Assignment – Subjective Questions

Table of Contents

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:

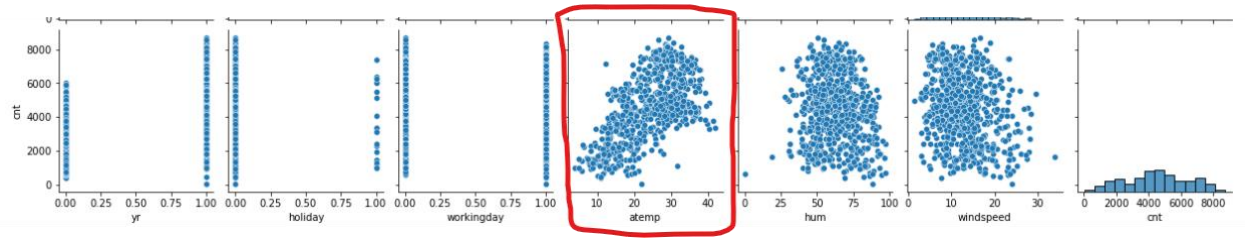
   a) Seasons 2, 3 and 4 showcase higher turnout for bike share
   b) Months 9 and 10 (Sep and Oct) showcase higher turnout for bike share
   c) Weekdays, 6 (Saturday) showcase a bit higher turnout for bike share
   d) Weather Situation, 2 and 3 have relatively lower turnout for bike share

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

   A: Dummy variables convert categorical values into binary and in the process, it creates a separate column for each category denoting values as 1/0. However, the same can be represented with one column less. For example, if we have "gender" column showing values as 1,2 (Male and Female), we can represent this in binary with single column only i.e. binary_female (1/0), which means if binary_female column has value 1 it means the actual data value is 2(Female), and if binary_female has value 0 it means the actual data value is 1 (Male).
   Hence to reduce the number of columns in overall dataset, it is good practice to use drop_first=True during dummy variable creation.

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

   A: Variable "atemp" seems to have a good strong correlation with "cnt" (target variable), as we can observe, as the "atemp" value increases we see increase in "cnt" values as well.



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

   A:

   a) Checked if all the variables in the model are significant (p-value should be less than 0.05)
   b) There should be no multi-collinearity, checked the VIF value all variables had VIF value less than 5
   c) Residual analysis, the error terms should showcase normal distribution
   d) After making predictions on the test data set, it should show linearity between y-test and y-pred

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

   A: Following 3 features are contributing significantly towards explaining the demand of the shared bikes –

	coef	P>|t|
weathersit_3	-0.3173	0
season_3	0.2941	0
season_2	0.2552	0

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

   A: It is a technique in which a dependent variable has a linear relationship with an independent variable. The main goal of Linear Regression is to learn the given data points and plot the trend line that best fits.

   If we have a dataset which contains information about x and y, where number of observations are recorded on x and y. Using this we need to find out a regression line which will give minimum error of a same model is applied on new dataset.

   Linear Regression Model is represented by following equation.

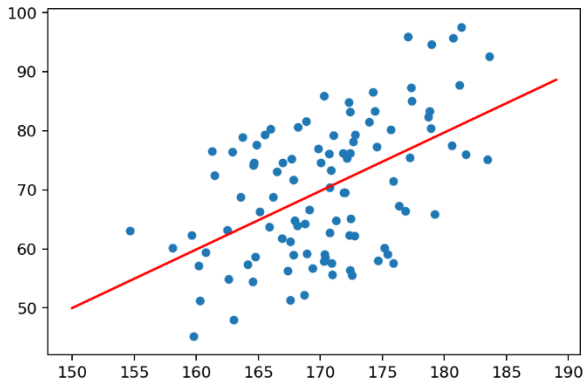
   y = b0 + b1X

   where –

   y = Dependent Variable

   b0 = Intercept

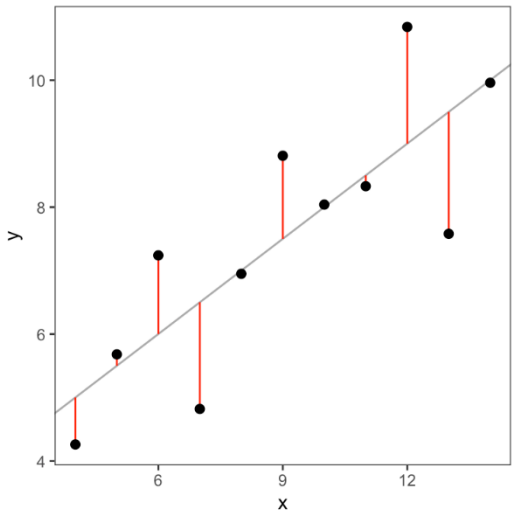
   b1 = Slope



In above image the red line denotes the best fit line.

