Assignment-based Subjective Questions

Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

- After plotting the categorical variables with target variable(Dependent variable) on boxplot and infer about their effect on the dependent variable
    - ✓ In Clear weather has high demands rental bikes.
    - ✓ Demands for rental bikes had been grown for year 2018 to 2019
    - ✓ Season 3 (fall) has highest demand for rental bikes.
    - ✓ Demand for rental bikes grows continuously till September month.
    - ✓ September month has highest demands for rental bikes and after September demand is decreasing as winter starts.
    - ✓ Working day, weekday and holiday is not affecting target variable or not giving any clear information of demands on rental bikes.
    - ✓ Bike rental is leant during Light rain and snow

Q2. Why is it important to use drop_first=True during dummy variable creation?

- drop_first equal to True is important during variable creation, because it helps in reducing extra column creation.
- Hence, it reduces the correlations created among the dummy variables.
- If we don't not used drop_first=True then it becomes redundant with dataset.

Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

- The ‘temp’ and ‘atemp’ variables has highest correlation with target variable (cnt).

Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

a) Linear relationship between independent and dependent variable.
b) Errors terms are normally distributed with mean 0.
c) Error terms are independent of each other.
d) Error terms have constant variance
e) Ensuring over-fitting by looking at R2 and Adjusted-R2 value.

Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

- Features ‘year’, month ‘atemp and week ‘Clear’ is highly correlated with target variable, so these are top features in building model.

General Subjective Questions

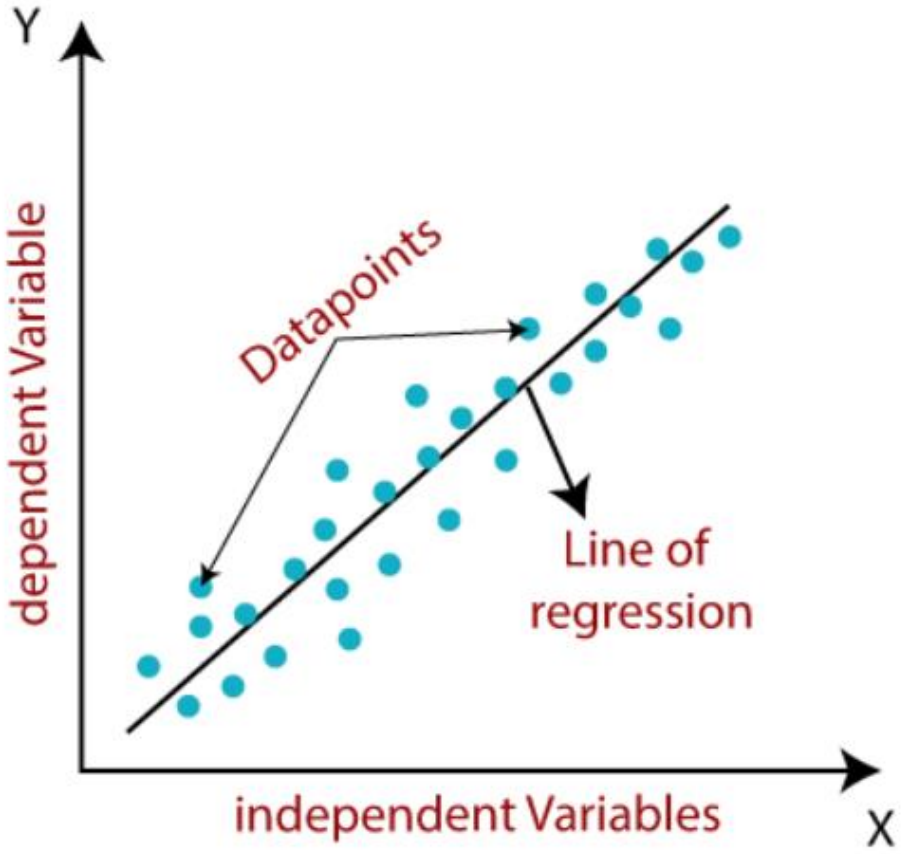
Q.1 Explain the linear regression algorithm in detail.

- Linear regression algorithm is a method of finding a linear relationship between a dependent (y) and one or more independent (x) variables, hence called as linear regression.
-T he linear regression model provides a sloped straight line representing the relationship between the variables.

Equation of Straight line:- $y= \theta 0+ \theta 1x$

Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
θ0= intercept of the line (Gives an additional degree of freedom)
θ1 = Linear regression coefficient (scale factor to each input value).

<u>Below Diagram Representing Linear Model</u>



- It fits the best line to predict the value of y for a given value of x.
- The model gets the best regression fit line by finding the best θ1 and θ2 values.
- For finding the best fitted line it uses Cost Function to find the best value of θ1 and θ2 so model aims to predict y value such that the error difference between predicted value and true value is minimum.

$$minimize \frac{1}{n} \sum_{i=1}^{n} (pred_i - y_i)^2$$

- **Cost function** of Linear Regression is **Root Mean Squared Error (RMSE)** between predicted y value (predi) and true y value (yi).

