Assignment-based Subjective Questions

Question1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer: Following are the inferences about the effect of categorical variables on the dependent variable:

1. There is huge bike demand especially in the fall season followed by summer as compared to winter and spring
2. Bike demand has shown a positive jump from last year.
3. on holidays, bike demand is low
4. bike demands is quite same over the weekdays, doesnt show any significant trend.
5. whether its working day or not, it has no effect on bike demand
6. bike demand is high during the mid of the year from may to oct.

Question2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer: Let's say we have a dummy variable with N levels, then that dummy variable can be explained by N-1 level as well. For a variable let's say season with four possible values (level) spring, summer, fall and winter, can be explained with (N-1) level that is spring, summer, winter (excluding the fall). If spring = 0, summer = 0, winter = 0, then naturally the seaon is fall. Hence its important to get rid of that redundant variable using drop_first= True.

Question3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: Looking at the pair-plot among the numerical variables, temp and atemp have the highest correlation (0.63) with the target variable (cnt).

Question4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer:

- Residual Analysis: Errors are normally distributed with a mean of 0. Actual and predicted results follow the same pattern. The error terms are independent of each other.
- Plot Test vs Predicted value test: The prediction for test data is very close to actuals.
- R2 value for test predictions: R2 score for predictions on test data (0.803) is close or similar to R2 score of train dataset (0.79). Which is good R score values, hence we can say our model is performing good on test data.
- Homoscedacity: We can observe that variance of the residuals (error terms) is constant across predictions. i.e., error term does not vary much as the value of the predictor variable changes.

Question5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: The top three features are:

1. yr (year) : having positive correlation with cnt
2. weathersit_bad: having noticeable negative relation with cnt
3. temp: having high positive correlation with cnt

General Subjective Questions

Question1. Explain the linear regression algorithm in detail. (4 marks)

Answer: Linear Regression is a machine learning algorithm based on supervised learning. Regression models a dependent variable (y) value based on independent variable (x). Linear Regression shows linear relationship between independent variable and dependent variable. If there is a single independent variable to consider then such linear regression is called Simple Linear Regression Model. If there are more than one independent variable to consider then that linear regression model is called Multiple Linear Regression Model. In Linear Regression the dependent variable (y) is called target or criterion variable and independent variable (x) is called predictor or feature variables.

The Linear Regression model can be represented by the following equation, which is the equation of the best fit regression line.

Y = a0 + a1*x1 + a2*x2

y = dependent or target variable

a1 and a2 = linear regression coefficent

a0 = intercept the linear regression line

x1 and x2 = independent or predictor variables

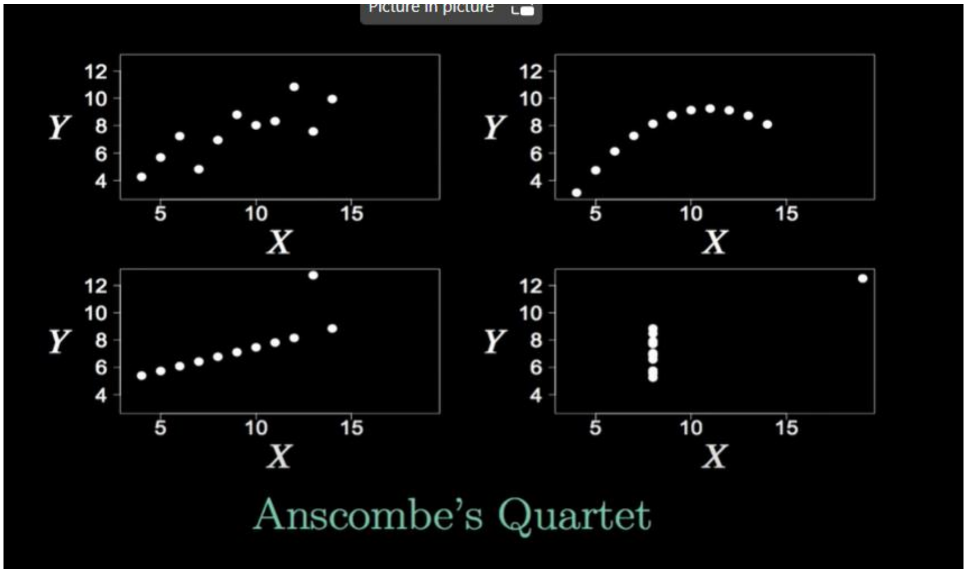
This equation can be found by minimising the the cost function, RSS (Residual sum of square) in case using Ordinary Least Square method.

