Assignment-based Subjective Questions

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

Ans: From analysis of categorical variables below are inference about their effect on dependent variable

1: Season :- Fall/Rainy season frequency of bike sharing demand is very high as compare to other seasons and for spring season frequency is very low.

2: Year(yr) :- Year 2019 bike sharing demand was very high as compare to demand in year 2018.

3: Month(mnth) :- August and September month are high demanded month whereas January and February is very low.

4: Holiday :- Whether it is holiday or not 75^{th} percentile is equal for both but when it was holiday then box area (25^{th} percentile to 75^{th} percentile) is equal to range of (50^{th} percentile to 75^{th} percentile) when it was holiday.

5: Weekday and Working day :- For both week day and working day box plots are pretty much same. 50^{th} percentile is pretty much equal.

6:- WeatherSit :- When weather situation is clear then bike sharing demand is pretty high and when it is light-snow then demand went to low.

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

Ans: When we have categorical variables in our dataset and we want to convert them into numeric in terms of binary then we use get_dummy() method which converts category into binary form in the way that for N category it creates N levels. So but we can manage these levels with (N-1).For this we use drop_first property of get_dummy() method which drop first column/level to get (N-1) labels.

For example : suppose we have three categories: yes, no and maybe

Then get_dummy() will create 1,0,0 and 0,1,0 and 0,0,1. So in this case we can drop first column so that if it is 0,0 then it is called ‘yes’.

Q 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?
Ans: temp and atemp have the highest correlation with the target variable.
Q 4. How did you validate the assumptions of Linear Regression after building the model on the training set?
Ans:

1. For Linear Regression we look for relationship between Target variable and Predictor Variables. We took help of Pair Plot, Heatmap to check: is there any linear relation between Target variable that is on y-axis and Predictor variable that is on x-axis.
2. Data Cleaning : Checked missing values in any column, Dropped unnecessary features (with respect to Target variable).Verified outliers, derived new features if required.
3. Create Correlation Matrix to know about correlation of each predictor variable with Target variable.
4. Create dummy variables from Categorical variables
5. Re-scaling the features to put all on same scale for any further calculation.
6. Split the data into train - test dataset.
7. Fit and transform the model using train dataset first.
8. After fitting model check all necessary metrics like coefficients, p-value, R-squared, Adjusted R-squared
    - p-value should be less than 0.05 consider be good for features selection for model.
    - R-squared and Adjusted R-squared should be high consider to be fit for explain most of the variance but not over/under fitted.
9. There is one more factor called VIF (Variance Inflation Factor) which can to determine the relation between predictor variables.
    - VIF > 10 : Definitely high VIF value and variable should be eliminated.
    - VIF > 5 : Cab be okay, but it is worth inspecting.
    - VIF < 5 : Good VIF value, No need to eliminate variable.
10. After fitting the model, validate the **assumptions**:
        - Error term should be normally distributed.
        - There should be linear relationship between the X and Y (Target and Predictors variable)
        - Error terms are independent of each other.
        - Error terms have constant variance.

