Title of the assignment : Predict the price of the Uber ride from a given pickup point on the agreed drop-off location.

Dataset Description : The project is about on world's largest taxi company Uber inc. In this project, we're looking to predict the fare for their future transactional cases.
Uber delivers service to lakhs of customers only. Now it becomes really important to manage their data properly to comeup with new business ideas to get best results.

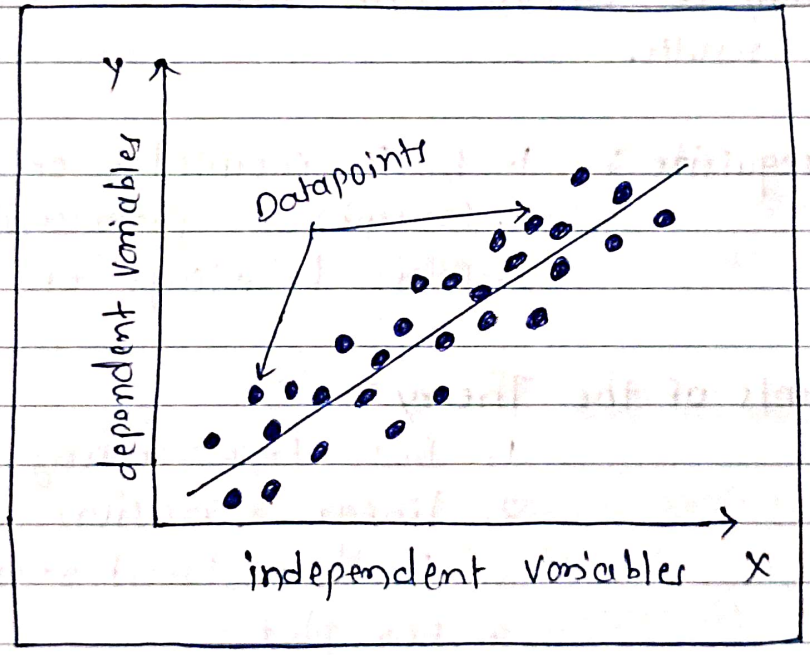
Prerequisite :
1. Basic knowledge of Python.
2. Concept of preprocessing data.
3. Basic knowledge of Data Science.

Contents of the Theory :
1. Data preprocessing
2. Linear regression.
3. Random forest regression models.
4. Box plot
5. Outliers
6. Haversine
7. matplotlib.

Data preprocessing : Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data.

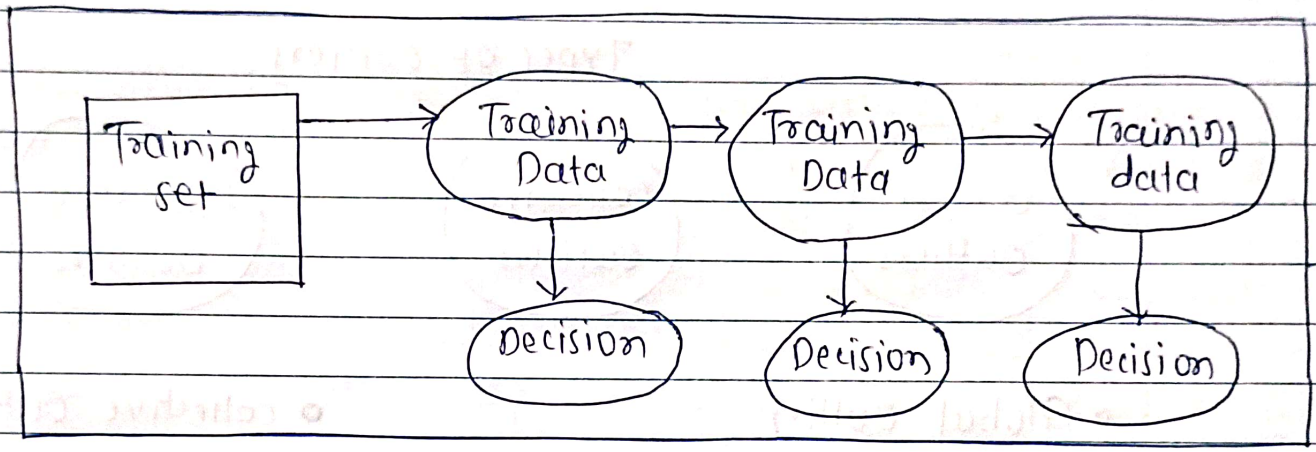
Linear Regression: Linear regression is one of the easiest and most popular machine learning algorithm. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.



