Assignment-based Subjective Questions

Que 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans.

Below is the equation of best fit line, it explains the predictors which are affecting the target variable positively or negatively along with their coefficient.

The categorical variables are underlined and highlighted with bold.

cnt = (0.130722 + 0.232563* **yr** - 0.096575*__**holiday**__ + 0.517336* temp - 0.149709* windspeed

+0.101217* **season_summer** + 0.137090*__**season_winter**__ + 0.054141*__**mnth_Aug**__

+0.116291* **mnth_Sep** - 0.081139*__**weathersit_2**__ - 0.281852*__**weathersit_3**__)

1) **year**: year and count show positive relation with count variable, which means the demand is going to rise over the coming years, although this result may be biased as we only had dataset of 2 years which was following the same trend.

2) **holiday**: holiday showed negative relation with count, as per our equation, on holiday the rental count would dip by 0.096575 units.

3) **temperature** - with unit increase in temp the count will increase by 0.517336. As the data had max temp 35 C, this interpretation might hold true till temp in the vicinity of 35 C only.

4) **windspeed**: windspeed and count shows negative relation. For a unit increase in wind speed the rental count would decrease by 0.149709.

5) **season_summer, season_winter**: Both the seasons shows positive relation with the count.

6) **mnth_Aug, mnth_Sep**: In month of Aug., Sept the demand for rental would increase.

7) **weathersit_2, weathersit_3**: For a unit increase in weather situation 2, the count will decrease by 0.081139 and for a unit increase in weather situation 3, the count will decrease by 0.281852

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

Ans.

When we apply "get_dummies" function to a categorical column, the column gets split into number of columns which are equal to the unique values present in it (number of categories) and assigns 1 & 0 where that category is True in each column. This is also knowns as "One Hot Encoding".

Now "drop_first = True" drops first column, first encoded category, from all the encoded columns. This results in 1 column less than number of categories in which were present in a column. i.e., if 5 unique categories were there in a column, then it would get split into 4 columns with first category removed.

This is the most important step to drop first column because, we have to make the dummy variables linearly independent of each other otherwise we will end up in dummy variable trap which is whenever one variable can be derived from rest of the variables then these variables are said to be multicollinear resulting in R-squared value 1 and ultimately VIF value infinite which is clearly the case of multicollinearity. When all the variables are correlated, it would be difficult for model to tell how strongly a particular variable is affecting target variable.

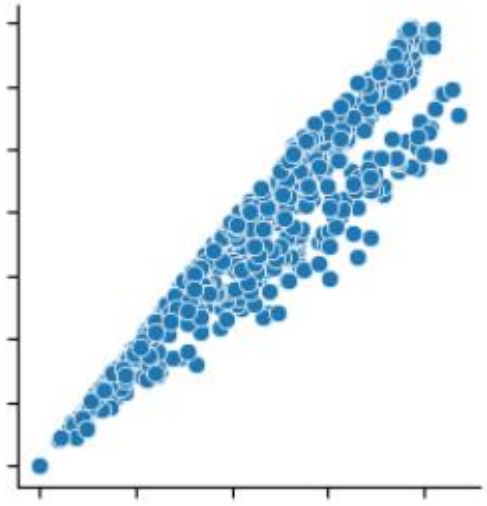
Example:

Let's take example from our bike dataset only, for season variable, there are 4 seasons and once we perform one hot encoding 1's and 0's will be assigned to each season. Consider summer season, the encoding will be 1 for summer and 0 for all other seasons, now the model can derive that if all other seasons are 0 then it is definitely summer season. In this case, model was able to predict one variable based on other remaining variables, this shows the multicollinearity between season variables. Now, once we drop any of the season variable, e.g., summer, then model will have no other option than to use rest of variables for prediction as we have already taken care of multicollinearity by dropping one variable.

Que 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans.

"Registered" numerical variable has the highest correlation with target variable.



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

Ans.

Assumption 1: Numerical predictors variables and target variable should have linear relationship.

This assumption can be validated by plotting pair plots of numerical variables and at least few of the variables should follow linear relationship with target variable. “temp”,” atemp”,” registered” & ”casual” variables were following linear relationship with “cnt”.

Assumption 2: Error terms are normally distributed

After model building, we can predict the target values using our model and the difference between the actual target value and predicted target value is the error and these error terms are assumed to be normally distributed, so plotting the distribution plot of error terms (actual – predicted), we can validate this assumption.

Assumption 3: Error terms are independent of each other.