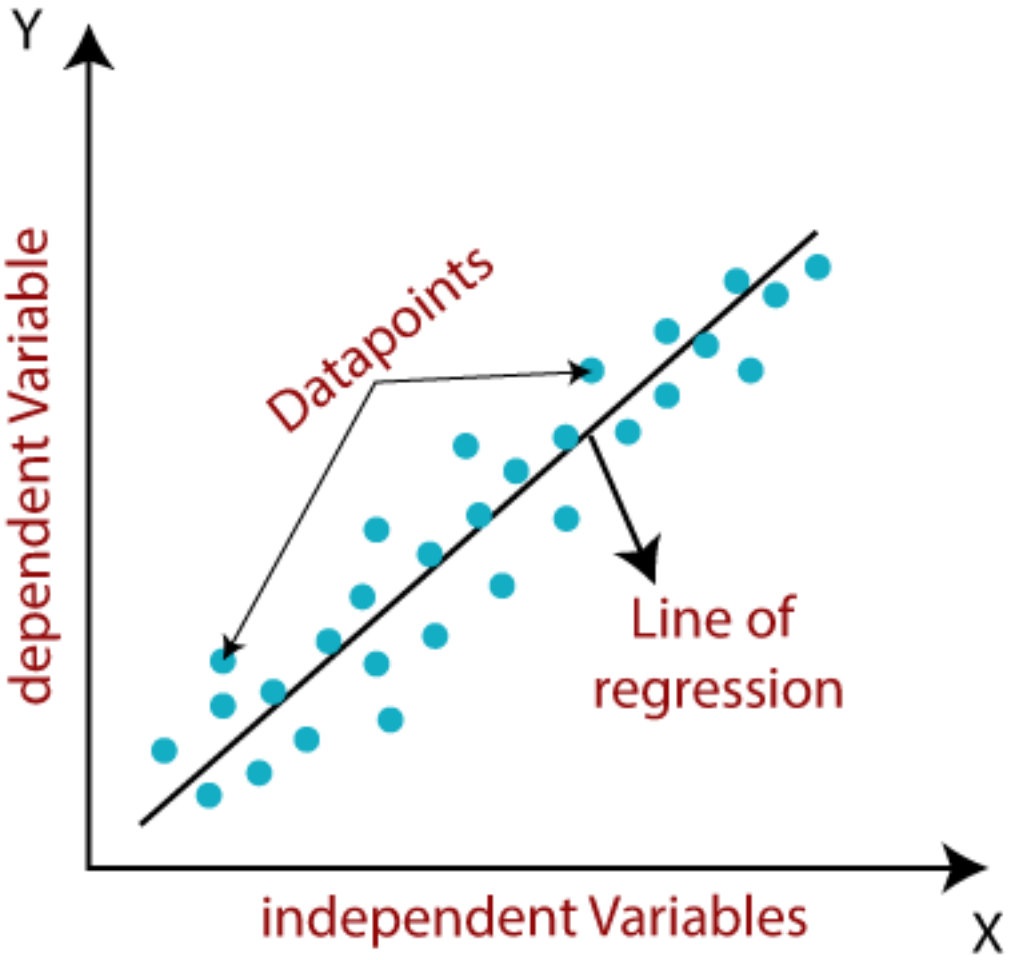
11. Transform the test dataset by predictors fit for linear model on train dataset. For scaling the test dataset we use only transform not fit_transform.
12. Then finally we calculate r2_score for both train and test dataset which should be high or near to 1 for model to be significant and also difference between these two should also be very less then we can say that our model is good fit linear regression model.

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

1. Temp : Positive Correlation and highest correlation value among all the features
2. Year (yr) :- Positive Correlation and also higher as compare to other features.
3. WeatherSit (Light Snow and Mist) : Negative Correlation, as it decreases bike sharing demand increases.

General Subjective Questions

Q 1. Explain the linear regression algorithm in detail.
Ans: Linear regression is one of the easiest and most popular Machine Learning algorithms. It shows how one dependent/target variable is linearly related to other independent/predictor variables. Linear regression helps us to predict about continuous/real or numeric variables such as **profit, price of any product, sales count** etc.
Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (X) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable/s.



Mathematically, we can represent a linear regression as :
y= c + mx + ε

Here,
y = Dependent variable (Target Variable)
X = Independent variable (Predictor Variable)
c = Intercept of the line (Gives an additional degree of freedom)
m = Linear regression coefficient (scale factor to each input value)
ε = Random error
In Linear regression model, we have to find the best value of m (Slope/ Coefficient) and c (intercept) so that we can fit best linearity using y predict values.
Steps to follow:

1. Draw Heatmap or Pair plot between target and predictor variables to know about relationship between them
2. Split the complete dataset into train and test dataset

