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?**
   a. **Season**

      i. It is obvious that count of bike sharing customers get vary with season. Fall (Season 3) shows the highest average of customers approx. 32% followed by summer and winter.
      ii. So, Season might be a good feature to analyze count
      iii. Bike Sharing varies with season as below:

      Fall > summer > winter > spring

   b. **Month**

      i. It is seen that the months May, June, July, August and September has nearly equal average of bike sharing customers and is of about 10% approx with a median value ranges between 4000 and 6000
      ii. Also, in final model Jul, Sep and Nov has some contribution towards count variable

   c. **Year**

      i. There is an increase in average customer count when compared between years 2018 and 2019 from 37.8% to 62.2%
      ii. Year variable has co-efficient of 0.24 (approx.) in final model. So, year has positive correlation with Count Variable.

   d. **Weather Situation**

      i. About 69% of bookings happened when the weather situation is of category 1 (Clear, Few clouds, partly cloudy, partly cloudy) with median of about 5000 booking.
      ii. The second situation of weather (Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist) has mean booking of about 31% with median count of about 4000 bookings.
      iii. So, the target variable count is influenced by the feature weather situation.

   e. **Holiday**

      i. It is seen that about 97.6% of customers booked for bike sharing if the day is not a holiday. Thus the data has imbalance in it and is biased for one category. So this might not be a good feature to explain the target (dependent variable)

   f. **Weekday**

      i. The variable weekday shows that the average booking happened in all days fall between 13.6% and 14.8% with median also ranges with 4000 to 5000.

ii. This indicates that this might not a good feature to predict dependent variable (count)

g. Working Day

i. About 68.5% of bookings happened on working day. This shows an underlying fact that bike sharing is happening the most on working days. So, working customers for their transportation are preferring bike sharing the most

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

a. In order to convert categorical variable to numerical encoding, get_dummies method is used in pandas. Drop_first is a parameter that needs to be set while creating dummies.

b. Drop_first parameter is used to drop the first categorical variable while doing one hot encoding

c. For Example,

i. If a categorical variable Season has 4 values – Summer, Winter, Autumn, Spring is dummied using pandas.get_dummies method we will have 4 different columns created in data frame with value either 0 or 1.

ii. If we use drop_fist=True, then we will have only 3 columns created.

Sample Data before
Creating dummies

Season
Summer
Winter
Spring
Autumn

Data After creating dummies
with drop_first=False

Season	Summer	Winter	Spring	Autumn
Summer	1	0	0	0
Winter	0	1	0	0
Autumn	0	0	0	1
Spring	0	0	1	0

Data after creating dummies
With drop_first=True

Season	Summer	Winter	Spring
Summer	1	0	0
Winter	0	1	0
Autumn	0	0	0
Spring	0	0	1

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

a. Registered is the numerical variable that has the highest correlation with target variable.

b. Since this is another count variable of different category, final model is built neglecting this feature. So, upon building model the variable temp has the highest correlation.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
    a. **Homoscedasticity**
        i. Scatter plot between residuals and predicted value will validate this assumption

    b. **Normal Distribution of Error terms**
        i. While plotting distribution plot or histogram of residuals or error terms we can prove the normality of Error terms

    c. **No Multi Collinearity**
        i. Calculating Variance Inflation Factor (VIF) for the features of final model will prove there is no multi collinearity between the independent variables

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

    a. Temperature
    b. Weather Situation (Light Snow – Weather sit 3)
    c. Year

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
    a. Linear Regression is a Machine learning algorithm which based on supervised learning. Supervised learning algorithm is one in which model is built by providing input data set along with the output that need to be predicted.
    b. Linear Regression is a regression modelling technique in which prediction is made by finding relations between the target variable with the independent variables.
    c. Linear regression is a basic form of regression where the model is built to predict the dependent variable (target) based on some independent variables (Features).
    d. There are two types of Linear regression:
        i. Simple Linear Regression (SLR) – where one target variable is predicted with one independent variable
        ii. Multiple Linear Regression (MLR) – where one target variable is predicted with more than one independent variables.
    e. As the name suggests that this algorithm aims at predicting the linear relationship between the target and predictor variables. This algorithm aims at fitting a straight line which is the best fit line by reducing the Cost Function.
    f. Let us consider an example where we need to predict the expected Sales of a company based on the Marketing budget. We have the past year's data on Sales and Marketing budget so that we have to predict the current year's Sales given the current year's marketing data. So, we found that there some linear relationship between these

variables as shown in the below plot.

