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?**

Answer:I have done analysis of categorical variables season,mnth,yr,weekday,weathersit,holiday and working day using boxplot .Their effect on the dependent on the dependent variable ‘cut’ is as below:

- Hotter season ‘summer’ attracts more booking of bikes.
- Months namely August & September sees more demand for bikes.
- Every year booking & demand for bikes is increasing.
- Sunday’s have more number of booking as compared to start of week.
- Misty & light snowrain attract good number of bike booking.
- People book less number of shared bikes on holiday as compared to non-holiday’s.

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

Answer:It is important to use drop_first=True for a categorical variable with ’n’ level we can create ’n-1’ new columns each indicating whether that level exists or not using zero and one.Therefore using drop_first=true makes sure that resultant can match up n-1 levels thereby reducing the correlation among dummy variables.

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

Answer: The ‘temp’ and ‘atemp’ variables have highest correlation with the target variable ‘cnt’.

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

Answer:I have validated the assumptions of Linear Regression based on below features:

- Normality of error terms :Error terms should be normally distributed.
- Validating Linear Relationship :Linearity should be observed among variables.
- Validating Multi-collinearity :Using heatmap we can check and multicollinearity among variables should be insignificant.
- Validating Homoscedasticity :No visible patterns should in residual values.
- Validating for Independence of residuals: using the Durbin-Watson test we can check & there should be no auto-correlation.

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

Answer: Top 3 features contributing significantly towards explaining the demand of the shared bikes are :

- temperature
- year
- season

General Subjective Questions

1.Explain the linear regression algorithm in detail.

Answer: Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

Mathematically the relationship can be represented with the help of following equation –

$$Y = mX + c$$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slope of the regression line which represents the effect X has on Y

c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

A regression line can be :

- Positive Linear Relationship -A linear relationship will be called positive if both independent and dependent variable increases.

- Negative Linear Relationship -A linear relationship will be called positive if independent increases and dependent variable decreases.

Assumptions of Simple Linear Regression –

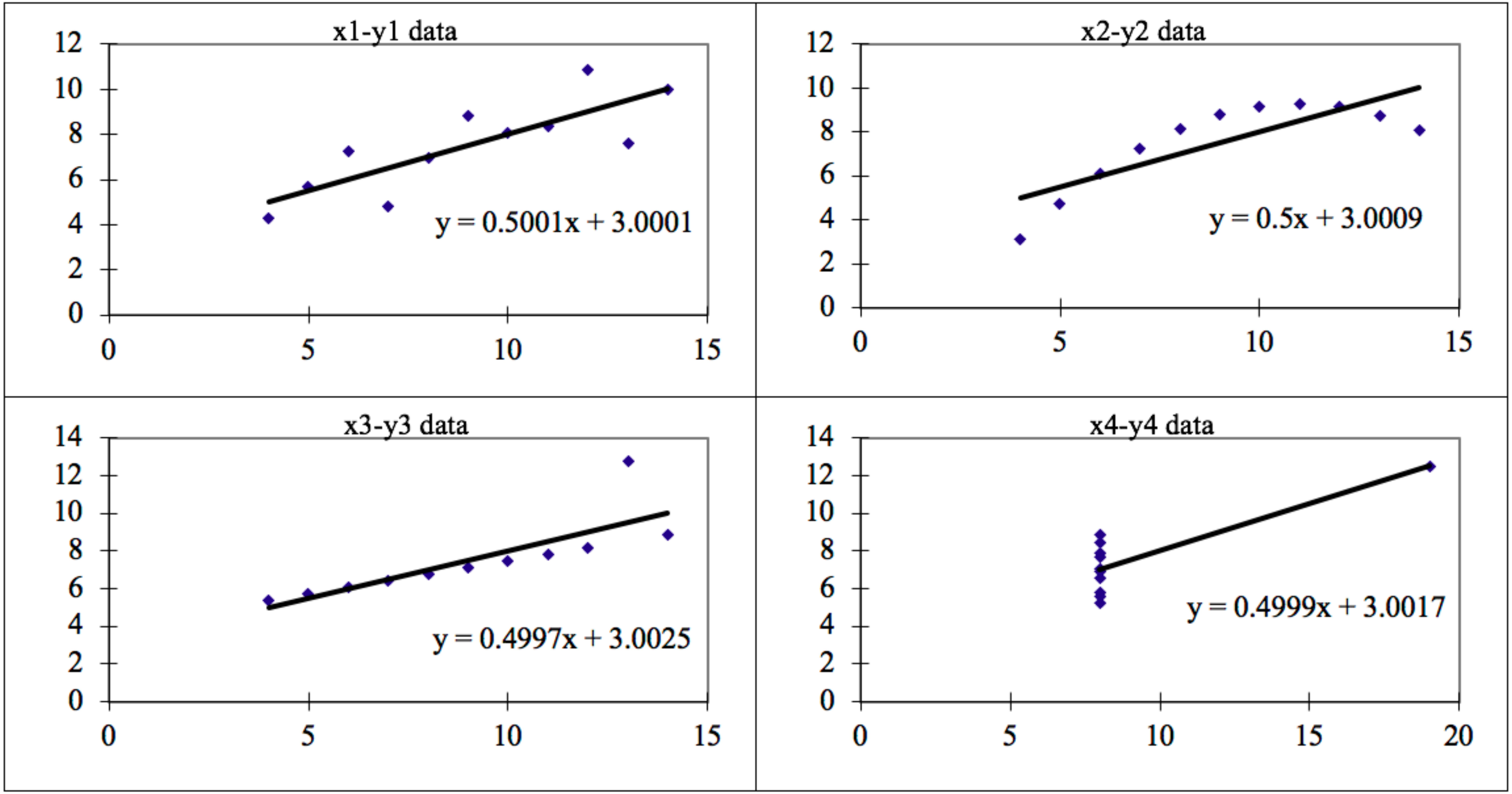
1. Linearity: The relationship between X and the mean of Y is linear.
2. Homoscedasticity: The variance of residual is the same for any value of X.
3. Independence: Observations are independent of each other.
4. Normality: For any fixed value of X, Y is normally distributed.

2. Explain the Anscombe's quartet in detail?

Answer: Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots. It was constructed to illustrate the importance of plotting the graphs before analyzing 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.

Graphs in next page:

Anscombe's Quartet



Anscombe’s quartet graph analysis:

- 1 st data set fits linear regression model as it seems to be linear relationship between X and y
- 2 nd data set does not show a linear relationship between X and Y , which means it does not fit the linear regression model.
- 3 rd data set shows some outliers present in the dataset which can’t be handled by a linear regression model.
- 4 th data set has a high leverage point means it produces a high correlation coeff.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3. What is Pearson’s R?

Answer: The Pearson correlation method is the most common method used for numerical variables. It assigns a value between – 1 and 1, where 0 is no correlation, 1 is total positive correlation, and – 1 is total negative correlation. Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson productmoment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations. Pearson’s R Formula is as follows: Here,

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

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

Answer: Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we will fit data in specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithm tends to weigh high values magnitudes and ignore other parameters which will result in incorrect modeling.

Difference between Normalizing Scaling and Standardize Scaling:

1. In normalized scaling minimum and maximum value of features being used whereas in Standardize scaling mean and standard deviation is used for scaling.
2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
3. Normalized scaling scales values between (0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
4. Normalized scaling is affected by outliers whereas standardized scaling is not having any effect by outliers.
5. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.
6. Normalized scaling is called as scaling normalization whereas standardized scaling is called as Z Score Normalization.

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

Answer:VIF(VarianceInflationFactor) basically helps explain the relationship of one independent variable with all the other independent variables.

The formulation of VIF is given below: A VIF value of greater than 10 is definitely high, a VIF of greater than 5 should also not be ignored and inspected appropriately.

A very high VIF value 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.

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

Answer :Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.

QQ plot can also be used to determine whether or not two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

Importance of QQ Plot in Linear Regression : In Linear Regression when we have a train and test dataset then we can create Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not. Advantages:

- It can be used with sample size also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot Q-Q plot use on two datasets to check
- If both datasets came from population with common distribution
- If both datasets have common location and common scale
- If both datasets have similar type of distribution shape
- If both datasets have tail behaviour.