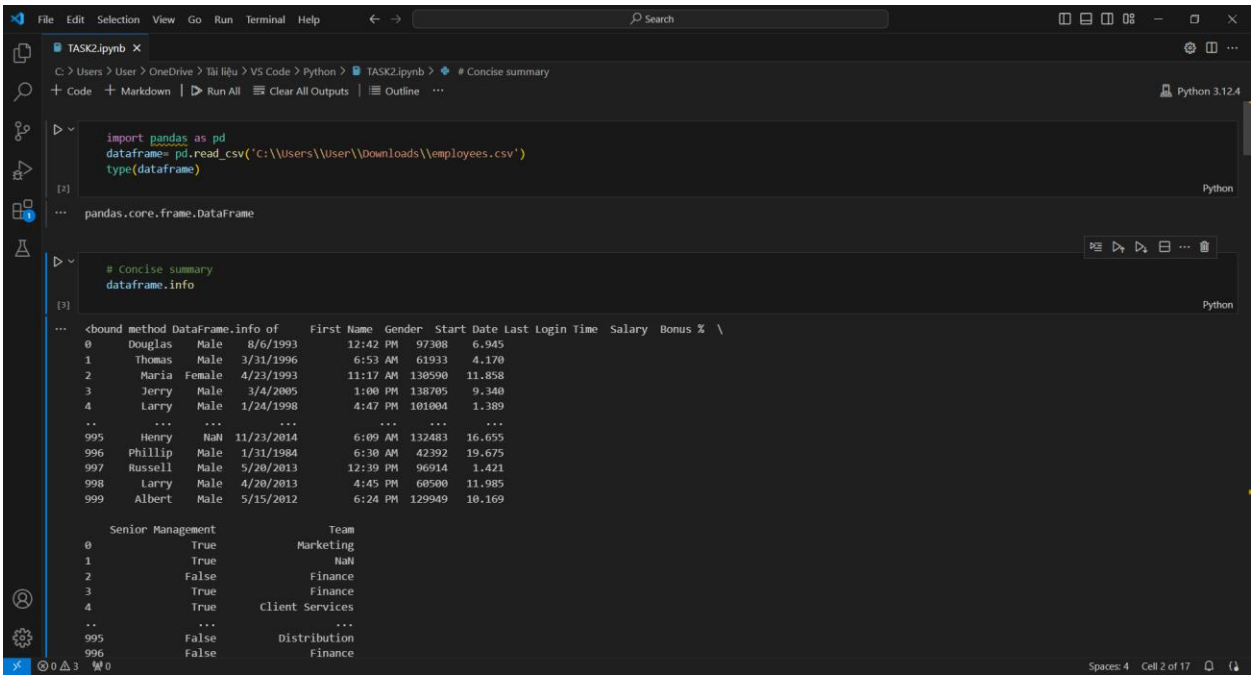
TASK 2

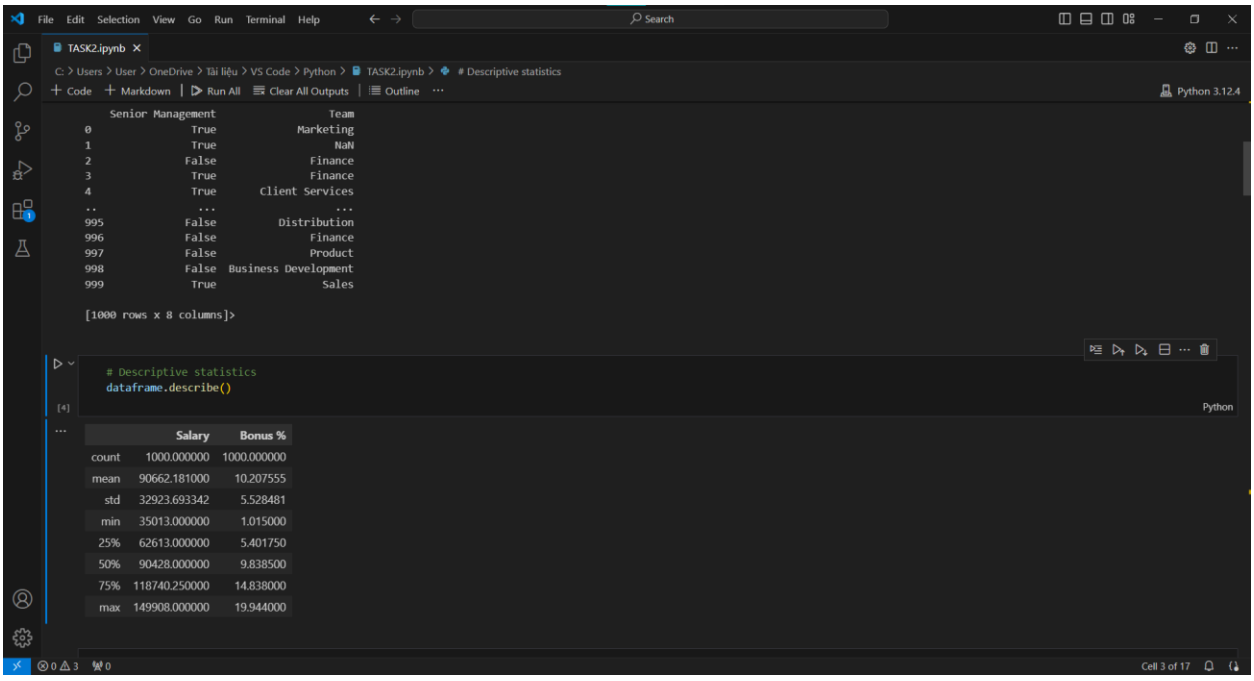


```
import pandas as pd
dataframe= pd.read_csv('C:\\Users\\User\\Downloads\\employees.csv')
type(dataframe)
```

```
pandas.core.frame.DataFrame
```

```
# Concise summary
dataframe.info
```

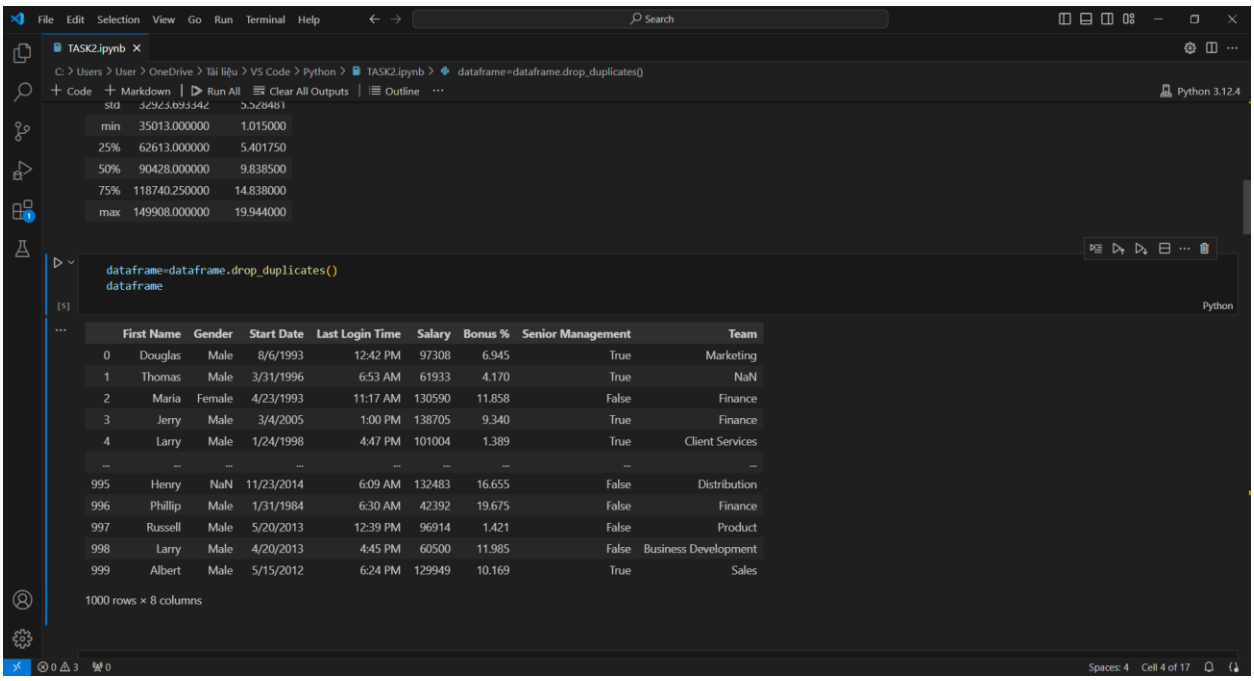
```
<bound method DataFrame.info of     First Name  Gender  Start Date Last Login Time  Salary  Bonus % \
0       Douglas    Male    8/6/1993       12:42 PM   97308    6.945
1        Thomas    Male   3/31/1996        6:53 AM   61933    4.170
2         Maria  Female   4/23/1993       11:17 AM  130590   11.858
3         Jerry    Male    3/4/2005        1:00 PM  138705    9.340
4         Larry    Male   1/24/1998        4:47 PM  101004    1.389
..          ...     ...         ...            ...     ...      ...
