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?

Ans: Not all categorical variables had a significant impact on the outcome of the prediction, but those who had are yr, holiday, light rain, misty, summer, winter and spring. A few of these had negative correlation with cnt (target variable).

Yr -> Denoted Pre-covid usage of the bikes. Helpful since the organization is targeting improvement post-covid. More bikes used in 2019 which may be sign of people being more enlightened towards global warming and personal health.

Holiday -> This variable has a negative coefficient depicting that most / more usage of these bikes is to commute to and from work.

Light Rain & Misty -> Also had negative coefficient which is expected as these are unsafe conditions for the usage of a bike.

And contrary to these, the summer had a positive coefficient.

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

Ans: drop_first allows you to either drop the first dummy variable or keep it. In most / all of the cases all of the information that a column with dummy variable has to convey can be conveyed using k-1 dummies, where k is the total number of variables in the column.
This also means a slightly more efficient computational process.

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

Ans: In an unfiltered dataframe 'registered' shows the highest correlation with the target variable. But post filtering the variable that shows the most correlation would be temp / atemp.

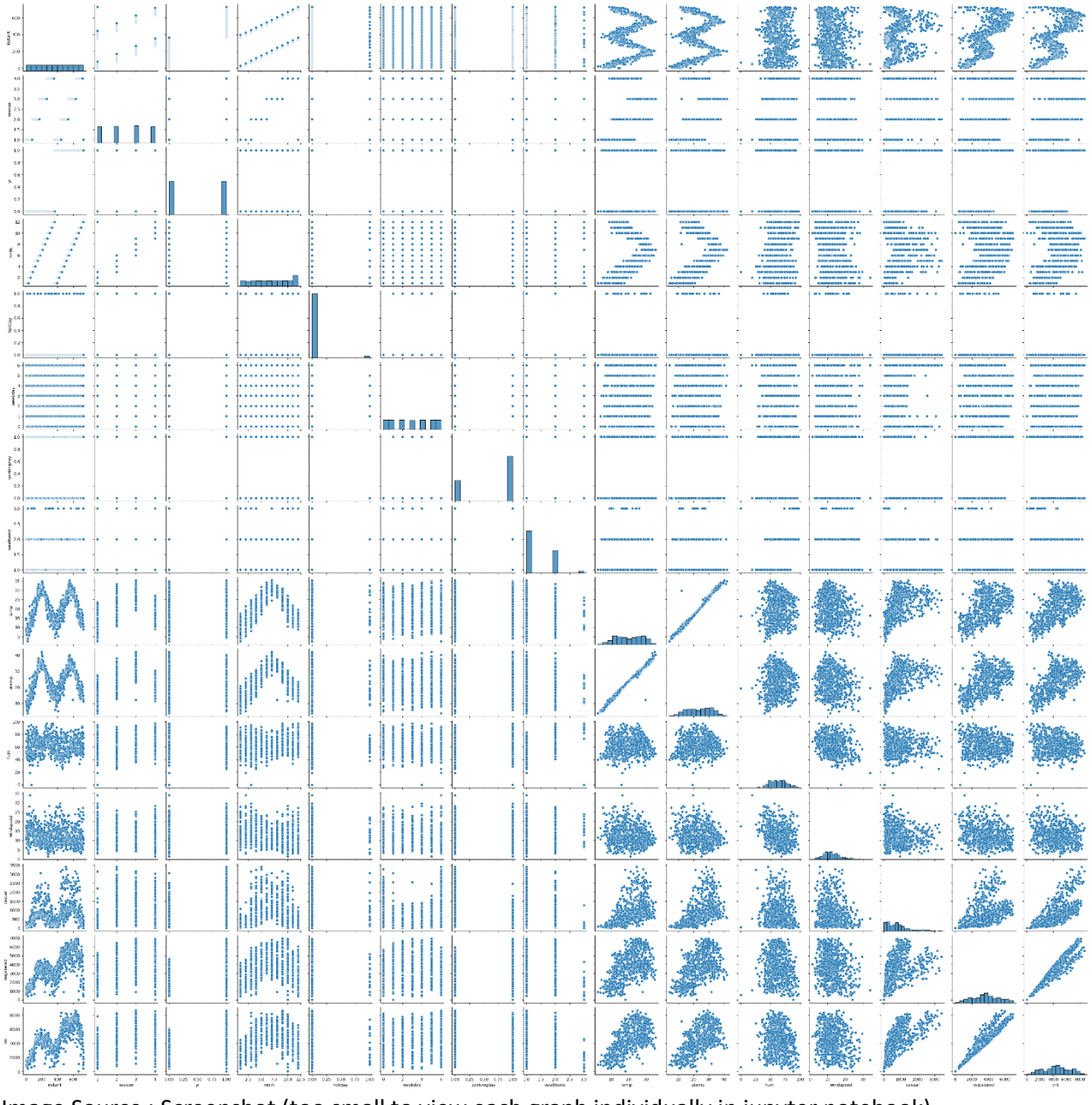


Image Source : Screenshot (too small to view each graph individually in jupyter notebook)

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

Ans: pValue, variance inflation factor. These two indicate the significance of the selected variable and the correlation of the selected variable with other variables. Along with a bit of intuitive thoughts such as -> riding a bike is not preferred in the rainy season, most of the people who use bikes use it to commute from or to work since the number of registered users is significantly higher than the casual ones.