Random Forest Regression models:

Random forest is a popular machine learning algo. that belongs to the supervised learning technique. It can be used for both classification & regression problems in ML.

It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.



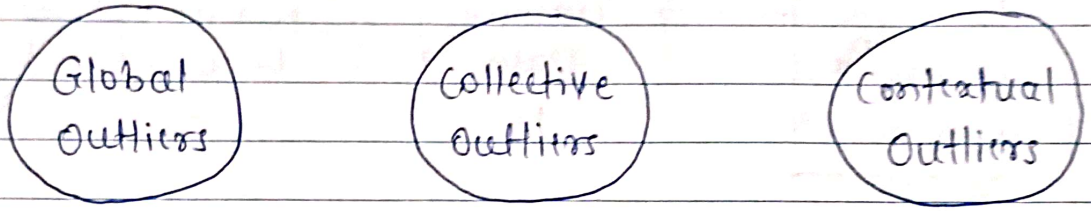
R improves a boxplot() function to create a boxplot. There is following syntax of boxplot () function.

boxplot (x, data, notch, varwidth, names, main).

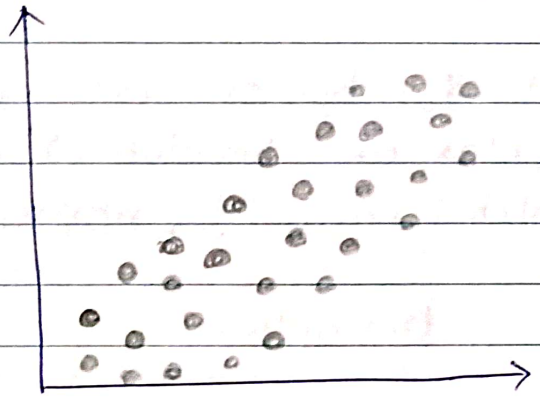
Sr.	Parameter	Description
1.	x	It is a vector or a formula.
2.	data	It is the data Frame.
3.	notch	It is a logical valueset as true to draw.
4.	varwidth	It is also a logical valueset as true to draw the width of the boxsame as the sample size.
5.	names	It is enough group of labels that will be printed under each boxplot.
6.	main	It is used to give a little.

Outliers : As the name suggests, "outliers" refer to the data points that exist outside of what is to be expected. The major thing about the outliers is what you do with them.

Types of outliers

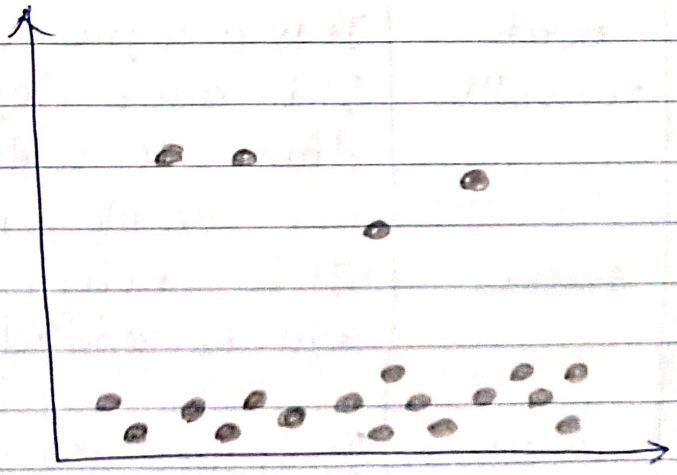


- Global Outlier



- collective Outlier

- Contextual Outlier



Haversine :- The haversine formula calculates the shortest distance between two points on a sphere using their latitude and longitudes measured along the surface. It is important for use in navigation.

