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.

- Bike demand in the fall is the highest.
- Bike demand takes a dip in spring.
- Bike demand in year 2019 is higher as compared to 2018.
- Bike demand is high in the months from May to October.
- Bike demand is high if weather is clear or with mist cloudy while it is low when there is light rain or light snow.
- The demand of bike is almost similar throughout the weekdays.
- Bike demand doesn’t change whether day is working day or not.

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

Ans.

• It is important in order to achieve k-1 dummy variables as it can be used to delete extra column while creating dummy variables.

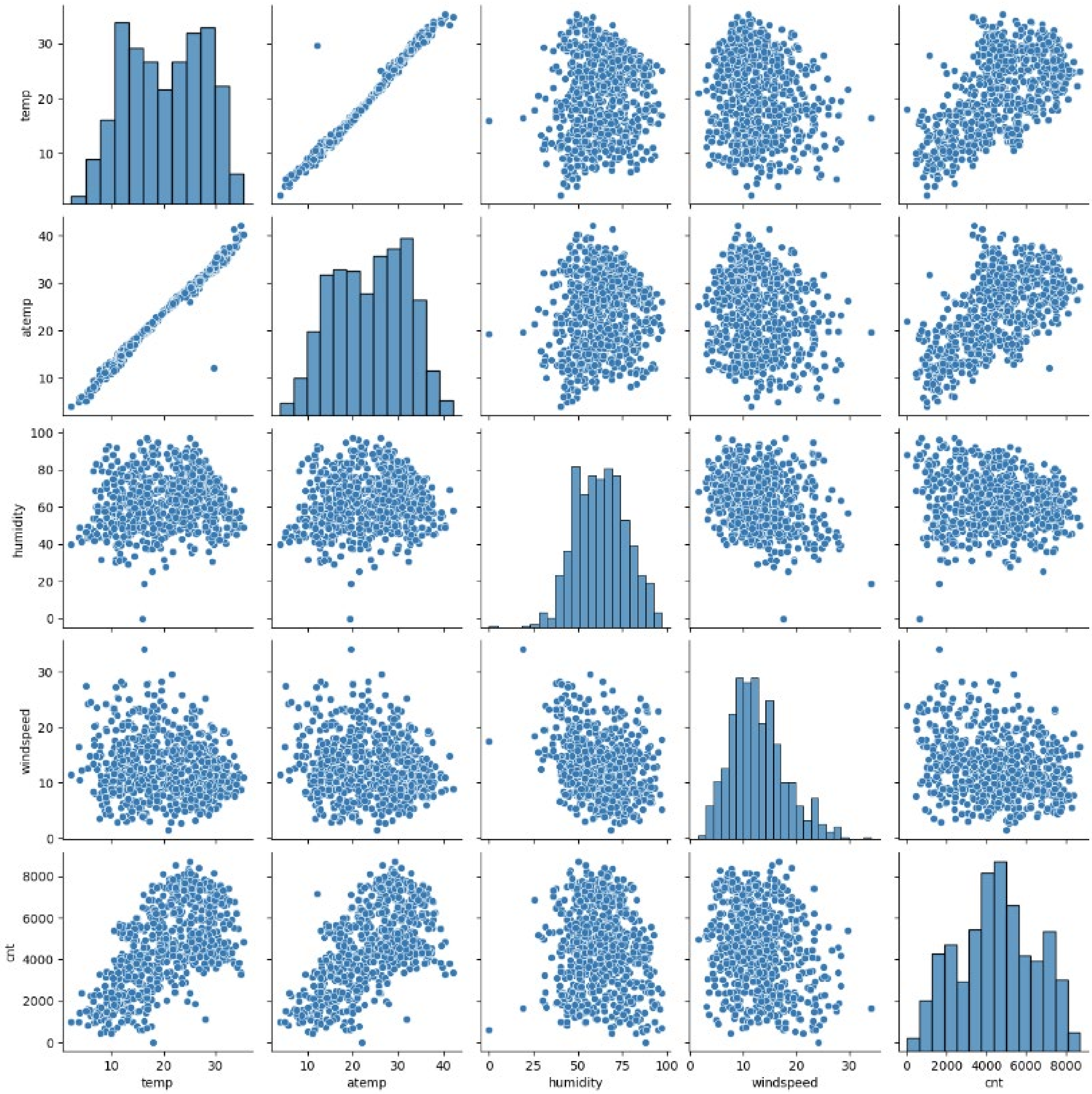
•For Example: We have three variables: Furnished, Semi-furnished and un-furnished. We can only take 2 variables as furnished will be 1-0, semi-furnished will be 0-1, so we don’t need unfurnished as we know 0-0 will indicate un-furnished. So we can remove it

•It is also used to reduce the collinearity between dummy variables .

