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? (3 marks)

Ans. I have found that when the weather is **Clear, Few clouds, Partly cloudy or Partly cloudy** there are more registrations for Bike-sharing and amongst all the seasons the registration is more when it is **fall**

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

Ans. It is important because any non-ordinal categorical variables can be described in at least n-1 columns where n is the size of the sample. Removing redundant information (multicollinearity) and providing precise variables to the model will help in better interpretation and the background gradient decent variable will also work more efficiently.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans. By looking at the pair-plot it can be observed that atemp (temperature) has the highest correlation with the target variable

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Ans. I validate the linear regression by verifying my below assumptions

- There is at least one coef which is not zero
- Residuals have mean zero and it is normally distributed
- Residuals are independent of each other
- Homoscedasticity of residuals
- No perfect multicollinearity

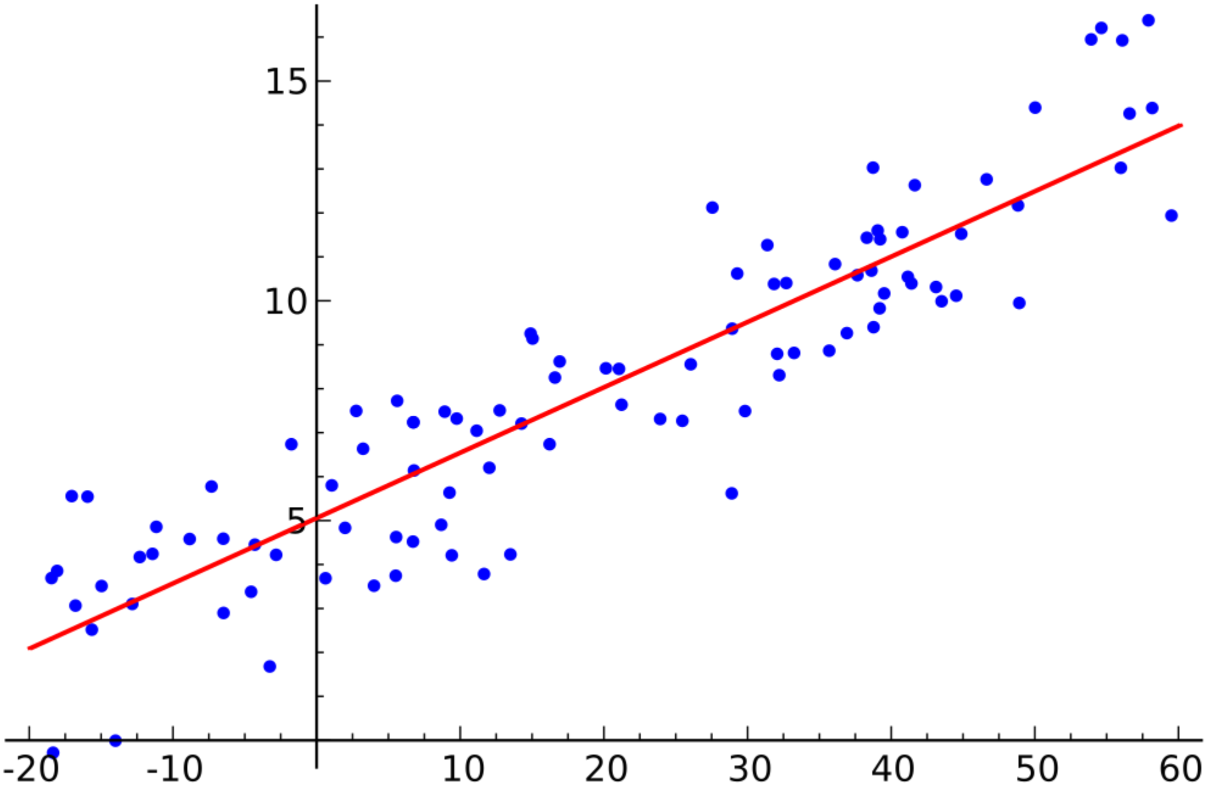
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans The top 3 features are yr (Year), holiday and spring

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Ans Linear regression is a statistical model used to estimate the relationship between the dependent variable and the independent variable by fitting a line through the data. The most common method of fitting a line is called the least square method. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). This deviation is also known as error.



As the above diagram explains data points are blue and the regression line is red.