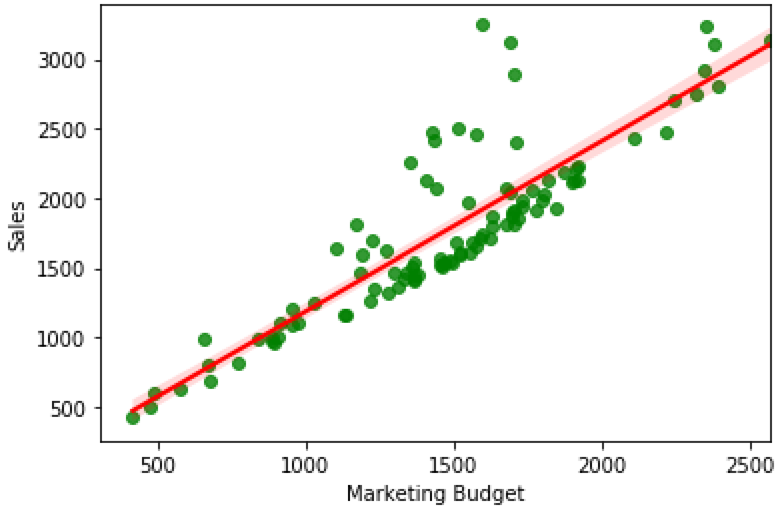


FIG: MARKETTING BUDGET VS SALES

g. The red line indicates that the best fit line which is having minimal error.

h. Linear regression follows below equation:

$$\boxed{Y = \beta_0 + \beta_1 X_1 + \beta_1 X_1 + \dots. + \beta_n X_n}$$

Where β_0, is the intercept.

$\beta_1, \beta_2, \beta_n$ are the slopes of predictor variables X_1, X_2, X_n respectively.

i. For achieving Best Fit Line, the algorithm aims to predict the Y value so that the difference between the predicted value and true value is minimum.

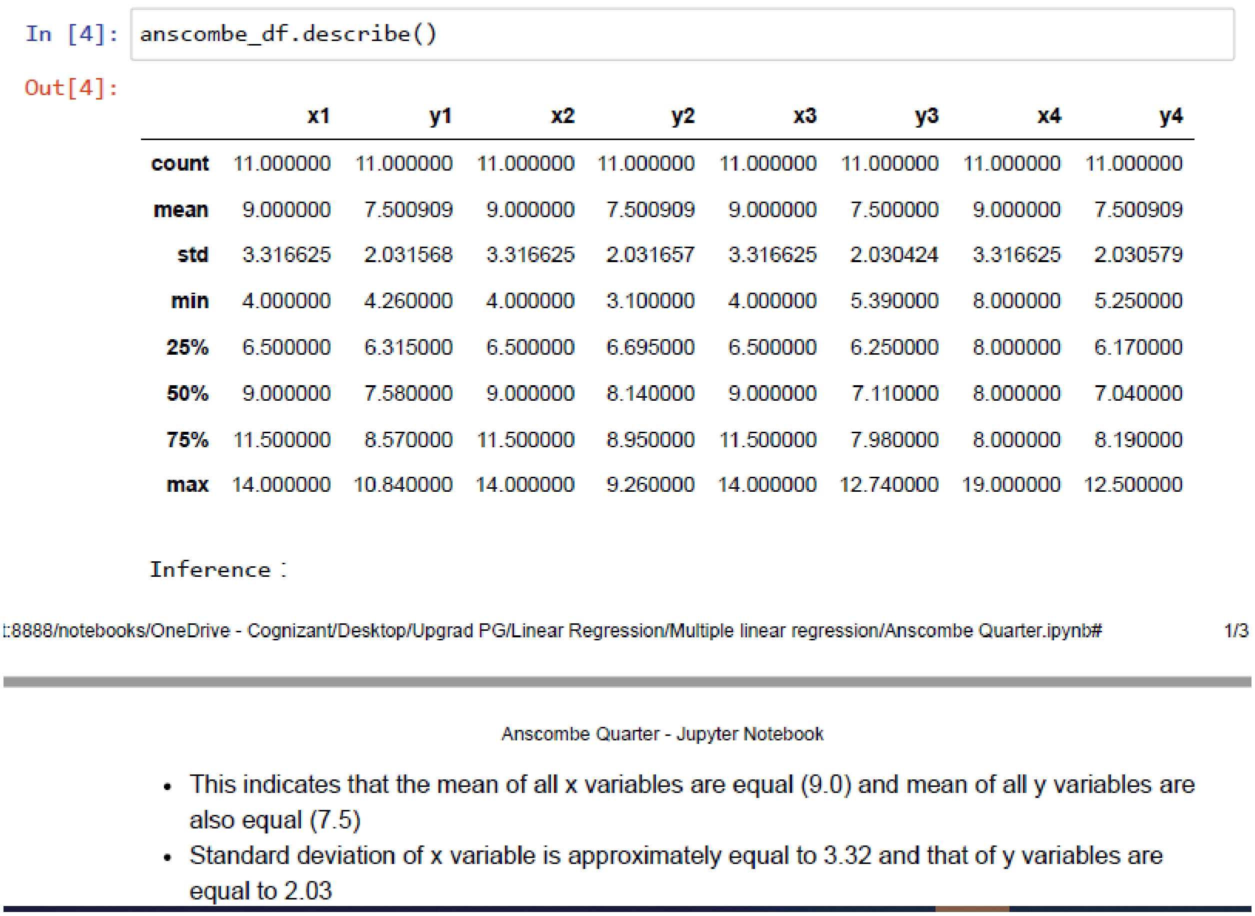
j. So, the Cost function for Linear regression is given by

