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)

Most high categorical variables seen is year, season and weather situation. In later year the count increased by a double factor. Season spring is with negative coefficient, determining, opposite season - fall to be favourable. Weather situation mostly clear to be more favourable. Comparative higher count observed for non-holiday days.

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

Drop_first true will remove the additional column, nth level in a column. The reason being, remaining n-1 columns already covers the value of the feature. Thus, eliminating multicollinearity.

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

temp

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

Created a histogram on the error terms which showed uniform distribution and model scatter plot showed constant variance – homoscedasticity. Also, plotted a line graph showing match in actual and predicted target variable.

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

Temperature, year and season

General Subjective Questions

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

Linear regression algorithm by using ordinary least square method we form calculated R-square value, which is minimized value of sum of square of error terms and total sum of square errors. Here, error terms are also called residual, difference in actual data point and predicted data point. Formula is given as –

RSS = Sum of (Actual point – Predicted point)^2

TSS = Sum of (Actual point – Average actual point)^2

Rsquare = 1 – RSS/TSS

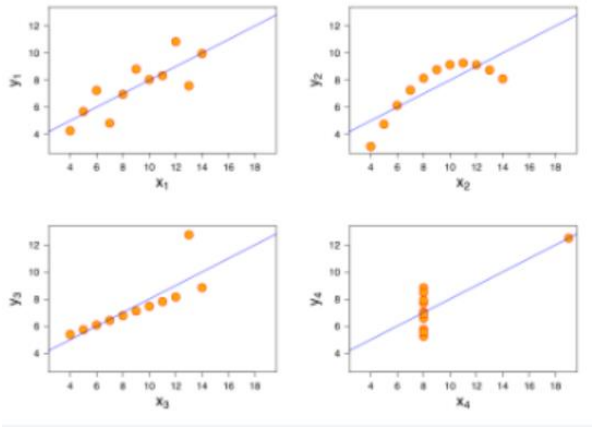
R-squared gives the significant increase value and Adjusted R-squared will apply penalty with increase in variables and show significant value.

Next the best line fit is predicted on a best selected model having high R-squared. The line equation is as shown below. Above cost function finds best value for all beta coefficients -

Y = Bo + B1X1+ B2X2+… BnXn

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

Linear regression has few shortcomings which can be best explained by using below phenomena, Anscombe's quarter -



The first graph shows good linear relationship in two variables, x and y. Whereas, second one doesn't have a good linear relation. Third graph shows a presence of outlier, distorting, linear strength. Last one shows a data point from rest with very high value producing high coefficient.

Thus, quartet shows the importance of visualizing dataset in graph before analysing and shows the limitation of basic statistics properties.

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

The linear relationship between features in statistics is Pearson's R, also known as, Pearson correlation. This correlation is used to measure the strength and direction of linear correlation.

The Pearson correlation coefficient value can be either positive or negative between [-1,1] . Also by squaring the r value will result in R-squared, which is coefficient of determination indicating amount of variance. If positive value, the affect will be increasing for both variables, if it is negative, then one variable increases, the other variable will decrease.

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

With scaling a feature's data range is transformed. Scaling is performed to bring different data ranges into a common scale. Given an example of a salary in currency feature like INR, USD or

EURO the direct comparison would give incorrect result for checking highest salary. In such a case a common measurement would be required that treats all in same scale.

Normalized scaling is the Min-Max Scaler which transforms the data into [0,1] or [-1,1] range, such that, Xnew = (X – Xmin)/(Xmax – Xmin) for a given independent variable feature.

Standardized scaling transforms the data such that the resulting mean is 0 and standard deviation is 1. It forms the result by subtracting mean and dividing by standard deviation, also called as Z-score - Xnew = (X – mean) /std. dev

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

VIF is variance inflation factor, in case of perfect collinearity between independent variable the value of VIF becomes infinity. The R-squared value becomes 1, thus VIF value, which is 1/(1-Rsquared) will come to infinity.

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

Q-Q plot or quantile-quantile plot is used to match probability distribution graph by using quantile, which, is a standard distribute variate on x-axis and random variable on Y-axis. This plot helps to find the type of distribution for the random variable, the value to find.

Importance of using Q-Q plot in linear regression can be useful to validate normal distribution of residuals. For example, it can be used to check if two population dataset, have a common distribution. However, as the sample size increases, n > 30, the reason to check normality of residual is not required, as it follows Central Limit theorem.