This assumption states that the error terms do not have any correlation with each other, this means that the next error term should not depend on the its previous error term or terms. This can be validated by plotting regplot of predicted y and error terms and look for any patterns in the graph, if we fail to identify any pattern then we can say that the error terms are independent of each other.

Also, calculating “Pearsonr” coefficient and checking if its value is almost 0, this proves the strength of correlation between predicted values and error terms is almost 0.

Assumption 4: Error terms have constant variance (Homoscedastic)

This assumption can be validated by plotting regplot of actual vs predicted values and observing the data points. If the data points are equally distributed along with the line and are not converging towards the line or diverging away from the line then we can say that the error terms have equal variance.

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

Ans.

Sr no	Feature	Description	Coefficient
1	Temp	Temperature in C	+0.517336
2	Weathersit_3	Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog	-0.281852
3	Yr	Year	+0.232563

General Subjective Questions

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

Ans.

Linear regression is finding the equation that describes the correlation between independent variables(predictors) and dependent variable (target). This is achieved by finding the best fit line which passes through maximum number of data points keeping the sum of squares of errors at its minimal (least square error). Error or residual is the difference between actual and predicted value of target variable.

There are two types is linear regression: 1) Simple Linear Regression.

2) Multiple Linear Regression.

The equation of **Simple linear regression** can be written as "y = a + b X" OR

in standard terms "$y = ß_o + ß_1 x + Ɛ$".

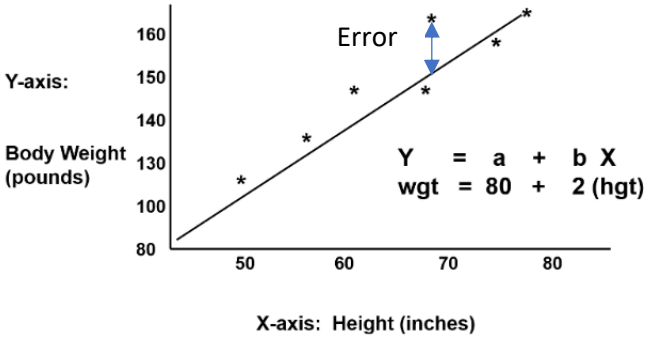
Where, y is target variable

x is predictor

$ß_o$ is y-intercept (constant)

$ß_1$ is slope of the line

Ɛ is error term.



The equation of **Multiple linear regression** can be written as "$y = ß_o + ß_1 X_1 + ß_2 X_2 + ... + ß_n X_n + Ɛ$".

This is extension of Simple LR where the target variable is correlated to the more than one predictor. The predictors can be numerical or categorical or both. ß denotes change target variable per unit increase in variable when all other variables are kept constant.

$Ɛ = sum((y-y\_pred)^2)$ → squared to remove "-ve" sign.

A linear regression model helps in predicting the value of dependent variable by finding out the coefficient of an independent variable, which signifies how strongly that variable is affecting the dependent variable. Also, the accuracy of prediction is denoted by the R-squared and p-value values. The R-squared value indicates how much of the variation in the dependent variable can be explained by the independent variable and the p-value explains how reliable that explanation is. The R-squared values range between 0 and 1. A value of 0.8 means that the explanatory variable can explain 80 percent of the variation in the observed values of the dependent variable. A value of 1 means that a perfect prediction can be made, which is rare in practice. A value of 0 means the explanatory variable doesn't help at all in predicting the dependent variable. Using a p-value, you can test whether the explanatory variable's effect on the dependent variable is significantly different from 0. Usually the dependent variables with p-value less than 0.05 (α is 5%) are considered significant, Therefore, dropping these variables will cause R-squared value to drop and if R-squared value does not drop significantly that means that the dropped variable was insignificant.

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

Ans.

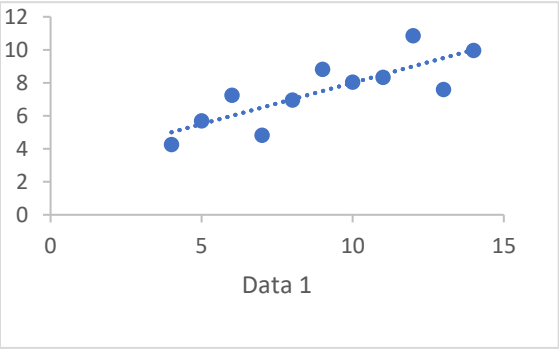
In 1973, a famous statistician Francis Anscombe constructed a group of four data sets which were identical in descriptive statistics but were very different from each other when plotted using scatter plot. The group of this 4 data set is known as Anscombe's quartet.