Matplotlib :- Matplotlib is an amazing visualization library in python For 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on Numpy arrays and designed to work with the broader scipy stack.
It was introduced by John Hunter in the year 2002.

mean squared Error :- The mean squared Error (MSE) or mean squared deviation of an estimator measures the average of error squares i.e. the average squared difference between the estimated values and true values.

- It is a risk function, corresponding to expected value of the squared error loss.

Conclusion :-
In this way we have explored concept correlation and implement linear regression and random forest regression models.

Title :- classify the email using the binary classification method. Email spam detection has two states :
1. Normal state - Not spam
2. AbNormal state - spam

- Use K-Nearest Neighbors and support vector machine classification. Analyze their performance.

Dataset Description :- The csv file contains 5172 rows, each row for each mail. There are 3002 columns. The first column indicates Email names. The name has been set with numbers and not recipients names to protect privacy. The last columns has the tables for prediction.

objectives :- student should be able to classify email using the binary classification and implement email span detection technique by using K-nearest neighbors and support vector machine algorithm.

Prerequisite :-
1. Basic knowledge of python.
2. Concept of K-nearest Neighbors & support V. machine.

Concept of Theory :-
1. Data preprocessing
2. Binary classification
3. K-nearest Neighbors
4. support vector machine, train, test & split procedure.

Data preprocessing : Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model when creating a machine learning project, it is not always a case that we come across the clean and formatted data. While doing any operation with data, it is mandatory to clean it and put in a formatted way, so for this, we use data pre-processing task.

A real-world data generally contains noises, missing values and may be in an unusable format which cannot be directly used for machine learning model. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increase the accuracy and efficiency.

- finding missing data.
- Encoding categorical Data
- splitting dataset into training and test set.
- Feature scaling.

Conclusion :-
In this way we have explore concept correlation and implement linear regression, K-nearest neighbors and support vector machine algorithm.

Title : Given a bank customer, build a neural network based classifier that can determine whether they will leave or not in the next 6 month.

Dataset Discription : The case study is from an open-source dataset from Kaggle. The dataset contains 10,000 sample points with 14 distinct features such as customer, credit score, Geography, Gender, Age, Tenur, Balance, etc.

Perform the following steps :

1. Read the dataset
2. Distinguish the feature and target-set and divide the dataset into training and test-sets.
3. Normalize the train and test data.
4. Initialize and build the model. Identify the points of improvement and implement the same.
5. Print the accuracy score and confusion matrix.

Objectives of assignment :

Students should be able to distinguish the features and target set and divide the data set into training and test sets and normalize them and students should build the model on the basis of that.

Prequisite : 1. Basic of knowledge of Python.
2. Concept of confusion matrix.

TSSM
Page No.