<u>Gradient Descent:</u>

- It is a method to update θ1 and θ2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line for the model.

Linear regression can be further divided into two types of algorithm.

- Simple Linear Regression:
  - If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

  Equation:- $y= \theta 0+ \theta 1x+ \varepsilon$
    - Y= Dependent Variable (Target Variable)
    - X= Independent Variable (predictor Variable)
    - θ0= intercept of the line (Gives an additional degree of freedom)
    - θ1 = Linear regression coefficient (scale factor to each input value)
    - ε = random error

- Multiple Linear regression:
  - If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

  Equation:-

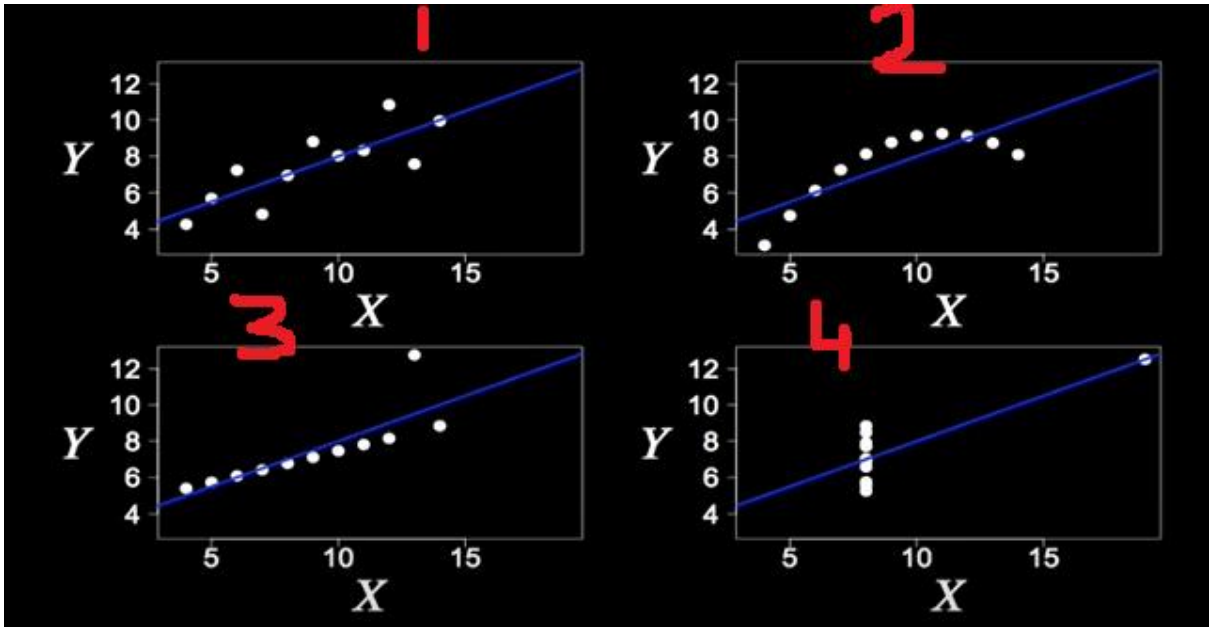
$$y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon$$

  - y = the predicted value of the dependent variable
  - β0 = the y-intercept (value of y when all other parameters are set to 0)
  - β1= the regression coefficient (β1) of the first independent variable (X1)
  - βn = the regression coefficient of the last independent variable
  - e = model error (how much variation there is in our estimate of y)

2. Explain the Anscombe's quartet in detail.

- It is a statistical observation which was published by a statistician Mr.Frank Anscombe in 1970s
- It describes the importance of visualisation(plot) of dataset before making a model to take a correct decision on the dataset as summary statistics of model doesn't gives the correct information at all the time.

- In the below given 4 different dataset were used and for all ,the fitted line has same summary statistics i.e. same mean of X and Y, variance of X and Y, same correlation and same slope of the line , same intercept.



1st Figure:-

- The relation between X and Y looks linear and we can use linear regression model to find the best fit line.

2nd Figure:-

- From the plot we can observe that the relationship between X and Y is not linear hence linear regression is not a good option to use and we need to choose another model.

3rd Figure:

- The relation between X and Y looks linear and we can use linear regression model to find the best fit line but outliers throws the slope and intercept.
- We should do the EDA to find the reason of the outlier.

4th Figure:-

- This dataset might also have the linear relationship between X and Y but from the plot we can conclude that we should try to acquire more data for intermediate X-values to make sure it really does the good job.

Q3. What is Pearson's R?

