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:

- Spring Season having lesser demand compared to other season and Fall season had more demand
- 2019 had more demand compared to 2018. Demand is growing over the year
- AUG, SEP, OCT months had high demand and demand increase gradually from JAN to SEP and then drops till end
- More demand observed on the non-holiday. Demand is less on holiday most people not using the services during holiday
- More demand observed on the clear weather situation. Demand falls on light snow whether situation. Cloudy weather situation has moderate demand
- Working day and non-working day has almost similar demand

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

Ans:

it is essential to handle these categorical variables appropriately in order to get a good model. One way to deal with them is creating dummy variables. The key idea behind creating dummy variables is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one.

Example: Season has 4 values (summer, winter, spring, fall)

drop_first=False creates the following table

summer	winter	spring	fall
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Summer can be explained by all zero in winter, spring and fall. So, we don't need column summer.

by using drop_first=True it drops first value and automatically create n-1 dummy variables. So, we don't have manually drop one for each categorical variable.

On using drop_first=True following dummies created

winter	spring	fall
0	0	0
1	0	0
0	1	0
0	0	1

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

Ans: Temp variable 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:

- Linear Relationship Between X and Y
  From the pair plot we can observe there is clear linear relationship between temp and demand
- Error terms are normally distributed
  Upon plotting error terms histogram we are able to observe error terms follow the normal distribution
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)
  When we plot a scatterplot y_pred against error we can observe there is no dependency between them and no visible patterns

- Multicollinearity
  Using VIF verified there is no multicollinearity and dropped all features having >5 VIF
- Overfitting

Model validated against the test data set to observed r^2(0.791), adjusted r^2(0.778) and MSE are still good which states no overfitting in the mode

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:

- temp
- weather situation Light Snow
- year 2019

General Subjective Questions

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

Ans:

Linear Regression is a supervised machine learning model helps to predict a continuous variable based on independent variable.

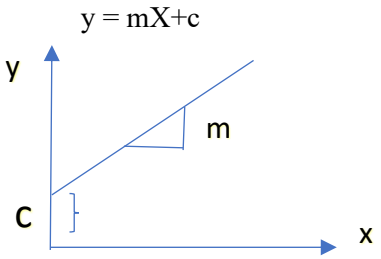
There are two types of linear regression

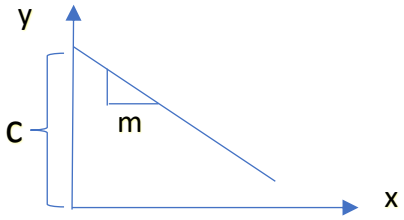
1. Simple Linear Regression
2. Multiple Linear Regression

Simple Linear Regression

Goal of the simple linear regression model a best fitting line that explains the linear relationship between the target variable and the independent variable by reducing the error between actual and predicted value to minimal.

It is explained by the equation





Positive relation Ship

Negative relationship

y is the outcome variable

X is independent variable

m is the slope

Assumptions:

1. Linear Relationship Between X and Y
2. Error terms are normally distributed
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

Multiple Liner Regression:

Multiple linear regression is an extension of the simple linear regression where best fitting relationship created to predict outcome variable based on multiple independent variables.

It explained by the equation

$$y = C+m_1X_1+ m_2X_2+m_3X_3 +\ldots+m_nX_n$$

Assumption as same as the simple linear regression.

1. Model now fits a 'hyperplane' instead of a line
2. Coefficients still obtained by minimizing sum of squared error (Least squares criterion)

Additionally, we need to consider the multi-collinearity and overfitting

multi-collinearity

model built using multiple independent variables, some of these variables are corelated.

Because of this co-efficient swing widely and p-values are not reliable

overfitting

Model fits too well to training data and doesn't generalize as a result accuracy drops when the evaluating against test data

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

Ans:

Anscombe's quartet is a group of datasets (x, y) that have the same mean, variance, standard deviation, and regression line, but which are qualitatively different.

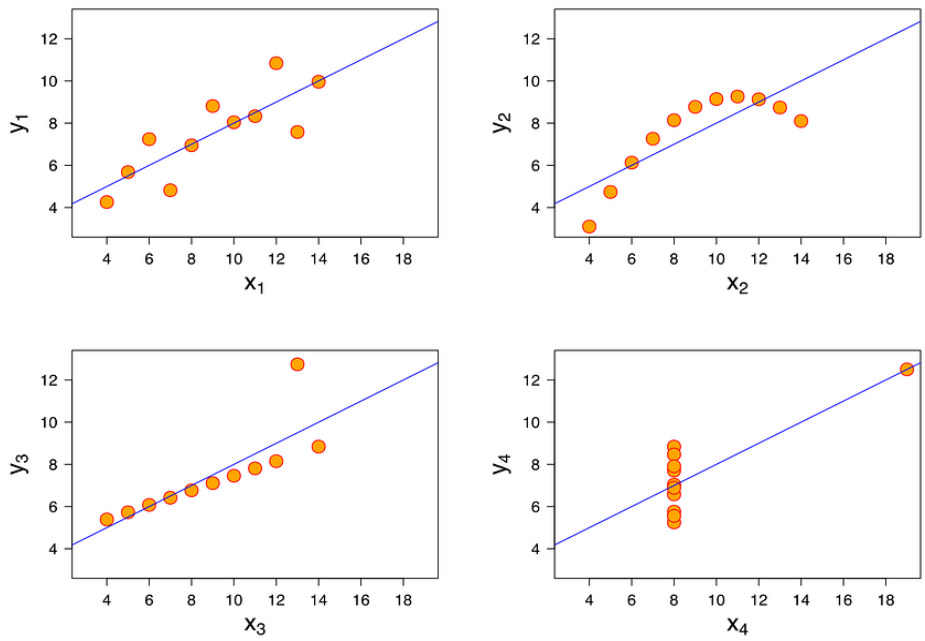
The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Data set and Visualization:

