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)**

I have done the analysis by using Boxplot,Barplot and Heatmaps are following are the observations from analysis

* fall has highest demand for rental bikes
* Demand is continuously growing each month till June,September month has highest demand. After September, demand is decreasing
* Thu,Fri,Sat and Sun have more number of bookings than week starting
* The good weathersit has highest demand
* When there is a holiday, demand has decreased,may be people want to spend time with family or at home
* can find demand has grown from 2018 to 2019

1. **Why is it important to use drop\_first=True during dummy variable creation? (2 mark)**

drop\_first=True is used for droping extra column created while dummy variable creation

if drop\_first is True it removes first column for examples if we have 4 columns and all columns are zeros then deleted one is 1 or wiseversa

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

We have temp variable with highest correlation with target variable

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

* Checking error terms normally distributed
* Linear Relationship between Actual Test Data values & Predicted Test Data values
* Multicollinearity check
* Independence of varibles

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

Top 3 features contributing significantly towards explaining demand of shared bike is

* Temp
* Sep
* Fall

**General Subjective Questions**

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

   Regression is a statistical technique that shows an algebraic relationship between two or more variables.Based on this algebric relationship (rather than a function), one can estimate the value of a variable, given the values of the other variables.

**Linear Regression** is one of the most fundamental algorithms in the Machine Learning world which comes under supervised learning, It is mostly used for finding out the relationship between variables and forecasting, Different regression models differ based on – the kind of relationship between the dependent and independent variables, they are considering and the number of independent variables being used.

  When we try to find out a relationship between a dependent variable (Y) and one independent (X) then it is known as Simple Linear Regression/ Univariate Linear regression.

The mathematical equation can be given as:

**Y = β0 + β1\*x**

Where

1. Y is the response or the target variable
2. x is the independent feature
3. β1 is the coefficient of x
4. β0 is the intercept

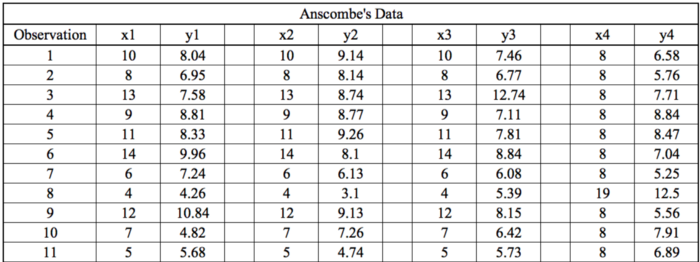
β0 and β1 are the model coefficients (or weights). To create a model, we must "learn" the values of these coefficients ,once we know the value we can use model to predict target variable

The main aim of the regression is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the points to the regression line.

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

It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the **importance**of **plotting the graphs**before analyzing and model building

There are these four data set plots which have nearly **same statistical observations**

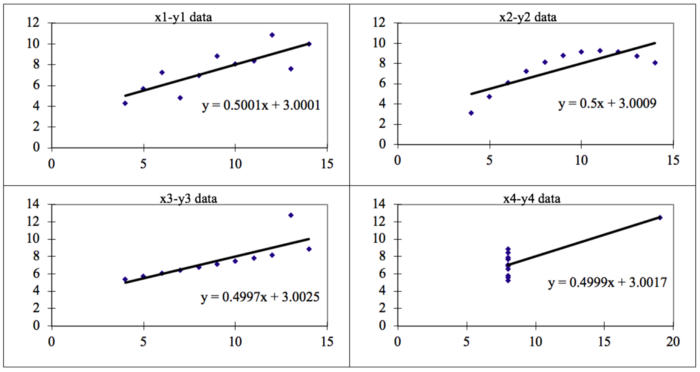


This provides same statistical information that have **variance**, and **mean**of all x,y points in all four datasets.

The statistical information for all these four datasets are approximately similar and can be computed as follows

Where mean of x is 9 for all,for y is 7.5

We can plot four data sets on x\y coordinate plane and find the observations



* **Data 1 fits**the linear regression model pretty well.
* Data 2 **could not fit**linear regression model on the data quite well as the data is non-linear.
* Data 3 shows the **outliers**involved in the dataset which **cannot be handled**by linear regression model
* Data 4 shows the **outliers**involved in the dataset which **cannot be handled**by linear regression model

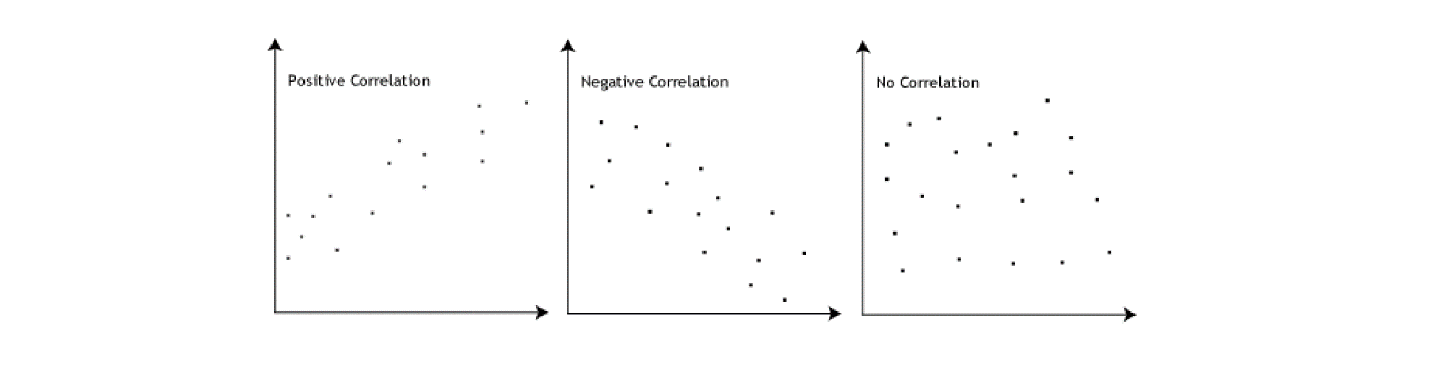
all the important features in the dataset should be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

**3. What is Pearson’s R? (3 marks)**

It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. if the variable tend to go up and down together ,the correlation coefficient will be high ,if variables go up and down in opposition with low value with one variable and other with high value correlation coefficient will be negative

However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.

**Pearson's correlation coefficient is the covariance of the two variables divided by the product of the standard deviations**.



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

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

Many times, data set contains features highly with magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units which leads to incorrect modelling. To avoid this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Difference between Normalized and Standardized scaling

| **S.NO.** | **Normalisation** | **Standardisation** |
| --- | --- | --- |
| 1. | Minimum and maximum value of features are used for scaling | Mean and standard deviation is used for scaling. |
| 2. | Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for standardization. |
| 3. | Scales values between [0, 1] or [-1, 1]. | It is not bounded to a certain range. |
| 4 | It is a often called as Scaling Normalization | It is a often called as Z-Score Normalization. |
| 5 | X=x-min(x)/max(x)-min(x) | X=x-mean(x)/sd(x) |

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

 VIF values have no upper limit, and that anything over 10 is usually bad news if you are trying to avoid multicollinearity especially for regression models such as multiple logistic regression.  perfect correlation between two independent variables and ie VIF =infinity, this value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

n 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.

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

Q-Q Plots (Quantile-Quantile plots) are one of plots to find data varies in differnt value of quantiles

That means the perpose of qq plots is to find if two set of data come from same set of distribution

The median in quantile lies in middle where 50% of data lies above and 50% of data lies below when 45 degree angle is plotted on qq plots if two data sets come from common distribution, the points will fall on reference line

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.