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

Ans.

- atemp and temp both have same correlation with target variable of 0.63 which is the highest among all numerical variables.



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

Ans.

- Linearity of relationship between response and predictor variables.
- Normality of the error distribution (Normal distribution of error terms).
- Constant variance of the errors or Homoscedasticity.
- Less Multi-collinearity between features (Low VIF)

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

Ans.

- temp
- light_rain_snow
- Sept

General Subjective Questions

1. Explain the linear regression algorithm in detail.

    Ans.

    It is a form of regression, where the target variable is continuous. It estimates the relationship between a target variable and one or more predictor variables.

    The Equation of linear Regression is y= m1x1 + m2x2 + m3x3 + ………. + m(n) x(n) + c.

    Where y is target variable and x1, x2 , x3 …… xn are predictor variables . And we have two unknowns, m, and c, and we need to choose those values of m and c, which provides us with the minimum error. We need to get the best fit line which is the line that has the minimum error. In linear regression, when the error is calculated using the sum of squared error, this type of regression is known as OLS, i.e., Ordinary Least Squared Error Regression.

    Error function is explained by ‘e = - y’, and error depends on the values of ‘m’ and ‘c’. Our aim is to build an algorithm which can minimize the error.

    And in order to do so we use cost function of Linear Regression, Which is:

    J (mi, c)= (1/2n) Σ (yi – yp)^2 Where yi and yp are expected values and predicted values.

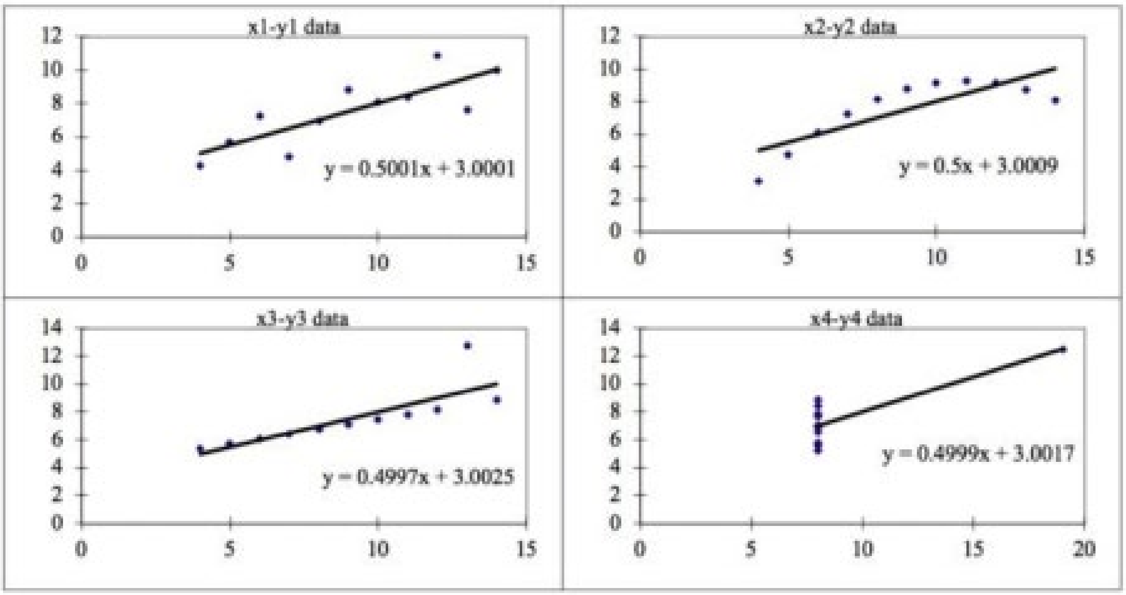
    Our main aim is to minimize J by changing m and c and it can be done using Gradient Descent Algorithm.

    Cost function measures the performance of a Machine Learning model for given data.

2. Explain the Anscombe’s quartet in detail.

    Ans.

    Anscombe’s Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics. They have very different distributions and appear differently when plotted on scatter plots. Anscombe’s Quartet tells us about the importance of visualizing 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 outliers and other features such as data diversity. Linear Regression can be only be considered a fit for the data with linear relationships.