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	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
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03



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

Ans:

Pearson's r in other words Pearson's correlation co-efficient, is the measure determines strength of the linear relationship between the two variables. It can be in range of -1 to 1

Sign defines which type of relationship

-1 perfect negative relationship

i. when one value goes up other goes down

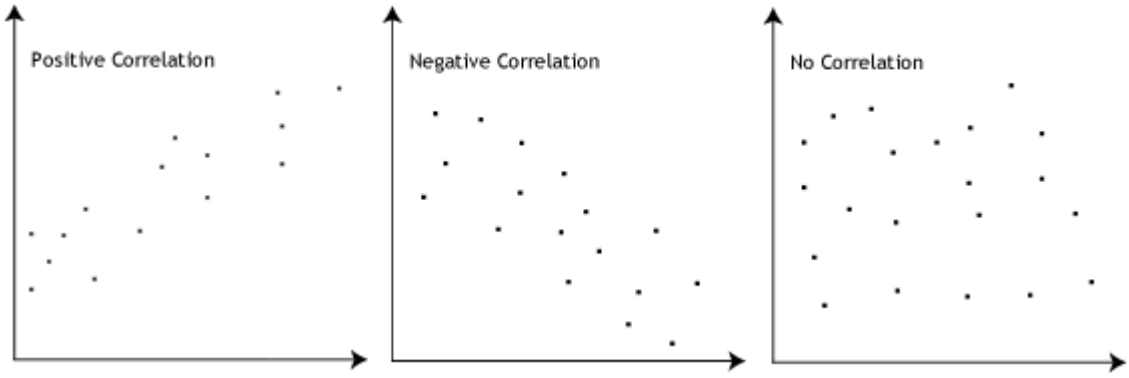
ii. when one value goes down other goes up

1 perfect positive relationship

i. when one value goes up other goes up

ii. when one value goes down other goes down

0 no relationship – change in one variable doesn't affect other variable



- 0.00-0.19: Very Weak
- 0.20-0.39: Weak
- 0.40-0.59 : Moderate
- 0.60-0.79 : Strong
- 0.80-1.0 : Very strong

The co-efficient r can be calculated from the below 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)

Ans:

When you have a lot of independent numerical variables. Some of them on the different scale. Which might lead weird co-efficient that might be difficult to interpret. Scaling need for the below reasons:

1. Faster convergence for gradient descent methods when they are the same scale
2. Ease of interpretation - Scaling makes it easier to compare the importance of different features in a model, particularly when interpreting coefficients or feature importance.
3. Avoiding Dominance – features with higher scale can dominate the lower scale variables during learning scale make it to give equal importance

There are two types of scaling

1. Standardizing: The variables are scaled in such a way that their mean is zero and standard deviation is one. Maintains the shape of distribution less sensitive to outliers

   $X_{standardized} = X - mean(X) / std(X)$

2. MinMax Scaling: The variables are scaled in such a way that all the values lie between zero and one using the maximum and the minimum values in the data. It preserves the relationship but may be sensitive to outliers

   $X_{normalized} = X - min(X) / max(X) - min(X)$

Difference between Standardizing and MinMax Scaling

1. Min Max Scaling transform with in range 0 to 1, but Standardizing does not impose any particular range
2. Min Max scaling sensitive to outliers and standardizing not sensitive to outliers
3. Min Max scaling perserves original distribution so good for interpretability and Standardizing centers around zero useful when comparing different unit or algorithm sensitive to scale

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

Ans:

multi-collinearity happens when model built using multiple independent variables, some of these variables are corelated. Because of this co-efficient swing widely and p-values are not reliable.

We can use scatter plot and heat map to see the correlation against one variable. But it is possible that just one variable might not completely explain some other variable but some of the variables combined might be able to do that. To check this sort of relations between variables, we use VIF.

VIF basically helps explaining the relationship of one independent variable with all the other independent variables.

$VIF(X_i) = 1/1 - R_i^2$

where R_i^2 is the R_i^2 value obtained by linear regression target variable X_i against all the other predictor variables.

VIF will be infinite when R_i^2 almost near 1 which makes VIF equation

VIF = 1/0

This happens there is perfect linear relationship between target variable X_i against all the other predictor variables.

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

Ans:

The quantile-quantile plot is a graphical method for determining whether two samples of data came from the same population or not. It helps to assess if two datasets came from some theoretical distribution such as a Normal, exponential or Uniform distribution

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value.

That is the 0.3 (or 30%) quantile is a point at which 30% of data falls below and 70% data falls above

For the reference purpose, a 45% line is also plotted, if the samples are from the same population, then the points are along this line. Further the distance it confirms they are from the different population or distribution.

The Quantile-Quantile plot is used for the following purpose:

- Determine whether two samples are from the same population.
- Whether two samples have the same tail
- Whether two samples have the same distribution shape.
- Whether two samples have common location behaviour.

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.

Importance of q-q plot in linear regression

One of the important assumptions in linear regression is error terms are linearly distributed. We can use the q-q plot to validate error terms follow normality. Q-Q plots help detect skewness, heavy tails, and outliers in the residuals. Non-normality or the presence of outliers can impact the validity of statistical inferences drawn from the regression model.