Assignment-based Subjective Questions

A1)

- Observations from above boxplots for categorical variables:
- The year box plots indicates that more bikes are rent during 2019.
- The season box plots indicates that more bikes are rent during fall season.
- The working day and holiday box plots indicate that more bikes are rent during normal working days than on weekends or holidays.
- The month box plots indicates that more bikes are rent during September month.
- The weekday box plots indicates that more bikes are rent during Saturday.
- The weathersit box plots indicates that more bikes are rent during Clear, Few clouds, Partly cloudy weather.

A2) drop_first=True is important t use, it helps in reducing extra columns created during dummy variable creation. Hence it reduces correlation created among dummy variables.

A3) the highest is “cnt” with correlation of 0.63

A4) The easiest way is using the Durbin Watson test . we can conduct this test using R’s built in function.

A5) the top 3 are

- Weathersit_Light_Snow is negative correlation
- yr_2019 is positive correlation
- temp is positive correlation

General Subjective Questions

A1)

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

A2)

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

A3) It means pearson correlation ecoefficiency

A4)

- scaling is a technique to standardize the the independent features present in data in a fixed range
- Scaling is performed cause it is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range
- Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance)

A5)if there is perfect correlation then VIF shows infinity

A6)

- a q-q plot tells about determining if two sets come from population with common distbution
- Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution

against quantiles of a theoretical distribution. Doing this helps us determine if a dataset follows any particular type of probability distribution like normal, uniform, exponential