995       Henry     NaN  11/23/2014        6:09 AM  132483   16.655
996     Phillip    Male   1/31/1984        6:30 AM   42392   19.675
997     Russell    Male   5/20/2013       12:39 PM   96914    1.421
998       Larry    Male   4/20/2013        4:45 PM   60500   11.985
999      Albert    Male   5/15/2012        6:24 PM  129949   10.169

    Senior Management                Team
0                True           Marketing
1                True                 NaN
2               False             Finance
3                True             Finance
4                True      Client Services
..                ...                 ...
995             False        Distribution
996             False             Finance
```

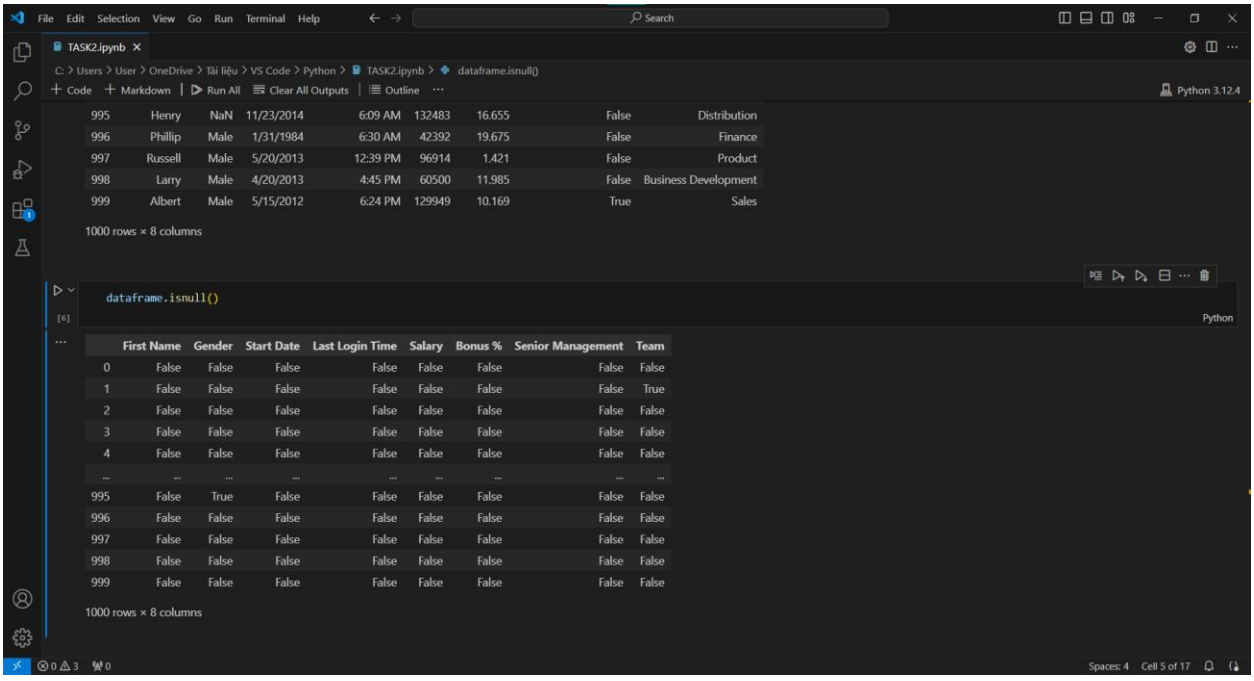


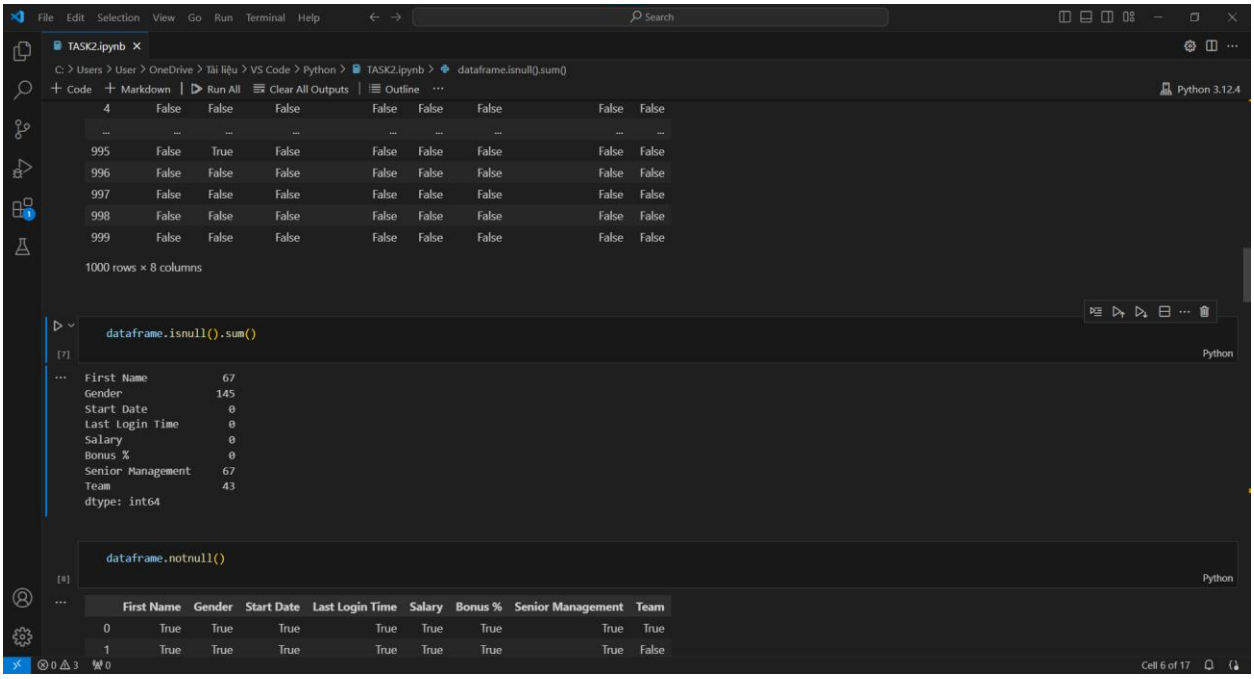
```
    Senior Management                Team
0                True           Marketing
1                True                 NaN
2               False             Finance
3                True             Finance
4                True      Client Services
..                ...                 ...
995             False        Distribution
996             False             Finance
997             False             Product
998             False  Business Development
999              True               Sales

[1000 rows x 8 columns]>
```

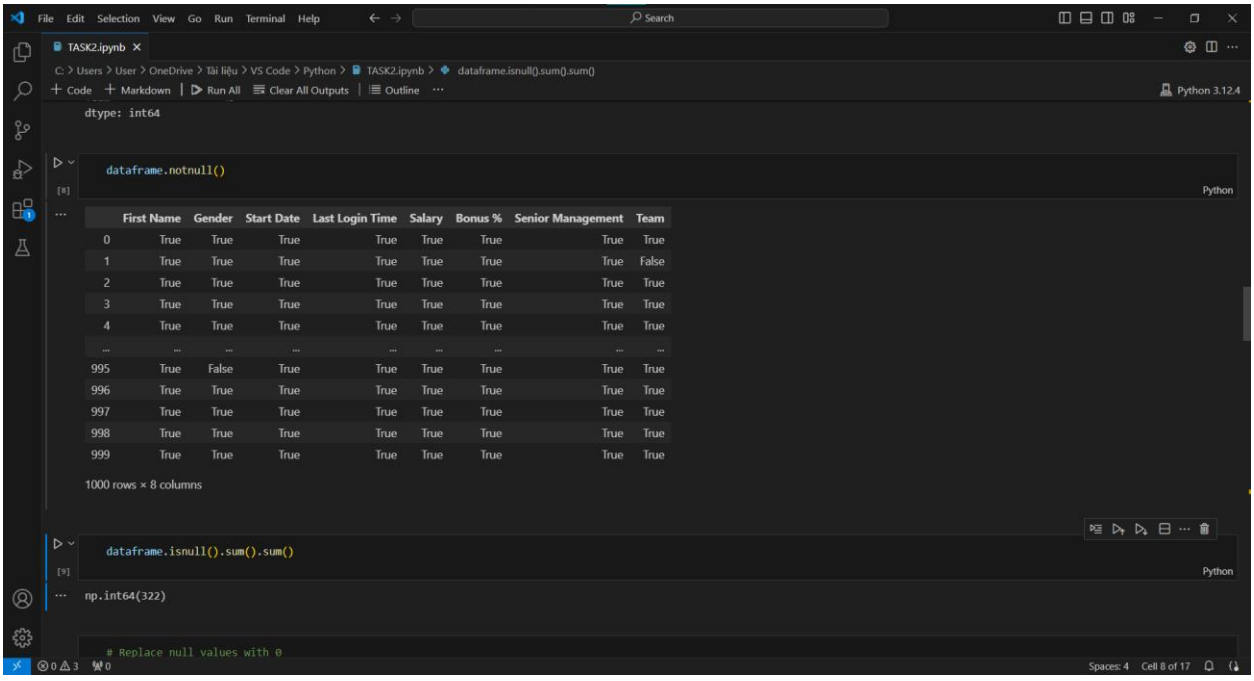
```
# Descriptive statistics
dataframe.describe()
```

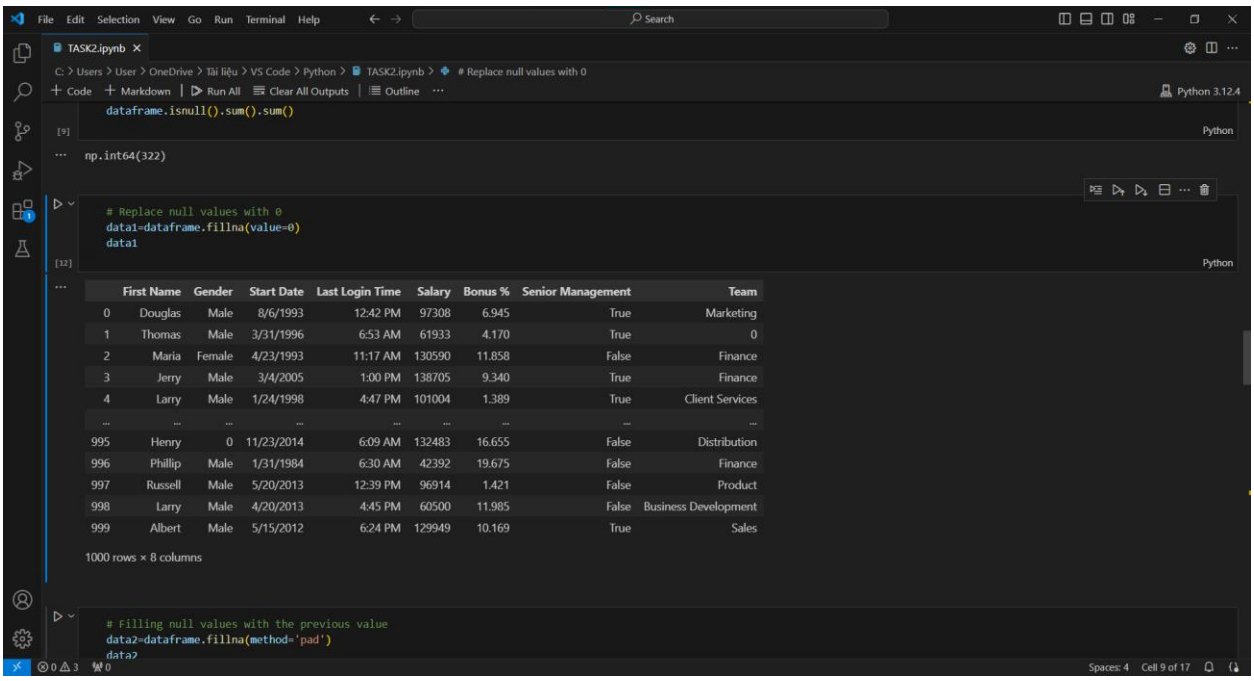