In the case of a model with p explanatory variables, the OLS regression model writes:

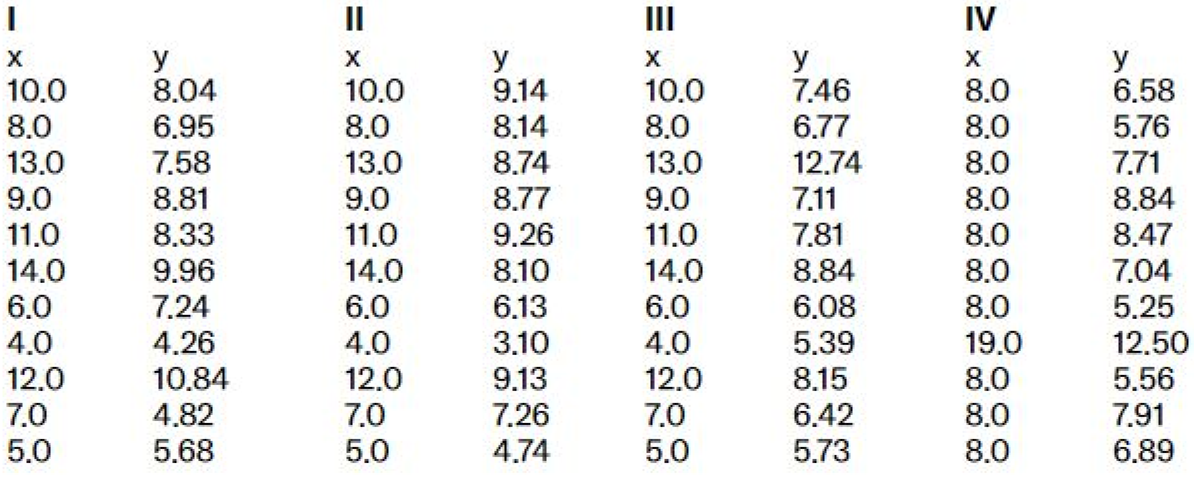
$Y = \beta 0 + \Sigma j=1..p \ \beta jXj + \varepsilon$

The OLS method corresponds to minimizing the sum of square differences between the observed and predicted values. This minimization leads to the following estimators of the parameters of the model:

[β = (X’DX)-1 X’ Dy σ^2 = 1/(W –p*) Σi=1..n wi(yi - yi)] where β is the vector of the estimators of the βi parameters, X is the matrix of the explanatory variables preceded by a vector of 1s, y is the vector of the n observed values of the dependent variable, p* is the number of explanatory variables to which we add 1 if the intercept is not fixed, wi is the weight of the ith observation, and W is the sum of the wi weights, and D is a matrix with the wi weights on its diagonal.The motive of this OLS algorithm is to minimize the sum of squares also known as errors.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Ans We can’t completely rely on summary statistics of data to understand it. We must also use graphs to understand different interpretations that can be derived from data. Anscombe’s quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.



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

All the summary statistics you’d think to compute are close to identical:

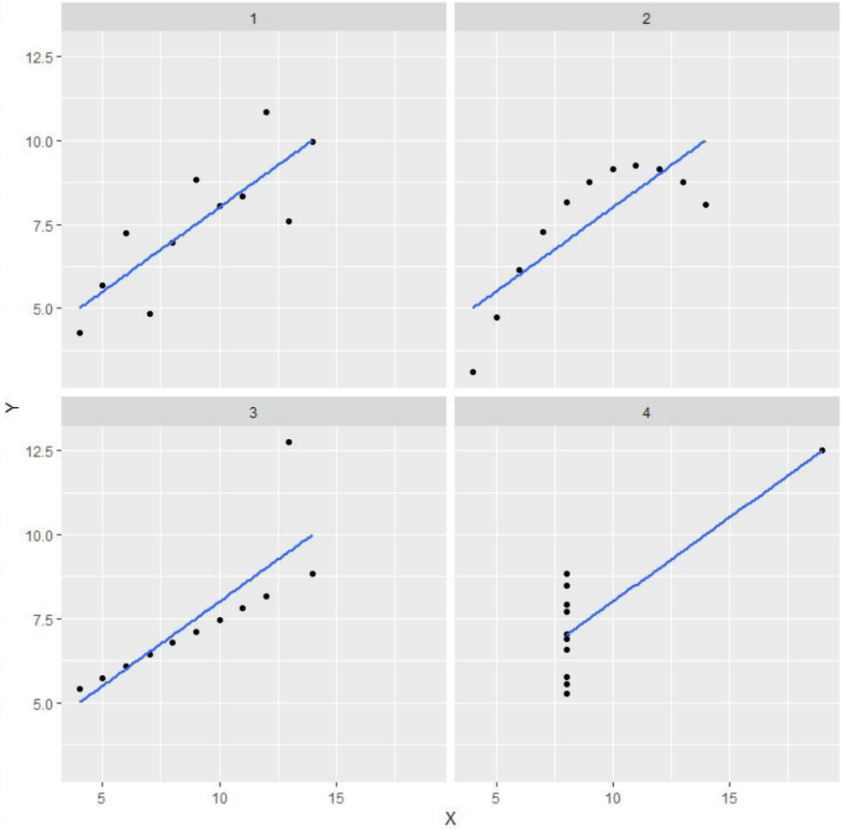
The average x value is 9 for each dataset

