Assignment-based Subjective Questions & Answers

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Ans:

Demand for shared bikes
1. Increased from 2018 to 2019
2. More during Holiday or Weekend compared to working day
3. More during fall season and least during spring
4. Least in January, February, December and more in June, September, August, July
5. More in June 2018 and Sep 2019
6. More in Monday and Friday and least in Tuesday and Wednesday
7. More in Clear, Few clouds, Partly cloudy or Partly cloudy weather and least in Light Snow, Light Rain + Thunderstorm + Scattered clouds or Light Rain + Scattered clouds and no demand at all in Heavy Rain + Ice Pallets + Thunderstorm + Mist or Snow + Fog

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

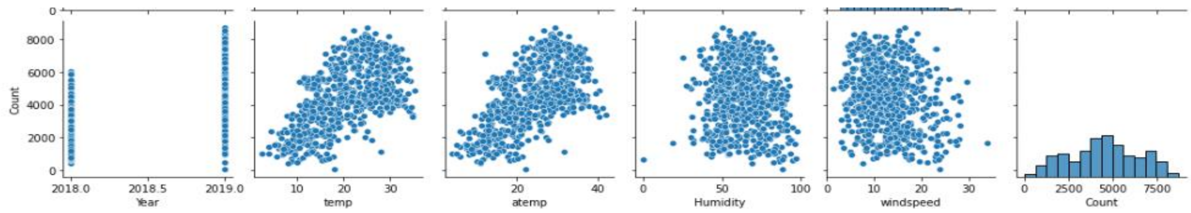
Ans :

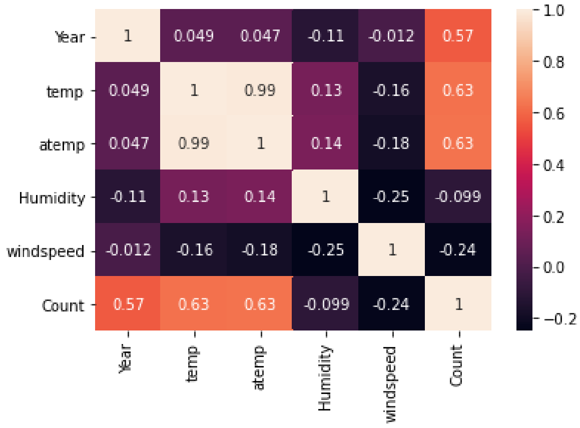
The drop_first parameter specifies whether or not to drop the first category of the categorical variable. If we set drop_first = True, then it will drop the first category. It will only produce K – 1 dummy variables for k categories. So it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans:

temp and atemp (both having same value correlation with Count target variable)

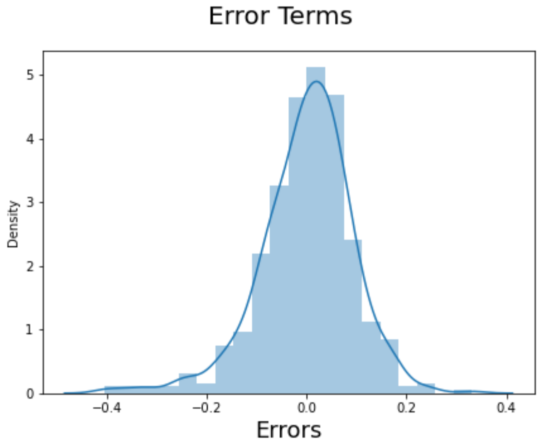




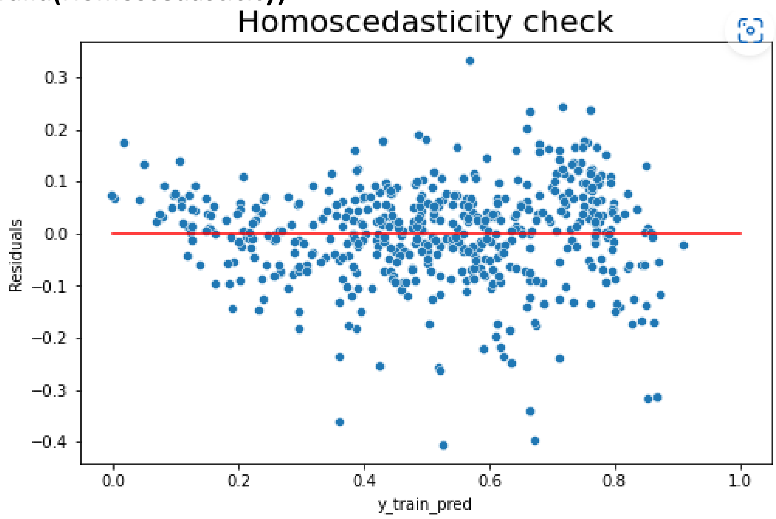
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans:

- **Plot the histogram of the error terms and confirmed that Errors are normally distributed and assumption is valid(Normality)**



- **Plot scatter with residuals and predicted values of train data and confirmed that Residuals have equal or almost equal variance across the regression line and assumption is valid(Homoscedasticity)**



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans:

- **temp = 0.5527(shared bike demand increase by 0.5527 for one unit temp)**
- **Light Snow(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) = -0.2785(shared bike demand decrease by 0.2785 for one unit Light snow)**
- **windspeed = -0.1552(shared bike demand decrease by 0.1552 for one unit windspeed)**

General Subjective Questions & Answers

1. Explain the linear regression algorithm in detail. (4 marks)

Ans :
Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent(y) and independent variable(x).