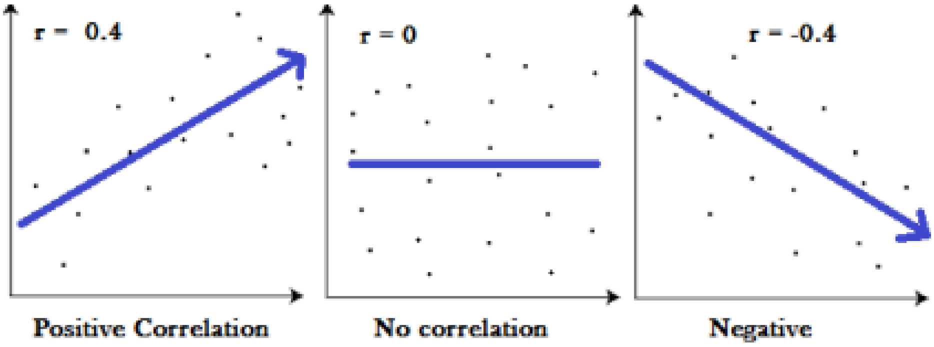
Data 1: Linear regression model can be fit.
Data 2. Linear regression model cannot be fit as data is non-linear.
Data 3: Linear regression model can be fit but still there are outliers that can't be explained by LR.
Data 4: Outliers in the data set that can't be explained by LR.

3. What is Pearson's R?

Ans.

Pearson's correlation coefficient is used to measure the strength of a linear relationship between data. Pearson's correlation is a correlation coefficient used in linear regression. It is used to find how strong a relationship is between data. It returns a value between -1 and 1, where:

- 1 indicates a strong positive relationship
- -1 indicates a strong negative relationship.
- 0 indicates no relationship at all.



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans.

It is a step of data preparation for modelling where scaling is applied to predictor variables to normalize the data within a particular range. It helps in speeding up the calculations in an algorithm like Linear Regression etc. It is used in order to scale the data into same units as algorithm only takes magnitude in account and not units resulting in faulty model. It is done to bring all the variables to the same level of magnitude.

Normalized Scaling also known as MinMax Scaling: MinMax Scaling means re-scales the values into a range of [0,1] .

- x = x – x(min) / x(max) – x(min) Standardized scaling: Standardization means re-scales data to have a mean of 0 and a standard deviation of 1.
- x = x – x(mean) / x(standard deviation)

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

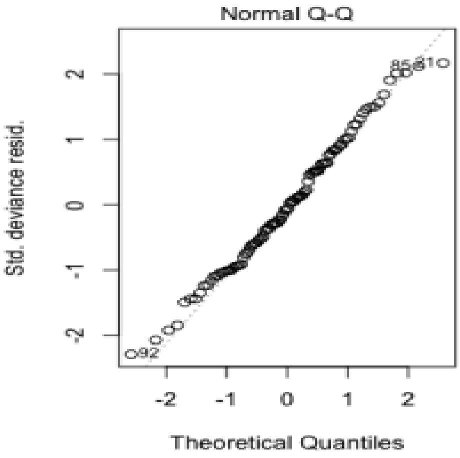
Ans.

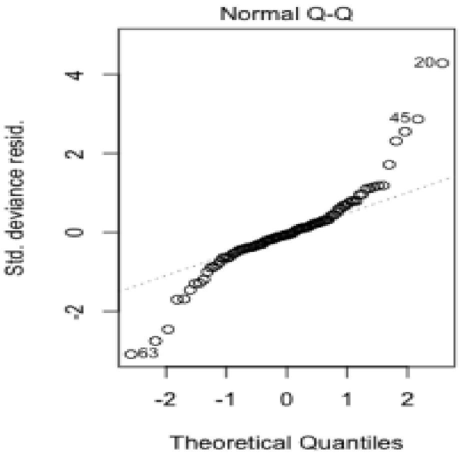
If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a high correlation between variables. If the correlation is perfect than R2=1, which in turn makes **VIF = 1 / (1- R2) infinity**

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

Ans.

Q Q Plots are plots of two quantiles against each other. The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A Q–Q plot is used to compare the shapes of distributions, providing a graphical view toward the given data. A 45 degree angle is plotted on the Q Q plot, if the sample data is normally distributed, it will fit on the Q-Q plot line. If not, then the data is skewed.


Normal Q-Q
Std. deviance resid.
Theoretical Quantiles


Normal Q-Q
Std. deviance resid.
Theoretical Quantiles