Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

- Season has an impact on the demand. Demand is observed to be the highest in the fall closely followed by summer. Demand is observed to be significantly reduced in the Spring.
- Monthwise demand data shows the similar trend as of seasnal data which is not surprising.
- There is a significant increase in the average demand in the year of 2019 compared to 2018 indicating solid year on year growth of the company.
- Weather on the given day seems to have an impact on the demand for that day. On a clear day the demand seems to be higher than the day with any other weather condition. Demand seems to be significantly dropped on the day with a poor weather.
- Average demand seems to be dropping on holidays.
- Average demand does not seem to be affected whether it is a working day or not.
- Average demand seems to be same on any given day of the week.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

We can represent the dropped variable with the rest of the variables.
Example: variable "season" has four dummy variables. We can convey the same information usig three dummy variables as follows:
000 -> fall
100 -> spring
010 -> summer
001 -> winter

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Variables "temp" and "atemp" have the highest correlation with the target variable.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Assumption about the form of the model:

Dependant variables like temp and temp showed a fairly linear relation with the target variable and also they showed high correlation with the target variable. This confirms the 'linearity

assumption'.

Assumptions about residuals:

1] Normality assumption: Error terms were found normally distributed confirming the assumption.
2] Zero mean assumption: Mean value of the residuals was found to be zero confirming the assumption.
3] Constant variance assumption: Scatter plot was drawn between residuals (temp, windspeed and hum) and fitted values to confirm the Homoscedatsticity. Plot was random in nature and distributed horizontally confirming the assumption.
4] Independent error assumption: Residual terms are independent of each other looking at the pair plot.

Assumptions about estimators:

There is no multicollinearity in the data as VIF values of all the residuals are within 5.
Also "Cond No" value is 16.2 which is less than 30 indicating no multicollinearity.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

1] temp (Temperature)
2] hum (Humidity)
3] windspeed (Wind Speed)

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Algorithm for Linear Regression using hybrid method (RFE+Manual)
1] Read the data
2] Understand and Clean the data
3] Visualize the data by plotting box plot and bar graph of categorical variables
4] Use pair plot to visualize the relation of all variables amongst each other (useful to check the linearity of numeric variables with respect to the target variable)
5] Continue the visualization by plotting heat map of all the variables to understand the correlation of the variables amongst each other.
6] Identify the significant variable base on the data visualization.
7] Create dummy variables for all the categorical variables whose value range is greater than 2
8] Drop one dummy variable from each category variable dummy variable set
9] Add the dummy variables to the original data set
10] Split the updated data set into Train data and Test Data (70:30 ratio)
11] Scale the numerical data from the Train data set between 0 to 1 (Min-Max scaling)
12] Plot the heat map again to understand the correlation amongst the Train data set.

13] Divide the Train data set into X_train and y_train. y_train contains only target variable and X_train contains rest of the variables

14] Run the RFE to restrict the number of variables to 10.

15] Store the significant variables indicated by RFE in X_train_lm

16] Build the model using statsmodel for the detailed statistics

17] Check the p-value of each variable

18] If p-value of any one or more variables is more than 0.05 then drop one such variable and go to step 16. Otherwise go to 19.

19] Check the R2 value of the model. If it is greater than 0.8 then go to step 20. Otherwise, add a new variable or replace the least significant variable with new variable which we think useful. Go to step 16.

20] Check the VIF of each variable. If any one or more variables have VIF more than 5 then drop a variable with highest VIF and go to step 16. If all VIFs are within 5, then got to next step.

21] Run the residual analysis of the train data and check the validity of all the assumptions. If any assumption is not found to be true then go to step 16 with new data set. Otherwise, go to next step.

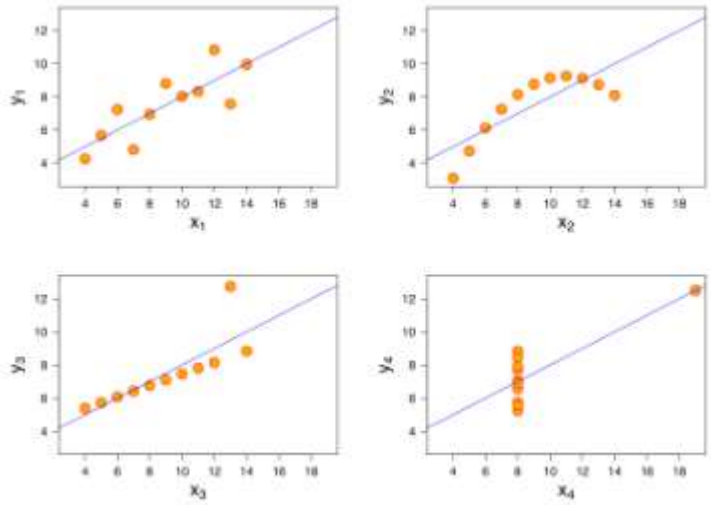
22] Run the prediction

23] Evaluate the model by ploting the scatter plot for test data and predicted data. If R2 is closer to the train data and graph is fairly linear then model can be considered as acceptable.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's quartet elaborates the limitations of simple linear regression when it has to handle outliers and non-linear data.



Linear regression works fairly well on the first dataset but unable handle the non-linear data in the second graph. Third and fourth graph shows sensitivity of the linear regression to the outliers. Model would fit well in absence of the outliers.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R is a correlation coefficient used to measure the strength of the relation between two variables. This coefficient is useful when the relationship is linear. This coefficient is not reliable when the relationship is non-linear. Hence, Pearson's R is used mainly in the linear regression model. Its range is -1 to 1. Value 0 indicates no correlation and 1 indicates strong linear relation. Direction indicates positive or negative linear relation.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling is a data pre-processing of the independent variable to normalize the data within a particular range.

Most of the times data set contains variables, which varies in magnitude, unit and range. Algorithm will correct incorrect model if scaling is not done as it will only consider magnitude. Scaling solves this problem. Scaling only affects the coefficient and no other model related parameter.

Normalized scaling maps the data between 0 and 1. Standardized scaling replaces the value with its z score. It brings all the data into standard normal distribution which has mean zero and standard deviation one.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

Infinite VIF indicates perfect collinearity between variables. For a perfect collinearity R^2 will be 1 and since VIF = 1/ (1-R^2). It will give infinite value. It means dataset contain one or more variable that can be entirely predicted by another independent variable. Identical or duplicate columns can also result infinite VIF.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or not. A Q-Q plot is a nice visual way to check for distributional assumptions. Q-Q plot is used is used in linear aggression for residual analysis. We can check whether the error distribution is normalized and centre to zero or not. This plot helps up to verify the assumptions about the residuals.