3.Scale train data set before fit the model

4.Use OLS (Ordinary Least Square) method to fit the model

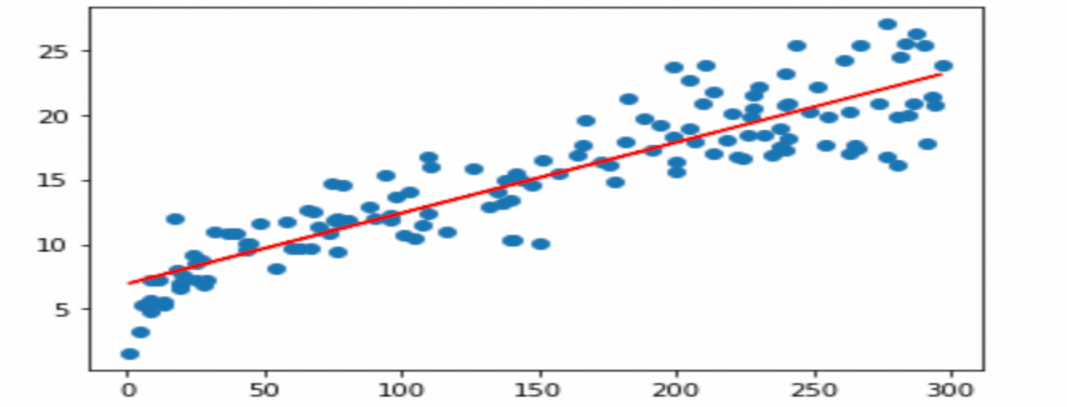
5.Get the summary after fit the model

6.From final summary (adding or removing variables from the model on the basis of metrics) get the value of **m** and **c .**

7.In the end fit the line using the formula y = mx + c

c = 6.948, m = 0.054

```
plt.scatter(X_train, y_train)
#plt.plot(X_train, 6.948+0.054*X_train, 'r')
plt.plot(X_train, y_train_predict, 'r')
plt.show()
```



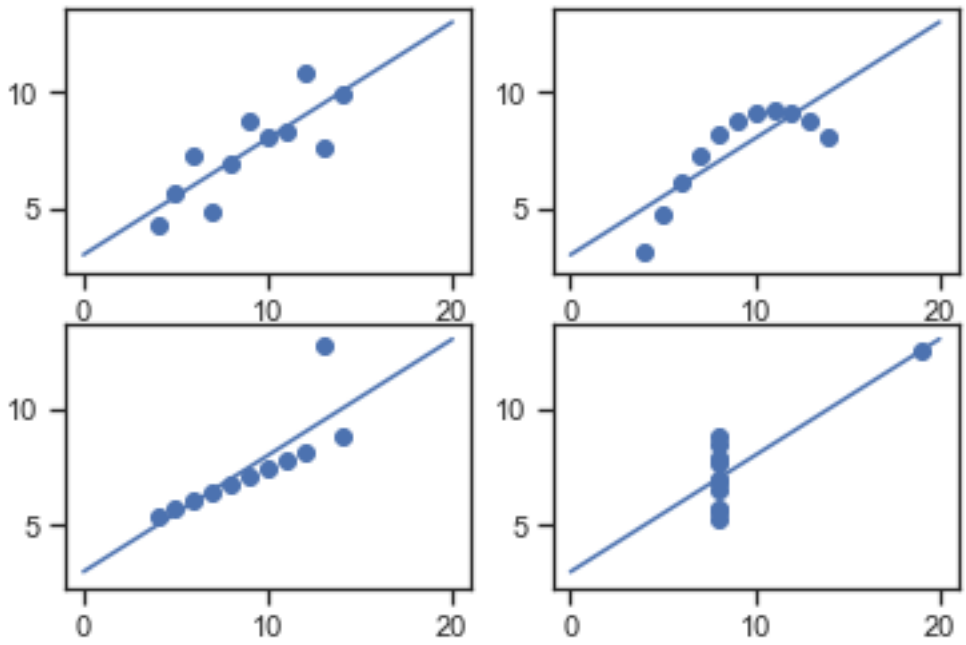
Q 2. Explain the Anscombe's quartet in detail.