- Correlation is a process of measuring the relationship between sets of data and how well they are related.
- It is the ratio between the covariance of two variables and the product of their standard deviation.
- It is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

Formula:-

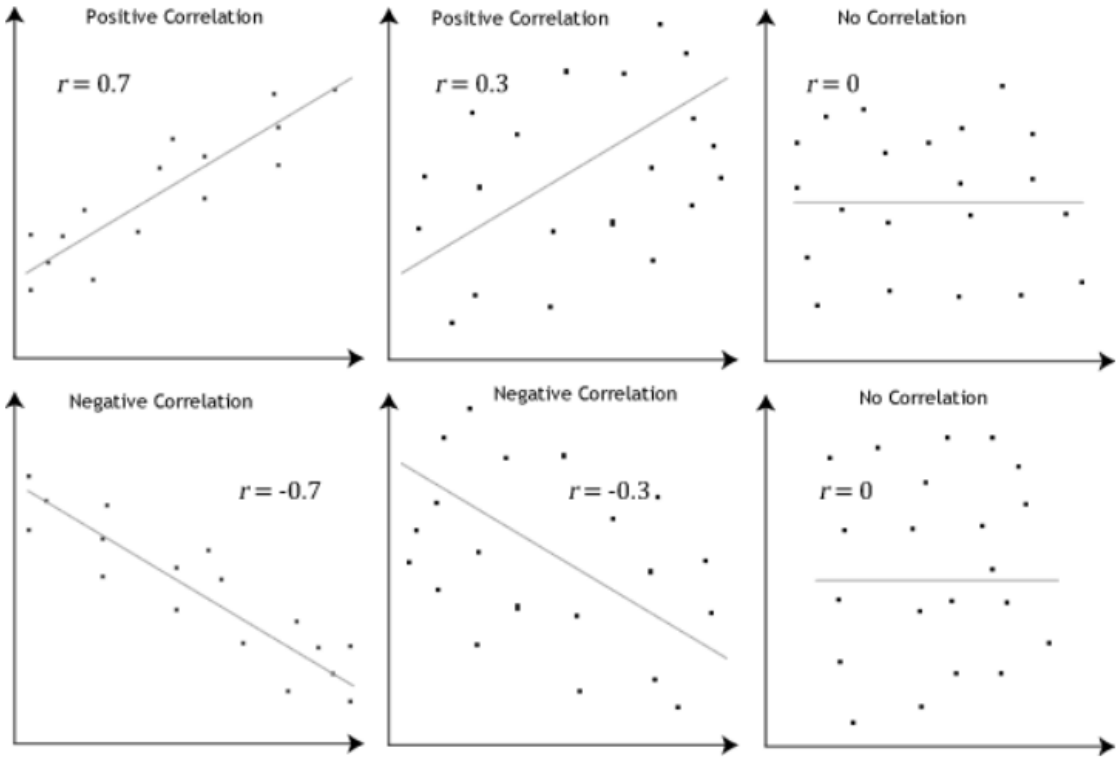
$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

σx and σy = population correlation coefficient

σxy = population covariance.

Limitation

1. It doesn't tell the difference between dependent variables and independent variables it only tells the nature of relationship.
2. It doesn't give any information about the slope of the line.



No Correlation

- R=0 indicates that there is no association between the two variables.

Positive Correlation

- R value greater than 0 indicates a positive association that is, as the value of one variable increases other also increases.

Negative Correlation

- R value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.

Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Scaling

- It is a process to bring the coefficient of the variables in a certain range by bringing all the variable in a comparable scale (interpretability)

Why is scaling performed?

- It will help us to better understand coefficient relation between then if not done then the variance in the coefficient of feature will be large.(ease of interpretation)
  Faster Convergence for Gradient Descent Method.

- If we rescale the variable in the range of 0-1 then the optimization happening behind the scene becomes much faster i.e minimisation routine becomes much faster.
- When we train a network using Gradient Descent function then it becomes very fast.

There are different method of Scaling.

1)Min Max Scaler (Normalisation)
2) Standard Scaler
3) Max Abs Scaler
4) Robust Scaler
5) Quantile Transformer Scaler
6) Power Transformer Scaler
7) Unit Vector Scaler

- MinMaxScaling is the most commonly form of scaling used in Linear Regression as it bring the variable in the range of -1 to 1 which is easy to understand.

Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

VIF

- VIF is a process to quantify the linear association and it is used when one variable might depend upon the combination of other variables.

VIF is infinite

- It happens because of correlation, means the information of one variable is perfectly represented by the combination of different independent variable.
- VIF value of infinity means that there is a perfect correlation which results as value 1 for R-Squared.
- VIF= 1/(1-R2)
- If R2=1 then VIF will be infinity so in this case we should drop that variable as it's information is already shown by the other set of variables.

Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression?

- Q-Q plot stands as Quantile-Quantile plots, as the name suggest as 2 quantiles, it plots 2 quantiles against each other.
- Quantile is a percentage where certain value falls below that percentage and certain above that.
- Suppose 25% quartile means 25% of data is below it and 75% of data is above it.
- Main purpose of Q-Q plot is to find the whether two set of data is coming from same distribution or not.
- It helps in linear regression ,when we received training and test data set separately, then we use Q-Q plot to confirm that both the data set are from populations with same distributions.
- If all points of quantile lies on or close to straight line at an angle of 45 degree from x –axis then it is called as similar distribution.