By S. Anil

Assignment-based Subjective Questions
Dataset - Bikes

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

Solution:

The list of categorical variables in the given dataset and their analysis are as follows:

1. Season
   (1: spring, 2: summer, 3: fall, 4: winter)
   Analysis –

- In Low in Spring, (Month Jan to March)
- In Summer slight increase (March to June)
- In Fall consistent (June to September)
- In Winter Decrease in cnt (September to December)

2. Year:
   2018 and 2019
   Analysis - The cnt increases in second year i.e. 2019

3. Month- Jan to Dec
   Analysis – Gradual Increase of cnt from January till September and then it declines till December upto some extent almost equivalent to March

4. Holiday –
   Analysis – when there is a holiday (0) there is an increase in cnt

5. Weekday - Sun to Sat
   Analysis - People have duly been scheduling quite uniformly on saturday, Sunday, thursdays and Fridays In Year 0 there is not much change in mean cnt on each weekday

6. Workingday (No and Yes)
   Analysis – Not much can be inferred from the workingday since the median is also quite same (from the boxplot)

7. WeatherSit –
   Analysis –

- Huge drop in cnt in Weather sit No. 3 i.e. Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- From weathersit boxplot it can be referred that the people mostly preferred the bikes during 1. Clear, Few clouds, Partly cloudy, Partly cloudy

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

Solution: If you don't drop the first column then your dummy variables will be correlated. This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example iterative models may have trouble converging and lists of variable importance may be distorted.

.

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

Solution: Registered Column Numerical Variable is highly correlated to cnt.

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

Solution:

a) Goodness of Fit : R Square = 0.792
b) Adjusted R Square = 0.788
c) F-Statistic = 189.9
d) Prob(F-Statistic) = 5.44e-163 ~= 0
e) P – Value of all the variables in the final model < .05
f) VIF Values of Variables <= 5

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

Solution:

1. Season Spring
2. Yr_1
3. Weathersit_clear

General Subjective Questions
Dataset - Bikes

1. Explain the Linear Regression Algorithm in Detail :

Solution:

- Linear Regression Algorithm is a machine learning algorithm based on supervised learning. A kind of Regression analysis, which is a technique of predictive modelling that helps you to find out the relationship between Input/ multiple input variables and the target variable.

- Linear Regression can be used to identify: a) Effect of featured variable to target variable b) Change in Target variable w.r.t input variables combined or keeping others constant. c) Prediction/Forecasting.

a) Simple Linear Regression :

The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.

b) Multiple Linear Regression :

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables (explanatory variables). The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

Methodology –

- **BEST FIT LINE :**
  Minimising the expression of Residual Sum of Squares which is equal to sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable:
  Residuals
  Y = B0 +B1X

B0 – Intercept
B1 – Slope

Using Ordinary Least Square method to minimize the RSS = The summation of (Y' – B0-B1X)Square from I = 1 to I = n
Where Y' is the predicted Targetted Variable.

The Strength of Linear Regression is assessed by following two metrics, namely :

1. Rsquare = Coefficient of Determination = (1 – RSS/TSS)

Where RSS – Residual Sum of Squares
TSS = Sum of Errors of the data from mean(Summation of (y_pred- y_mean)
This explains what portion of the given data variation is explained by the developed model. The value Lies between 0 – 1.