Ans: An Importance of Data Visualization: Most of the people believe that "numerical calculations are exact, but graphs are rough or even though it's completely wrong". But this is not true.

Anscombe's Quartet is the model example to demonstrate the necessity of data visualization along with statistical analysis. This model is developed by the statistician Francis Anscombe in 1973. This model comprises of four data-set and each dataset consists of (x,y) points. For these datasets we need to analyse, even though all these four datasets are sharing same statistics (mean, variance, standard deviation etc.) but when we plot graphical representation of these datasets they look totally different. Each graph shows different behaviour irrespective their statistics.

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

For above 4 datasets - after applying statistical formula we have :
Average value of x = 9
Average value of y = 7.5
Variance of x = 11
Variance of y = 4.12
Correlation Coefficient = 0.816
Linear Regression Equation : y = 0.5x + 3
However we have same statistical data for these 4 datasets but when we tried to plot graphical representation across x and y, we got below graphs:



Graphical Representation of Anscombe's Quartet

- Data-set I – consists of set of (x,y) which represents linear relationship with little variance.
- Data-set II – shows non-linear relationship between (x,y)
- Data-set III – looks like a tight linear relationship for (x,y), except one large outlier
- Data-set IV – shows that value of x remains constants, except for one outlier

To conclude, datasets which are identical over a number of statistical properties might have different visuals. So this Anscombe's Quartet model is pretty important to visualize our data before getting into statistical data and help to revisit the statistics output and re-introduce it as per need.

Q 3. What is Pearson's R ?

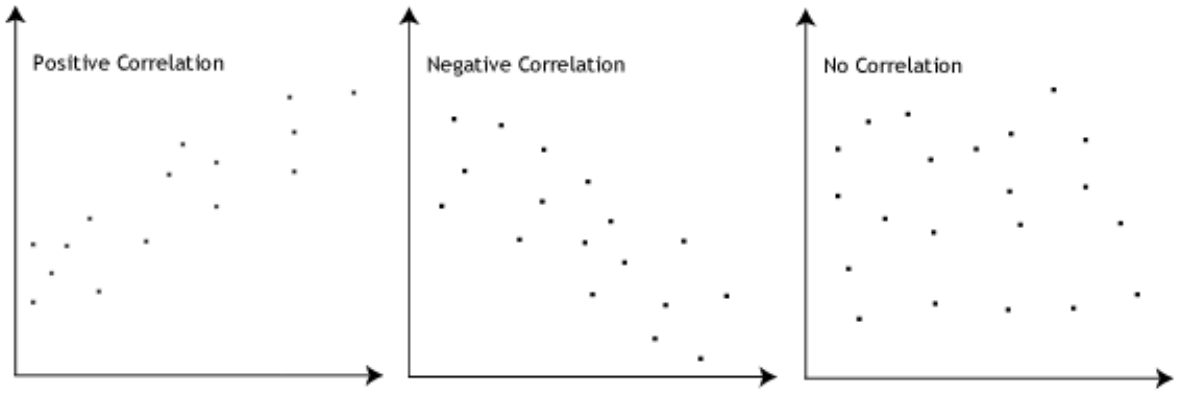
Ans: In statistics, Pearson's R also known as Pearson Correlation coefficient, or we can say bi-variate correlation, is a way to measure the linear correlation between two variables. Basically it is a co-variance of two variables, divided by product of their standard deviation.

Its value always lies between -1 to 1, means:

- r = 1, means very strong positive linear relationship between both the variables, or we can say both variables tends to change in same direction.
- r = -1, mean very strong negative linear relationship, or we can say both variables tends to change in different direction.
- r = 0, it means there is no relationship between variables.