$$\boxed{J = \frac{1}{n} \sum_{i=1}^{n} (pred_i - Y_i)^2}$$

k. The above equation represents the Root Mean Squared Error in case of Linear Regression and the algorithm finds the best fit line to minimize the above equation. So, the slope co-efficient are selected in such a way that the Cost Function J is minimum.

l. Linear regression uses Gradient Descent Algorithm to minimize the cost Function. In short, Gradient Decent is an optimization algorithm which is a first order iterative optimization method. This randomly assign values to co-efficient iteratively and reaches the optimal value for Cost function.

2. **Explain the Anscombe's quartet in detail.**
    a. Anscombe's quartet was constructed by Statistician Francis Anscombe in the year 1973. This quartet comprises of four data set which has approximately identical descriptive statistics but a completely different distribution appeared when they are visualized via plots.
    b. Each Data set consists of 11 points
    c. This demonstrates the importance of describing the data using graphs before starting the analysis on the data set and the influence of outliers in statistical properties.
    d. Short analysis on Anscombe's quartet is done and is attached in below screenshot



```
In [4]: anscombe_df.describe()
```

Out[4]:

	x1	y1	x2	y2	x3	y3	x4	y4
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	9.000000	7.500909	9.000000	7.500909	9.000000	7.500000	9.000000	7.500909
std	3.316625	2.031568	3.316625	2.031657	3.316625	2.030424	3.316625	2.030579
min	4.000000	4.260000	4.000000	3.100000	4.000000	5.390000	8.000000	5.250000
25%	6.500000	6.315000	6.500000	6.695000	6.500000	6.250000	8.000000	6.170000
50%	9.000000	7.580000	9.000000	8.140000	9.000000	7.110000	8.000000	7.040000
75%	11.500000	8.570000	11.500000	8.950000	11.500000	7.980000	8.000000	8.190000
max	14.000000	10.840000	14.000000	9.260000	14.000000	12.740000	19.000000	12.500000

