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)

1. “season” seems to be affecting our ”cnt” => “season” : “summer” and “fall” have center mass high
2. “yr” seems to be affecting our”cnt” => “yr” 2019 have increased “cnt” from 2018
3. “mnth” seems to be affecting our “cnt” => middle month of the year seems to have high “cnt” like “apr, may, jun, july, aug, sept”
4. “holiday” we don't see much impact on “cnt”
5. “weekday” no impact on “cnt
6. “workingday” no impact on “cnt
7. “weathersit” seems to be affecting our “cnt” => “weathersit” : “lightsnow” have decrease our “cnt”

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

When we have 4 categorical variable, if we use drop_first=True, then we **can reduce one column** instead of 4 column.

e.g.

House type : un-furnished, semi-furnished, fully-furnished

Lets say we have 3 house, A(un-furnished), B(semi-furnished), C(fully-furnished)

To represent this categorical column if we can add 3 dummy columns like this

A - 1 - 0 - 0

B - 0 - 1 - 0

C - 0 - 0 - 1

instead of adding 3 column we can remove the first column if both other column in 0 that it indicates the first column

A - 0 - 0

B - 1 - 0

C - 0 – 1

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

"temp" column has higest correlation with target column 'cnt'

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

Assumption of 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.

- We can draw distribution graph of residual(y_train-y_test) and make sure the mean is 0 and the graph is normalized distributed**.**
- We can also draw X_train vs residual to make sure the residual Is spread across the graph.

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

Top 4 contributing features:

1. Snow => -2604
2. 2091 =>
3. windspeed => -1422
4. jan => -1429

- In the year 2019, we have positive correlation with the target variable.
- This indicates as the snow increase the use of cycles decrease.
- As the windspeed increase, the use of cycles again decrease.
- In the january month, the use of cycles again decreases.

General Subjective Questions

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

Linear Regression Algorithm is a machine learning algorithm based on supervised learning. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Mathematically, we can write a linear regression equation as:

y = mx + c

where m => slope and c => y-intercept

Assumption in 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.

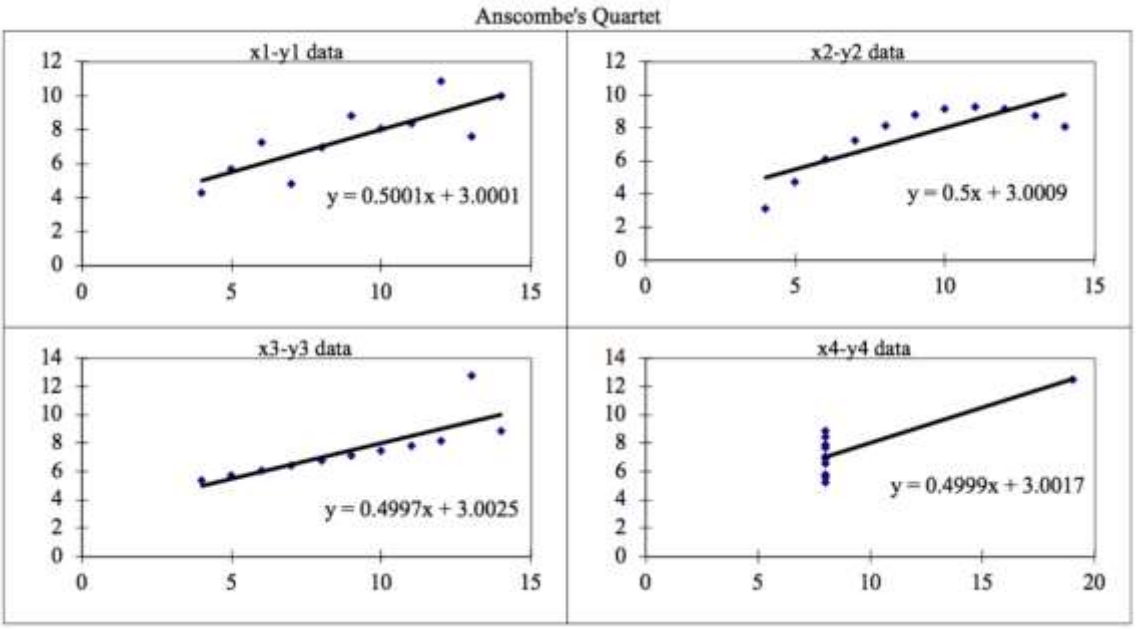
2. Explain the Anscombe's quartet in detail. (3 marks)

