Project - Sprint-1

You will have:

- Data Set
- Data Dictionary (what does each column in the data or what is the meaning of each column in data)

Required Python Package:

- Pandas
  - Data Transformation or reshaping the data
  - Data Cleaning
  - Data Analysis
  - Data Mining
  - Handling the missing value
  - Grouping, Aggregating, Statistics, Window Function, Apply, Merge, Join, Concat, Pivot, indexes etc
- Numpy
  - Numerical Computation
  - Array Operations
  - Integration with Other Python Package like Pandas, seaborn, and matplotlib
  - Use the power of numpy with Looping and Data Structure
- Data Visualization Python Package:
  - Matplotlib
  - Seaborn
  - Plotly.express and graph_objects

- Intermediate concept of Pandas
- Complete understanding of Numpy interms of numbers, and its functions
- Function (UserDefined Function)
- System define function (map, filter, zip, enumerate, reduce (itertools)

In [1]:
```
import pandas as pd
```

In [2]:
```
pd.__version__
```

Out[2]:
```
'2.2.2'
```

In [12]:
```
lst = [100, 99, 10,34,56,78,90,34,5,89,90,56,34,70,68,45,34,12,78,90,8,34
```

In [13]:
```
lst.index(34)
```

Out[13]:
```
3
```

In [14]:
```
lst.index(34,lst.index(34)+1)
```

Out[14]: 7

In [15]:
```
lst.count(34)
```

Out[15]: 8

In [20]:
```
lst.index(34,lst.index(34,lst.index(34)+1)+1)
```

Out[20]: 12

In [16]:
```
lst.count.__doc__
```

Out[16]: 'Return number of occurrences of value.'

In [17]:
```
import numpy as np
```

In [18]:
```
arr = np.array(lst)
arr
```

Out[18]:
```
array([100,  99,  10,  34,  56,  78,  90,  34,   5,  89,  90,  56,  34,
        70,  68,  45,  34,  12,  78,  90,   8,  34,  12,  56,  87,  56,
        23,  34,  56,  78,   9,  23,  56,  34,  67,  89,  34])
```

In [19]:
```
np.where(arr == 34)
```

Out[19]:
```
(array([ 3,  7, 12, 16, 21, 27, 33, 36]),)
```

In [23]:
```
lst[27]
```

Out[23]: 34

In [24]:
```
np.where(arr % 2 == 0)
```

Out[24]:
```
(array([ 0,  2,  3,  4,  5,  6,  7, 10, 11, 12, 13, 14, 16, 17, 18, 19,
20,
        21, 22, 23, 25, 27, 28, 29, 32, 33, 36]),)
```

In [26]:
```
arr[arr % 2 != 0]
```

Out[26]:
```
array([99,  5, 89, 45, 87, 23,  9, 23, 67, 89])
```

In [34]:
```
# can you delete all the even number from array
np.delete(arr, np.where(arr % 2 == 0))
```

Out[34]:
```
array([99,  5, 89, 45, 87, 23,  9, 23, 67, 89])
```

In [28]:
```
arr
```

Out[28]:
```
array([100,  99,  10,  34,  56,  78,  90,  34,   5,  89,  90,  56,  34,
        70,  68,  45,  34,  12,  78,  90,   8,  34,  12,  56,  87,  56,
        23,  34,  56,  78,   9,  23,  56,  34,  67,  89,  34])
```

In [32]:
```
np.delete(arr,[0,-1]) # You are understanding that, delete function worki
```

Out[32]:
```
array([99, 10, 34, 56, 78, 90, 34,  5, 89, 90, 56, 34, 70, 68, 45, 34, 12,
       78, 90,  8, 34, 12, 56, 87, 56, 23, 34, 56, 78,  9, 23, 56, 34, 67,
       89])
```

DATA Project

- Loading | Reading the data
- Identify the shape of the data
- Data type of each column
- Filter the columns based on its data type
- Handing the missing value
- Perform the summary statistics and descriptive statistics
- EDA : Exploratory Data Analysis

In [38]:
```
import pandas as pd
import sketch
```

In [39]:
```
EmployeeDB = {"EmpID" : [101,102,103,104,105],
              "EmpName" : ["Abhishek","Xiteej","Samiskha","Adya","Anirudh
              "Salary" : [10000,12000,230000,45000,56666]}
```

In [40]:
```
df = pd.DataFrame(EmployeeDB)
```

In [41]:
```
df
```

Out[41]:

	EmpID	EmpName	Salary
0	101	Abhishek	10000
1	102	Xiteej	12000
2	103	Samiskha	230000
3	104	Adya	45000
4	105	Anirudh	56666

In [42]:
```
df.sketch.ask("What is the EDA and what we can do with the help of EDA, p
```

