**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?

There were 4 categorical variables namely, ‘yr’, ‘holiday’, ‘weekday’ and ‘workingday’. Looking at the data, we can see that only ‘yr’ has a good relation with the dependent variable. This is also true logically as the business itself says that over the years, the company has been doing well over the years and are gaining popularity and demand year on year

1. Why is it important to use drop\_first=True during dummy variable creation?

For instance, let us say we have a variable gender, which has Male, Female and Others. We know that one of the values here can be inferred if we have information about the other two. Therefore, we can use ‘drop\_firs=True’ to make sure that out of the 3 dummy variables created, we can drop any one because only 2 of those variables are enough to represent the gender. If both these values are 0, it means it is the left over gender

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

‘temp’ and ‘atemp’ are the numerical variables that have the highest correlation with the target variable

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

Calculated the residuals using the formula, ytrue – ypred for the training set and looked at the histogram to realize that it was normally distributed

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

‘Holiday’, ‘weathersit\_light\_snow’(dummy variable) and ‘mnth\_mar’(dummy variable) were the most significant variables

**General Subjective Questions**

1. Explain the linear regression algorithm in detail

Linear regression is a machine learning and statistical model which is used to predict continuous variables. For eg, price of a house, price elasticity of demand, roi of marketing spends etc.

It is represented by the equation, y = mx + c where m is the slope and c is the intercept

The algorithm fits a best fit line to predict the unknown. There can be multiple fit lines in a regression model. But the line where the sum of errors is minimum, is the best fit line to choose

Linear regression uses an optimization algorithm called Gradient Descent that helps in find the best fit line.

Regression uses R2 (rsquared), which is a statistical measure that represents the proportion of variance explained by the independent variables to see if the model if doing good or not

Some of the performance metrics that can be used to check if the predictions are correct or not are mean squared error, root means squared error, mean absolute error, mean absolute percentage errors etc

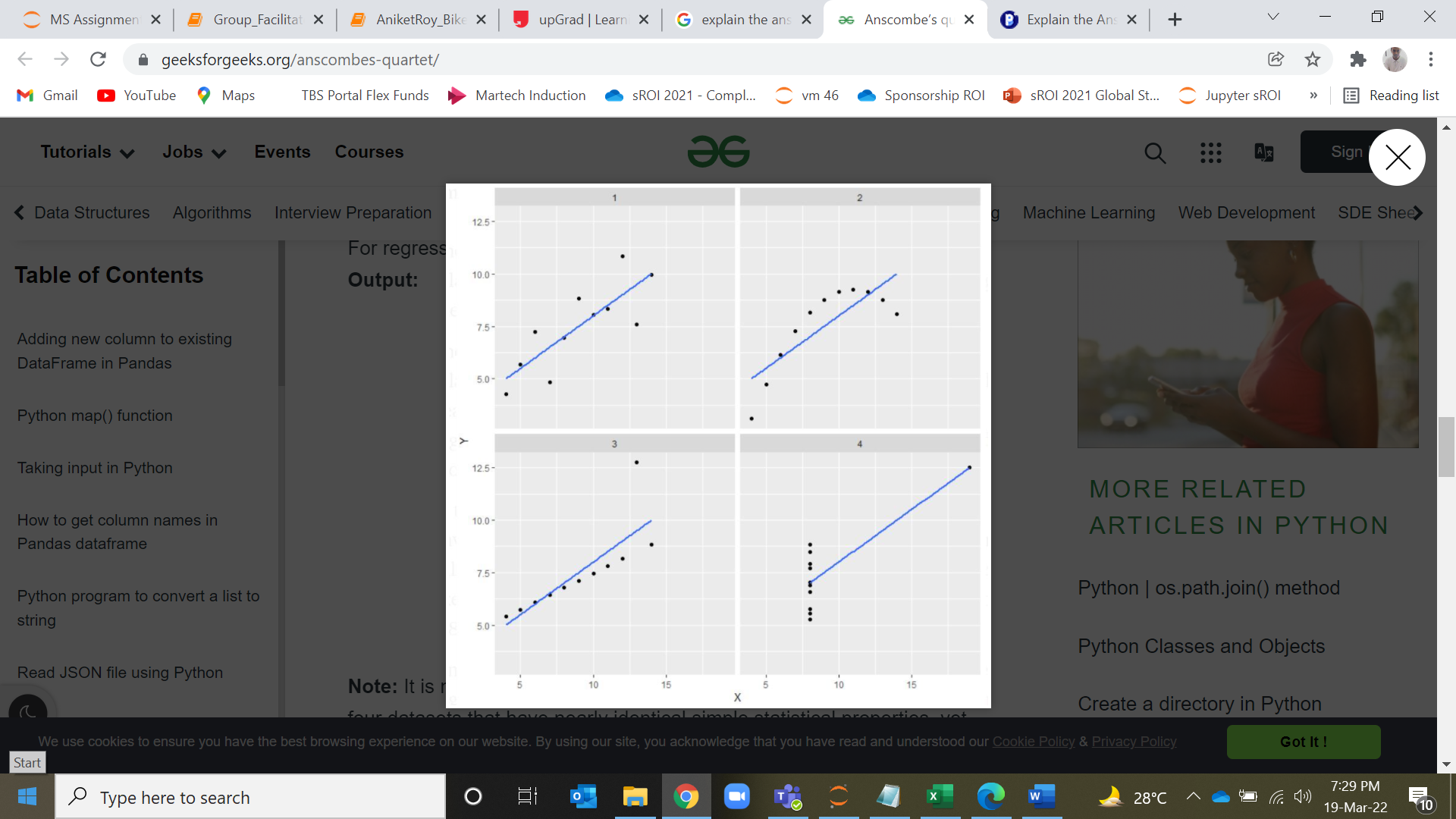
Some of the assumptions of a linear regression are,

