Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: In given data set our categorical variables are: season, month, weekday and weathersit. Below are the inference of these variables:

Season: Spring season has low demand as compared to other season and in fall has high demand.

Month(mnth): From Jun to Dec month there is high demands and Jan & Feb are very low.

Weekday: Demand looks like same in all the weekdays.

Weathersit: In clear few clouds weather situation is high demand and in light snow and thunder storm has low demand.

Year(yr): In 2018 bike sharing demand is low but in 2019 it is increased.

Holiday: In holidays there is less bike demand as compare to not holiday.

Q2: Why is it important to use **drop_first=True** during dummy variable creation?

Ans: drop_first=True is required because once we can identified the the category by n-1 then there is not need to keep all n category variables in data set.

For eg: We have n = 4 category for season(fall, spring, summer and winter) and below is data frame for season:

Fall	Spring	Summer	Winter
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Now, we don't need four column. We can drop the fall column, as the type of season can be identified with just last three columns:

- **100** will correspond to **spring**
- **010** will correspond to **summer**
- **001** will correspond to **winter**

Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation
with the target variable?

Ans: temp and atemp have the highest correlation with each other.

Q4: How did you validate the assumptions of Linear Regression after building the model on the
training set?

Ans: We have validate the below assumptions in given data set:

- Linear relationship between x and y.
- Error terms are normally distributed
- Error terms have constant variance - Homoscedasticity

Q5: Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: Below are the top three features:

1. Temp - Positive correlation
2. Year(yr) - Positive correlation
3. Weathersit - Negative correlation, as it will decrease bike demand will increase.

General Subjective Questions

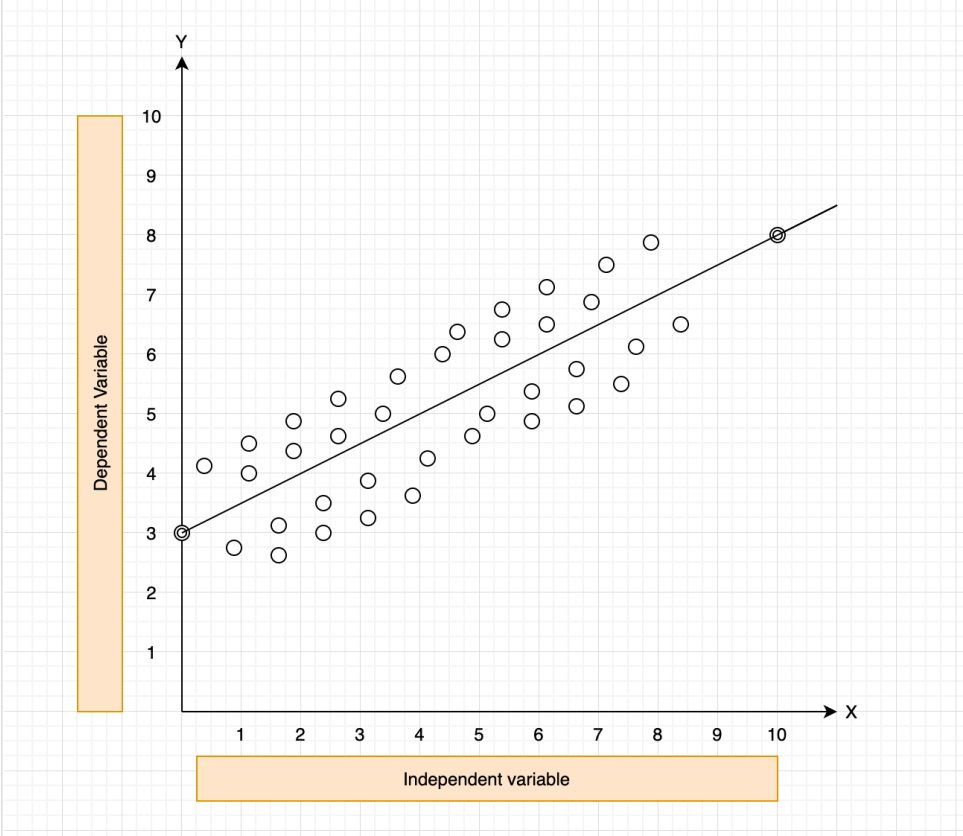
Q1: Explain the linear regression algorithm in detail.

Ans: Linear regression is used for linear relationship between continuous variables. It is mostly used in industry for exploration. It make the prediction based on the past data learning.

It is based on the Straight line equation:

y = mx + c

Where **m: slope** & **c: intercept**



Intercept: - In the above image when x = 0 then y = 3. So y=3 would be the intercept in this case. **i.e. c = 3**

Slope: The slope we calculated by (y2-y1)/(x2-x1), where (x1, y1) & (x2, y2) are two points through the line passes.