It was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties.

Anscombe's Data									
Observation	x1	y1	x2	y2	x3	y3	x4	y4
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.1	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.1	4	5.39	19	12.5
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89
			Summary Statistics					
N	11	11	11	11	11	11	11	11
mean	9.00	7.50	9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94	3.16	1.94	3.16	1.94	3.16	1.94
r	0.82		0.82		0.82		0.82	

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.

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.



Dataset 1: this fits the linear regression model pretty well.

Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.

Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model

Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

3. What is Pearson's R? (3 marks)

In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.[1] First, the set of intervals for the quantiles is chosen

Formula

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

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

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

What is scaling:

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.

Why is scaling performed?

1. if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model.

2. Another reason why feature scaling is applied is that few algorithms like Neural network gradient descent converge much faster with feature scaling than without it.

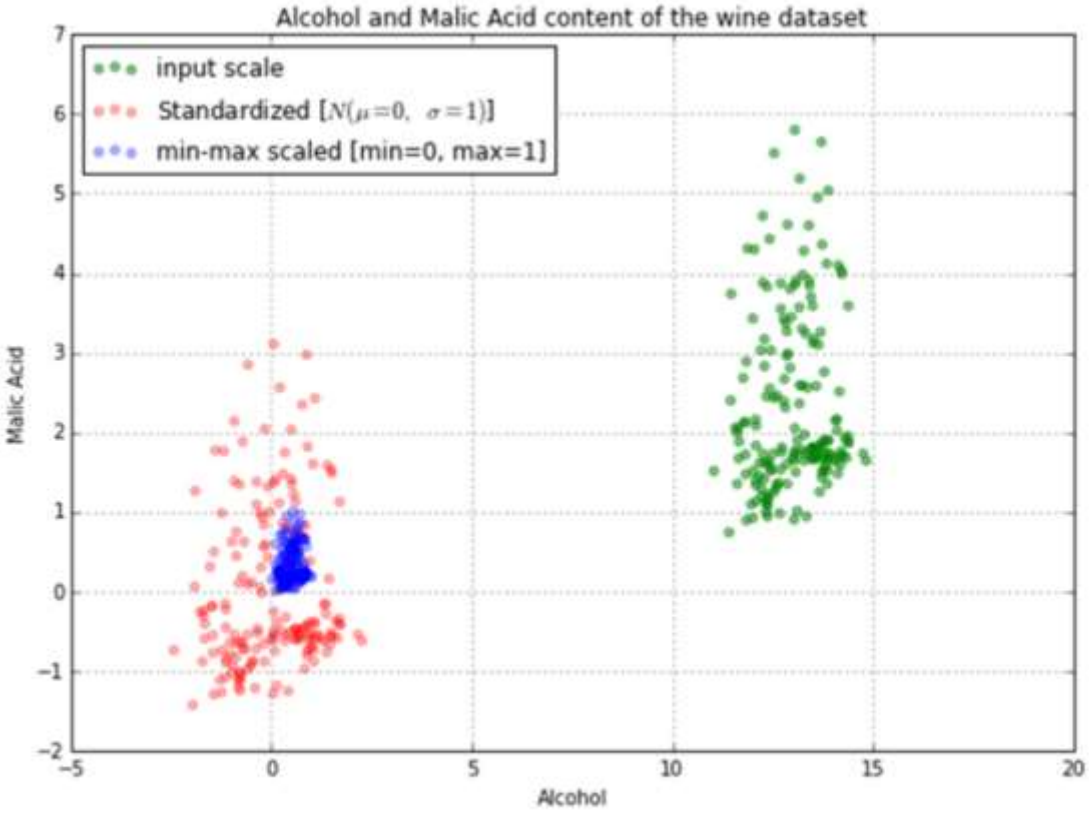
What is the difference between normalized scaling and standardized scaling?