1. Little or no multicollinearity amongst the independent variables
2. Residuals must be normally distributed
3. Homoscedasticity must not exist
4. Independent variables must have a linear relationship with the dependent variable
5. Explain the Anscombe’s quartet in detail

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed

* In the first one(top left) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
* In the second one(top right) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
* In the third one(bottom left) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
* Finally, the fourth one(bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient



1. What is Pearson’s R?

It helps us understand the strength and direction of two numerical variables. Its values ranges from -1 to +1. It is given by the formula, Cov(x,y) / (sd of x \* sd of y)

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

Each feature in the data has 2 main components, unit – how it is measured and magnitude – value of the feature

When we use algorithms that uses distances or gradient descent, we must normalize the data. When the distance between 2 points is very high due to different magnitude, then the algorithm takes a lot of time to calculate the distance. So, we must normalize. We know that linear regression using gradient descent to get the best fit line. We must normalize the data as it may not reach global minima if magnitude are high. Another reason to use scaling is that the feature may have different units. One may have kg and the other may be cm and so on. To make all these features into one unit of measure, we can use feature scaling

Normalization scales down between range 0 and 1. Standardization scales down such that its mean is 0 and standard deviation is 1

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

Formula for VIF is / (1-R2). Infinity can happen only when R2 is 1, which means the variable is perfectly correlated or perfectly explains the other variable. So, VIF may be infinity

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

They plot the quantiles of a sample distribution against quantiles of a theoretical distribution Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential

We regularly make the assumption of normality in our distribution as we perform statistical analysis and build predictive models. Machine learning algorithms like linear regression and logistic regression perform better where numerical features and targets follow a Gaussian or a uniform distribution.

It’s an important assumption as normal distribution allows us to use the empirical rule of 68 – 95 – 99.7 and analysis where we can predict the percentage of values and how far they will fall from the mean.

In regression models, normality gains significance when it comes to error terms. You want the mean of the error terms to be zero. If the mean of error terms is significantly away from zero, it means that the features we have selected may not actually be having a significant impact on the outcome variable.