To divide this Pearson's r value in different segment we have below:

- 0 < r < 5, there is weak relationship
- 5 < r < 8, there is moderate or average association between variables
- r > 8, there is strong linear relationship between variables



PEARSON'S r Formula :

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Here,
r = correlation coefficient
xi = values of x-variable in the sample
x-bar = mean of the values of x-variable
yi = values of y-variable in the sample
y-bar = mean of the values of y-variable

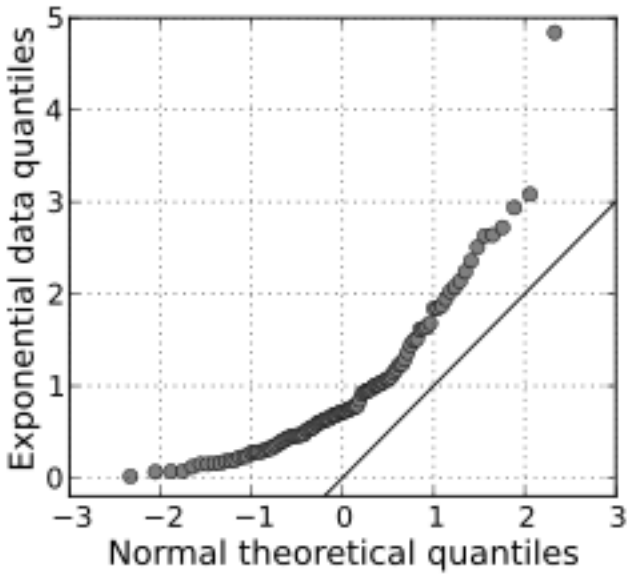
Q 4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Ans: Scaling is a pre-processing step which converts all the independent features into a normalize/standardize/comparable scale. It comes after train-test dataset split.
For example suppose we have salary (having values in thousands or lacs) and rank (having values from 1 to 10) features in dataset which are not on the same scale, so it might misinterpret our model prediction. That's why we use scaling the features of the dataset.
The main reason to perform this step is : we can interpret the coefficients of the features very easily after fitting the model with the help of scaling. We can prevent our model for any misleading.
Scaling only affect coefficient values, other metrics like r-squared, p-value, adjusted r-squared etc. remains unaffected.
There are basically two types through which we can perform scaling:

- Min-Max Scaling/Normalization (range from 0 to 1)
    - It compresses the data between 0 to 1.
    - It will not affect the values of dummy variables, which already has values 0 and 1.
    - Highly affected by outliers
    - Min-Max Scaling : x = (x – min(x))/(max(x)-min(x))
- Standardize Scaling (mean = 0, std = 1)

- It brings all data into standard normal distribution with mean = 0 and standard deviation = 1
- It will definitely distort the values of dummy variables.
- Less affected by outliers.
- Standardize Scaling : x = (x – mean(x))/standard deviation(x)

Q 5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Ans: VIF – Variance Inflation Factor calculates how well one independent variable is explained by all other independent variables combined. It measures the multicollinearity among the independent variables while doing Multiple Linear Regression.
When there is perfect correlation between two independent features then we get **VIF = infinity.**
In perfect correlation we have r-squared = 1
So, by the formula:
VIF = 1/(1-r2) = infinite (1/0 = infinite)
An infinite VIF value denotes that the corresponding feature may be expressed exactly by linear combination of other features which shows an infinite VIF as well.
So to come over from this issue we can drop one of the feature from the dataset causing this perfect multicollinearity.

Q 6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Ans: As the name suggests, Q-Q plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is fraction where certain values fall below that quantile. For example, median is a quantile where 50% data lie above this and 50% data fall below this quantile. The purpose of Q-Q plot is to determine whether two datasets come from same distribution. Whenever we interpreting a Q-Q plot , we shall concentrate on the ‘y = x’ line. We also call it the 45-degree line is statistics.
A Q-Q plot showing 45 degree reference line :



A Q-Q plot is used to compare the shapes of distribution, provide a graphical view of how different properties like scale, location and skewness are similar or different in two distributions.