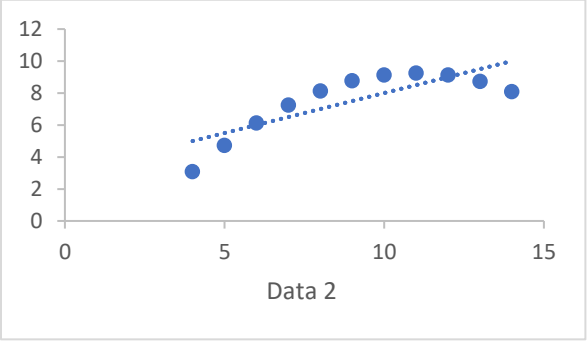
The motive behind Anscombe's quartet is to illustrate the importance of plotting the graphs before analyzing and model building, the effect of outliers and other observations on statistical properties. Anscombe's quartet mainly counters the impression that "numerical calculations are exact, but graphs are rough". Let's take a look at it in detail. Below are the 4 data sets with x and y and their statistical properties.

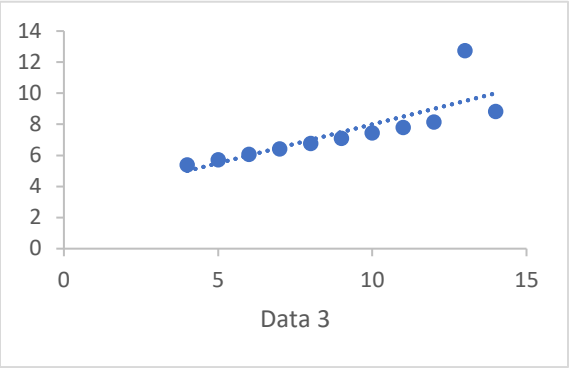
Anscombe's quartet							
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	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

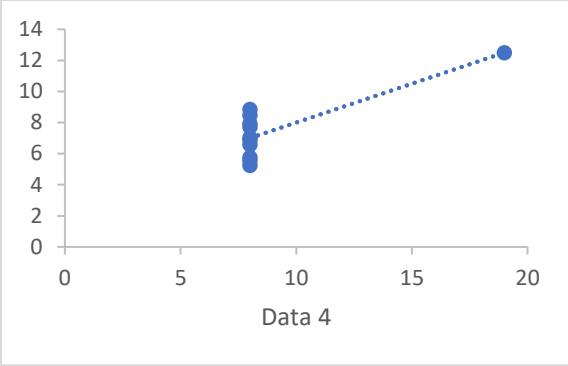
Mean	9.000	7.501	9.000	7.501	9.000	7.500	9.000	7.501
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
Pearson r	0.82		0.82		0.82		0.82	

We can observe that all 4 data sets have near identical statistics, so if we were to find the bets fit line for the above data set it will be "y = 3.0 + 0.5 x", but the catch is all the four datasets have different graph. So, a linear regression model can be easily fooled by this dataset if we relied only on statistical data only and ignored the graphs of dataset. Let's see the graphs of each data set.









Dataset 1 fits the linear regression model pretty well.

Dataset 2 could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model.

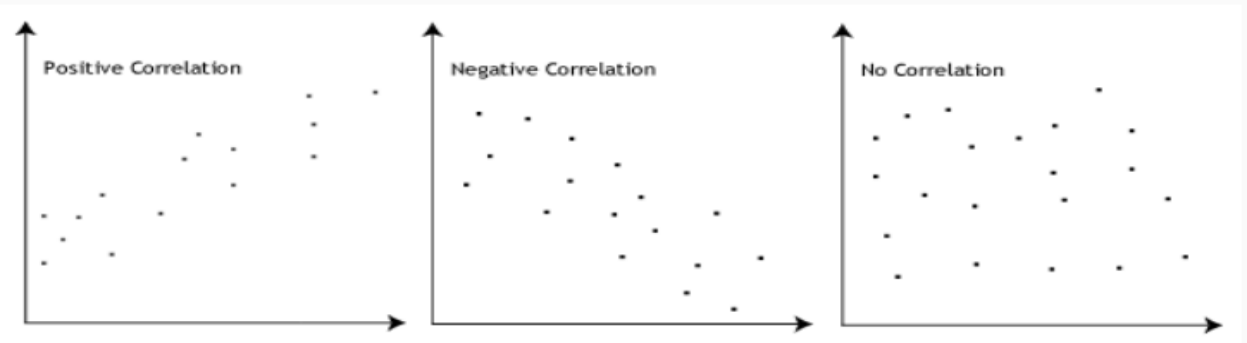
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model.

