Introduction

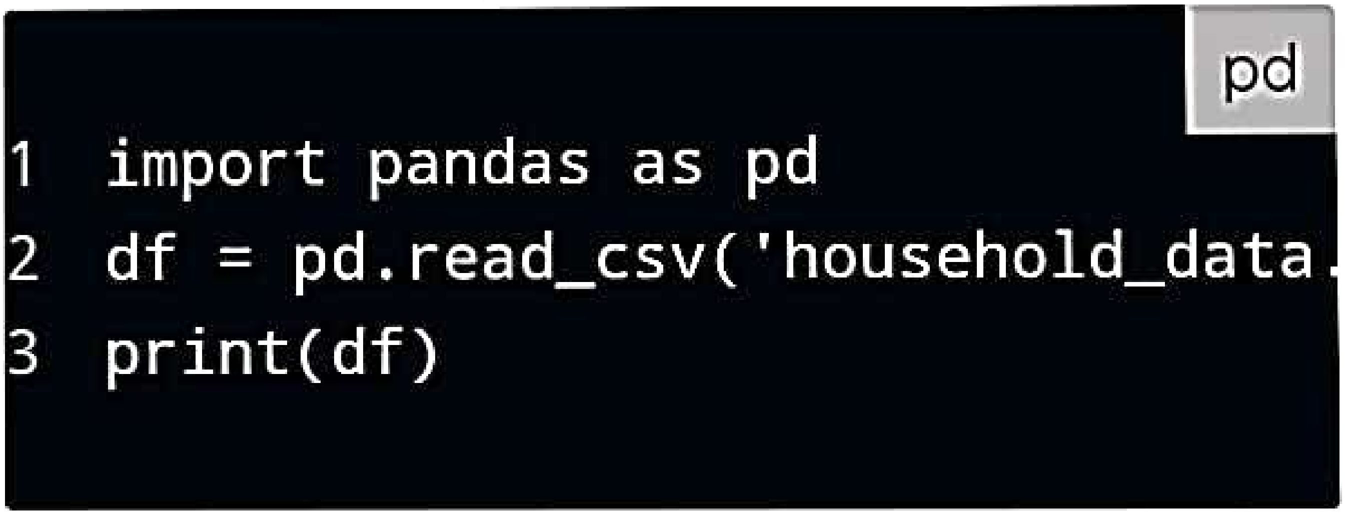
Importing data from different sources is fundamental to data science and machine learning. The abundance of good quality data not only eliminates a lot of pre-processing steps but also determines how likely your model is going to succeed in predicting plausible outcomes. The Python Panda library is the workhorse of a

Panda library is the workhorse of a data scientist when dealing with table or matrix forms of data. Panda is written on top of NumPy and provides the additional level of abstraction. This helps users focus more on solving the problem statement by hiding the elaborate implementation details. It takes the input in the form of csv, txt or sql file and converts it into the dataframe object which is then available for splicing and analysis.

Importing Data from Various Sources

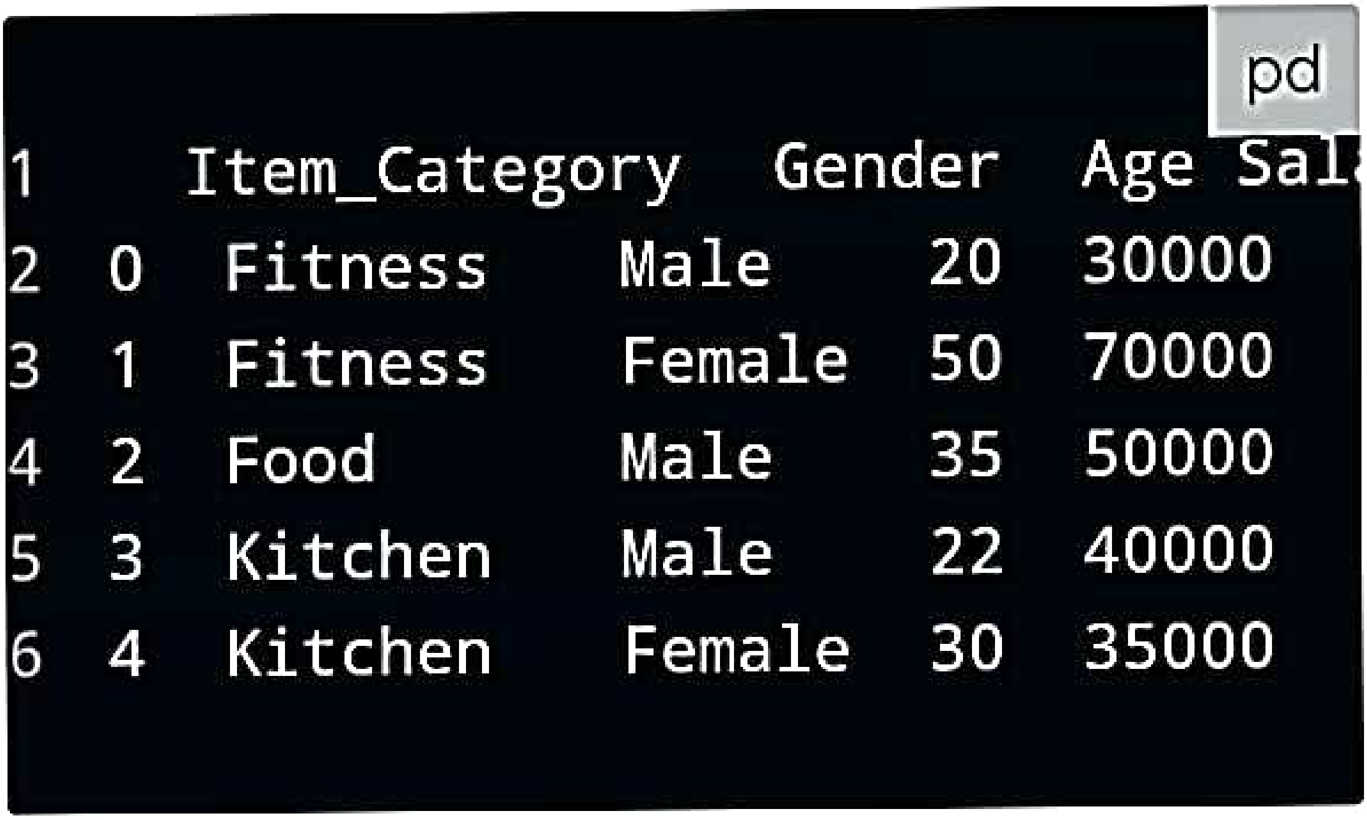
In this guide, we are going to work with household_data.csv; the contents of which are displayed below. Unless explicitly mentioned, the data of file will remain throughout this guide.

Reading the household_data.csv



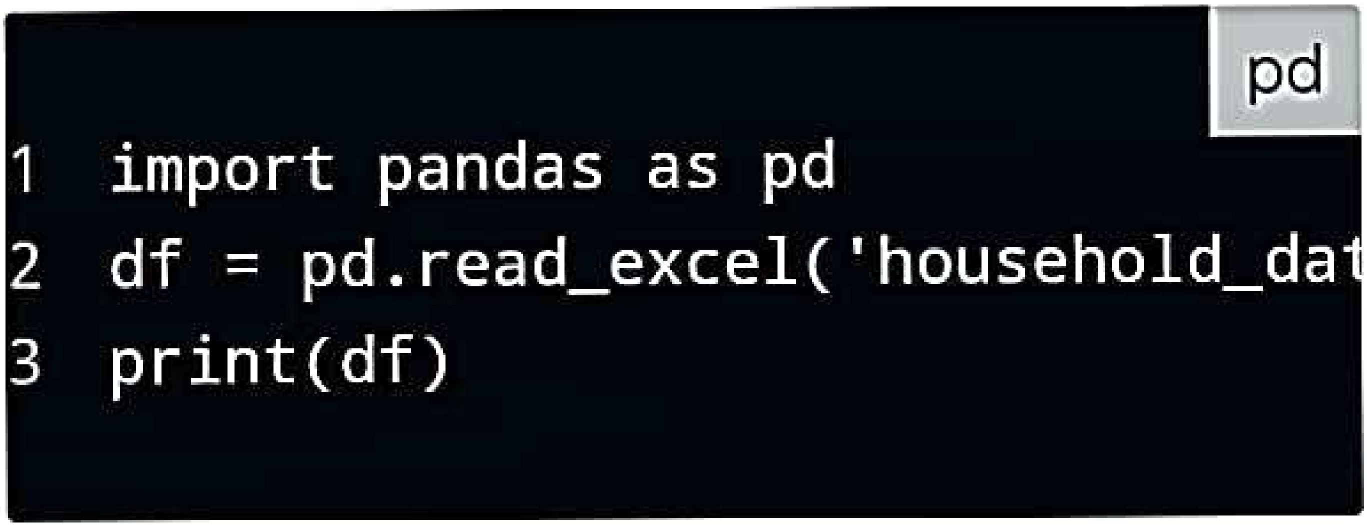
```
import pandas as pd
df = pd.read_csv('household_data.
print(df)
```

Output:



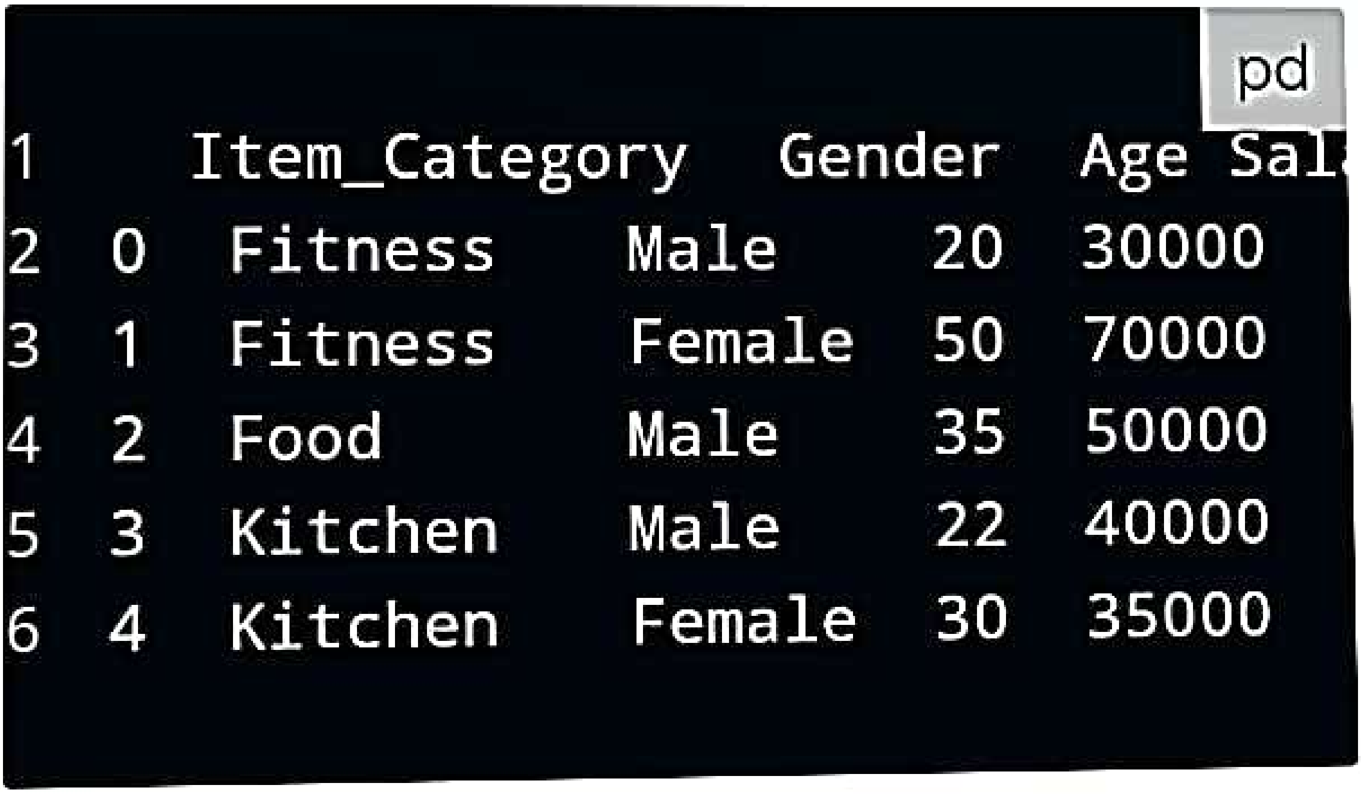
```
     Item_Category  Gender  Age  Sal
0  Fitness       Male      20  30000
1  Fitness       Female    50  70000
2  Food          Male      35  50000
3  Kitchen       Male      22  40000
4  Kitchen       Female    30  35000
```

Reading Excel Files



```
import pandas as pd
df = pd.read_excel('household_dat
print(df)
```

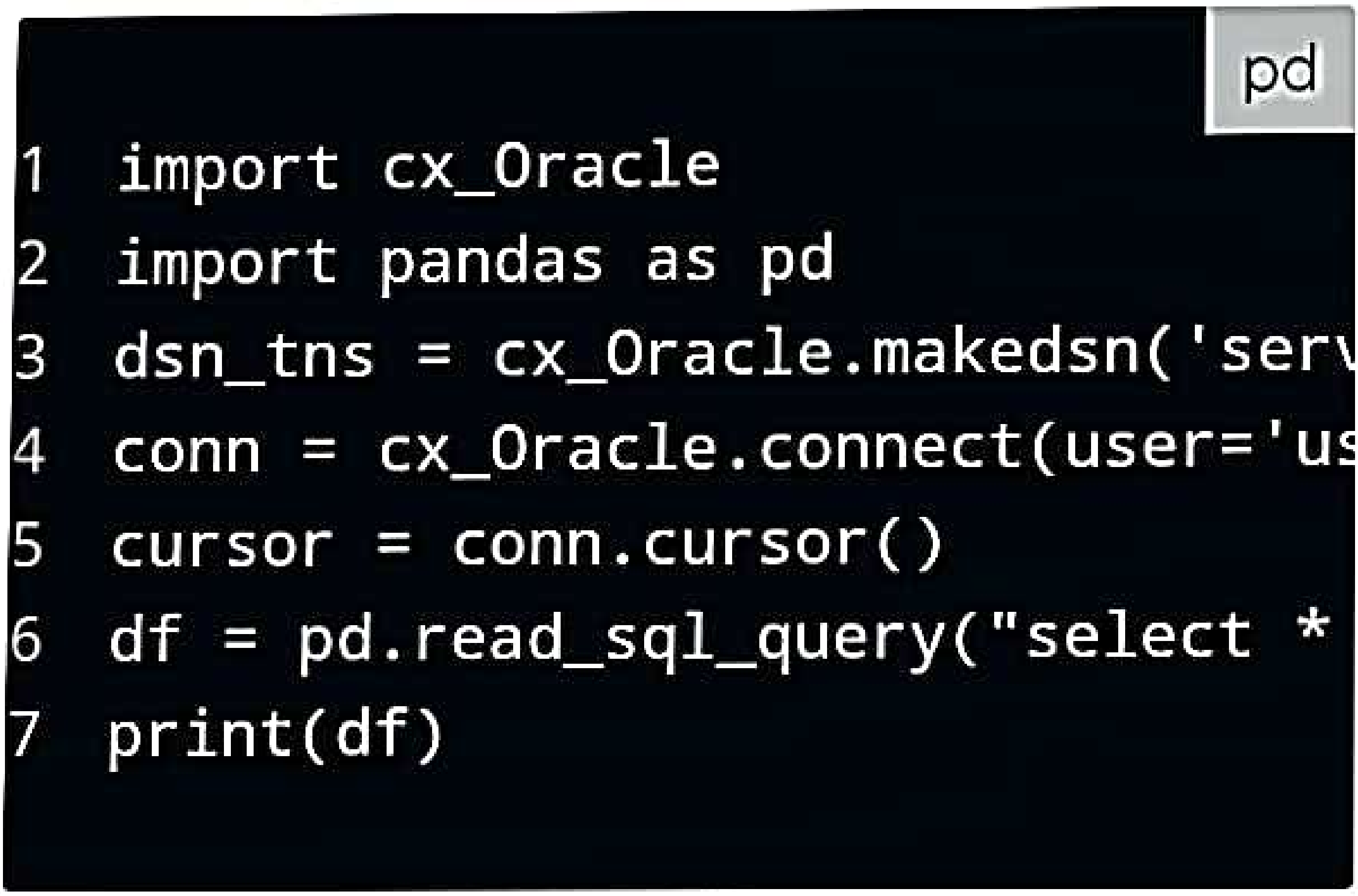
Output:



```
    Item_Category  Gender  Age Sal
0   Fitness        Male     20  30000
1   Fitness        Female   50  70000
2   Food           Male     35  50000
3   Kitchen        Male     22  40000
4   Kitchen        Female   30  35000
```

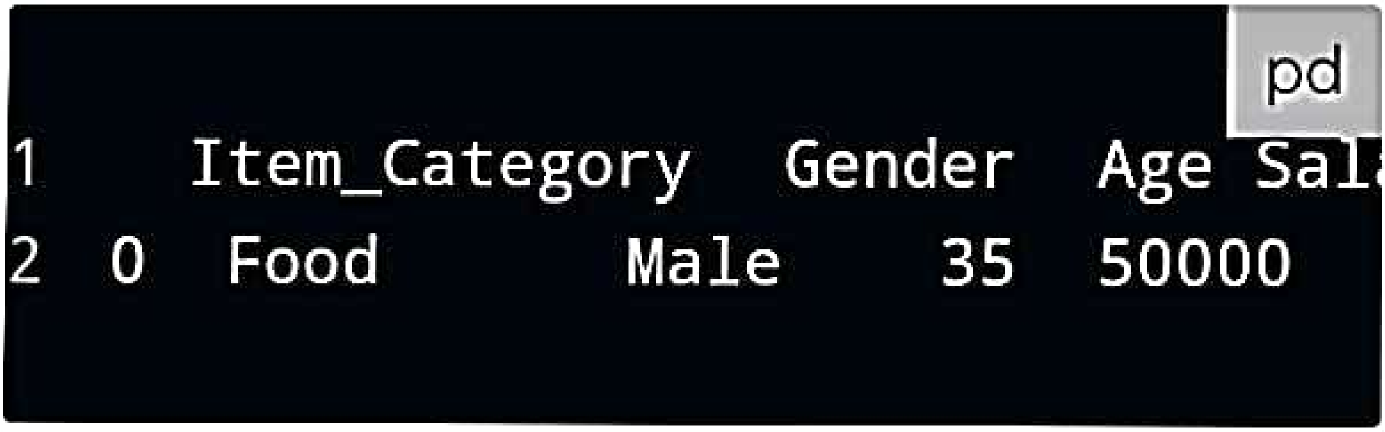
Reading the SQL File and Putting the Contents of It to Dataframe

We are going to see how to read the contents returned by the select statement to the dataframe. The below snippet is for Oracle but the idea remains same for other databases. Only the connection details should change.



```
import cx_Oracle
import pandas as pd
dsn_tns = cx_Oracle.makedsn('serv
conn = cx_Oracle.connect(user='us
cursor = conn.cursor()
df = pd.read_sql_query("select *
print(df)
```

Output:

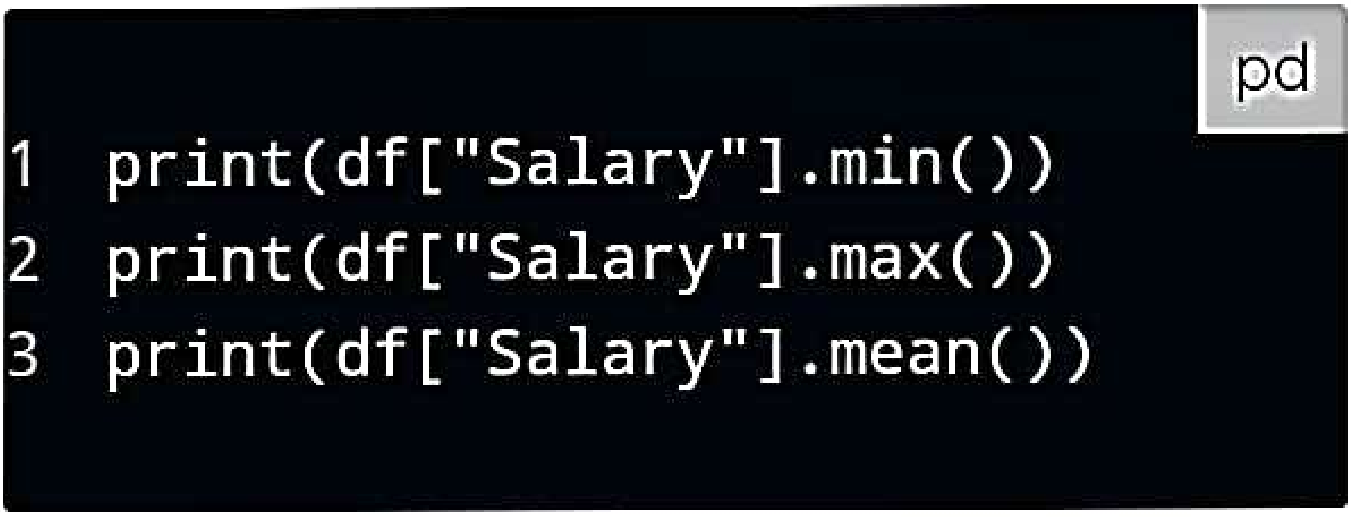


```
   Item_Category  Gender  Age Sala
0  Food         Male    35  50000
```

Splitting, Splicing, and Analysis of Data Using Dataframes

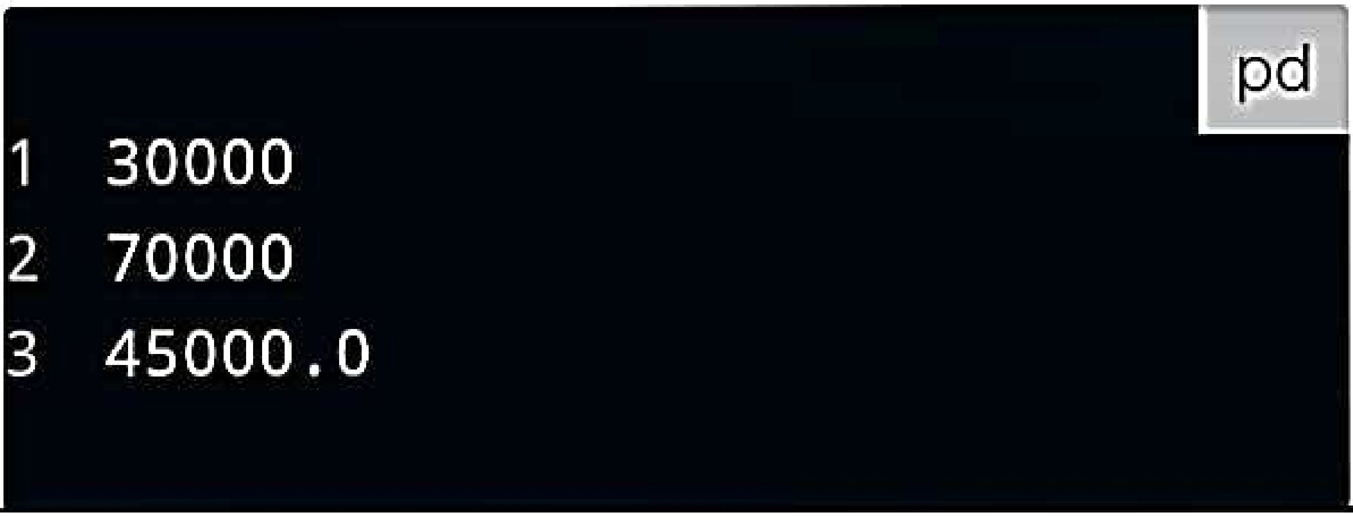
Panda provides various in-built functions that come in handy when dealing with the data set.

Getting the Minimum, Maximum and Average of a Column



```
print(df["Salary"].min())
print(df["Salary"].max())
print(df["Salary"].mean())
```

Output:

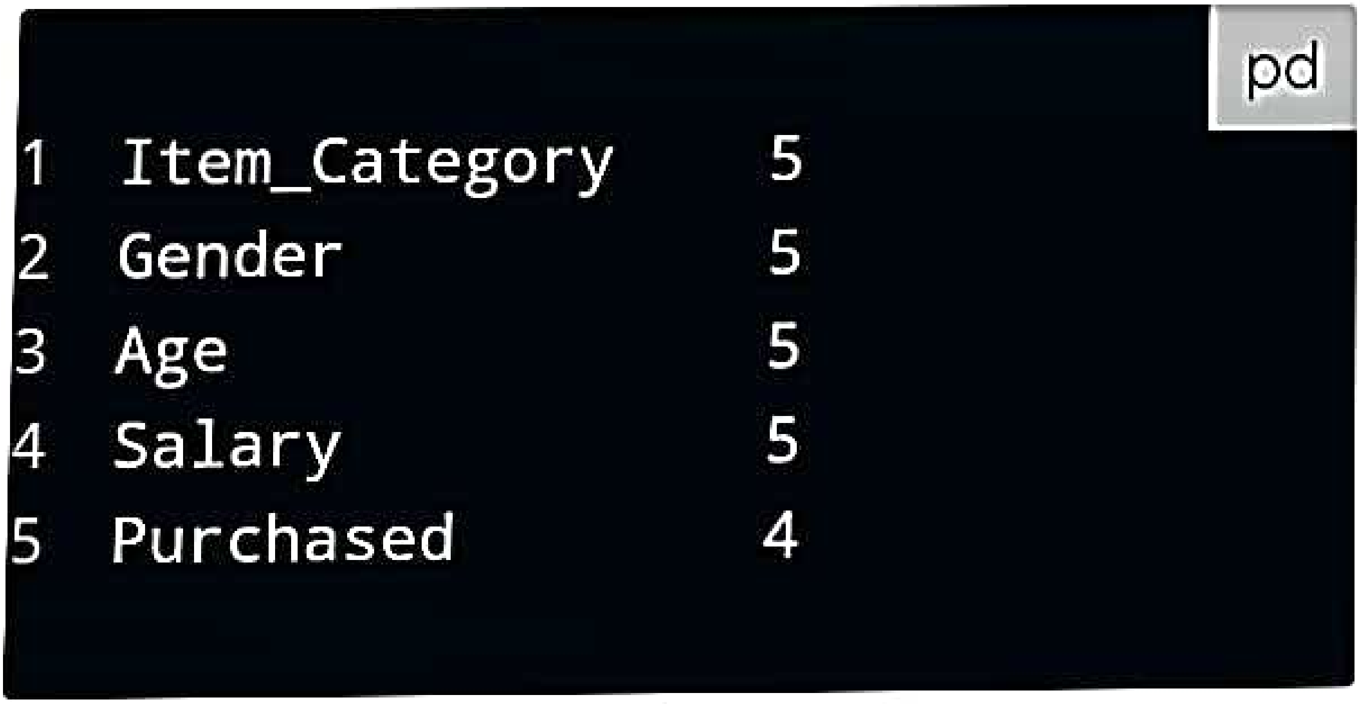


```
30000
70000
45000.0
```

Getting the Count for the Column

Count: This method is useful when the user is interested in getting the number of elements present per column. If there is any value that is left null than that is eliminated from the count. Assume if the value of purchased is left blank for one of the rows then following would be the output.

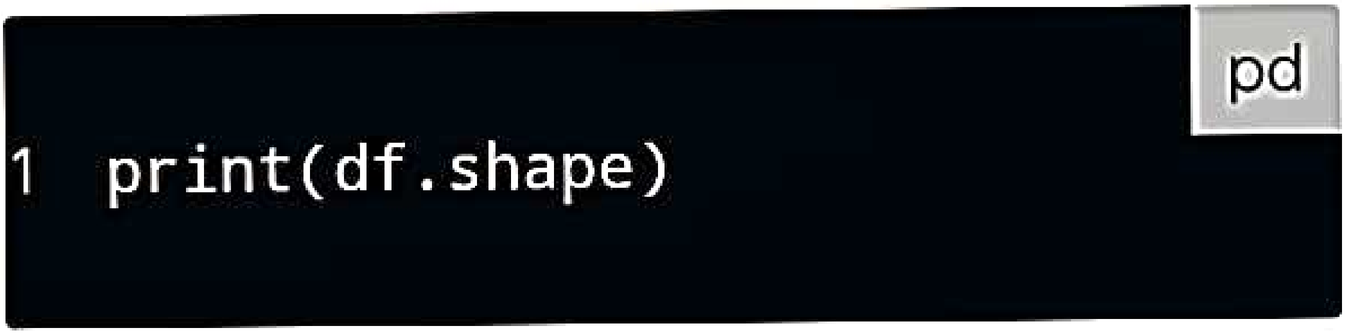
Output:



```
Item_Category    5
Gender           5
Age              5
Salary           5
Purchased        4
```

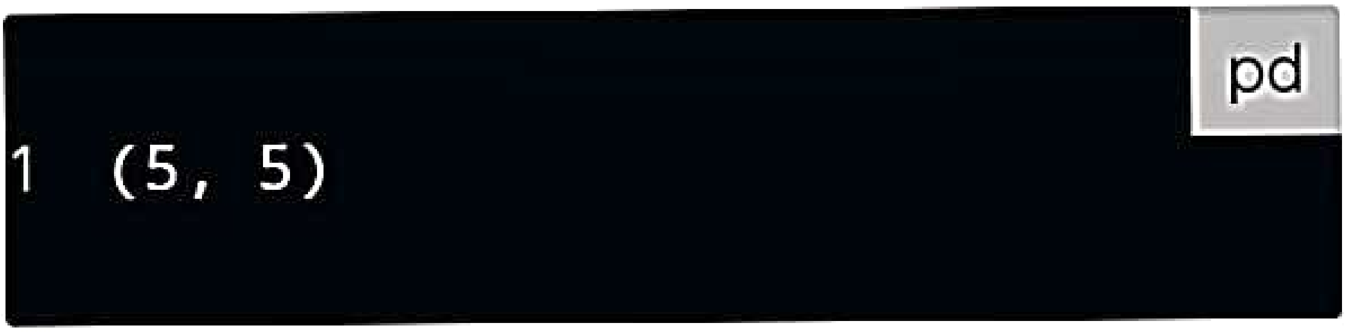
Shape and Size of the Dataframe

Shape is used to get the dimensions of the dataframe.



```
print(df.shape)
```

Output:



```
(5, 5)
```

Size is used to get the number of elements in the dataframe.

pd

```
print(df.size)
```

Output:

pd

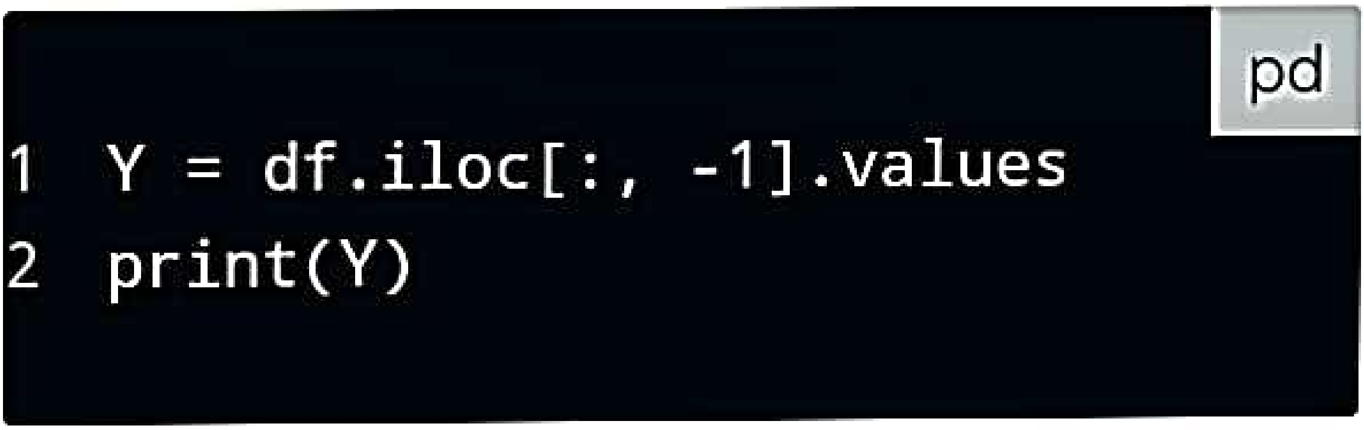
```
25
```

See the below equation:

```
y = 10a + 2b - 4.3c
```

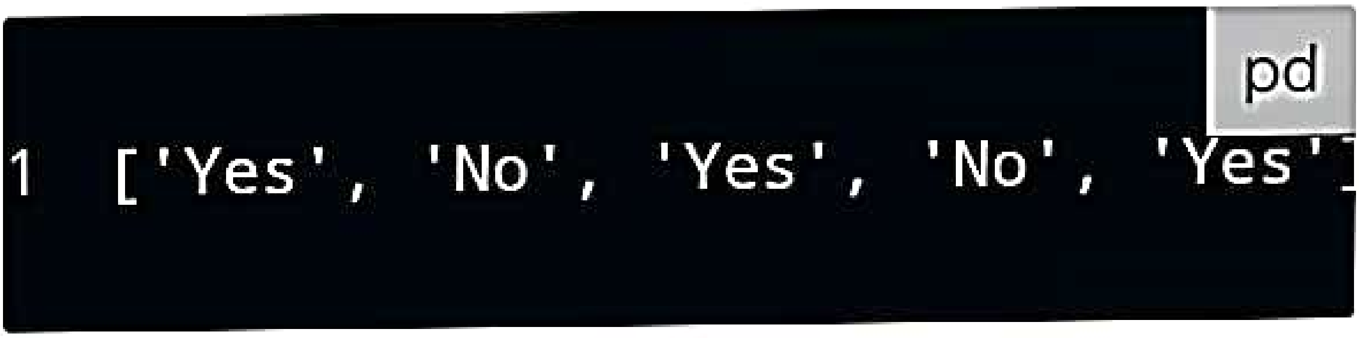
It demonstrates that the value of y is dependent on the value of a, b, and c. So, y is referred to as dependent feature or variable and a, b, and c are independent features or

Extracting the Dataset to Get the Dependent Vector



```
Y = df.iloc[:, -1].values
print(Y)
```

Output:



```
['Yes', 'No', 'Yes', 'No', 'Yes']
```

Conclusion

There are many other sophisticated methods available in Python Pandas that can help the user to import data from different sources to its dataframe. Once you have the data in the dataframe, it can then be used for various kinds of analysis. We also saw how to segregate the data into dependent and independent variables. In the next guide, we will see how to carry on a few more pre-