Linear Regression is of two types:
Simple Linear Regression is where only one independent variable is present and the model has to find the linear relationship of it with the dependent variable
Whereas, In **Multiple Linear Regression** there are more than one independent variables for the model to find the relationship.
Equation of Simple Linear Regression, where bo is the intercept, b1 is coefficient or slope, x is the independent variable and y is the dependent variable.

$$y= b_o + b_1x$$

Equation of Multiple Linear Regression, where bo is the intercept, b1,b2,b3,b4…,bn are coefficients or slopes of the independent variables x1,x2,x3,x4…,xn and y is the dependent variable.

$$y= b_o + b_1x_1 + b_2x_2 + b_3x_3 \dots . +b_nx_n$$

A Linear Regression model’s main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.
Error is the difference between the actual value and Predicted value and the goal is to reduce this difference.

Mathematical Approach:
Residual/Error = Actual values – Predicted Values
Sum of Residuals/Errors = Sum(Actual- Predicted Values)
Square of Sum of Residuals/Errors = (Sum(Actual- Predicted Values))2
i.e

$$\sum e_i{}^2 = \sum (Y_i - \hat{Y}_i)^2$$

Assumptions of Linear Regression –

1. Linearity: It states that the dependent variable Y should be linearly related to independent variables. This assumption can be checked by plotting a scatter plot between both variables.

Normality: The X and Y variables should be normally distributed.

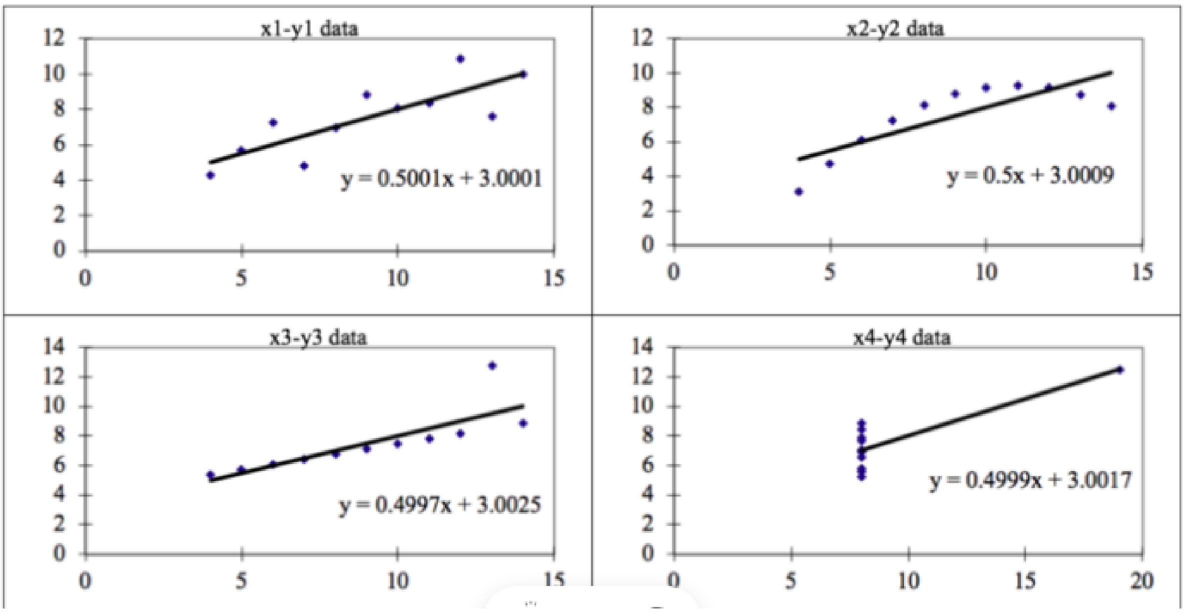
2. Homoscedasticity: The variance of the error terms should be constant i.e the spread of residuals should be constant for all values of X. This assumption can be checked by plotting a residual plot. If the assumption is violated then the points will form a funnel shape otherwise they will be constant.
   Error Term : y act – y pred
3. Independence/No Multicollinearity: The variables should be independent of each other i.e no correlation should be there between the independent variables. To check the assumption, we can use a correlation matrix or VIF score. If the VIF score is greater than 5 then the variables are highly correlated.