In above graph line passes (0, 3) and (2, 4). So the slope = (4-3)/(2-0) = 1/2. i**.e. m = 1/2**

In graph we have x and y axis.

In x-axis - we have independent variable.

In y-axis - we have dependent variable.

Q2: Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet is a data set of four group, in which have identical simple descriptive statistics but whenever we plot this model then we got very different distributions and appear. As you can see that the data set have very different distributed so they look entirely different from one another when you visualise the data on the scatter plots.

It is constructed in 1973 by the statistician Francis Anscombe to describe both the importance of data visualisation, and the impact of the statistical properties. This model comprises of four data-set and each dataset consists of (x, y) points. While the analyses of these four data-set, even though all these four datasets are being sharing the same statistics (mean, variance, standard deviation etc.) but when we plot graphical representation of these data-sets they look entirely different. Each graph shows different behaviour irrespective their statistics.

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04

6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Mean of x = 9.0

Mean of Y = 7.50

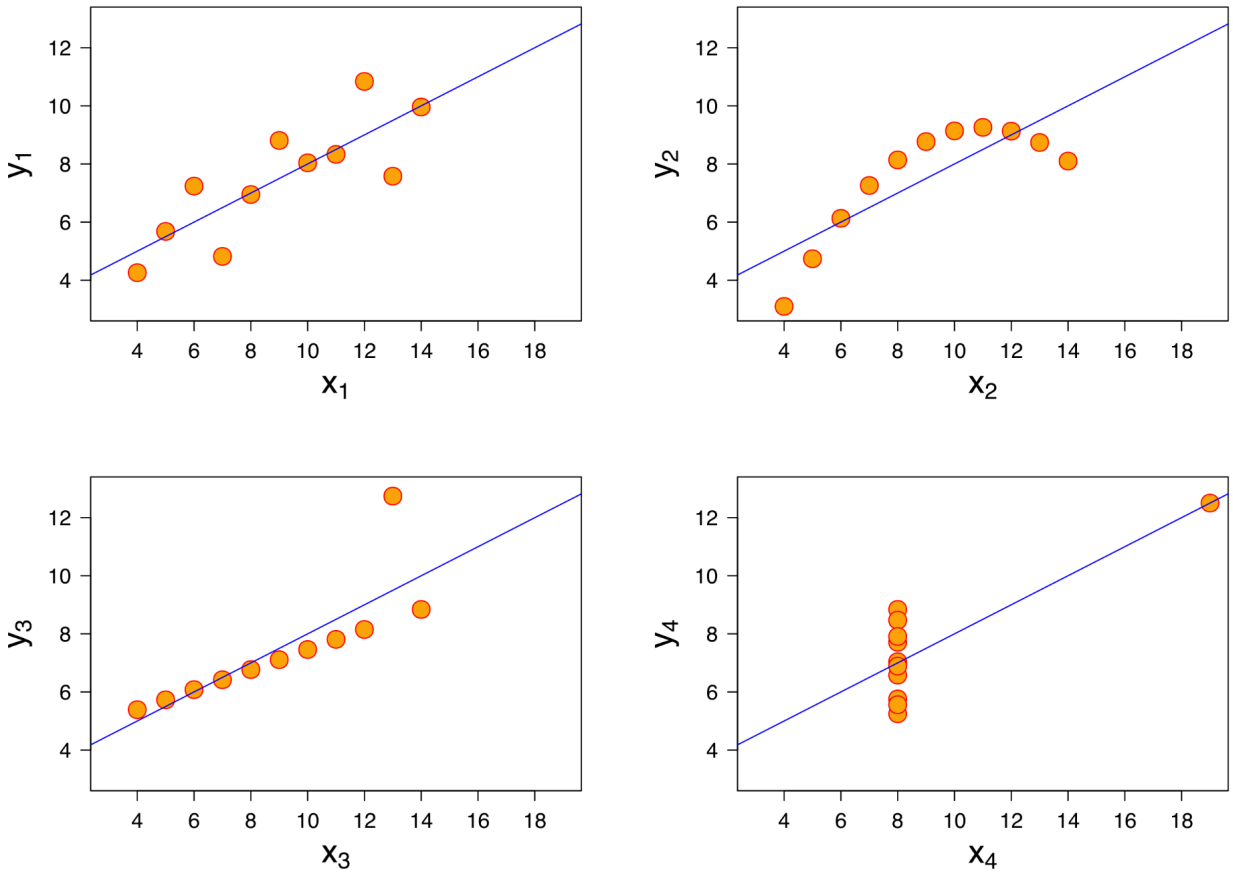
Variance of x = 11

Variance of y = 4.12

Correlation coefficient = 0.816

Linear regression line : y = 0.5*X + 3

As the the above statistic data of these 4 data-set look same but when we plot the graphical representation these, we got below graphs:



- Data-set I : Represent the linear - pattern
- Data-set II : Represent the non-linear - pattern
- Data-set III : Represent the linear - pattern except one outlier's
- Data-set IV : Represent that the values of x remains constant with one outlier's

To conclude, with the above analysis which is having identical data-set of statistics properties but might have different visuals and it is shown how any regression modelling can be fooled by the same. So before implementing the machine learning model, it is very important to have the visualisation on the important features.