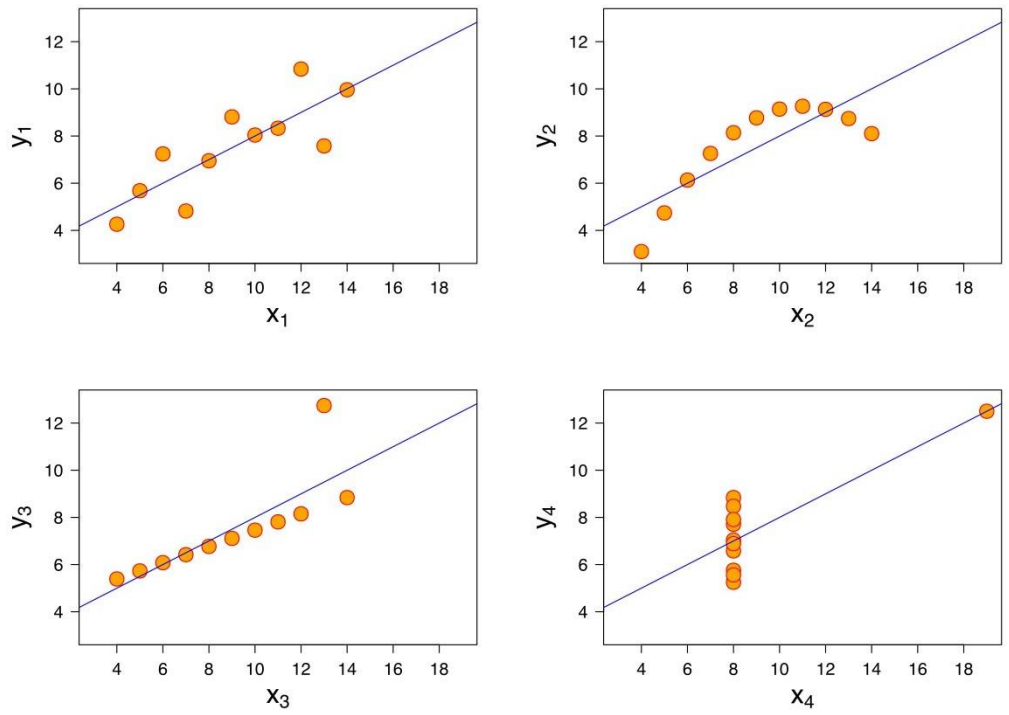
2. Performing the steps for linear regression model :
   a) Importing data set
   b) Understanding the dataframe
   c) Preparing Features and Target Variable
   d) Splitting of dataset – Training and Test data
   e) Applying model Linear Regression
   f) Coefficients Calculation
   g) Making Predictions
   h) Model Evaluation (Actual Vs. Predicted)
   i) Model Evaluation (Error Terms)
   j) Checking Mean Square Error and R Square
3. In Multiple Linear Regression
   a) Equation : Y_Pred = B0+B1X1+B2X2 ---------- +BnXn
   b) Regression Results from OLS
   c) Checking P-Value
      - H0 = Variable not significant(P >0.05)
      - H1 = Variable Significant (P<0.05)
   d) Summary of the model
   e) Creating Dummy variables
   f) R-Squared Vs. Adjusted R-Squared, the latter is better metric than R-Squared to assess how the model fits the data.
   g) Checking the Assumption if Multicolinearity is absent :
      - VIF – Variance Inflation Factor
   h) Model Validation (Rsquared between Predicted value and actual Value in the test set should be high)
   i) Variable Selection Methods - RFE

2. Explain the Anscombe's Quartet in Detail:

Solution:

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed.

However despite of the same statistics there lies a difference when plotted.



- The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

- The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

- In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

- Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

Anscombes Quartet is to illustrate the importance of looking data graphically before starting to analyse according to a particular type of relationship.

3. What is Pearsons'R?

Solution:

<u>The Pearson product-moment correlation coefficient (PPMCC), or the bi-variate correlation:</u>

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.

- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.

- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

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

Solution:

Scaling is to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

- **Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

- **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

5. You Might have observed that VIF value is infinite. Why does it happen ?

Solution:

VIF i.e Variance Inflation Factor is to identify the co linearity between features within a multiple regression model. IF there is a perfect correlation then the VIF value is infinity

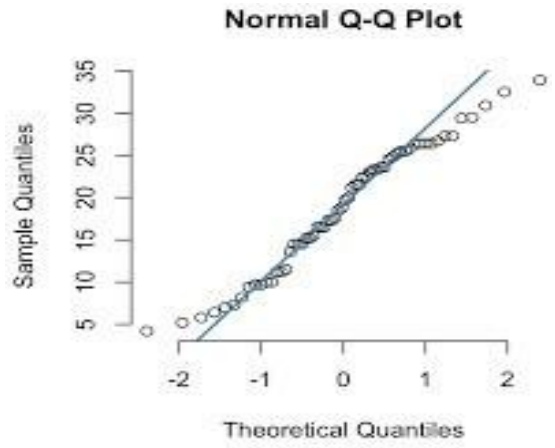
as VIF is the ration of variance between all given models beta divided by single variance beta.

6. What is a QQ plot ? Explain the use and importance of a QQ Plot in linear Regression ?

Solution:

A Q–Q plot is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions.

The points plotted in a Q–Q plot are always non-decreasing when viewed from left to right. If the two distributions being compared are identical, the Q–Q plot follows the 45° line y = x. If the two distributions agree after linearly transforming the values in one of the distributions, then the Q–Q plot follows some line, but not necessarily the line y = x. If the general trend of the Q–Q plot is flatter than the line y = x, the distribution plotted on the horizontal axis is more dispersed than the distribution plotted on the vertical axis.



We can investigate further in three ways: a density plot, an empirical CDF plot, and a normality test. Note that one should generally do the former two *after* the qq plot, as it's easiest to see that there are departures from normality in a qq plot, but it is sometimes easier to characterize them in density or empirical CDF plots.