4. The error terms should be normally distributed. Q-Q plots and Histograms can be used to check the distribution of error terms.
5. No Autocorrelation: The error terms (yact – ypred) should be independent of each other. Autocorrelation can be tested using the Durbin Watson test. The null hypothesis assumes that there is no autocorrelation.

2. Explain the Anscombe’s quartet in detail. (3 marks)

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.
It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing 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.



The four datasets can be described as:
Dataset 1: this fits the linear regression model pretty well.
Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets

3. What is Pearson’s R? (3 marks)

Ans:

In statistics, the Pearson’s R(Pearson correlation coefficient),the bivariate correlation or colloquially simply as the correlation coefficient — is a measure of 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 the covariance, such that the result always has a value between −1 and 1

Pearson correlation coefficient formula:

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$

Where:
N = the number of pairs of scores
Σxy = the sum of the products of paired scores
Σx = the sum of x scores
Σy = the sum of y scores
Σx2 = the sum of squared x scores
Σy2 = the sum of squared y scores

Examples of Pearson’s r

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans:

Scaling -Transforming a data so that it fits within a specific scale, like 0-100 or 0-1.
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. Most of the times, collected data set contains features highly varying in magnitudes, units and range
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.
There are 2 types feature scaling available and are normalized and standardized scaling

- Normalization: Normalization is a technique often applied as part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.
    - ➢ Min Max Scaling: Min-max normalization is one of the most common ways to normalize data. For every feature, the minimum value of that feature gets transformed into a 0, the maximum value gets transformed into a 1, and every other value gets transformed into a decimal between 0 and 1.
    - ➢ Mean Normalization: It is very similar to Min Max Scaling, just that we use mean to normalize the data. Removes the mean from the data and scales it into max and min values.
    - ➢ Max Absolute Scaling: Scale each feature by its maximum absolute value. This estimator scales and translates each feature individually such that the maximal absolute value of each feature in the training set will be 1.0. It does not shift/center the data, and thus does not destroy any sparsity. This scaler can also be applied to sparse CSR or CSC matrices.
    - ➢ Robust Scaling: This Scaler removes the median and scales the data according to the quantile range (defaults to IQR: Interquartile Range). The IQR is the range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile).
- Standardization:
    - ➢ Standard Scaler: Standardization is a scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation

Differences between Normalization and Standardization

SL.No	Normalization	Standardization
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling
2	It is used when features are of different scales	It is used when we want to ensure zero mean and unit standard deviation.
3	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4	It is really affected by outliers	It is much less affected by outliers.
5	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans:

Yes. I have observed and that shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Below are the some corrective actions to eliminate large/infinite VIF:

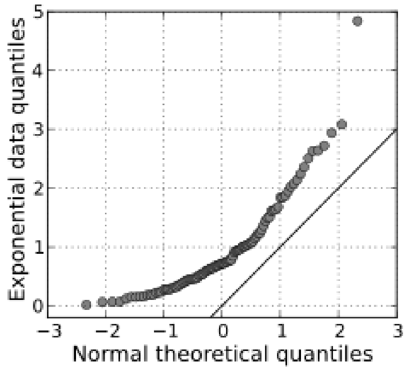
- One approach is to review the independent variables and eliminate terms that are duplicates or not adding value to explain the variation in the model.
- A second approach is to use principal component analysis and determine the optimal set of principal components that best describe your independent variables. Using this approach will get rid of your multicollinearity problem but it may be hard for you to interpret the meaning of these “new” independent variables.
- The third approach is to increase the sample size. By adding more data points to our model, hopefully, the confidence intervals for the model coefficients are narrower to overcome the problems associated with multicollinearity.
- The fourth approach is to transform the data to a different space like using a log transformation so that the independent variables are no longer correlated as strongly with each other.
- Finally, you can use a different type of model call ridge regression that better handles multicollinearity.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

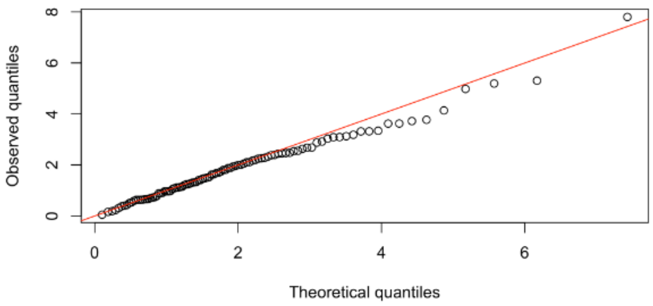
Ans:

The Q-Q plot is a graphical plotting of the quantiles of two distributions with respect to each other. In other words we can say plot quantiles against quantiles. A quantile is a fraction where certain values fall below that quantile. The purpose of Q-Q plots is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q-Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

A Q-Q plot showing the 45 degree reference line:



If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x.



If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.