Assignment-based Subjective Questions

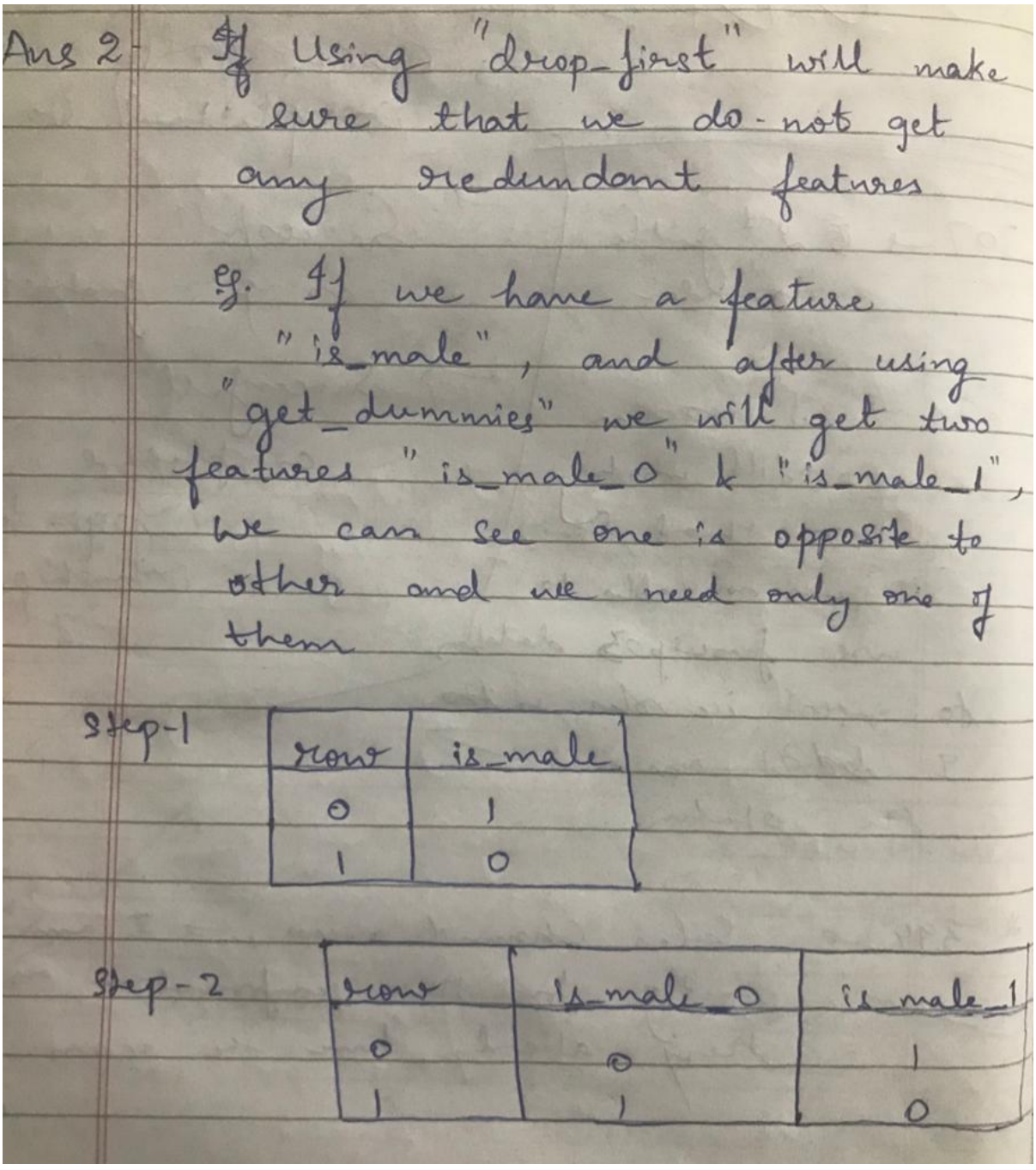
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans : Among the categorical variable the most influencing variable for the count is 'yr_1' which is Year 2019.

And the variable 'weathersit_Light_snow_rain' i.e. when the weather condition is Light Snow and Light Rain is relatively high negetively correlated among all other feature variables

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

Ans :



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

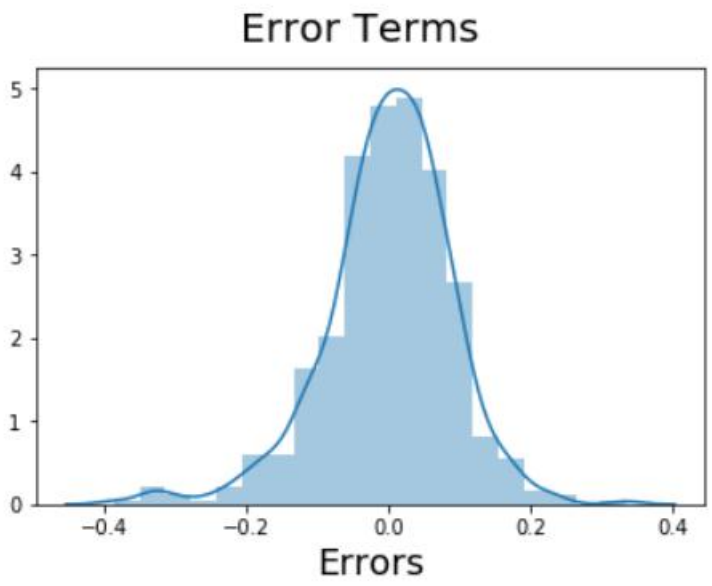
Ans : 'temp' feature, which is the temperature feature for the data is having the highest correlation with the target variable

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

Ans : Getting the linear equation of form y = m1.x1 + m2.x2 + m3.x3 + … + mn.xn + c in the final dataset.