Q3: What is Pearson's R?

Ans: Pearson's R is also called **Correlation Coefficient** and is used to measures the linear correlation between two variables. Its value lies between -1.0 to +1.0.

Pearson's R Correlation Coefficient is the co-variance of the two variables divided by their standard deviations. Pearson's Correlation Coefficient is named after **Karl Pearson.** He formulated the correlation coefficient from a related idea by Francis Galton in the 1880.

Below formula is used to calculate the measurement the strength of two variables.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Here, r = Correlation coefficient

xi, yi are the sample values with respect to i.

x-bar, y-bar are the mean value of sample x & y.

Real world example of Pearson's R:

- **Positive linear relationship(r = 1):** Mostly we saw the income of the person increased as the age increases .
- **Negative linear relationship(r = -1):** If the vehicle increases it's speed then the time taken to travel will decreases & vice-versa.

Q4: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: Scaling is pre-processing step which converts all the independent variables into a normalise or standardise and comparable scale. It will done after train-test the data set .

The main reason to perform this step is : we can interpret the coefficients of the features very easily after fitting the model with the help of scaling. We can prevent our model for any misleading.

So it is extremely important to rescale the variables so that they have a comparable scale. If we don't have comparable scales, then some of the coefficient as obtained by fitting the regression model might be very large or very small as compared to the other coefficient. This might become very annoying at the time of model evaluation.

So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale.

As you know, there are two common ways of rescaling:

1. Min-Max scaling (Normalisation):

    - Compress / Convert data - Between 0 and 1

- Highly affected by outliers

- Min-Max Scaling : x = (x – min(x))/(max(x)-min(x))

2. Standardisation(mean-0, sigma-1):

- Convert your data so that it has mean-0, sigma-1.

- It will definitely distort the values of dummy variables.

- Less affected by outliers.

- Standardize Scaling : x = (x – mean(x))/standard deviation(x)

Q5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: Variance Inflation Factor(VIF), given a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model. The formula for calculating **VIF** is:

VIFi = 1/1-(Ri)^2.

When there is perfect correlation between two independent features then we get VIF = infinity. In the perfect correlation we have r-squared = 1 so by the formula of VIF :

VIF = 1/(1-r2) = infinite (1/0 = infinite)

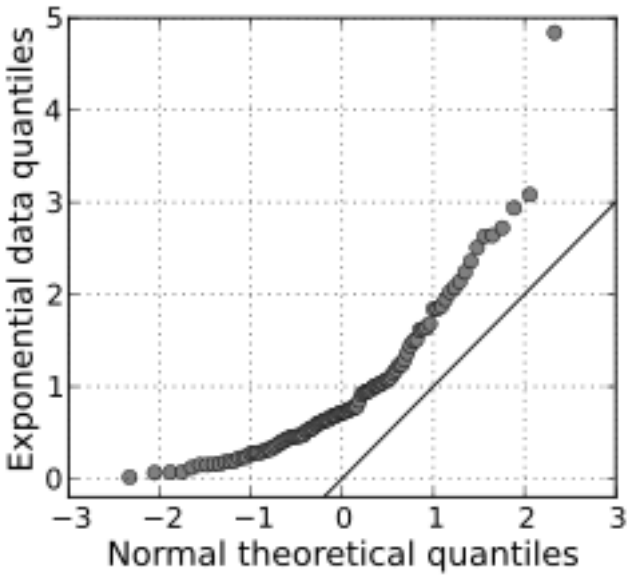
An infinite VIF value denotes that the corresponding feature may be expressed exactly by linear combination of other features which show an infinite VIF as well.

So to come over from this issue we can drop one of the feature from the dataset causing this perfect multicollinearity.

Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: A Q-Q plot (quantile-quantile plot) is a probability plot, and a graphical tool for comparing two probability distributions by plotting their quantiles against each other. A quantile is fraction where certain values fall below that quantile. For example, median is a quantile where 50% data lie above this and 50% data fall below this quantile. The purpose of Q-Q plot is to determine whether two datasets come from same distribution. Whenever we interpreting a Q-Q plot , we shall concentrate on the 'y = x' line. We also call it the 45-degree line is statistics.

A Q-Q plot showing 45 degree reference line :



A Q-Q plot is used to compare the shapes of distribution, provide a graphical view of how different properties like scale, location and skewness are similar or different in

two distributions.