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

Ans: Temp, Rain and the year(2018 / 2019).

Temperature has a significantly high coefficient and is a significant variable. Clearly people don't like cycling in a hot day since the coefficient is negative.

Rain, has a similar outcome as that of the temperature and it is a bit obvious as a rainy day is not the safest day to go riding on your bike.

Year, this is a confusing one. But it happens to be the variable with third highest coefficient and equally high significance. I believe this is the cause of people (population) noticing climate change whilst the fitness industry has been growing for a while so, that may add up to it as well. Or maybe even gas prices (inflation).

General Subjective Questions

Explain the linear regression algorithm in detail.

Ans: Linear Regression is the process of finding an equation that best describes the relation between the dependent and independent variable. This is achieved by fitting the line between the variables. It would treat the values of that line as the outcome (prediction) for the given set of variables it is already trained for with varying values.

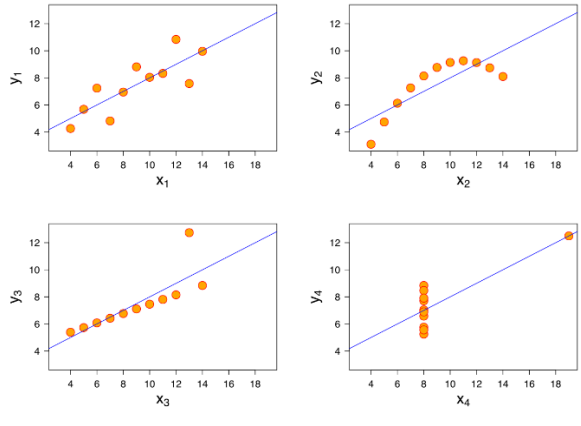
It is used in predictive modelling and has mainly two types SLR (Single Variable) and MLR(Multiple Variables).

The formula :-

Y = B0 + B1X1 + B2X2 ….. + BnXn (where n is the number of features). This for MLR forms a hyperplane while for a single variable forms a normal line.

2. Explain the Anscombe's quartet in detail.

Ans: Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven x and y points. This was done to depict the importance of graphing the data and the effect of outliers



. Image Source : Wikipedia

3. What is Pearson's R?
Ans: It is a linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations thus, it is essentially a normalized measurement of covariance. Thus, the result always has the value between -1 and 1.
One of the examples could be the altitude and room temperature -> As the altitude increases (landscape) the normal room temperature decreases.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans: Scaling in lay-man terms is used to put all the variables (continuous) into the same range. For example lets consider our bikes sharing data, the range of temp, cnt, casuals and registered are pretty different and this may lead to the model considering one of these with a higher range to be an influential variable but that might not be the case. To avoid these kind of errors we scale all these variables to the same range so the model treats them equally and hands out the actual variable with most impact on the target variable.

Normalized scaling (Scales everything down to the range of -1 to 1): If the distribution is not gaussian this method works the best. If there are outliers in the data, it would create problems by messing with the actual scaling that should've happened.

Standardization : This method rescales the features so that they behave like a gaussian distribution .Mean = 0 and std. deviation = 1. This method usually takes care of the outliers since it treats the data to be or act like a gaussian distribution.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: The formula for VIF = 1 / 1 – r^2
The value of VIF being infinite means the denominator = 0 ie. R^2 = 1
This means that the variable is highly correlated to other variables in the dataset (active / included). This could be the case while employing multicollinearity and if the variables manage to create perfect multiple regression on other variables. This could also be an outcome of the case when we include two binary variables for the same feature for example holiday and working day (possibility). But when working day is further elaborated in to monday, Tuesday, etc it should be fine.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression
Ans: Quantile – Quantile plot is a probability plot that aids in the comparison of the two different probability distributions by plotting their quantiles against each other.
The major use case of a QQ plot is to identify if two data sets come from the same distribution. This is done by plotting the first data set's quantiles on the X – axis and the other data set's quantiles on the Y – axis. It also helps us determine if the data set follows any probability distribution like normal, uniform or exponential.