Moreover, the strength of Linear Regression is mainly explained by R2 where R2 = 1 - (RSS/TSS). Where RSS is Residual Sum of Squares and TSS is Total Sum of Squares.

Linear Regression model is used in many fields for example, economics, medical, phsychology etc.

Question2. Explain the Anscombe’s quartet in detail. (3 marks)

Answer: Anscombe’s quartet is a group of four dataset which are similar in terms of simple statistics like best fit line, mean, variance etc. But shows very different nature or spread when visualized or plotted. The four datasets are as follows.



All these datasets show or follow the same best fit regression line and have the same statistics. It is always important to visualize or plot the data before making a model out of it. As it tells different story if compared with the model built on this kind of dataset. Also, the Linear Regression is only considered fit for the data with linear relationships and is incapable of handling any other kind of datasets.

Question3. What is Pearson’s R? (3 marks)

Answer: Pearson’s R which is also referred to as Pearson Correlation Coefficient or Person product-moment correlation coefficient (PPMCC) or bivariate correlation. It is the measure of linear correlation between two sets of data. It gives the sense of direction and strength between the relation of two sets of data.

Basically it is the covariance of two variable divided by the product of their standard deviation. It is a normalised measurement of covariance such that the result always has a value between –1 and 1

It is given by the formula:

Pearson’s R or P(x,y) = cov(x,y)/sd(x)*sd(y)

Where cov(x,y) is the covariance of x and y sd(x) = standard deviation of x

Sd(y) = standard deviation of y

Quaestion4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?   (3 marks)

Answer: During the pre-preprocessing of the data, scaling is an important step before we move to modelling. It is a technique which is applied on independent variable so that they have common scale. In other words, it is a way to normalize the data in a particular range for the independent variable.

When we collect the data, it is common that data has values with different magnitude, units etc. If the scaling is not performed, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to other coefficients. Which will cause trouble or will be annoying during model evaluation. So, it is advised to use standardization or normalization so that the units of coefficients obtained are all on the same scale.

Normalized Scaling (MinMax scaling) : brings all the data withing a range of 0 and 1. It is given by

x = (x-min(x)) / (max(x) - min(x))

Standardize Scaling: brings all the data into standard normal distribution of with mean of 0 and standard deviation of 1. It is given by

x = (x – mean(x)) / sd(x)

One disadvantage that normalization has over standardization is that it loses some information in the data about the outliers especially.

Question5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks)

Answer: VIF is infinite, when there is perfect correlation. As we know that VIF is given by the following formula

VIF = 1 / (1 – R2) ; where R2 is the correlation between two variables under consideration.

Now if there exists a perfection correlation between the two variable then R2 will be 1 in that case. Now if calculate VIF with R2 = 1. Then this will lead to 1/0, which is nothing but infinite value or infinity.

To solve this problem, we drop one those feature variable which is causing such perfect collinearity. An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

Question6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)

Answer: Quantile – Quantile plot (Q-Q) are plots of two quantiles against each other. A quantile is a fraction where a certain value falls below that quantile. It is a graphical tool that helps to determine if two sets of data possibly came from the same theoretical distribution such as uniform or normal distribution. In another words it helps to assess if two data sets came from population with common distribution.

This helps in cases of Linear Regression when we have training and test dataset received separately and we want to know if the two datasets are from population with same distribution.

If two distributions being compared are similar, then points in the Q-Q plot will approximately lie on the line y=x. If distributions are linearly related, then points in the Q-Q plot will approximately lie on a line but not necessarily y=x.