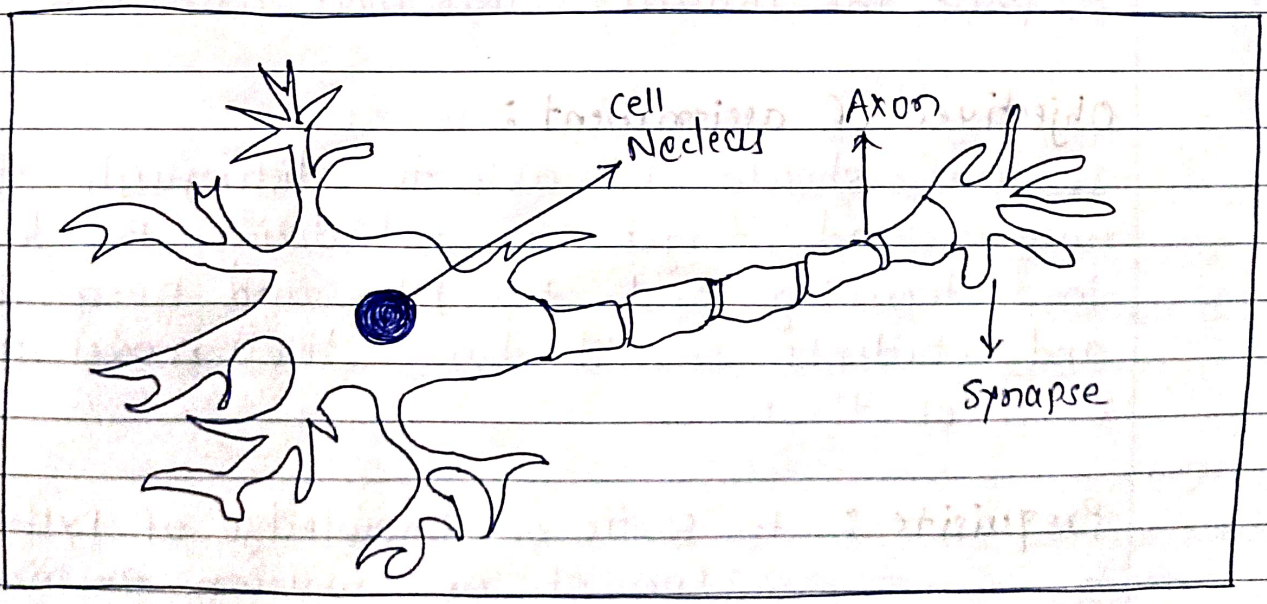
Contents of Theory :

1. Artificial Neutral Network.
2. Keras
3. tensorflow
4. Normalisation
5. Confusion matrix

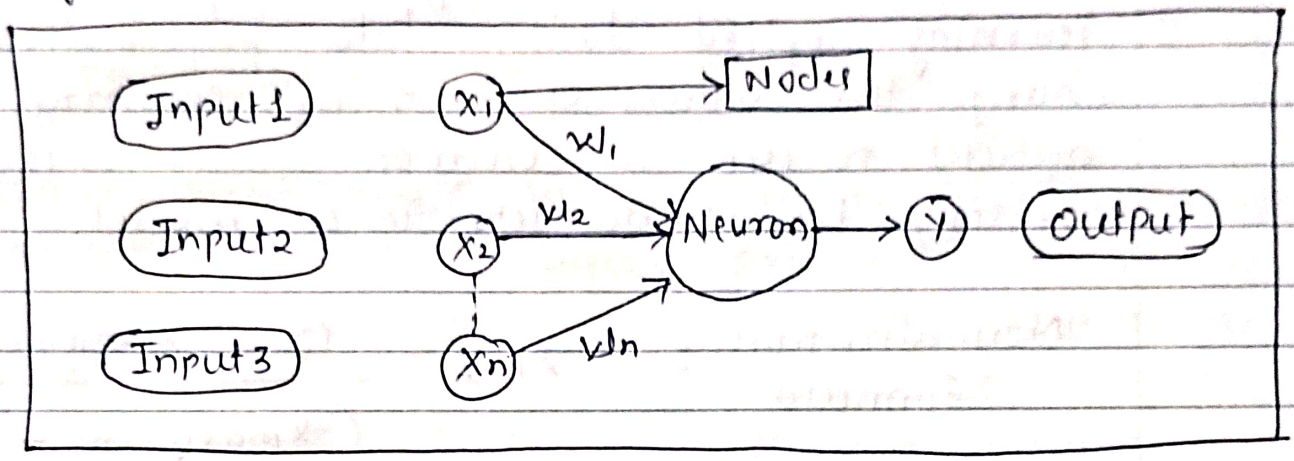
Artificial Neural Network :

The term "artificial Neural Network" is derived from biological neural Networks that develops the structure of a human brain.

- similar to the human brain tyt that has neurons interconnected to one another, artificial neural Network also have neurons that are interconnected to one another in various layers of the networks. These neurons are known as nodes.



The given figure illustrate the typical diagram of biological Neural Network.



Keras :

Keras is an open-source high-level Neural Network library, which is written in python is capable enough to run on Theano, Tensorflow, or CNTK. It was developed by one of the Google engineers, Francois chollet. It is made user-friendly, extensible and modular for facilitating faster experimentation with deep neural network.

Tensorflow :

Tensorflow is a Google product, which is one of the most famous deep learning tools widely used in the research area of a machine learning and deep neural network.

It come into the market on 9th November 2015 under the Apache License 2.0. It is built in such a way that it can easily run on multiple CPUs and GPU's as well as on mobile operating systems.

Normalisation :
Normalization is a scaling technique in machine learning applied during data preparation to change the values of numeric columns in the dataset to use a common scale. It is not necessary for all datasets in a model.

Normalization Formula : $X_n = \frac{(x - X\ minimum)}{(X maximum - X minimum)}$

Confusion matrix : Actual values

Predicted values	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Error Rate} = \frac{FP + FN}{TP + FP + FN + TN}, \quad \text{Precision} = \frac{TP}{Tr + FP}$$

Conclusion :
We clearly understand ANN, Keras, tensorflow.

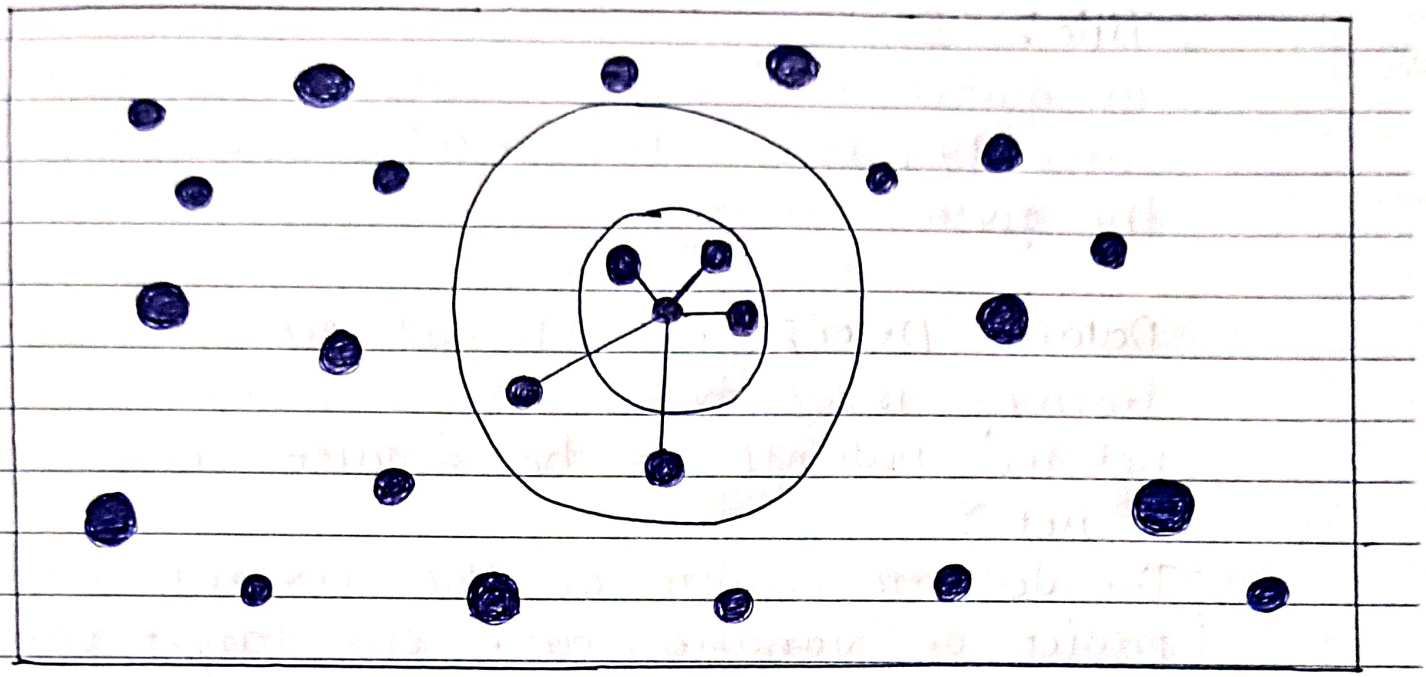
TSSM
Page No.

