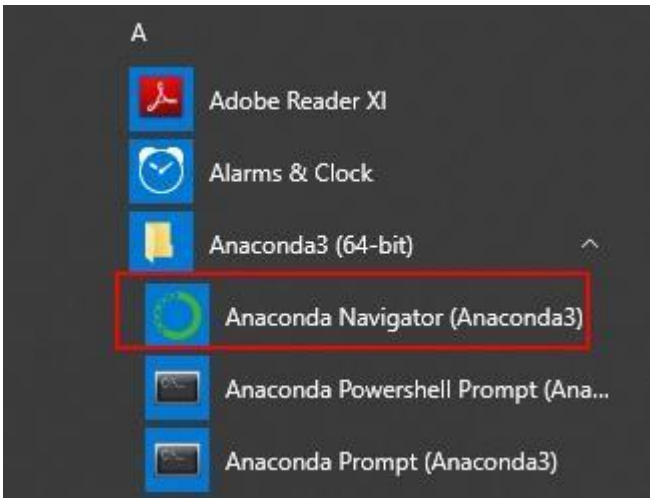
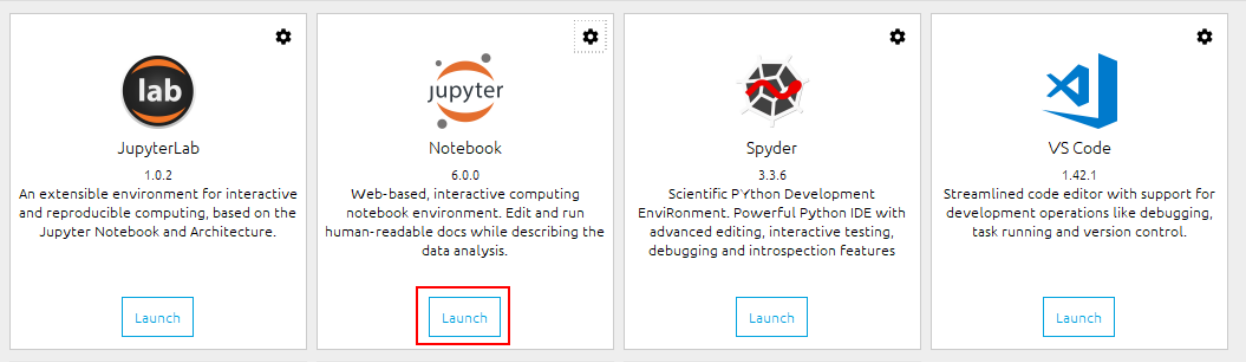
Module 7: Hands-On: 6

Data Cleaning.

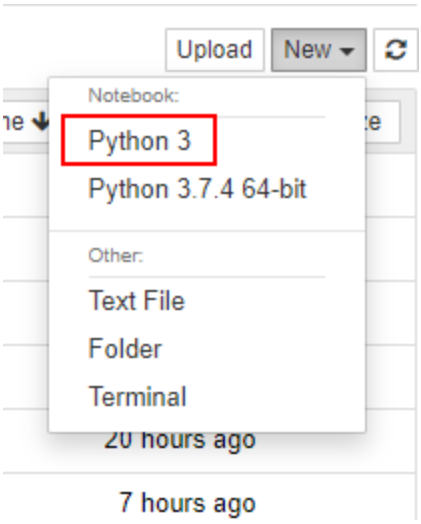
Step 1: Open Anaconda Navigator



Step 2: Click on Launch button under jupyter notebooks.



Step 3: After the notebook opens click on new and Python 3.



Step 4: Import the required packages and read data from time_seriest.csv in a dataframe.

```
In [13]: import pandas as pd
         import matplotlib.pyplot as plt
         import seaborn as sns

In [26]: data = pd.read_csv('time_series.csv')

In [15]: data.head()
```

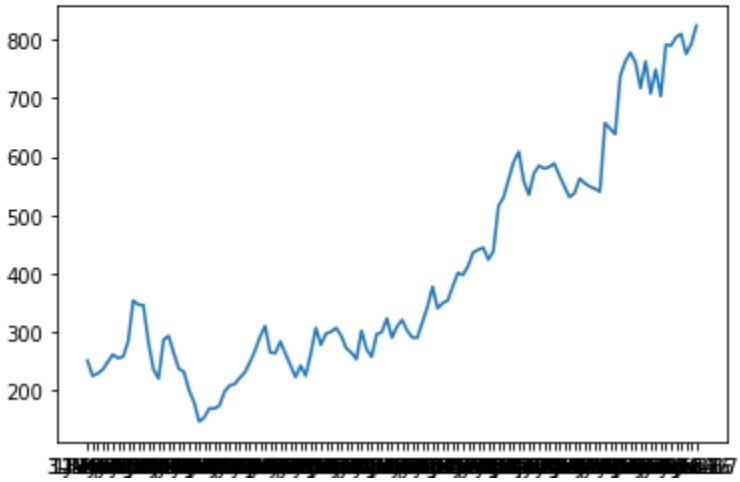
Out[15]:

	Date	AAPL	ADBE	CVX	GOOG	IBM	MDLZ	MSFT	NFLX	ORCL	SBUX
0	3-Jan-07	11.107141	38.869999	50.777351	251.001007	79.242500	17.519524	24.118483	3.258571	15.696321	15.752188
1	1-Feb-07	10.962033	39.250000	48.082939	224.949951	74.503204	16.019426	22.092464	3.218571	15.028588	13.930813
2	1-Mar-07	12.037377	41.700001	51.900383	229.309311	75.561348	16.009354	21.857189	3.312857	16.583584	14.138198
3	2-Apr-07	12.930043	41.560001	54.588032	235.925919	81.934280	16.924608	23.480597	3.167143	17.196436	13.984914
4	1-May-07	15.701322	44.060001	57.598267	249.204208	85.786057	17.111704	24.146753	3.128572	17.726965	12.988567

Step 5: Plot a line graph and take a look at Google's historical data about its stock price.

```
In [16]: plt.plot(data['Date'], data['GOOG'])
```

```
Out[16]: [<matplotlib.lines.Line2D at 0x2414bb620b8>]
```



Step 6: Read data from national parks and take a look at first 5 rows.

```
In [17]: data = pd.read_csv('national_parks.csv')
```

```
In [18]: data.head()
```

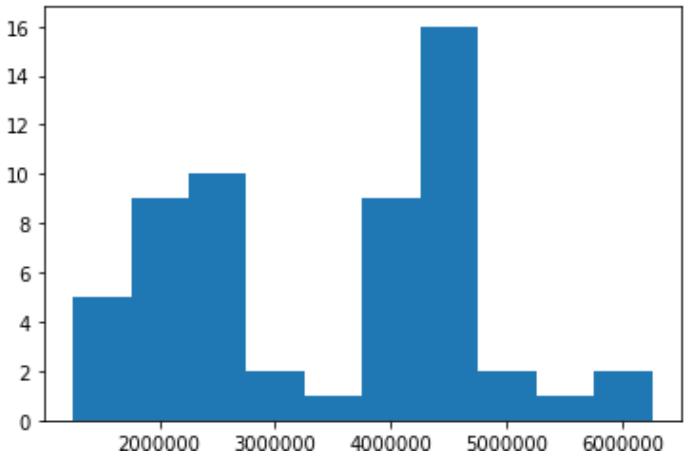
```
Out[18]:
```

	Year	Badlands	GrandCanyon	BryceCanyon
0	1961	833300	1253000	264800
1	1962	1044800	1447400	251000
2	1963	1074000	1539500	289500
3	1964	1079800	1576600	300300
4	1965	1091300	1689200	366800

Step 7: Plot a histogram based on the 'GrandCanyon' column.

```
In [19]: plt.hist(data['GrandCanyon'])
```

```
Out[19]: (array([ 5.,  9., 10.,  2.,  1.,  9., 16.,  2.,  1.,  2.]),
          array([1253000. , 1753123.8, 2253247.6, 2753371.4, 3253495.2, 3753619. ,
                 4253742.8, 4753866.6, 5253990.4, 5754114.2, 6254238. ]),
          <a list of 10 Patch objects>)
```



Step 8: Read data from 'types_movies.csv'and take a look at first 5 rows.

```
In [30]: data = pd.read_csv('types_movies.csv')
```

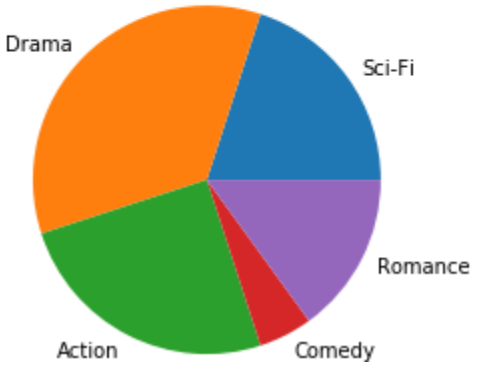
```
In [31]: data.head()
```

Out[31]:

	Sector	Percentage
0	Sci-Fi	20
1	Drama	35
2	Action	25
3	Comedy	5
4	Romance	15

Step 9: Plot a pie chart based on percentage of movies and set labels to be sector column.

```
In [32]: plt.pie(data['Percentage'], labels=data['Sector'])
         plt.show()
```



Step 10: Create and visualize a correlation matrix on time_series data using heatmaps.

```
In [27]: data = pd.read_csv('time_series.csv')
         matrix = data.corr()

In [29]: sns.heatmap(matrix, annot=True)

Out[29]: <matplotlib.axes._subplots.AxesSubplot at 0x2414ccd1c50>
```