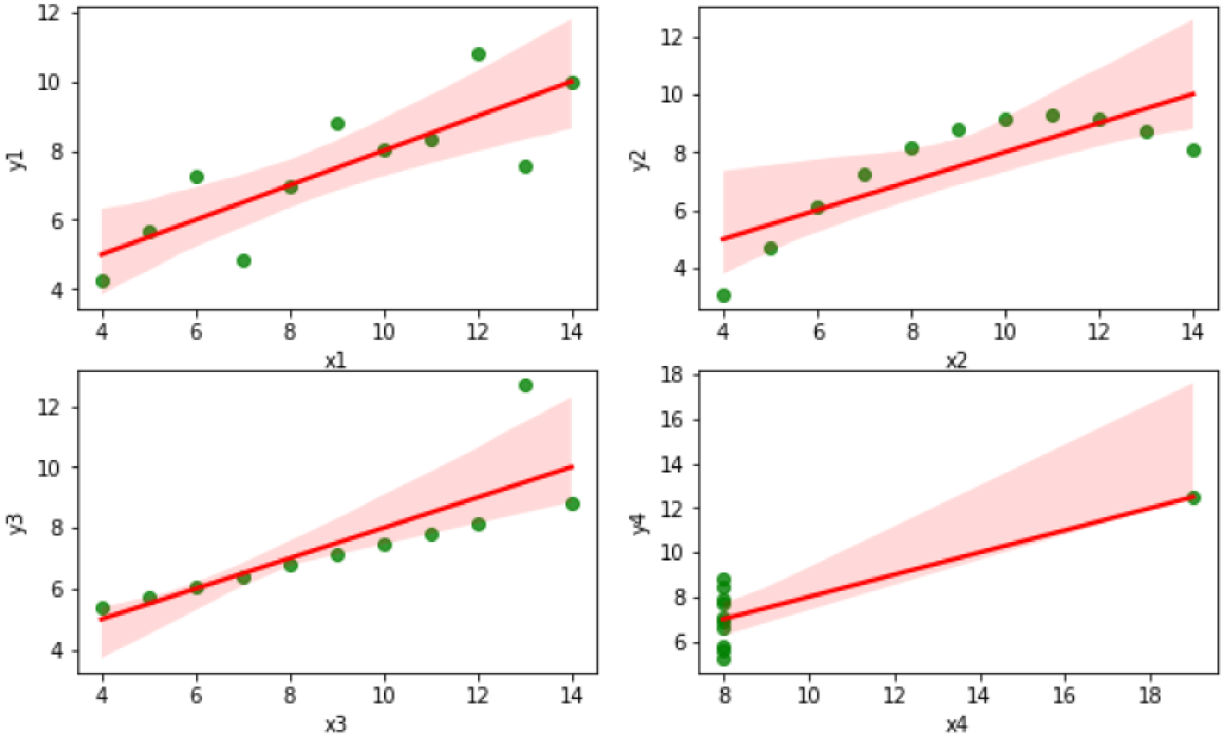
Inference :

:8888/notebooks/OneDrive - Cognizant/Desktop/Upgrad PG/Linear Regression/Multiple linear regression/Anscombe Quarter.ipynb# 1/3

Anscombe Quarter - Jupyter Notebook

- This indicates that the mean of all x variables are equal (9.0) and mean of all y variables are also equal (7.5)
- Standard deviation of x variable is approximately equal to 3.32 and that of y variables are equal to 2.03

e. Some of the inferences as below:



- ✓ The first plot shows that there is some linear relation between x1 and y1. Also, it is evident from the heat map that the Pearson-R for x and y for all 4 data sets are same (0.82)
- ✓ The second plot shows that there is no significant linear relation between X2 and Y2.
- ✓ In the third graph, the first 9 points are having strong linear relation with each other. But the presence of outliers in the data set reduces the Pearson - R to 0.82. This indicates the influence of outliers in the data set.
- ✓ The fourth graph indicates that y value varies without any changes in x value. But there is no significant decrease in correlation coefficient.

3. What is Pearson's R?

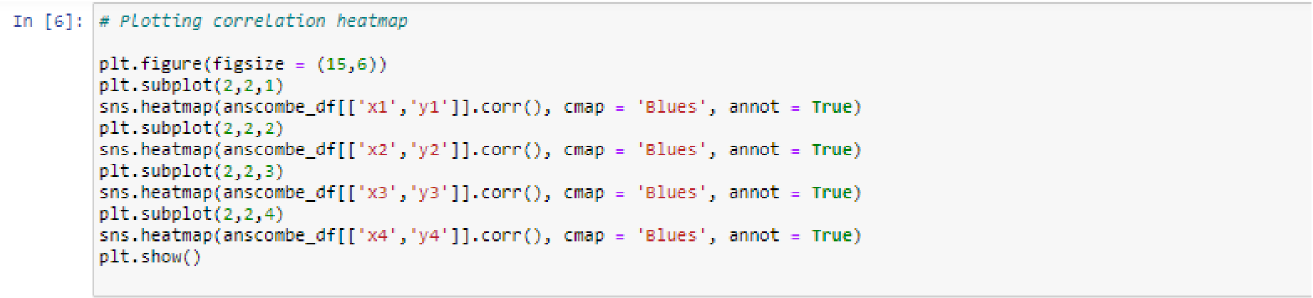
a. Pearson's R is also called as Pearson Correlation Coefficient which is used to measure the linear relationship between two numerical variables.

b. Pearson's R is very useful in Linear Regression

c. The value of Pearson's R ranges from -1 to 1.

