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?

- *The demand of bike is less in the month of spring when compared with other seasons*
- *The demand bike increased in the year 2019 when compared with year 2018.*
- *Month Jun to Sep is the period when bike demand is high. The Month Jan is the lowest demand month.*
- *Bike demand is less in holidays in comparison to not being holiday.*
- *The demand of bike is almost similar throughout the weekdays.*
- *There is no significant change in bike demand with working day and non-working day.*
- *The bike demand is high when weather is clear and Few clouds*
- *We don't have any data for Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog*

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

- *drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.*
- *Suppose, we have 5 unique values in a column called "Fav_genre"- "Rock", "Hip hop", "Pop", "Metal", "Country" This column contains value While dummy variable creation, we usually generate 5 columns. In this case, drop_first=True is not applicable. A person may have more than one favorite genres. So dropping any of the columns would not be right. Hence, drop_first=False is the default parameter.*

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables. In our assignment we have done the drop_first = 1 as there is only one value In column and we just need n-1 columns

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

From the above pairplot we could observe that, temp (Continuous variable) and also atemp (Continuous variable) has highest positive correlation with target variable cnt.

For the corr or heat map if you would have plot then you can see that
Instant/Registered are have high corelation which is expected which is a sum of the cnt
Yr categorical has the strong correlation as well for categorical variable

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

1. *The distribution plot of error term shows the normal distribution with mean at Zero.*
2. *Predicted vs Actual are independent or randon*

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

1. *Temperature (0.539) (positive impact)*

2.Weathersit_3(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) (-0.298) [negative impact]

3. year (0.231) positive impact)

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables *getting used. There are many names for a regression's dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.*

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model. Hypothesis function for Linear Regression : While training the model we are given : x: input training data (univariate – one input variable(parameter)) y: labels to data (Supervised learning) When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values. θ1: intercept θ2: coefficient of x Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

There are mainly 2 types of regression Single Linear regressing when the X is 1 feature and more than 1 is Multi linear regression

There are multiple way to find the regression in python like using statmodel and sklearn and even neural networks

1. *Linear Regression*

2. *Decision Tree*

3. *Support Vector Regression (Kernel Linear)*

4. *Lasso Regression*

5. *Rigid Regression*

6. *Random Forest*

Etc..

2. Explain the Anscombe’s quartet in detail.

Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analyzing and model building, and the effect of other observations on statistical properties.There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets

3.What is Pearson’s R?

In statistics, the Pearson correlation coefficient (PCC, pronounced /ˈpɪərsən/) – also known as Pearson's r, the Pearson *product-moment correlation coefficient (PPMCC), the bivariate correlation,[1] or colloquially simply as the correlation coefficient[2] — is a measure of linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. As with covariance itself, the measure can only reflect a linear correlation of variables, and ignores many other types of relationships or correlations. As a simple example, one would expect the age and height of a sample of teenagers from a high school to have a Pearson correlation coefficient significantly greater than 0, but less than 1 (as 1 would represent an unrealistically perfect correlation).*

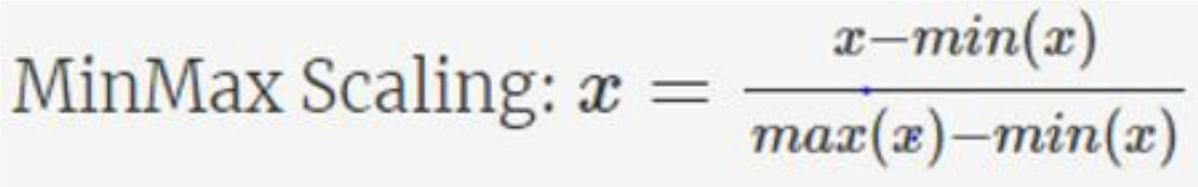
4 . What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

It is always possible that collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that ***scaling just affects the coefficients*** *and none of the other parameters like* ***t-statistic, F-statistic, p-values, R-squared****, etc.*

Normalization/Min-Max Scaling:

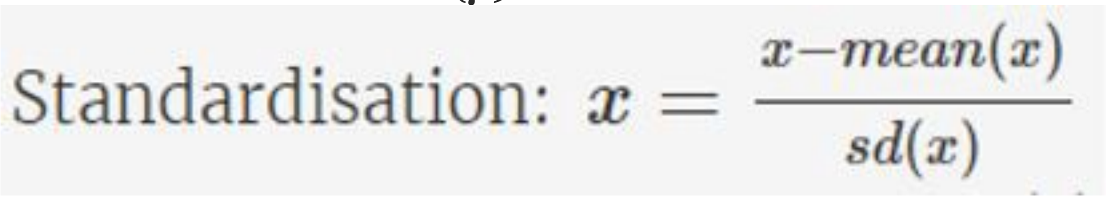
- *It brings all of the data in the range of 0 and 1.* ***sklearn.preprocessing.MinMaxScaler*** *helps to implement normalization in python.*



$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

- *Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (**μ)** zero and standard deviation one (**σ**).*



$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

- ***sklearn.preprocessing.scale** helps to implement standardization in python.*
- *One disadvantage of normalization over standardization is that it **loses** some information in the data, especially about **outliers**.*

Minmax the Value will be between 0 and 1 .

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

If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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).

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

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

This helps in a scenario 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.

Few advantages:

a) It can be used with sample sizes also *b) 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.*

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior

We can use seaborn to do a qqplot and matplotlib