Assignment-based Subjective Questions

1. The categorical variables ‘season’ show higher range of riders using bikes in the fall season and lowest in spring season. It shows lower extreme of outliers in winter season showing it being the season with the least used bike sharing service. We can view the similar variance in the categorical column of month as well which corresponds with the season category. The ‘weathersit’ columns suggests that clear/misty weather is most preferable by riders. In regard to climate conditions, the temperature variable seems to have highest correlation with the number of riders. The year 2019 shows the higher range of consumers.

2. It helps in better correlation of the variables overall as the new data frame with the dummy variables can be inferred accurately without the first column as the values in the other signify the existence of the entered category for a required data point. For example, in the housing case study furnished would mean the values are 0,0 in the remaining two columns and semi-furnished being 1,0 and so on, hence there would not be a need of the first column and while predicting the linear regression line, this arrangement would be more helpful.

3. The variables ‘temp’ and ‘atemp’ have highest correlation with the target variable.

4. By first checking the R-squared, Adjusted R-squared values and the p values for the coefficients. As the Adjusted R-squared value came out to be 0.83, the model tested out to be significant. Upon further checking the VIF for every feature and dropping features with VIF higher than 5 we can go forward with repeating the model evaluation.

5. The top 3 features contributing significantly to the demand in shared bikes are temperature, holiday and windspeed.

General Subjective Questions

1. Linear regression comes under the supervised learning method of the machine learning algorithms. A target variable which is a continuous variable is being predicted through a linear regression model. Past data is used in the model building in this algorithm. An equation of a simple linear equation is used to map the relationship between a dependant and an independent variable which is also the target variable, in which the slope and intercept play an important role in the regression, this line is called the regression line. The y coordinate of actual data – y coordinate of the predicted data is calculated as residuals and the sum of these residual squares is used to create what is known as the best fit line on a plot between the dependant and target variables. The residual sum of squares and total sum of squares (which is calculated from the difference between the y coordinate of actual data to the average of the actual data) are used to calculate the R-squared value which gives the metric for how accurate the best fit line is in predicting the variance from the actual data. Ex. If the R- squared value turns out to be 0.6, it means that we are able to predict 60% of the variance in the data

accurately. Higher the value, the better prediction we have made in regards to analysing a particular data set and predicting the effect of variables in it for required results. With the above steps we build the model first, then the variance inflation factor is checked in order to reduce multicollinearity, values <5 are considered safe and the variables are not dropped. The variables with high p-value (higher than 0.05) are considered insignificant and dropped. Finally in a multiple linear regression model we use the value of Adjusted R – squared as it takes into account multiple variables and the R-squared value might change as we edit the variables based on the model's accuracy.

2. Anscombe's Quartet was created to give emphasis to visualisation of data and its analysis so that all the details such as outliers and other closely observable properties that may have an effect on the statistical process are considered. The quartet consists of four data sets with basically the same or similar statistical results overall but have completely different graphical representations, with all four data sets having exactly eleven data points plotted.

3. Pearson's r is a correlation coefficient that measures the linear correlation or association between two variables. Its value ranges from 1 to –1 showing the absolute value as a coefficient of correlation between the variables. If the value is 0 it means that there is no correlation and 1 would mean the perfect possible correlation. The higher the value the higher the relationship between the variables.

4. The numerical values of the variables are on a different scale; hence scaling is used to interpret the data properly and bring all the variables on the same scale for the model. Normalized scaling is used to bring all the variables on the same scale by bringing the range of the scale to 0 to 1 or –1 to 1. Standardized scaling is when the values are subtracted from the mean and divided by standard deviation, this is used when we want to ensure a zero mean and standard deviation 1.

5. VIF is infinite when the R-squared value is 1 i.e., the correlation between variables is perfect. There also might be perfect multicollinearity in which case a feature has to be dropped.

6. The Q-Q plot is used to plot two quantiles against each other to see if the data of the two variables come from the same distribution. In case of a linear regression problem, it helps to check if both the training and test data that is received separately come from the data sets with the same distributions.