Title: Implement K-nearest neighbors algorithm on diabets.csv dataset. Compute confusion matrix, accuracy, error rate, precision and recall on the given dataset.

Dataset description: We will try to build a machine learning model to accuracy predict whether or not the patients in the dataset have diabetes or not?
The dataset consists of the several medical predict or variables and one target variable, outcome. predictor variables includes the number of pregnancies has the patient had, their BMI, insulin level, age, and so on.

Objectives:
students should be able to preprocess dataset and identify outliers, to check correlation and implement KNN algorithm and random forest classification models. Evaluate them with respective scores like confusion-matrix, accuracy-score, mean-squared-error, r2-score, roc-auc-score, roc-curve, etc.

Prerequisite:
1. Basic knowledge of Python
2. Concept of confusion matrix
3. concept of roc_auc_curve
4. concept of Random forest & KNN algorithm.



K-nearest Neighbours (KNN) is a supervised machine learning model. Supervised learning is when a model learns from data that is already labeled. A supervised learning model takes in a set of input objects and outputs values.

The model then trains on that data to learn how to map the inputs to the desired output so it can learn to make predictions on unseen data. KNN models work by taking a data points. The data points is then assigned the labels of the majority of the 'k' closest points.

Conclusion: In this way we build a neural network based classifier that can determine whether they or not.

Title : Implements K-means clustering / hierarchical clustering on sales-data-sample.csv dataset. Determine the number of clusters using the elbow method.

Dataset Description : The data includes the following features :

1. Customer ID
2. Customer Gender
3. Customer Age
4. Annual income of the customer (in Thousand Dollars).
5. Spending scores of the customers (based on customer behaviour & spending nature).

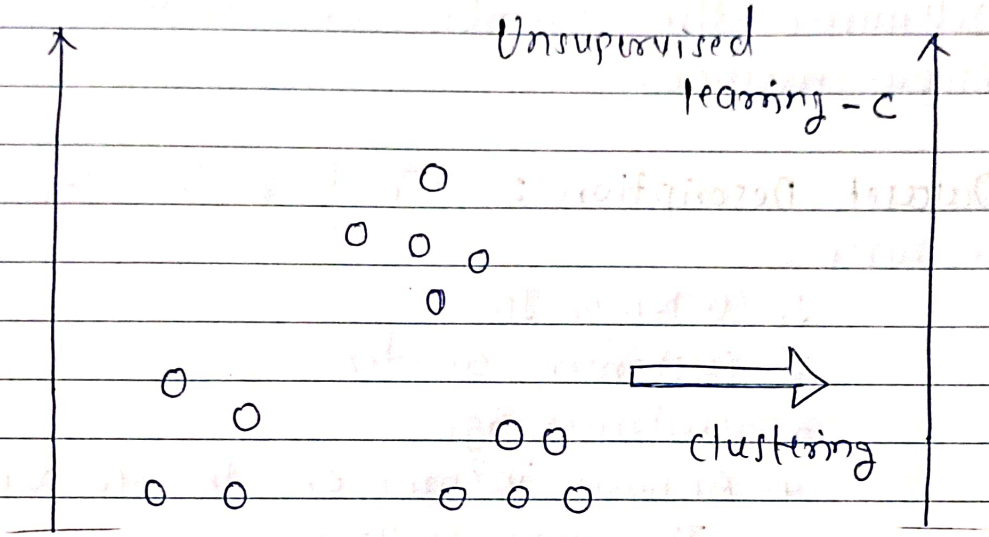
objective of Assignment : (Prerequisites)