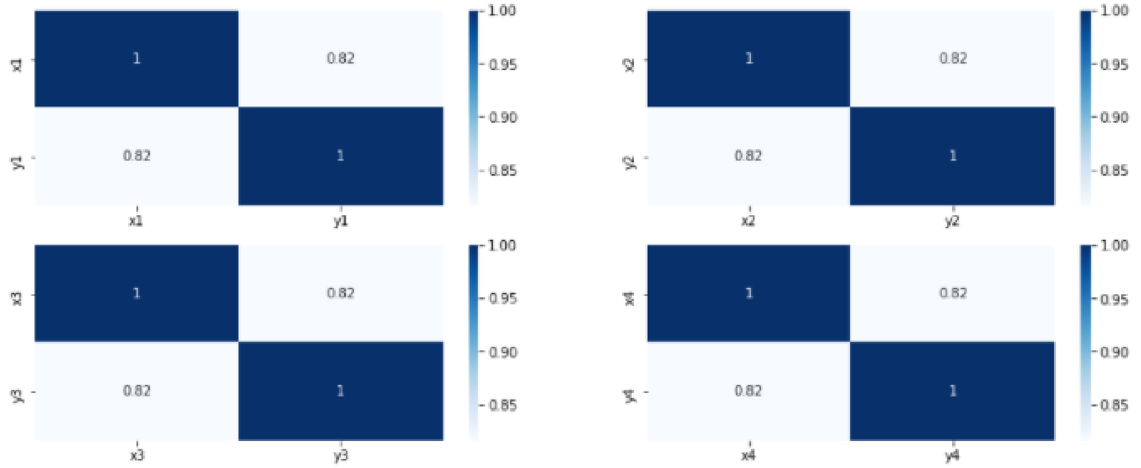
Pearson's R Value	Insight
-1	There is a strong negative linear correlation between the two variables. So, if one variable increase the other variable decreases
0	There is no linear correlation between the variables
1	There is a strong positive linear correlation between the two variables. So, if one variable increase the other variable also increases.

d. In Pandas, Pearson's R can be found by using corr() function. A sample code and the corresponding heat map is shown below.



```
In [6]: # Plotting correlation heatmap

        plt.figure(figsize = (15,6))
        plt.subplot(2,2,1)
        sns.heatmap(anscombe_df[['x1','y1']].corr(), cmap = 'Blues', annot = True)
        plt.subplot(2,2,2)
        sns.heatmap(anscombe_df[['x2','y2']].corr(), cmap = 'Blues', annot = True)
        plt.subplot(2,2,3)
        sns.heatmap(anscombe_df[['x3','y3']].corr(), cmap = 'Blues', annot = True)
        plt.subplot(2,2,4)
        sns.heatmap(anscombe_df[['x4','y4']].corr(), cmap = 'Blues', annot = True)
        plt.show()
```



4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
    a. Scaling or Feature scaling is a method of normalizing the range of independent variables or feature variables.
    b. Scaling is performed to convert all the independent variables to converge to a common scale.
    c. **Why Scaling is performed?**
        i. Ease of Interpretation
        ii. Faster convergence for Gradient Descent algorithm
    d. There are 2 common methods of Scaling
        i. Standardization
        ii. Min-Max scaling

e. Standardization converts the data such that the resulting distribution has mean value of 0 and standard deviation value 1.

   i. Formula: $X^1 = \frac{X - Mean(X)}{Std.Dev}$

f. Min – Max Scaling coverts the data within the range of 0 to1 or -1 to 1

   i. Formula : $X^1 = \frac{X - min(X)}{Max(X) - Min(X)}$

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   a. VIF stands for Variance Inflation Factor, which is a measure of Multi collinearity. If VIF value is greater than 5, then the variable is collinear with other feature and the variable has to be dropped or treated to lower VIF value

   b. $VIF = \frac{1}{1 - R^2}$
   c. If VIF = ∞ , then this represents that the variable is exactly linear with another variable
   d. In order to resolve this, we have to drop any of the variables from our model.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?**
   a. Q-Q Plot or Quantile-Quantile plot is a plot that helps us to identify whether any two data sets follows the same distribution like Normal, Uniform or Exponential Distribution.
   b. In Linear Regression, Q-Q plot is used to find whether the train and test data sets follow the same distribution
   c. A Q-Q Plot is plot of quantiles of first data set against the quantiles of second data set.
   d. In Python, statsmodel.api has qqplot and qqplot_2samples to plot Q-Q Plot for same and two different data sets respectively.