The average y value is 7.50 for each dataset

The variance for x is 11 and the variance for y is 4.12

The correlation between x and y is 0.816 for each dataset

A linear regression (line of best fit) for each dataset follows the equation y = 0.5x + 3So far these four datasets appear to be pretty similar. But when we plot these four data sets on an x/y coordinate plane, we get the following results:

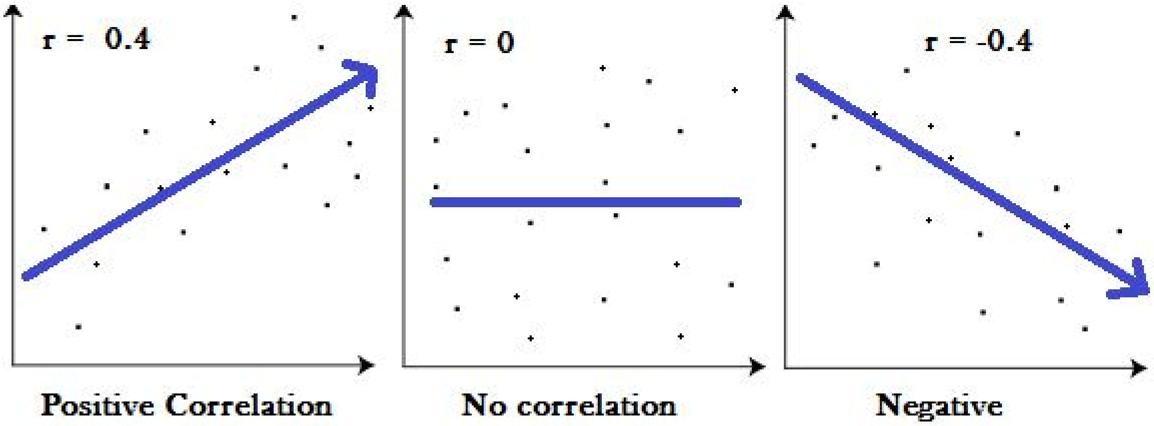


3. What is Pearson’s R? (3 marks)

Ans Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of the correlation coefficient, but the most popular is Pearson’s. Pearson’s correlation (also called Pearson’s R) is a correlation coefficient commonly used in linear regression.

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



One of the most commonly used formulas in stats is Pearson's correlation coefficient formula

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Machine learning algorithms like linear regression, logistic regression, neural network, etc. that use gradient descent as an optimization technique require data to be scaled. Take a look at the formula for gradient descent below:

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

The presence of feature value X in the formula will affect the step size of the gradient descent. The difference in ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that

the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling. Here’s the formula for standardization:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation. Here’s the formula for standardization:

$$X' = \frac{X - \mu}{\sigma}$$

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans Variance inflation factors show the degree to which a regression coefficient will be affected because of the variable's redundancy with other independent variables. As the squared multiple correlations of any predictor variable with the other predictors’ approach unity, the corresponding VIF becomes infinite.

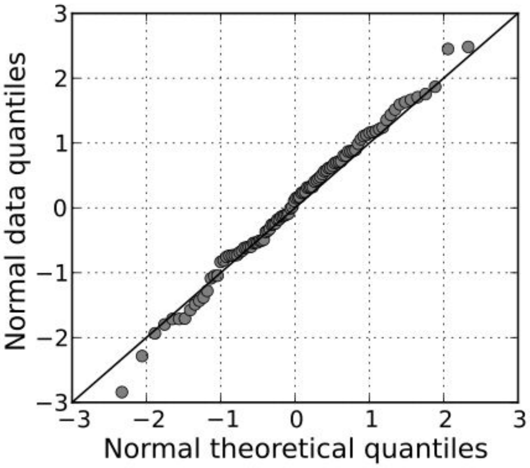
For any predictor orthogonal (independent) to all other predictors, the variance inflation factor is 1.0. VIF thus provides us with a measure of how many times larger the variance of the ith regression coefficient will be for multicollinear data than for orthogonal data. An advantage of knowing the VIF for each variable is that it gives a tangible idea of how much of the variances of the estimated coefficients are degraded by the multicollinearity.

If the VIF value of a variable is 10 or greater, it can be removed from the analysis as it is too much collinear as compared to other variables and redundant information. If the VIF of a variable is between 5 to 10 then it should be thoroughly inspected and decided with the help of business knowledge whether that variable is important or not. Any variable with less than 5 can be considered for further analysis.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

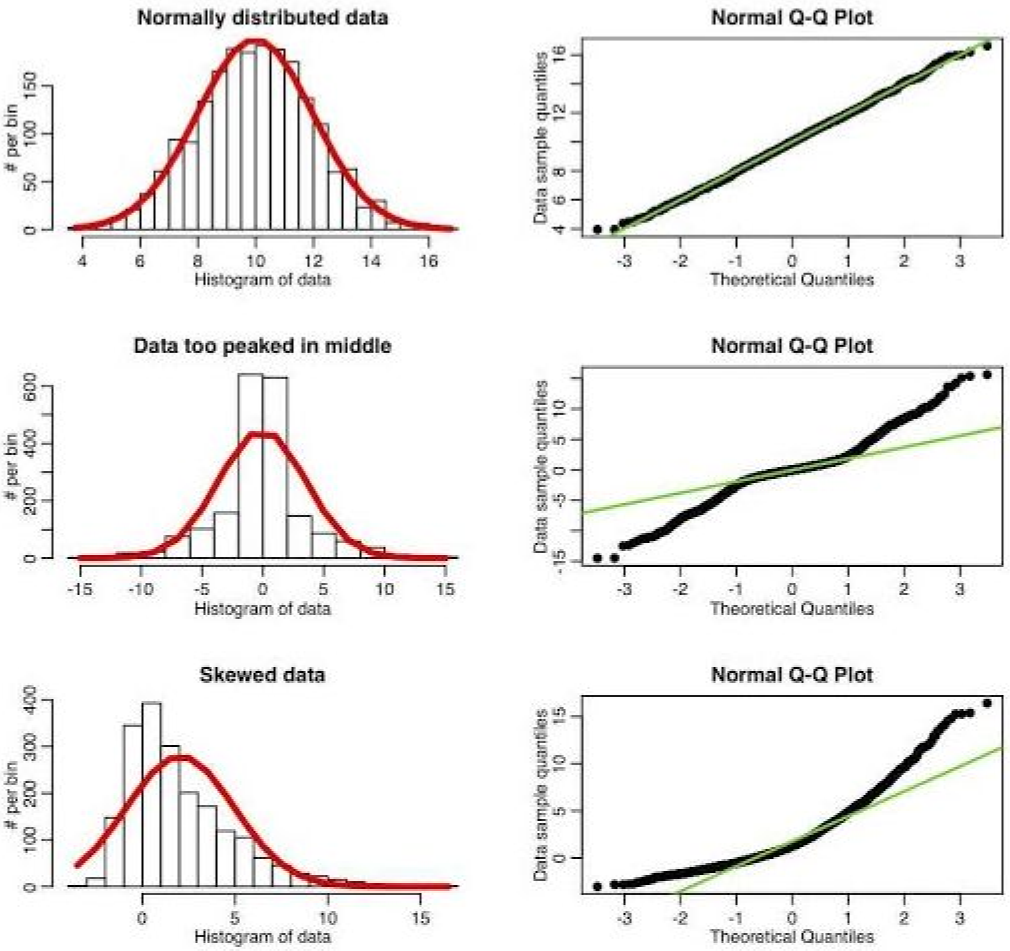
Ans Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.

We plot the theoretical quantiles or basically known as the standard normal variate (a normal distribution with mean=0 and standard deviation=1)on the x-axis and the ordered values for the random variable which we want to find whether it is Gaussian distributed or not, on the y-axis. Which gives a very beautiful and a smooth straight line like structure from each point plotted on the graph.



Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc. You can tell the type of distribution using the power of the Q-Q plot just by looking at the plot.

Below are the types of output that you can expect.


Normally distributed data
# per bin
0
50
100
150
4
6
8
10
12
14
16
Histogram of data
Normal Q-Q Plot
Data sample quantiles
4
6
8
12
16
-3
-2
-1
0
1
2
3
Theoretical Quantiles
Data too peaked in middle
# per bin
0
200
400
600
-15
-10
-5
0
5
10
15
Histogram of data
Normal Q-Q Plot
Data sample quantiles
-15
-5
0
5
10
-3
-2
-1
0
1
2
3
Theoretical Quantiles
Skewed data
# per bin
0
100
200
300
400
0
5
10
15
Histogram of data
Normal Q-Q Plot
Data sample quantiles
0
5
10
15
-3
-2
-1
0
1
2
3
Theoretical Quantiles