Subjective Questions and answers

Submitted by - Sudeep Singh

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?

Answer: After analysing on categorical variables using the boxplot and bar plot. We can infer these points-

- Fall season has more bookings and spring season has least.
- Aug, June and September months attracts more and jan has least bookings.
- There is no significant variation in bookings in respect of weekdays.
- Clear weather has the maximum bookings and Light_snowrain has least.
- Year 2019 has more bookings than 2018
- Booking seemed to be almost equal either on working day or non-working day.

2) Why is it important to use drop first=True during dummy variable creation?

Answer : drop_first = True helps in reducing the extra column created during dummy variable creation. It saves space for one column. As we know if we have x different values in any categorical variable we can represent those with x-1 dummy variable.

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

Answer: The variable 'temp' has highest correlation with the target variable

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

Answer: The validations are done based on these assumptions-

- Error terms are normally distributed.
- In the final model, there are no significant multicollinearity among variables
- There are no visible patterns found in residual analysis.
- No auto-correlation

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

Answer: These 3 features contributing significantly towards explaining the demand of bikes-

- Temp
- Winter
- September

Note: "year" is also contributing, but there is only 2 years so we can't take it as significant contribution.

General Subjective Questions

1) Explain the linear regression algorithm in detail.

Answer : It is the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship means once a variable increase/decrease the dependent variable will also increase or decrease.

It can we represent with an algebraic equation as - $Y = mX + c$

- Y - is the dependent variable that we are going to predict and X is independent variables that will be use to make predictions.
- m - is the slope.
- c - Y-intercept, a constant value.

The linear relationship can be positive or negative in nature−

- **Positive Linear Relationship:** If dependent variable increases as independent variable increases.
- **Negative Linear relationship:** If dependent variable decreases as independent variable increases.

Linear regression is of two types −

- **Simple Linear Regression -** Based on single independent variable.
- **Multiple Linear Regression -** Based on multiple independent variables.

Assumptions - Using the linear regression, we verify these assumptions on the dataset.

- **Multi-collinearity –** Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
- **Auto-correlation –** Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- **Relationship between variables –** Linear regression model assumes that the relationship between response and feature variables must be linear.
- **Normality of error terms –** Error terms should be normally distributed
- **Homoscedasticity –** There should be no visible pattern in residual values.

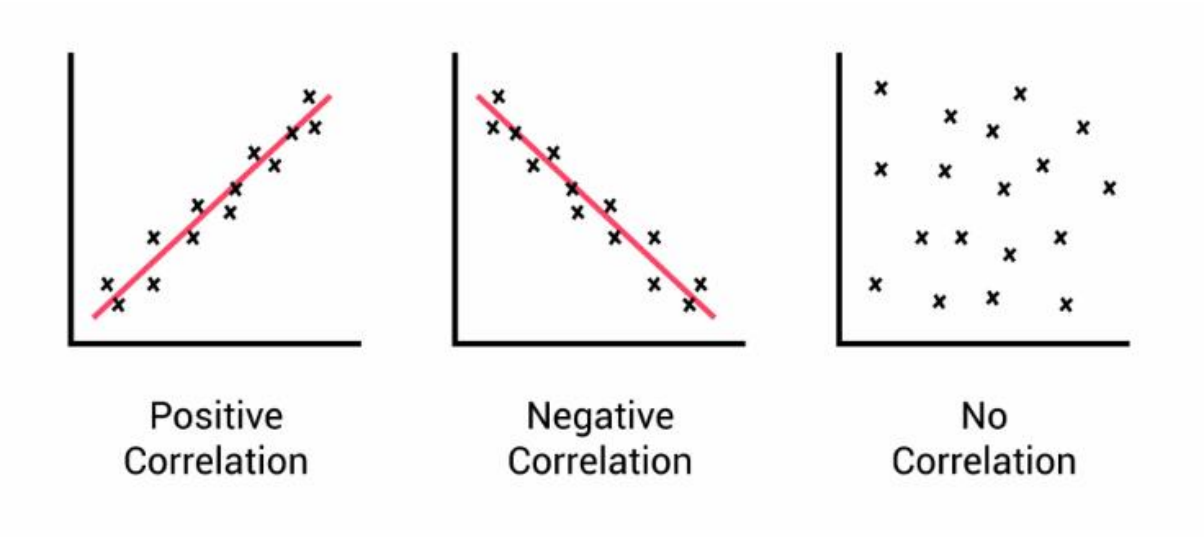
2) Explain the Anscombe's quartet in detail.

Answer : 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.

3) What is Pearson's R?

Answer : Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Answer :

Scaling : It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why is it performed : Most of the times, 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.

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

Standardized 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). sklearn.preprocessing.scale helps to implement standardization in python.

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

Answer: It happens for the perfect correlation. If two independent variables shows strong correlation, the VIF tends to infinity.

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

Answer: Q-Q is an abbreviation for quantile-quantile. Q-Q plot is a graphical technique for determining if two data sets come from the same population or not.

If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

Why do we generally use Q-Q plot-

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behavior.

How to Draw Q-Q plot

- Step1: Data Collection for Q-Q plot.
- Step2: Sort the data. It can be in Ascending or descending
- Step3: Draw a normal distribution curve.
- Step4: Find the z-value for each segment.
- Step5: Plot the dataset values against the normalizing cut-off points.