The best fit line is then found by minimising the RSS (Residual Sum of Squares), which is the residual for any data point with respect to the predicted line. It can be calculated with the help of following formula.

$$RSS=\sum_{i=1}^{N}(Y_i-B\hat{}_0-B\hat{}_1X_i)_2$$



Similarly, TSS is also calculated, which is the sum of errors of the datapoints from mean of response variable.

The strength of this regression model is then evaluated with R^2 Or Coefficient of Determination, which can be calculated with formula: $R^2 = 1 – (RSS / TSS)$

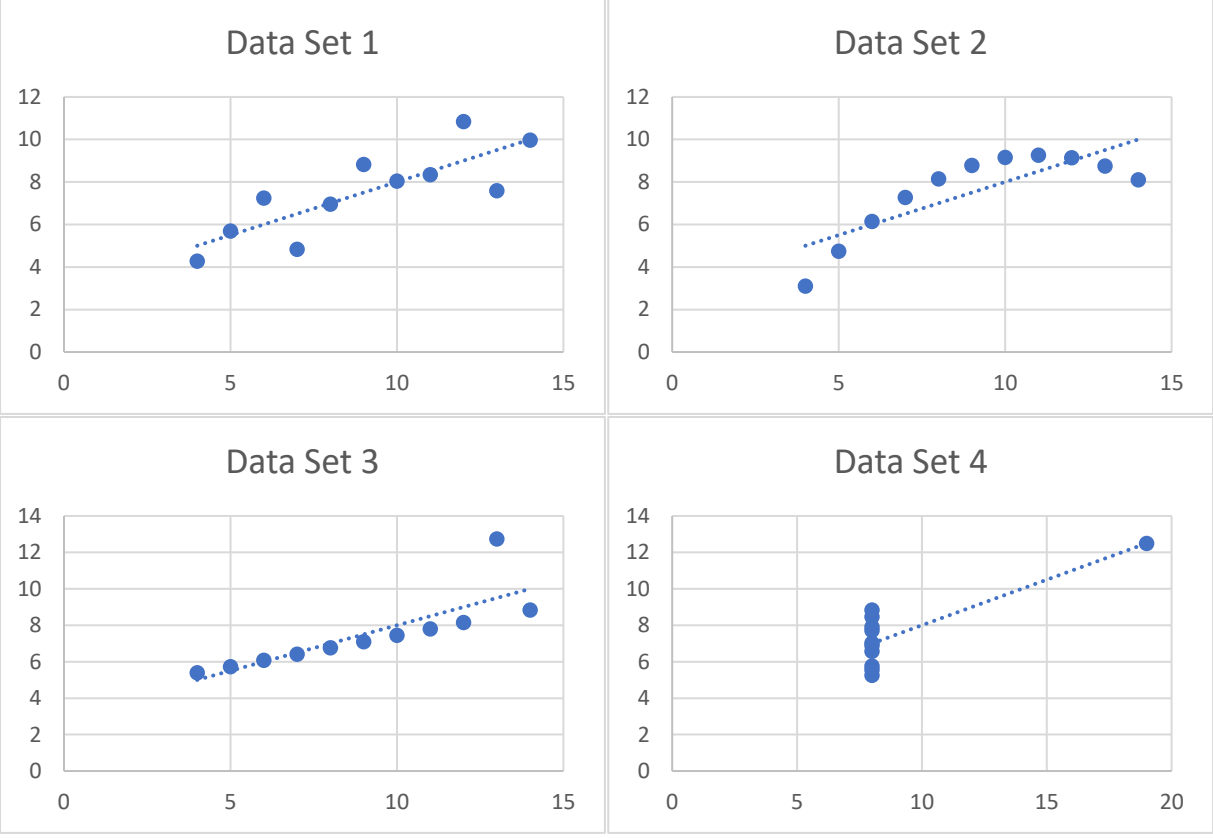
Basically, the R^2 indicates the how well the regression model fits the observed data. For example, R^2 of 60% reveals that 60% of the data fit the regression model.

2. Explain the Anscombe's quartet in detail.

A: These are basically four datasets which appears to have very identical descriptive statistics (mean, standard deviation etc.), however when plotted on a graph these 4 data sets show different trends. Each of the dataset consists of 11 coordinates, and these were originally constructed to demonstrate the importance of graphs in data analysis.

	Data Set 1		Data Set 2		Data Set 3		Data Set 4	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
S.D.	3.2	1.9	3.2	1.9	3.2	1.9	3.2	1.9

If we observer the above table, the mean and standard deviation for each data set points are identical, however if we observer the same data in charts, we can see different trends.



3. **What is Pearson's R?**

A: It is standard measure of correlation in statistics, it shows the linear relation between two sets of data. It is usually denoted by letter r. It's value ranges from -1 to 1 where –

-1: means strong negative correlation

0: means no relationship at all

1: mean strong positive correlation

Formula for R is given as follows -

$R = n(\sum xy) - (\sum x)(\sum y) / \sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2}$

Where -

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx2 = the sum of squared x scores

Σy2 = the sum of squared y scores

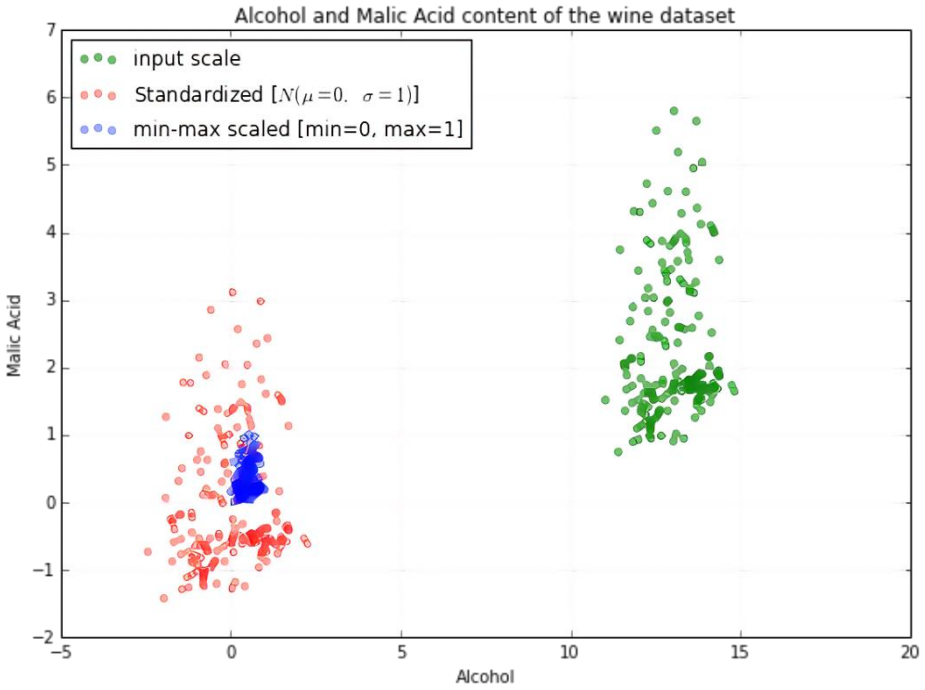
4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

A: It is a technique to transform numerical data into a single scale. Since each numerical data could have its own range and it will impact the coefficients and will create incorrect interpretation of the importance of the variable. Hence it is recommended to scale all numerical variables into single unified scale.

Standardize Scaling: This technique scales all data in standard normal distribution with mean 0 and standard deviation as 1.
Formula: x = x – mean(x) / sd(x)

Normalized Scaling (MinMax Scaling): In this technique all the data is fit into range of 0 to 1.
Formula: x-min(x) / max(x) – min(x)

The normalized scaling is used generally, as it fits data in range of 0-1 which works with other binary variables as well.



In the above image we can see the difference between standardized and normalized(min-max) scaling, as for the standardize scaling data is fit into scale of ~3.25 - -1.25, whereas for the normalized scaling data is fit into scale of 0-1.

Most important thing to note in both the scaling techniques the shape of the original data remains the same.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   A: VIF is used to detect the multicollinearity amongst the predictor variables. The formula for VIF is – VIF = 1 / (1-R^2)
   Where, R^2 indicates the how much variance of a dependent variable is explained by other independent variables. So, in case where one variable is perfectly correlated with other variable, we will get R^2 as 1 which means VIF = 1 / 0 = infinity
   Hence when we get VIF as infinity, we should understand that the given variable is in perfect correlation with the other.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
   A: It is a graphical method to determine whether two samples of data came from the same distribution or not.
   It helps in a scenario where we get training and test data set received separately. In such case we can confirm using Q-Q plot that both the data sets are from populations with same distributions