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)

Spring, lightsnow contributes to less demand.

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

When there are P values then P-1 dummy variables can explain the feature well.

For example if there are 3 possible values like rainy, spring, summer if value is not rainy, spring, it reflects summer.

So dropFirst would drop one of the dummy column without which the data could stil be interpreted correctly.

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

Temperature has the highest correlation with the target variable cnt.

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

Multicolinearity: Made sure that VIF is less than 5.

Homoscedasticity: variance of residuals is within constant variance at every level.

Error terms are normally distributed with test & train data.

Linear dependency of target variable on few variables like temp.

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

Temp, year, light snow are the top 3 factors.

General Subjective Questions

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

The term regression means fitting a curve against a known set of data points to predict a target variable when one or more known predictor variables are available.

Linear regression is the process of fitting a line on a known set of predictor variables to predict a target variable which can explain the good variance in the data and with minimized error terms.

A linear equation is of the form:

y = B0 + B1X1 + B2X2 +…… + BiXi +…..BnXn

where y is the continuous target variable.

X1, X2…Xn are the known predictor variables which have linear dependency on y.

B0 is the y intercept.

Coefficient Bi can be defined as the number of units of increase in y a unit increase in Xi contributes to when all other predictor variables are kept constant. But in the real time there exists multi collinearity hence minimizing multi collinearity is an important factor.

The following are the assumptions made by Linear regression:

i. Relationship b/w predictor & target variable is linear.

ii. Error terms are normally distributed.

iii. Error terms are independent of each other.

iv. Homoscedasticity: variance of residuals is within constant variance at every level.

v. Multi collinearity: there exists no/least dependency of predictor variables on other predictor variables.

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

Anscombe have created 4 data sets with nearly identitical simple statistics yet appears to be completely different when graphed.

This experiment shows the effect of outliers on the statistics of data and also importance of graphing the data.

In his experiment, all 4 data sets have given a similar model of straight line but which wasn't the case when verified graphically.

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

Corelation is a measure of how two variables associated to each other.

Pearsons R is a bivariate corelation coefficient to measure linear corelation between two features.

Value is calculated as follows:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2- (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

The values are in range of [-1,1] these values indicate the direction and strength of corelation.

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

Some variables have very low range in their values, for few categorical values, it can be low as 0, 1 only. And some continuous variables have values as high as millions. So it becomes a problem for the model to estimate various variables at different scales. Hence scaling is performed to represent entire data at a same scale.

Normalization/ Min max scaling is used to transform variables onto a similar scale.

It is calculated as

Xnew = (X – Xmin)/(Xmax – Xmin)

Geometric definition can be: N dimensional data is being squished to N dimensional unit hyper cube.

Value ranges can be [-1,1] or [0,1]

Standardization/ Z score normalization is obtained by subtracting mean and diviting by standard deviation.

Xnew = (X – mean)/std

Geometrically, it transforms origin vector to mean data and squish/expand points based on std.

It is used when feature distribution is normal/ gaussian. It is least impacted by outliers.

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

Variance inflation factor identifies to what extent preditor variables are corelated to each other (multi collinearity measure)

VIF = 1/(1 – R2)

R2 values are within a range of [0,1] so when R2 is ~1 denominator would be close to ZERO which would result in an infinite VIF

Which inturn means that this variable can be explained perfectly by other predictor variables. Hence this can be dropped.

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

Quantile-quantile plot is a graphical method for comparing two probability distributions. It is a plot of quantiles of two distributions against each other.

For a uniformly distributed data, q-q plot would be perfect straight line through many points.

If we are comparing 2 different sample groups, it explains how one is different to other.