Assignment-based Questions

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

Analysis of categorical variables using boxplot reveals following observations.

1. There are significantly lesser number of service users in the spring season.
2. Total users have increased significantly from 2018 to 2019
3. Months 11,12,1,2 reports significantly lesser number of users
4. Casual user appears to use the service more on holiday however it is opposite for the registered users
5. Similar trend is followed by casual and registered users for working day.
6. Casual users tend to use the service more on the weekend. It is opposite for registered.
7. Service users decline with worse weather situation. Weather Situation 4 (Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog) reported no users.

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

Use of drop_first reduces total number of features involved in model building. It helps in reducing complexity of model. It also helps deal with correlation between multiple variables and help deal with multicollinearity.

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

Analysis using pair-plot shows that *temp* and *atemp* have the highest correlation with the target variables – *casual, registered and cnt.*

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

There are 4 assumptions of linear regression –

1. Linearity
2. Normal distribution of error terms
3. Homoscedasticity
4. No multicollinearity

Residual Analysis is done after building models to test assumptions.
Linearity can be tested using scatter plot of independent variables with dependent variables. Distribution plot of error term is used to check if they are normally distributed. Homoscedasticity is tested using scatter plot of residuals against predicted values. Multicollinearity is tested using variance inflation factor.

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

1. For casual users, features with highest absolute coefficient value in final model are – temperature, humidity, windspeed, weathersit_3, season_4
2. For registered users – temperature, humidity, weathersit_3
3. For all users – temperature, humidity, weathersit_3, season_4

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Linear Regression is a supervised machine learning models that tries to fit a line through a set of data.

Say y is a dependent variable and X is independent variable

The goal of regression is to find b_0, b_1 such that, the line denoted by the below equation is best fit line.

$$y = b_0 + b_1 x$$

The best fit line is the line for which the value of cost function is minimum

Cost function for LR is given by

$$\sum_{i=0}^{n}(y_{actual} - y_{predicted})^2 = \sum_{i=0}^{n}(y_{actual} -)^2$$

Gradient Descent is used to minimize the cost function. It works by iteratively finding b0,b1 values that decreases the cost function till it reaches minimum and no changes occurs.

General steps involved in Building LR model is

1. Data Preparation
2. EDA
3. Creating dummy variables for categorical data
4. Feature scaling
5. Train test split
6. Building model with one feature and then adding feature iteratively or building model with all features and reducing features by evaluating P-score and VIF
7. Residual analysis and Linear Regression assumptions testing
8. Model Evaluation and Prediction

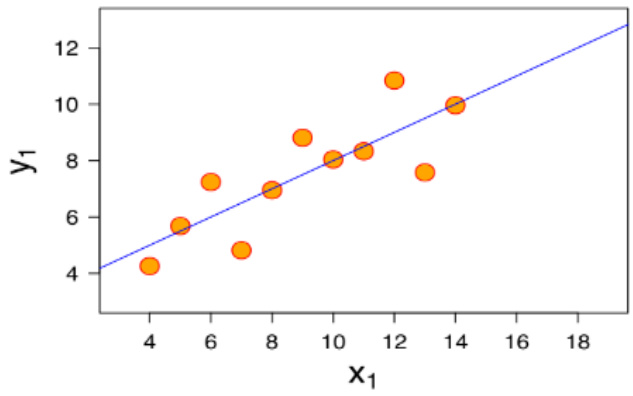
Q2. Explain the Anscombe's quartet in detail.

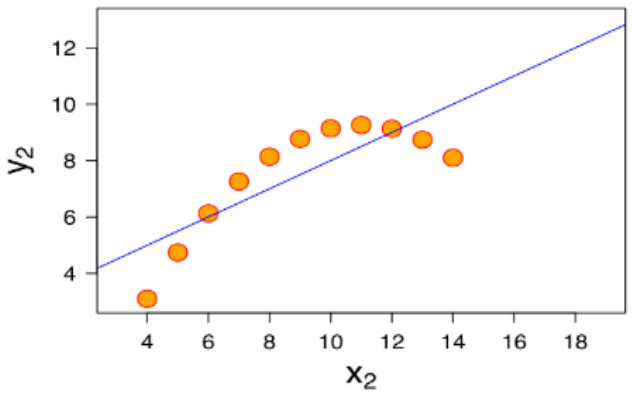
It is a set of four dataset. The dataset has identical simple descriptive statistics but are significantly different when plotted. It is used to emphasize the importance of visualizing data before analysis. The data set are as follows.

