Coding challenge

Python

Name: Rohan Vinayak Chaudhari

Batch: Data Engineering 1

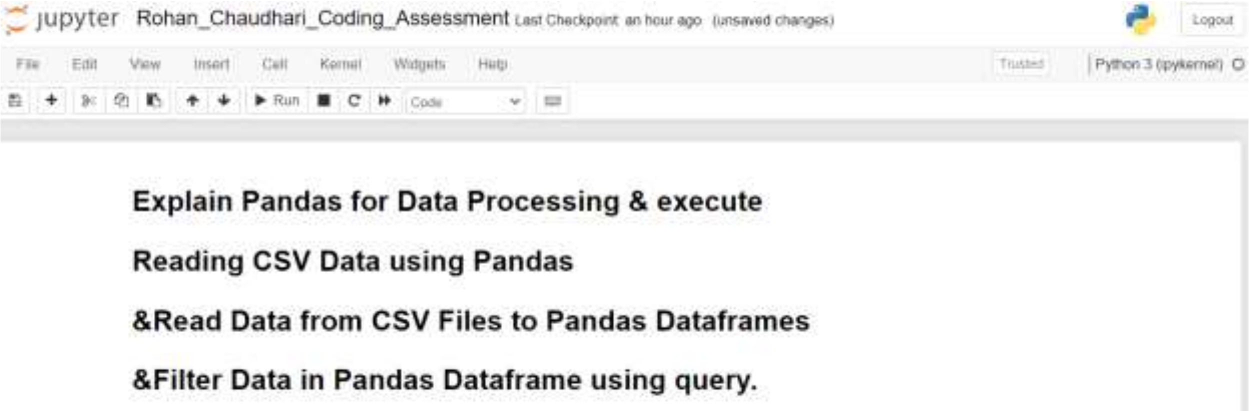
Question1: Explain Pandas for Data Processing & execute Reading CSV Data using Pandas
&Read Data from CSV Files to Pandas Dataframes
&Filter Data in Pandas Dataframe using query.

1)Pandas for data processing:

Pandas is an open-source data manipulation and analysis library for Python. It provides data structures such as Series and DataFrame for efficiently manipulating large datasets.Some Points include:

1. **DataFrame:** A two-dimensional table with rows and columns.Similar to sql tables.
2. **Series:** A one-dimensional labeled array containing of holding any data type.
3. **Data Cleaning:** Pandas provides functions for handling missing data and cleaning the dataset.
4. **Data Filtering and Selection:** Easy ways to filter, select, and manipulate data.
5. **Merging and Joining:** Combine different datasets using various methods.
6. **Aggregation:** Perform operations on data grouped by certain criteria.

Question1:


jupyter
Rohan_Chaudhari_Coding_Assessment Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

