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? (3 marks)

Ans:

- Dependent variable is highly influenced by the Categorical variables that is they are both positively and negatively influenced.
- In the proposed model, variables like months and seasons and also year 2019 have positive coefficient in the model equation, which implies that those are linearly related with the dependent variable (count) which aided the business of the company.
- Fall season have highest demand for rental bikes whereas spring has the lowest and also there is a gradual increase in demand till the month of June and September has the highest and then the demand decreases.
- Not all the categorical variable shows a positive relation with the target variable. Some independent variable like light snow, misty & cloudy climate is negatively impacted the business of the company.

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

Ans: It is important because it helps in reducing extra columns while creating dummy variables that is only p-1 variables are needed to define p variables and also helps to eliminate redundant variables.

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

Ans: temp and atemp has the highest correlation with the target variable(count) after removing irrelevant columns.

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

- Plotting the error function, from the graph it's clear that the error function is normally distributed with mean zero.
- Actual and predicted result shows same pattern.
- Since the error terms does not possess any certain pattern, it is independent of each other.
- Error terms have constant variance across predictions, implies that it holds homoscedasticity.

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

The top 3 features contributing significantly towards explaining the demand of the shared bikes are:

1. Temperature: positively correlated
2. Year (2019) : positively correlated

3. Weathersit-3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) : negatively correlated.

General Subjective Questions

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

Ans: Linear regression can be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables.

Machines uses its previous experience and data for predicting the future. One of the sophisticated way of doing so is called LINEAR REGRESSION. It is based on supervised learning (learning from the previous data) method and the target variable is called supervisor.

Based on the number of independent variables it is classified into 2:

1. Simple linear regression (only one independent variable) The general equation is given by Y= B0 + B1X Where B0= Y intercept B1= coefficient of X

2. Multiple variable linear regression (more than one independent variable) The general equation is given by Y= B0 + B1X1+B2X2+B3X3……..BnXn

B1,B2,B3….Bn are coeffect of X's

Steps:

- EDA , Analysing the data clean the data
- Creating dummy variables for the categorical columns
- Split the given data into test and train data , generally in the ratio of 7:3/8:2
- Building the model using statsmodel / sklearn
- Check for the multicollinearity and Pvalue
- If the model has high VIF and Pvalue drop columns having high correlation, again rebuild the model
- Stop where desirable accuracy is obtained (check for r2 values).
- Residual analysis of data. check whether the model holds the general assumptions of linear regression.
- Test the data using the model.
- If it holds the accuracy.
- Make predictions using the mode.

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

Ans: This is constructed by Francis Anscombe in 1973 to counter the famous belief that numerical calculations are exact and fair whereas the graphical representations are rough . This model consists of 4 data sets that are almost identical in calculations but have a different distribution when it is plotted. Each dataset consisted of 11 (x,y) points. . By this he announced that before applying algorithms for training models, we should visualize them and understand the distribution of data and find out the outliers etc This is still used to visualise the data graphically before analysing it.

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

Pearson's R is a correlation coefficient is a statistic that measures the linear correlation between two variables. It is the covariance of two variables, divided by 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.. It gives the details about the nature of relations as well as the intensity of their relationship too.

The coefficient values can be ranged from -1 to +1

Where :

▪ +1 : prefect positive relation

▪ -1 : prefect negative relation

▪ 0 : no relationship at all

There are certain requirements for Pearson's Correlation Coefficient:

- Scale of measurement should be interval or ratio
- Variables should be approximately normally distributed
- The association should be linear
- There should be no outliers in the data

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

Ans: Scaling is the process converting the independent features present in the data to a fixed range.

This is used to:

- handle the highly varying values. In the case of machine learning if the scaling is not done, the machine will consider the higher value as higher and the lower value as small, irrespective of the unit.
  eg. In some cases the distance is given as 2 km and 100 m , but the algorithm doesn't able to differentiate the units, so will make a huge impact on output prediction.
- Apart from this scaling is also helps to deals with slow or unstable learning process.
- Ease of interpretation.
- Faster convergence for gradient descent method.

Normalisation and standardisation are the most commonly followed approaches for scaling.

Normalisation is the scaling technique in which the values are rescaled between 0 and 1. while the standardisation helps to redistribute the values around the mean with a unit standard deviation.

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

Ans: VIF is a parameter used to measure the collinearity between the independent variables in the dataset. High VIF means that the variables are highly depended with one another while infinite value of VIF indicate that the correlation between the variables is ‘1’ , ie there is a perfect correlation.

This can be happened due to the exact replication of columns or due to some derived variables.

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

Q-Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.