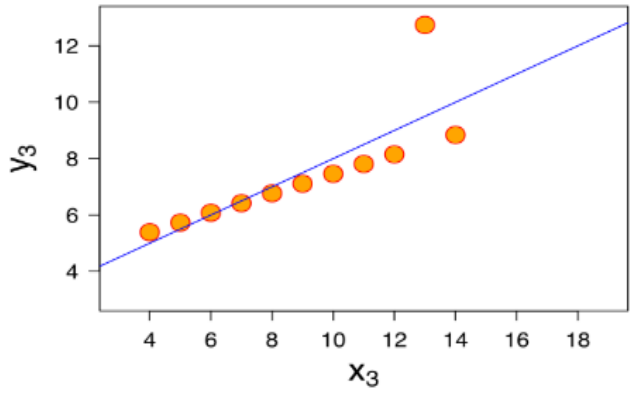
Data 1		Data 2		Data 3		Data 4	
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

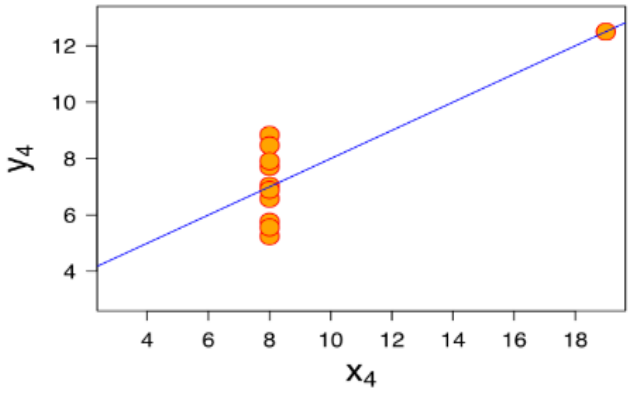
Identical metrics	*Values*
Mean(x)	*9*
Mean(y)	*7.5*
Sample-Variance(x)	*11*
Sample-Variance(y)	*4.125*
Correlation(x, y)	*0.816*
Coefficient-of-determination	*0.67*
Linear regression equation	*y = 3.00 + 0.500x*

Plots of data are as followed









Q3. What is Pearson's R?

Pearson's R simply called correlation coefficient is measure of linear correlation between two variables. Its value ranges between -1 to 1.
Its formula is given as below:

$$\rho(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{\sum_{i=0}^{n}(x_i - x_u)(y_i - y_u)}{\sqrt{\sum_{i=0}^{n}(x_i - x_u)^2}\sqrt{\sum_{i=0}^{n}(y_i - y_u)^2}}$$

The correlation coefficient is positive when the both the variables lie on same side of their mean and negative otherwise. In other words, the sign of Pearson's R determines tendency of the variables lying on same or opposite of their means. The absolute value determines how greater this tendency is.

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

Scaling is the process of taking a range of numeric values and transforming into a range of different values. It can be thought as changing the units of variables.

When building models with multiple independent variables, some variables may have significantly higher range of values. E.g., A and B are two features. A has range (1000,3000). B has Range (10,30). ML algorithms may emphasize feature A over B more due to this. Hence scaling is required. Scaling also helps in optimizing performance of some algorithms.

Two common scaling methods are:

Standardization: $x_{standardized} = \frac{x - \mu}{\sigma}$

Normalization: $x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}$

There is no hard and fast rule to determine which is better and depends on application. Standardization is suitable when we know that data follows normal distribution. Normalization helps with outlier treatment. Standardization does not have bounded range. Normalized data will lie between [-1,1].

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

Infinite VIF means that there is perfect correlation between, the feature and remaining feature. All the variance in the said feature is explained by other feature.

VIF is given by below formula:

$$VIF(X) = \frac{1}{1 - R_{squared(X)}}$$

When R-Squared or coefficient of determination for the x feature on remaining features is equal to 1 VIF is infinity.

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

Q-Q plot plots the quantiles of a sample data against quantiles of some theoretical data that follows some distribution (like normal). If both the data form a line as shown in chart, both datasets follow same distribution.