So, it is crucial to visualize data before implementing any machine learning algorithms and above four data sets were created intentionally to demonstrate the same.

3. What is Pearson's R? (3 marks)

Ans.

Pearson's correlation coefficient also referred as Pearson's R is the measure of linear correlation between two data sets which ranges between -1 to 1. the magnitude of coefficient implies the strength of linear correlation, 1 implies perfect linear correlation and 0 implies no linear correlation. The sign implies the whether the variables are directly proportional or inversely proportional. i.e., if sign of coefficient is +ve then with increase in x, the value of y also increases and vice-versa. if sign of coefficient is -ve then with increase in x, the value of y decreases and vice-versa.



The formula for Pearson's R is as below

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = Pearson's correlation coefficient,

x_i = values of x variable, $\bar{x}$ = mean of the values of x variable

y_i = values of y variable, $\bar{y}$ = mean of the values of y variable

When to use Pearson's Correlation coefficient:

1) The variables are quantitative.
2) The variables are normally distributed.
3) The data have no outliers.
4) The relationship between variables is linear.

We can use Spearman's rank correlation coefficient which is a better measure than Pearson's R if any of the above condition is not satisfied by the variables.

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

Ans.

If a dataset has data points far from each other, scaling is a technique to make them closer to each other or in simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower.

In the machine learning algorithms if the values of the features are closer to each other then there are chances for the algorithm to get trained well and faster instead of the data set where the data points or features values are farther away from each other or have high differences with each other will take more time to understand the data and the accuracy will be lower. Scaling the values is a good idea in regression modelling because scaling of the data makes it easy for a model to learn and understand the problem. Normalization and Standardization are the two main methods for the scaling the data.

Normalization: Normalization is the method of rescaling data where we try to fit all the data points between the range of 0 to 1 so that the data points can become closer to each other and it is performed when the data is not following normal or gaussian distribution. In linear regression, most of the numerical data is not in a specific range, performing normalization or using MinMaxScalar brings the data in range 0 to 1 which makes variables comparable in terms of their coefficients in equation of best fit line. But it may not preserve the relationship between variable but preserves relationship in a variable and is sensitive to outliers. So, if outliers are present in the data set or relationship between data points of different variables is important aspect we should avoid using or use normalization by taking care of these aspects.

Standardization: Standardization is to make data points centered about the mean of all the data points presented in a feature with a unit standard deviation. This means the mean of the data point will be 0 and the standard deviation will be 1. Standardization is performed when the data set is following gaussian or normal distribution and is less sensitive to outliers and also preserves the relationship between the data points of different variables

Formulae:

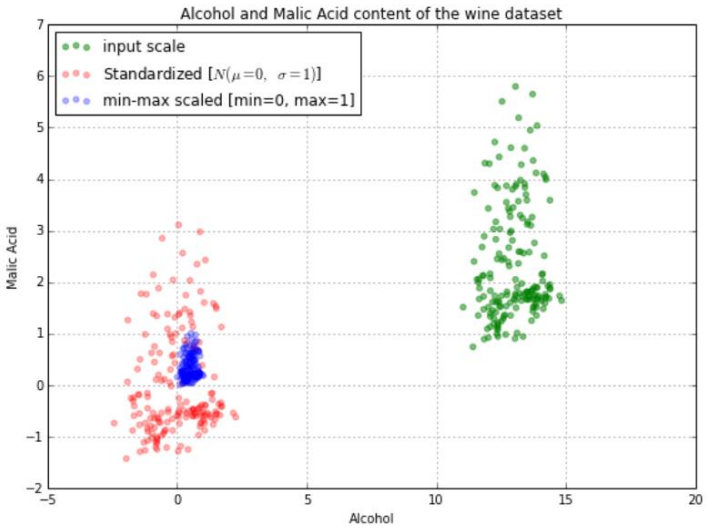
$X_{norm} = (X - X_{min}) / (X_{max} - X_{min})$

$X_{stand} = (X - \mu) / \sigma$

Where, μ is mean of sample and σ is standard deviation of sample.

Effect of scaling:

We can visualize the effect of scaling from graph. The green data points are of default data set, blue data points are of normalized data and red ones are of standardized data.