1. Knowledge of Python.
2. Onsupervised learning.
3. Clustering
4. Elbow method.

student should be able to understant how to use unsupervised learning to segment different-different clusters or groups and use to them to train your model to predict future things.

Clustring Algorithm try to find natural clusters in data, the various aspects of how the algorithms to based on the principle that items with within the same cluster.

The data is grouped in such a way that related elements are close to each other.



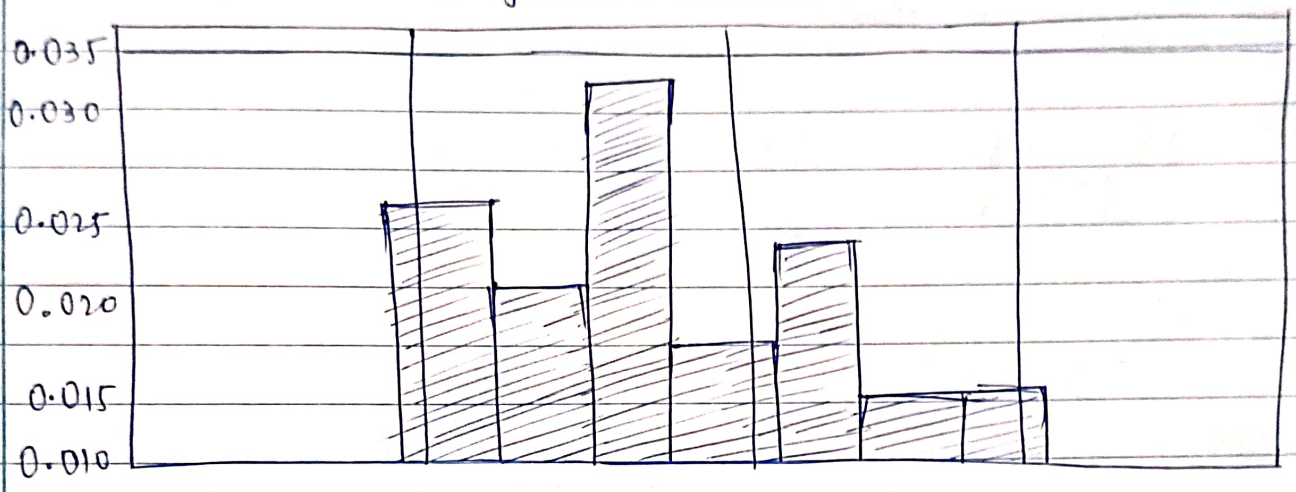
Uses of clustering :

marketing - In the field of marketing clustering can be used to identify various customer groups with existing customer data. Based on that, customers can be provided with discounts, offers, etc.

The data has 200 entries, that is data from 200 custom
data.head()

	Custom. ID.	Gender	Age	Annual Income	Spending score
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	06
3	4	Female	23	16	77
4	5	Female	31	17	40

Distribution of Age :



data.corr()

- The data seems to be intresting. Let us look at the data distribution.

Annual Income Distribution :

```
plt.figure (figsize = (10, 6))
sns.set (style = 'whitegrid')
sns.distplot (data ['Annual Income (k$)'])
plt.ylabel ('count').
```

Conclusion : Implemented k-means clustering / hierarchical clustering on sales_data_sample.csv dataset.
Determined number of clusters using the elbow method.