Also getting the Normally Distribution Curve for Error Terms.

```
Text(0.5, 0, 'Errors')
```



No multicollinearity should be their which can be checked by the respective VIF values, which should be less than 5 for the respective features.

	Features	VIF
0	const	44.99
2	hum	1.86
9	weathersit_Cloudy	1.55
1	temp	1.27
5	season_winter	1.25
10	weathersit_Light_Snow_Rain	1.23
4	season_summer	1.19
3	windspeed	1.18
8	mnth_September	1.11
7	yr_1	1.03
6	holiday_1	1.02

From these ways and inferences we can validate the assumptions of Linear Regression.

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

Ans : 'temp', 'yr_1' and 'mnth_September' i.e.Temperature, Year-2019 and September month.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans : *Linear Regression Algorithm is a machine learning algorithm which is based on supervised learning. Linear regression is a part of regression analysis. Regression analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.*

Regression analysis is used for three types of applications:

1. *Finding out the effect of one or more Input variables on Target variable.*
2. *To find out upcoming trends.*

It is one of the basic forms of machine learning where we train a model for prediction of the behaviour of your data based on some variables. As the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b are the y-intercept of the line and the slope of the line respectively.

eg. Predicting the price of the house for a particular area with the given factors like area_measurements, No. Of rooms, on_the_main_road and so on with other factors.

2. Explain the Anscombe's quartet in detail.

Ans : ***Anscombe's quartet*** *consists of four datasets which have nearly identical simple statistical properties, yet they appear very different when they are graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.*

Simple understanding:

Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points.

Those 4 sets of 11 data-points are given below.

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

3. What is Pearson's R?

Ans : Correlation is a way or a method for finding out the relationship between two quantitative, continuous variables, for example, age and blood pressure, height and weight of a person etc.

Hence, we can say that Pearson's correlation coefficient (r) is a **measure of the strength of the association** between the two variables.

The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear. For correlation only purposes, it does not really matter on which axis the variables are plotted. However, conventionally, the independent (or explanatory) variable is plotted on the x-axis (horizontally) and the dependent (or response) variable is plotted on the y-axis (vertically).

We can also check the correlation between the variable using the heatmaps graphs as well.

Values of the correlation coefficient lies in the interval of -1 to +1

* Positive correlation between two variables tells that the variables are directly proportional to each other.

* Negative correlation between the two variables tells that one variable decreases on increasing the other variable and vice-versa i.e. they are inversely proportional to each other.

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

Ans : *Feature Scaling is a method of standardizing the independent features present in the data in a fixed range. It is generally performed during the data pre-processing technique to handle highly varying values or units.*

If feature scaling technique is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values, which is the reason this scaling technique is performed.

Difference in the Normalization and Standardization scaling is :

- **Min-Max Normalization:** *This technique re-scales a feature or observation value with distribution value between 0 and 1.*

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **Standardization:** *It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.*

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

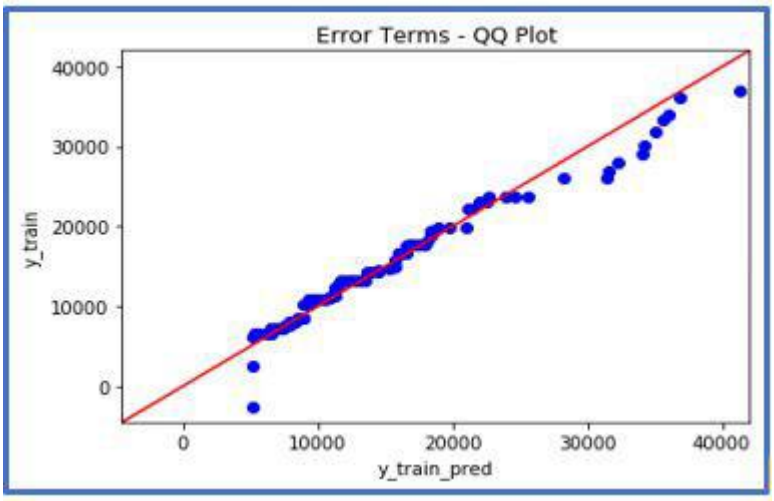
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans : *First of all, VIF stands for Variance Inflation Factor. While performing regression analysis, VIF assesses whether factors are correlated to each other (multicollinearity), which could affect p-values and the model isn't going to be as reliable.*

An ***infinite VIF*** *value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an* ***infinite VIF*** *as well). In other words we can say that we can remove that variable whose VIF value is infinite as, that can be expressed in the same way with with other variable's linear combination as well.*

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

Ans:



Quantile-Quantile (Q-Q) plot, is one of the graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or could be Uniform distribution as well. Also, it helps to determine if two data sets come from populations with a common distribution.

This usually helps in a situation of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Advantages of Q-Q plot :

* *It can be used with sample sizes also.*
* *Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.*