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

Ans.

$$VIF = 1 / (1 - R^2)$$

VIF stands for variance inflation factor, it explains how much times the variance of coefficient of a variable is greater than the variance if the variable were entirely independent of any other variable.

This means that if a variable X1 has VIF value 3.2, then it implies that the variance of X1 is currently 3.2 times greater than variance of X1 if the variable X1 would have been non-correlated with any of the other variable. In simple terms, if variable X1 is non-correlated with any other variable then its R-squared value would become 0 and VIF 1, which shows perfect non-correlation. And if there is some correlation between X1 and other variables then R-squared value would be greater than 0 and accordingly VIF would increase.

In a dataset, it might be possible that some of the predictor variables have perfect correlation with each other, means one predictor can be perfectly explained by other predictor variables which is called as multicollinearity, this causes R-square value to be 1 and resulting in VIF value infinite. To overcome this issue, we need to drop the one variable causing multicollinearity.

The case where we encounter VIF to be infinite is when we use one hot encoding on a categorical variable for dummy columns and forget to drop any of the encoded column per variable. For e.g., if a variable has 3 categories and we performed one hot encoding but did not drop any one of the 3 dummy columns then any one of the three dummy variables would be perfectly correlated with the remaining two. This will raise issue of infinite VIF.

Also, when two numerical variables have correlation with each other such that one can be derived from another and we fail to drop the less important variable then we will definitely get VIF value as infinite. e.g., suppose we have three columns, marks, out of, and percentage then from percentage and out of, marks can be derived, from marks and out of, percentage can be derived and from marks and percentage, out of can be derived. So, all three variables are perfectly correlated with each other. Therefore, we will have to drop either of three variables to take care of multicollinearity and infinite VIF value.

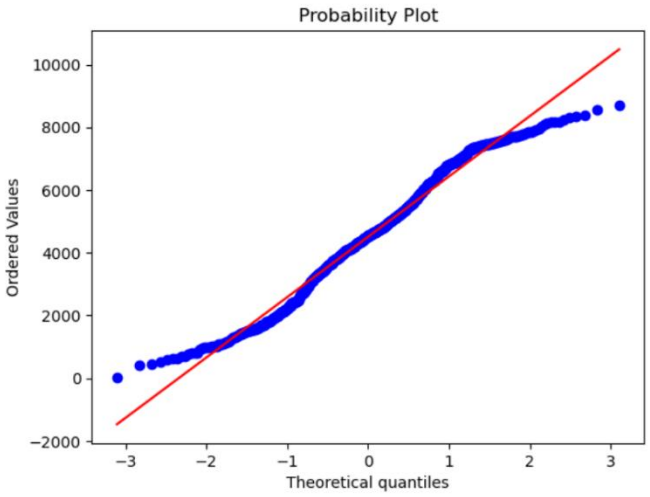
6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Ans.

Q-Q plots are also known as Quantile-Quantile plots. The Q-Q plot plots the quantiles of a sample distribution against quantiles of a theoretical distribution. This helps us to determine if the input dataset follows the given theoretical distribution like normal, uniform, exponential etc.

Q-Q plot is basically a regression plot with quantiles of the theoretical dataset on x axis and the quantiles of dataset whose distribution we want to check on y axis. The output of QQ plot is a scatter plot with each quantile of both the dataset plotted as an individual points and a straight line at 45° angle. If the line is passing through most of the points, then we can say that the given dataset follows the theoretical distribution. We can test the given data set against normal, uniform, exponential distributions.

For e.g., the given graph is of count variable which is target variable from our data set plotted against the theoretical normal distribution and we can see that most of the points are passing through the red line which implies normal distribution.



Q-Q plot is used to check if two data sets

1. come from populations with a common distribution
2. have common location and scale
3. have similar distributional shapes
4. have similar tail behavior

Importance of a Q-Q plot in linear regression

1. We can check if the distribution of actual values is similar to the distribution of predicted values.
2. By plotting QQ plot of predicted vs actual values from training dataset, we can even conclude if X values are greater or Y values are greater by observing the region in which most of the data points are lying. If the most of the data points are above the line, then we can say that Y values are greater and if the most of the data points are below the line, then we can say that X values are greater.
3. We can also check if the distribution between train and test dataset is similar or test data set is normal so that we would get an idea about whether to expect significant variation in R-squared value of train and test data set results or not. For e.g., from QQ plot, if we know that train and test data set have great variation in their distribution, then we can also expect the same results in R-squared value.