highly varying in magnitudes, units, and range. If scaling is not done, then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we must do scaling to bring all the variables to the same level of magnitude.

Normalization/Min-Max Scaling:

- It brings all of the data in the range of 0 and 1.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

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

- One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression 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.

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile.

The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.