Explain Pandas for Data Processing & execute

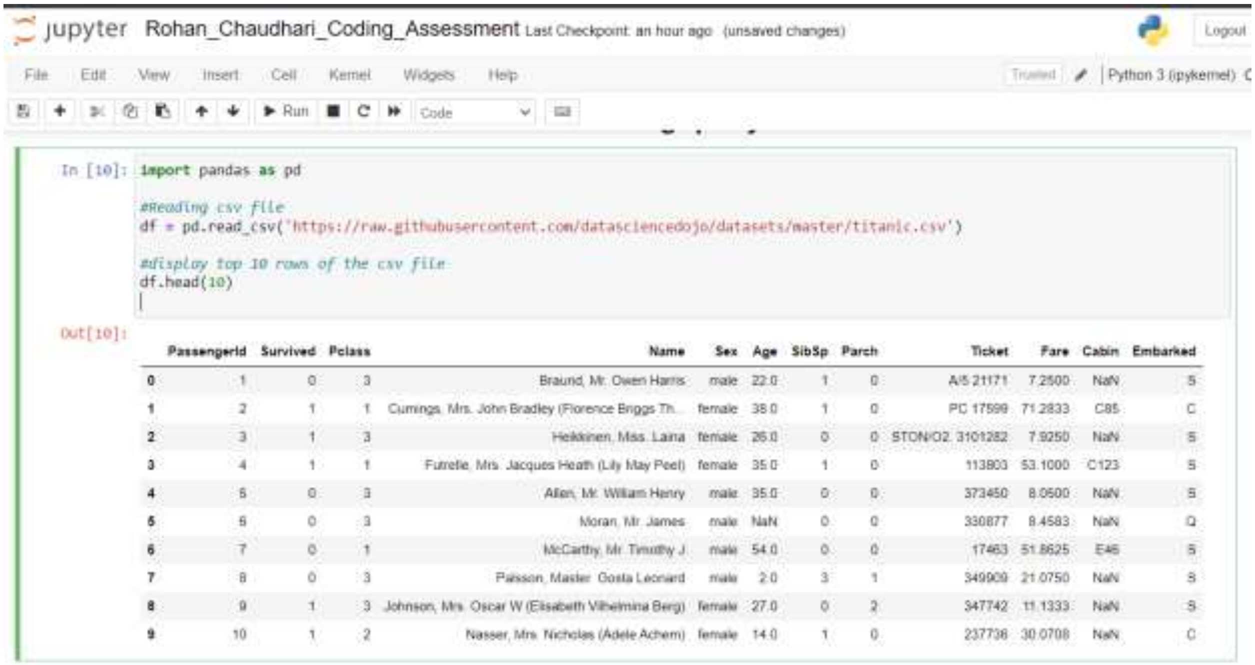
Reading CSV Data using Pandas

&Read Data from CSV Files to Pandas Dataframes

&Filter Data in Pandas Dataframe using query.

2) Reading CSV Data using Pandas:

Here I have taken a csv data name titanic where read the data using read_csv and display the top 10 data from the csv file.


jupyter
Rohan_Chaudhari_Coding_Assessment Last Checkpoint: an hour ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code

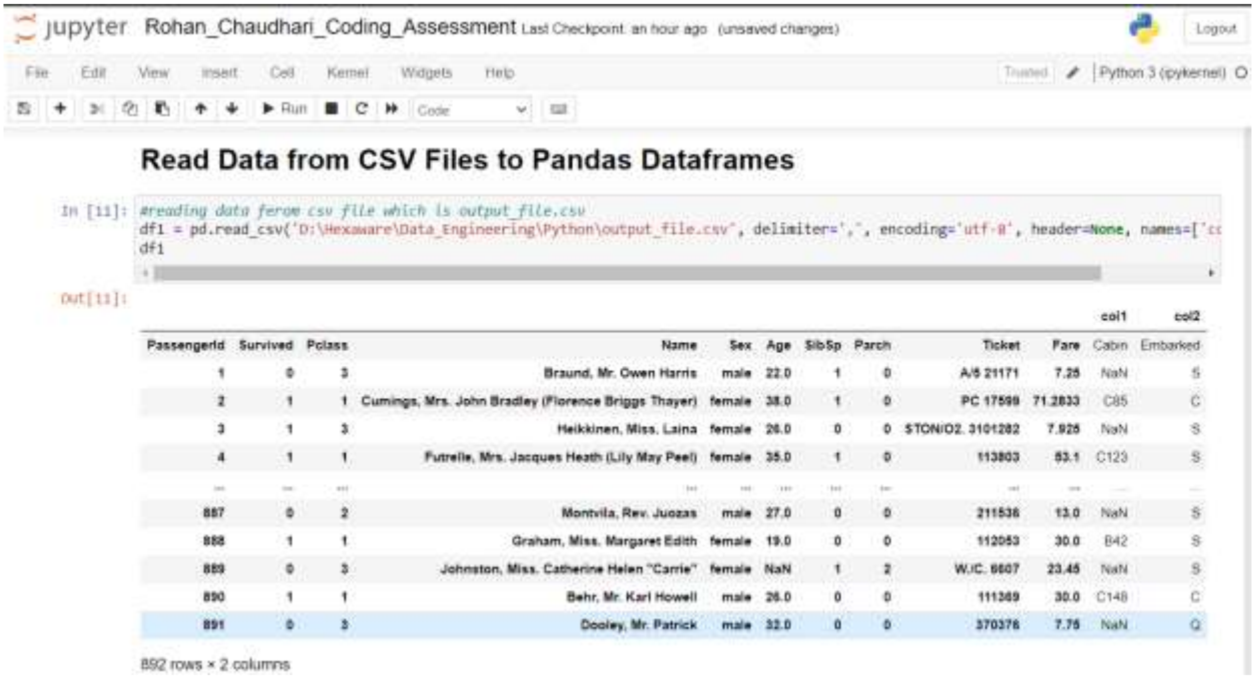
```
In [10]: import pandas as pd

         #Reading csv file
         df = pd.read_csv('https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv')

         #display top 10 rows of the csv file
         df.head(10)
```

Out[10]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

3)Reading CSV data from file using pandas

Here used the same dataset and read the data into dataframe from a already existing file.



Read Data from CSV Files to Pandas Dataframes

```
In [11]: #reading data ferom csv file which is output_file.csv
         df1 = pd.read_csv('D:\Hexaware\Data_Engineering\Python\output_file.csv', delimiter=',', encoding='utf-8', header=None, names=['cc
         df1
```

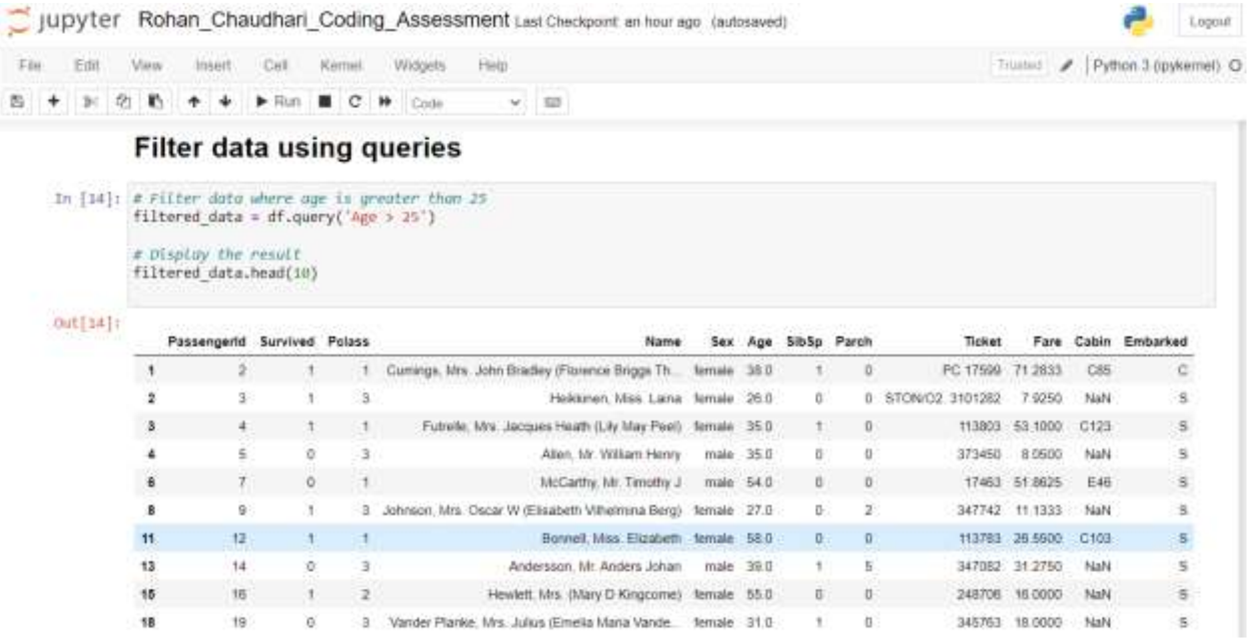
Out[11]:

										col1	col2
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25	NaN	S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925	NaN	S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1	C123	S
...	...	...	...	...	...	...	...	...	...	...	...
887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0	NaN	S
888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0	B42	S
889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	S
890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0	C148	C
891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q

892 rows × 2 columns

4)Filter data using query

Numeric Comparison:



Filter data using queries

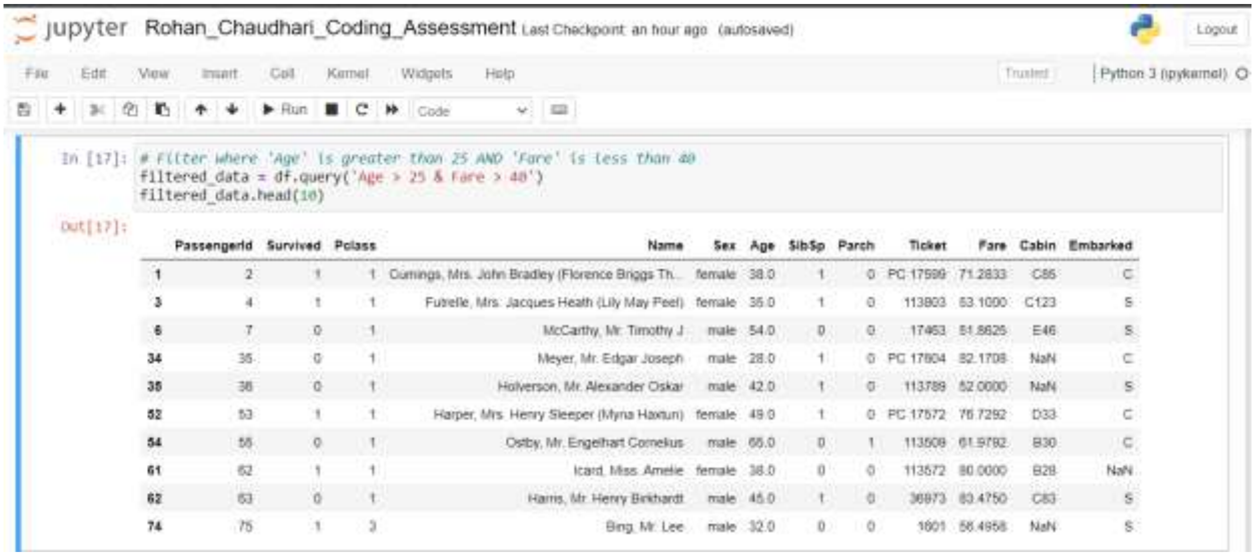
```
In [14]: # Filter data where age is greater than 25
         filtered_data = df.query('Age > 25')

         # Display the result
         filtered_data.head(10)
```

Out[14]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
13	14	0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.2750	NaN	S
15	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.0	0	0	248706	16.0000	NaN	S
18	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vande...	female	31.0	1	0	345763	18.0000	NaN	S

Logical And:

Extracted the data where age is greater than 25 and fare is grater than 40



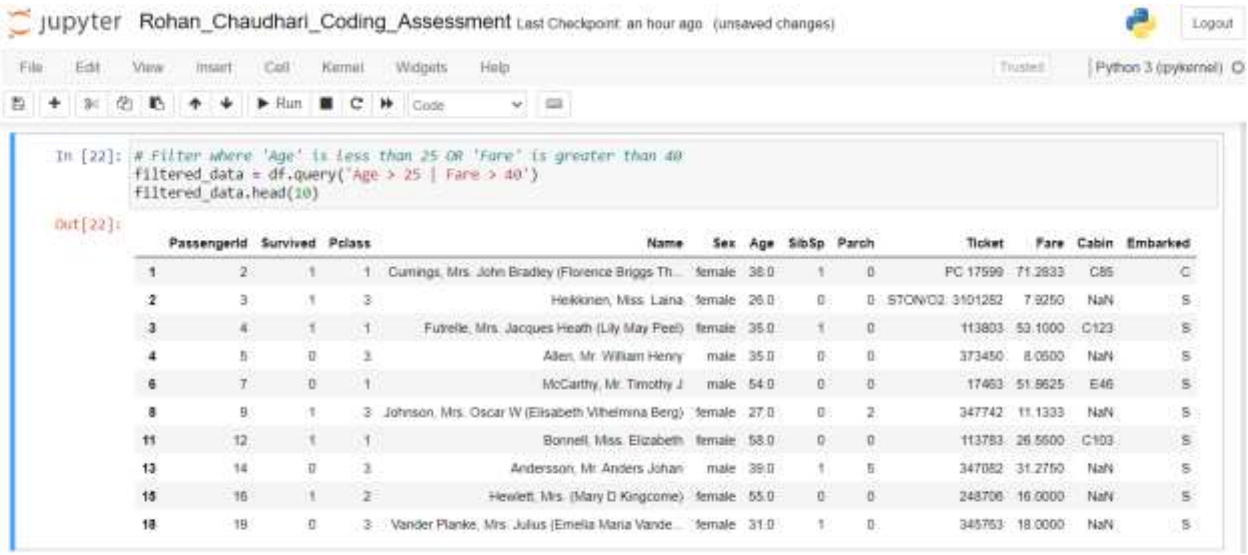
```
In [17]: # Filter where 'Age' is greater than 25 AND 'Fare' is less than 40
         filtered_data = df.query('Age > 25 & Fare > 40')
         filtered_data.head(10)
```

Out[17]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
34	35	0	1	Meyer, Mr. Edgar Joseph	male	28.0	1	0	PC 17604	82.1708	NaN	C
35	36	0	1	Holverson, Mr. Alexander Oskar	male	42.0	1	0	113789	52.0000	NaN	S
52	53	1	1	Harper, Mrs. Henry Sleeper (Myna Haxtun)	female	49.0	1	0	PC 17572	76.7292	D33	C
54	55	0	1	Ostby, Mr. Engelhart Cornelius	male	65.0	0	1	113509	61.9792	B30	C
61	62	1	1	Icard, Miss. Amelie	female	38.0	0	0	113572	80.0000	B28	NaN
62	63	0	1	Harris, Mr. Henry Birkhardt	male	45.0	1	0	36973	83.4750	C83	S
74	75	1	3	Bing, Mr. Lee	male	32.0	0	0	1601	56.4958	NaN	S

Logical OR:

Extracted the data where age is greater than 25 or fare is greater than 25.



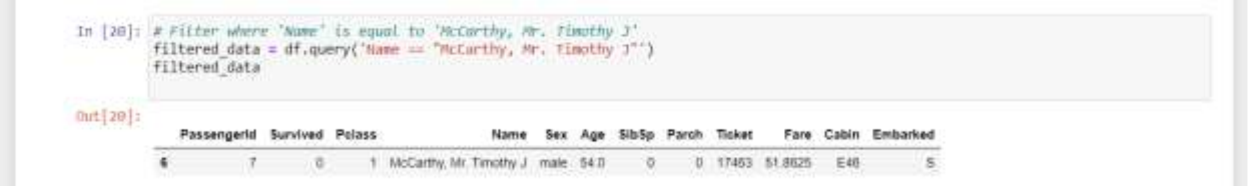
```
In [22]: # Filter where 'Age' is less than 25 OR 'Fare' is greater than 40
         filtered_data = df.query('Age > 25 | Fare > 40')
         filtered_data.head(10)
```

Out[22]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
11	12	1	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.5500	C103	S
13	14	0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.2750	NaN	S
15	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.0	0	0	248706	16.0000	NaN	S
18	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vande...	female	31.0	1	0	345763	18.0000	NaN	S

String Filtering:

Extracted the data where the name is similar to ‘McCarthy, Mr. Timothy J’



In [20]:
```
# Filter where 'Name' is equal to 'McCarthy, Mr. Timothy J'
filtered_data = df.query('Name == "McCarthy, Mr. Timothy J"')
filtered_data
```

Out[20]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S