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?

   There are 7 categorical variables in the dataset,

   a. Holiday , Wokingday, Weekday
      Visualization of these three categorical variables with number of bike rentals taken shows that,
      - Bike rentals are more on holidays ~4500, whereas its below 4000 on non-holidays.
      - Bike rentals are bit more on working days.
      - Bike rentals are almost same on all weekdays.

   b. Year
      Bike rentals have increased decently in year 2019, avg bike rentals were 4000 in year 2018 and it jumped to 6000 in year 2019.

   c. Months
      Bike rentals increased from Jan to Jun gradually and then it was almost constant in June, July, August, September, October and then gradually decreased in November and December.

   d. Seasons
      Bike rentals we low in spring (around 2000) whereas in summer and fall it increased to around 5000 and then slightly reduced to 4500 in winter.

   e. Weather situation:
      Bike rentals were high when weather was clear, decreased when there was moisture and decreased further when there were light rains.

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

   When a category variable is converted to dummies, one of the column can be inferred from all other dummy variables. One of the variable is dropped to avoid multicollinearity. Drop_first = True indicates to drop first dummy variable.

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

   Temperature variable has highest correlation with number of bike rentals.

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

   After creating a model using training data set, residuals are calculate and then distribution plot on that residuals are drawn. Mean of the residuals is zero and the shape of distribution is almost same as normal distribution.

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

   1. Temperature on that day. – Impacts positively.
   2. Weather situation light rains – Impacts negatively.
   3. Year – Rentals increased in year 2019.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

A linear regression is a supervised algorithm that learns to model a dependent variable y as a linear function of multiple independent variables that best fits the data.

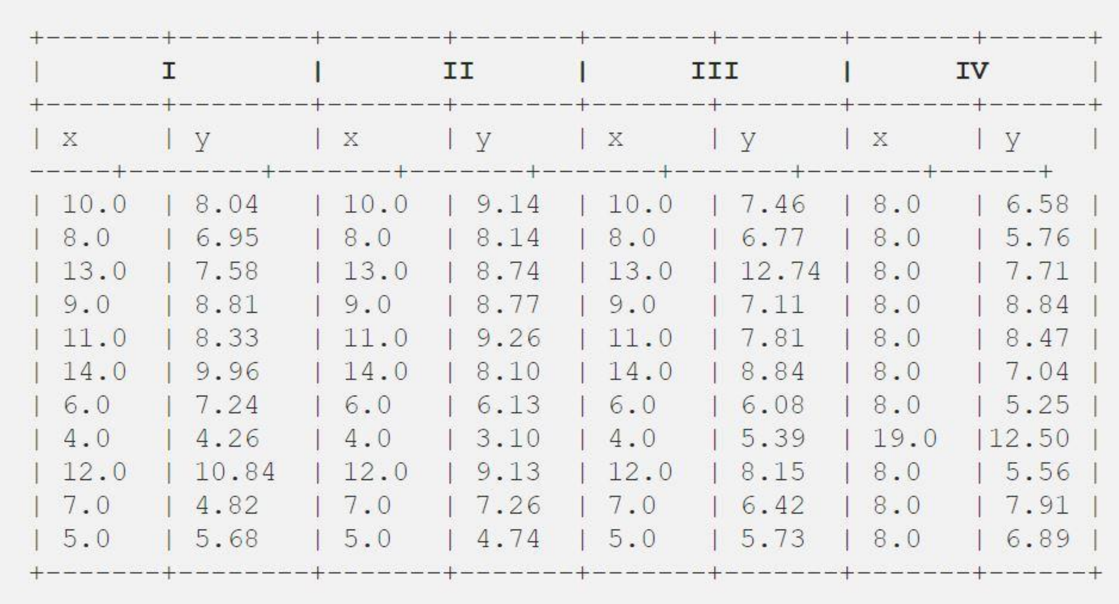
A multiple linear regression is expressed as

$yi=\beta 0+\beta 1Xi1+\beta 2Xi2+\cdots+\beta pXip+\varepsilon i,\quad i=1,\cdots,n,$

Below steps are followed normally to create a linear regression algorithm

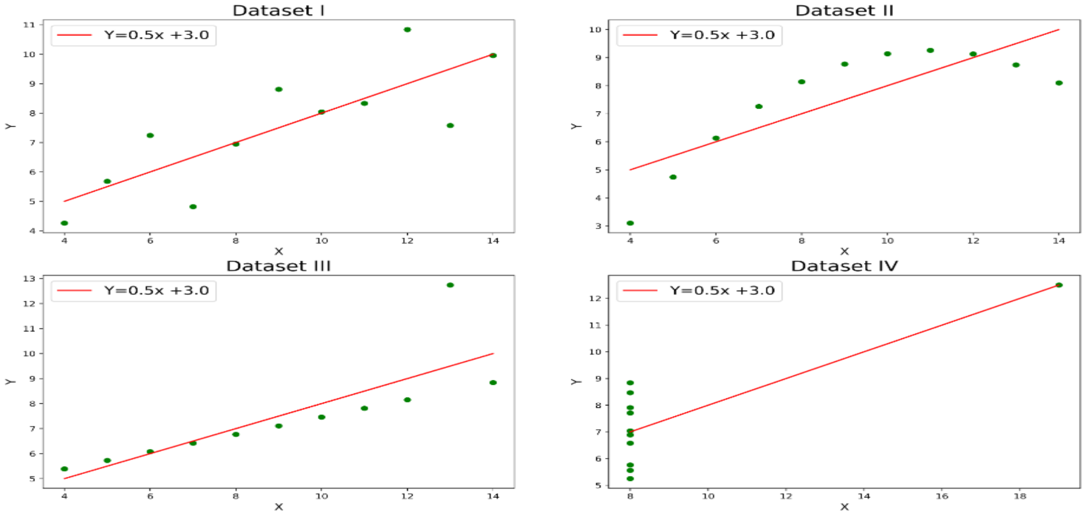
1. Reading, understanding and visualizing the data.
2. Preparing the data for modelling (train test split, rescaling)
3. Training the model.
4. Residual Analysis
5. Predictions and evaluation of the test set.

2. Explain the Anscombe's quartet in detail.
   Anscombe's quartet is a group of 4 data sets which have same mean, standard deviation, correlation, regression line slope and intercept but have different sets of data points.
   These data sets were created to demonstrate the importance of data visualization and show that summary of statistics alone could be misleading.
   Below is the list of quartets with different data sets.



```
+-------+--------+-------+-------+-------+-------+-------+------+
|       I        |       II      |      III      |      IV      |
+-------+--------+-------+-------+-------+-------+-------+------+
| x     | y      | x     | y     | x     | y     | x     | y    |
-----+--------+-------+-------+-------+-------+-------+------+
| 10.0  | 8.04   | 10.0  | 9.14  | 10.0  | 7.46  | 8.0   | 6.58 |
| 8.0   | 6.95   | 8.0   | 8.14  | 8.0   | 6.77  | 8.0   | 5.76 |
| 13.0  | 7.58   | 13.0  | 8.74  | 13.0  | 12.74 | 8.0   | 7.71 |
| 9.0   | 8.81   | 9.0   | 8.77  | 9.0   | 7.11  | 8.0   | 8.84 |
| 11.0  | 8.33   | 11.0  | 9.26  | 11.0  | 7.81  | 8.0   | 8.47 |
| 14.0  | 9.96   | 14.0  | 8.10  | 14.0  | 8.84  | 8.0   | 7.04 |
| 6.0   | 7.24   | 6.0   | 6.13  | 6.0   | 6.08  | 8.0   | 5.25 |
| 4.0   | 4.26   | 4.0   | 3.10  | 4.0   | 5.39  | 19.0  |12.50 |
| 12.0  | 10.84  | 12.0  | 9.13  | 12.0  | 8.15  | 8.0   | 5.56 |
| 7.0   | 4.82   | 7.0   | 7.26  | 7.0   | 6.42  | 8.0   | 7.91 |
| 5.0   | 5.68   | 5.0   | 4.74  | 5.0   | 5.73  | 8.0   | 6.89 |
+-------+--------+-------+-------+-------+-------+-------+------+
```

Though data sets are different, mean, variance and regression line remain same as shown in the figure below.



3. What is Pearson’s R?

Pearson correlation coefficient denoted by ‘r’ is correlation coefficient that measures the relation between two sets of data.

It is calculated using

r = Sigma(Xi – Xm)*Sigma(Yi -Ym)/SQRT (Sigma(Xi-Xm)^2) * SQRT(Sigma(Yi-Ym)^2)

Where Xi,Yi represent each element in the data set and Xm, Ym represents mean of the data.

Correlation coefficient ranges from -1 to 1.

- If correlation coefficient is negative, it implies that when one variable increase other decreases or vice versa.
- If correlation coefficient is positive, it implies that when one variable increases or decreases other variable will also increase or decrease proportionately.
- If correlation coefficient is zero, that means these two variables are not related at all.

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

Scaling is a method to normalize the range of independent variables or features of a data set.

Scaling is performed to bring all features to a similar scale so that no single feature dominates the learning algorithm. Scaling ensures that the features contribute equally to the models performance.

Normalized scaling is scaling between max and min values of the data set. Maximum value is represented as 1 and minimum value is represented as 0.

Normalized scaling : Xi – Xmin/Xmax - Xmin

Standardized scaling is done using mean of the sample and standard deviation. It is calculated as

Xi – u/sigma, where u is mean and sigma represents standard deviation.

In normalized scaling , range always lies between 0 and 1 even if there are outliers. Whereas in standardized scaling range varies as per the distribution.

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

Infinite VIF was observed for variables like holiday, workingday, weekday_Monday, weekday_Tuesday. These variables had perfect correlation with other predictor variables.

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

   The quantile-quantile(Q-Q) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or it follows some other distribution.

   In linear progression it is assumed that error terms are normally distributed. Q-Q plot can be be used to confirm the errors are following normal distribution. When Q-Q plot is drawn for residuals of train data and test data they should appear similar.