EDA stands for Exploratory Data Analysis, which is a process of analyzing and summarizing data in order to gain insights and understanding about the data. It involves using statistical and visualization techniques to explore the data and identify patterns, trends, and relationships. The steps involved in EDA are as follows: 1. Data Collection: The first step in EDA is to collect the data from various sources such as databases, spreadsheets, or web scraping. 2. Data Cleaning: Once the data is collected, it needs to be cleaned to remove any errors, missing values, or irrelevant data. This step is crucial as it ensures that the data is accurate and reliable for analysis. 3. Data Exploration: In this step, various statistical techniques are used to explore the data and understand its characteristics. This includes calculating summary statistics such as mean, median, mode, standard deviation, etc., and visualizing the data through charts, graphs, and plots. 4. Data Preprocessing: This step involves transforming the data into a format that is suitable for analysis. This may include converting categorical variables into numerical variables, handling missing values, and normalizing or standardizing the data. 5. Feature Engineering: Feature engineering involves creating new features from existing ones to improve the performance of machine learning models. This can include feature selection, dimensionality reduction, and creating new variables based on domain knowledge. 6. Statistical Modeling: In this step, various statistical models are applied to the data to identify patterns and relationships between variables. This can include regression analysis, clustering techniques, or classification algorithms. 7. Interpretation and Communication: The final step in EDA is to interpret the results of the analysis and communicate them effectively to stakeholders. This involves creating visualizations and reports that clearly convey the insights gained from the data. With the help of EDA, we can gain a better understanding of our data and make informed decisions based on the insights obtained. It can help us identify trends, patterns, and relationships in the data, which can be used to make predictions and improve business processes. EDA is also useful for identifying outliers and anomalies in the data, which can help in detecting errors or fraud. Overall, EDA is a crucial step in the data analysis process and can provide valuable insights that can drive decision-making and lead to better outcomes.

In [43]:
```
df.sketch.ask("How can I learn Data Analysis")
```

1. Start with the basics: Before diving into data analysis, it is important to have a strong foundation in statistics and programming. Make sure you have a good understanding of concepts like mean, median, mode, standard deviation, and correlation. Also, learn a programming language like Python or R, which are commonly used for data analysis. 2. Take online courses: There are many online courses available that can teach you the fundamentals of data analysis. Some popular platforms include Coursera, Udemy, and DataCamp. These courses will provide you with hands-on experience and real-world examples to help you understand the concepts better. 3. Practice with real datasets: The best way to learn data analysis is by practicing with real datasets. You can find datasets on websites like Kaggle, UCI Machine Learning Repository, and Google Dataset Search. Start with simple datasets and gradually move on to more complex ones. 4. Learn data visualization: Data visualization is an important aspect of data analysis. It helps in understanding the data better and communicating insights effectively. Learn how to use tools like Tableau, Power BI, or Matplotlib to create visualizations. 5. Join online communities: Joining online communities like Reddit's r/dataisbeautiful or LinkedIn groups can help you connect with other data analysts and learn from their experiences. You can also ask for feedback on your analysis and get valuable insights from experts in the field. 6. Read books and blogs: There are many books and blogs written by experts in the field of data analysis that can help you learn new techniques and stay updated with the latest trends. Some popular books include "Data Science for Business" by Foster Provost and Tom Fawcett and "Python for Data Analysis" by Wes McKinney. 7. Attend workshops and conferences: Attending workshops and conferences is a great way to network with other data analysts and learn from their experiences. You can also attend talks and workshops by industry experts to gain insights into the latest tools and techniques used in data analysis. 8. Practice, practice, practice: The key to becoming a good data analyst is to practice regularly. Keep challenging yourself with new datasets and try to solve different types of problems. This will help you improve your skills and become a better data analyst.

In [45]:
```
np.random.rand(1,100,500)
```

Out[45]:
```
array([[[0.48415453, 0.25582993, 0.36145899, ..., 0.39729679,
         0.1387662 , 0.68492889],
        [0.39864451, 0.73476683, 0.68392676, ..., 0.01111537,
         0.92681382, 0.41307513],
        [0.08848152, 0.14600662, 0.42050222, ..., 0.66781855,
         0.29226186, 0.45572675],
        ...,
        [0.32090486, 0.21397446, 0.54298166, ..., 0.24402261,
         0.3824978 , 0.8058306 ],
        [0.22405795, 0.93825255, 0.78692805, ..., 0.55699144,
         0.48621085, 0.54991404],
        [0.67303429, 0.14000894, 0.74212944, ..., 0.00938655,
         0.0404665 , 0.09952553]]])
```

In [54]:
```
import numpy as np
np.random.seed(12)
np.random.randint(low = 10, high = 100, size = 20)
```

Out[54]:
```
array([85, 37, 16, 12, 13, 77, 86, 58, 32, 59, 62, 15, 23, 99, 44, 85, 8
4,
       10, 86, 23])
```

In [ ]: