

**Advance Software Engineering (SE-406)**

**LAB A1-G3**

**Laboratory Manual**



**Department of Software Engineering**

**DELHI TECHNOLOGICAL UNIVERSITY(DTU)**

Shahbad Daulatpur, Bawana Road, Delhi-110042

**Submitted to: -**

Ms. Parul Sharma

**Submitted by:-**

Name: ASHISH KUMAR

Roll number: 2K18/SE/041

# EXPERIMENT 6

- ASHISH KUMAR

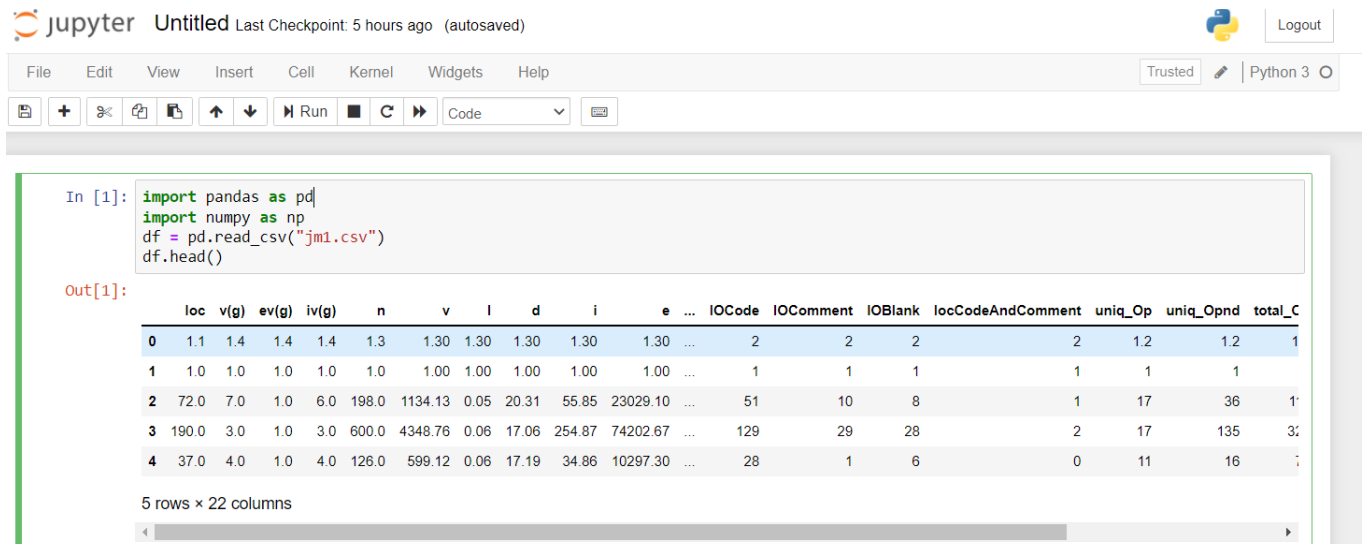
- 2K18/SE/041

**Aim:-** Preprocessing and cleaning of data collected in experiment 5.

**Introduction:-** Data Preprocessing is the most important step when we are building our model. In **Data Preprocessing** step, the data is transformed into a form where it becomes suitable for model ingestion. **Data Cleaning** is the process of analyzing data for finding incorrect, corrupt, and missing values and ablating it to make it suitable for input to data analytics and various machine learning algorithms.

## Code & Output:-

```
import pandas as pd
import numpy as np
df = pd.read_csv("jm1.csv")
df.head()
```



The screenshot shows a Jupyter Notebook interface. The top bar includes the Jupyter logo, the text "Untitled", and "Last Checkpoint: 5 hours ago (autosaved)". On the right, there is a "Logout" button. Below the top bar is a menu bar with "File", "Edit", "View", "Insert", "Cell", "Kernel", "Widgets", and "Help". To the right of the menu bar is a "Trusted" status indicator, a pencil icon, and "Python 3". Below the menu bar is a toolbar with icons for saving, adding cells, running, and other notebook functions. The main area of the notebook shows a code cell with the following code:

```
In [1]: import pandas as pd
import numpy as np
df = pd.read_csv("jm1.csv")
df.head()
```

Below the code cell is an output cell showing the first five rows of the data frame. The output is a table with 22 columns and 5 rows. The columns are: loc, v(g), ev(g), iv(g), n, v, l, d, i, e, ..., IOCode, IOComment, IOBlank, locCodeAndComment, uniq\_Op, uniq\_Opnd, total\_C. The rows are indexed 0 to 4.

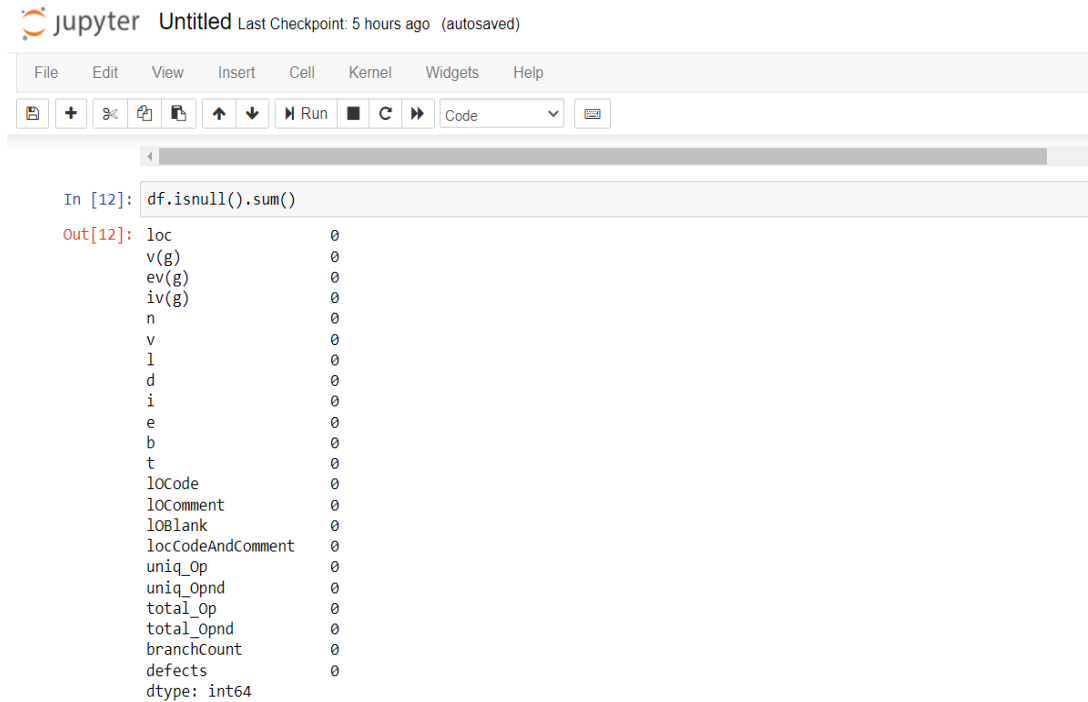
	loc	v(g)	ev(g)	iv(g)	n	v	l	d	i	e	...	IOCode	IOComment	IOBlank	locCodeAndComment	uniq_Op	uniq_Opnd	total_C
0	1.1	1.4	1.4	1.4	1.3	1.30	1.30	1.30	1.30	1.30	...	2	2	2	2	1.2	1.2	1
1	1.0	1.0	1.0	1.0	1.0	1.00	1.00	1.00	1.00	1.00	...	1	1	1	1	1	1	1
2	72.0	7.0	1.0	6.0	198.0	1134.13	0.05	20.31	55.85	23029.10	...	51	10	8	1	17	36	1
3	190.0	3.0	1.0	3.0	600.0	4348.76	0.06	17.06	254.87	74202.67	...	129	29	28	2	17	135	3
4	37.0	4.0	1.0	4.0	126.0	599.12	0.06	17.19	34.86	10297.30	...	28	1	6	0	11	16	1

Below the table, it says "5 rows x 22 columns".

df.isnull()

[illegible]

`df.isnull().sum()` #shows how many of the null values  
#it shows zero missing values.



The screenshot shows a Jupyter Notebook interface with the following components:

- Header: Jupyter logo, "Untitled", and "Last Checkpoint: 5 hours ago (autosaved)".
- Menu bar: File, Edit, View, Insert, Cell, Kernel, Widgets, Help.
- Toolbar: Icons for saving, adding cells, zooming, copying, pasting, undo, redo, and running code.
- Code cell: `In [12]: df.isnull().sum()`
- Output cell: `Out[12]:` followed by a list of variables and their corresponding sum of null values (all are 0).

Variable	Sum of Null Values
loc	0
v(g)	0
ev(g)	0
iv(g)	0
n	0
v	0
l	0
d	0
i	0
e	0
b	0
t	0
locCode	0
locComment	0
locBlank	0
locCodeAndComment	0
uniq_Op	0
uniq_Opnd	0
total_Op	0
total_Opnd	0
branchCount	0
defects	0
dtype: int64	

**Result:-** Since there are no missing value, no **data cleaning needed** because the data is all important.

**Learning from experiment:-** We are successful in Preprocessing and cleaning of a dataset collected in experiment 5 i.e. NASA Metrics Data Program.