normalized scaling:

min-max scaling or min-max normalization, it is the simplest method and consists of rescaling the range of features to scale the range in [0, 1].

standardized scaling:

Feature standardization makes the values of each feature in the data have zero mean and unit variance.



The impact of Standardization and Normalisation on the Wine dataset

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

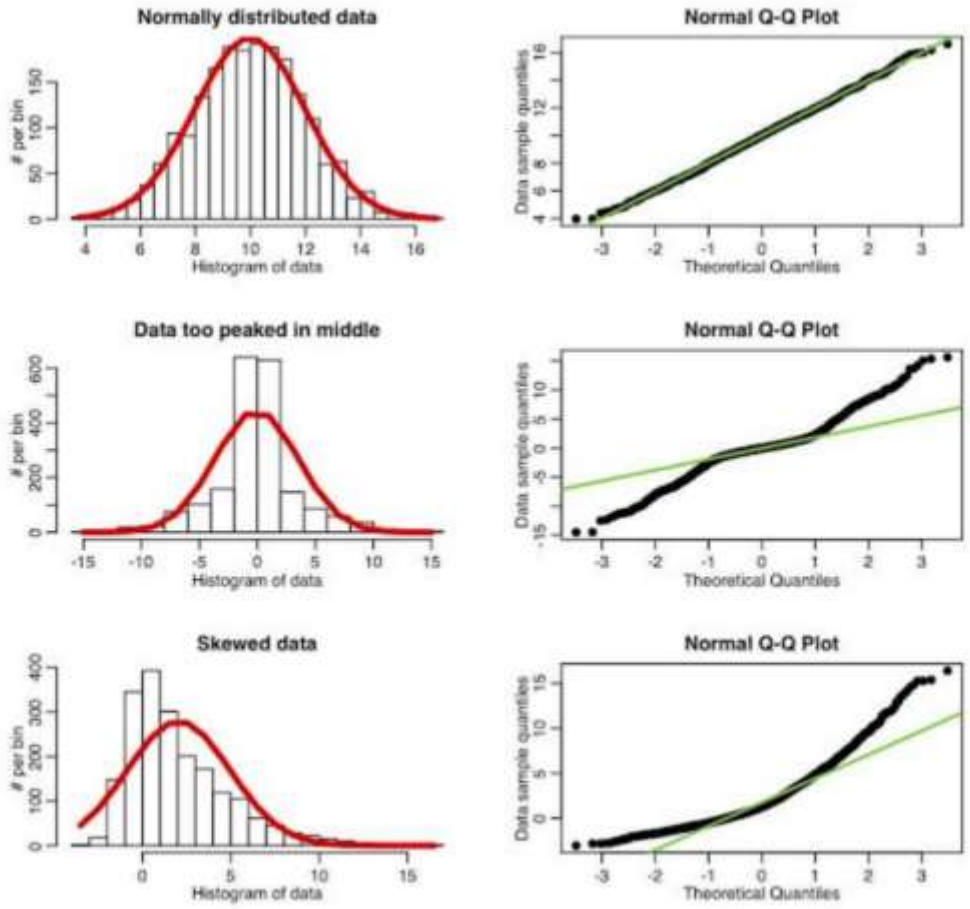
(3 marks)

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

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

(3 marks)

In statistics, a Q–Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.[1] First, the set of intervals for the quantiles is chosen



Source: Sherrytowers Q-Q plot examples