	Salary	Bonus %
count	1000.000000	1000.000000
mean	90662.181000	10.207555
std	32923.693342	5.528481
min	35013.000000	1.015000
25%	62613.000000	5.401750
50%	90428.000000	9.838500
75%	118740.250000	14.838000
max	149908.000000	19.944000

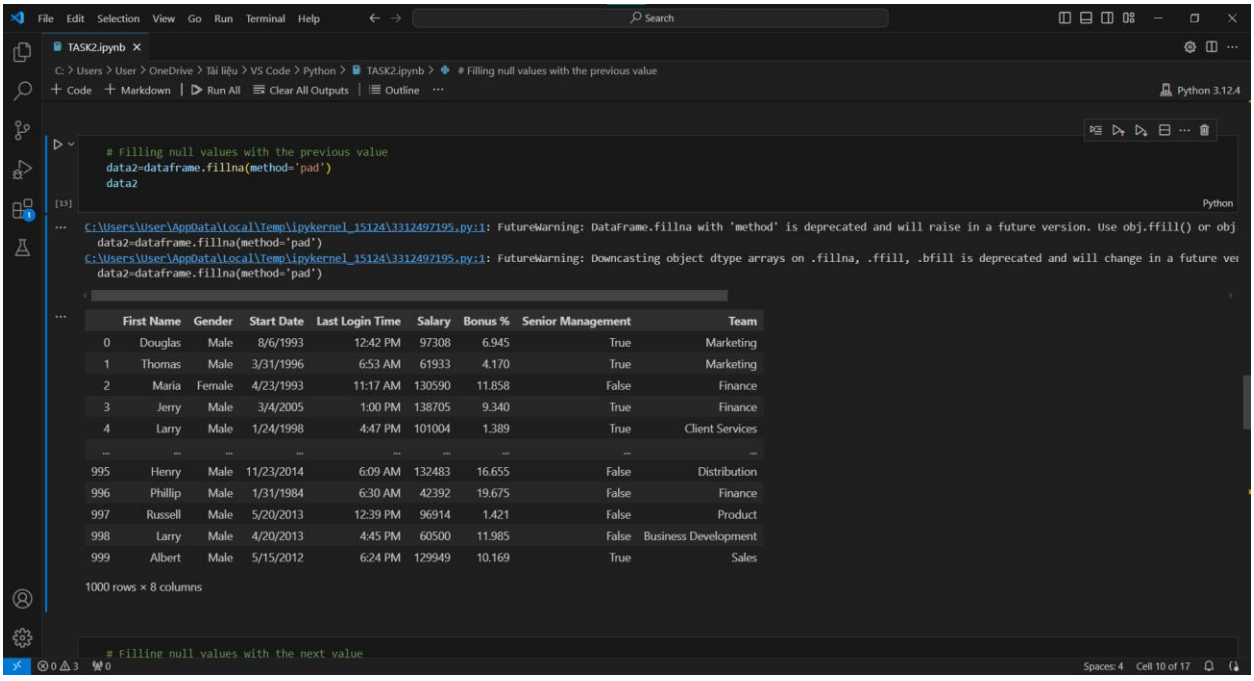

File Edit Selection View Go Run Terminal Help
Search
TASK2.ipynb
C: > Users > User > OneDrive > Tài liệu > VS Code > Python > TASK2.ipynb > dataframe=dataframe.drop_duplicates()
Code Markdown Run All Clear All Outputs Outline
Python 3.12.4
std 32923.693342 5.528481
min 35013.000000 1.015000
25% 62613.000000 5.401750
50% 90428.000000 9.838500
75% 118740.250000 14.838000
max 149908.000000 19.944000
dataframe=dataframe.drop_duplicates()
dataframe
[5]
Python
First Name Gender Start Date Last Login Time Salary Bonus % Senior Management Team
0 Douglas Male 8/6/1993 12:42 PM 97308 6.945 True Marketing
1 Thomas Male 3/31/1996 6:53 AM 61933 4.170 True NaN
2 Maria Female 4/23/1993 11:17 AM 130590 11.858 False Finance
3 Jerry Male 3/4/2005 1:00 PM 138705 9.340 True Finance
4 Larry Male 1/24/1998 4:47 PM 101004 1.389 True Client Services
995 Henry NaN 11/23/2014 6:09 AM 132483 16.655 False Distribution
996 Phillip Male 1/31/1984 6:30 AM 42392 19.675 False Finance
997 Russell Male 5/20/2013 12:39 PM 96914 1.421 False Product
998 Larry Male 4/20/2013 4:45 PM 60500 11.985 False Business Development
999 Albert Male 5/15/2012 6:24 PM 129949 10.169 True Sales
1000 rows × 8 columns
Spaces: 4 Cell 4 of 17

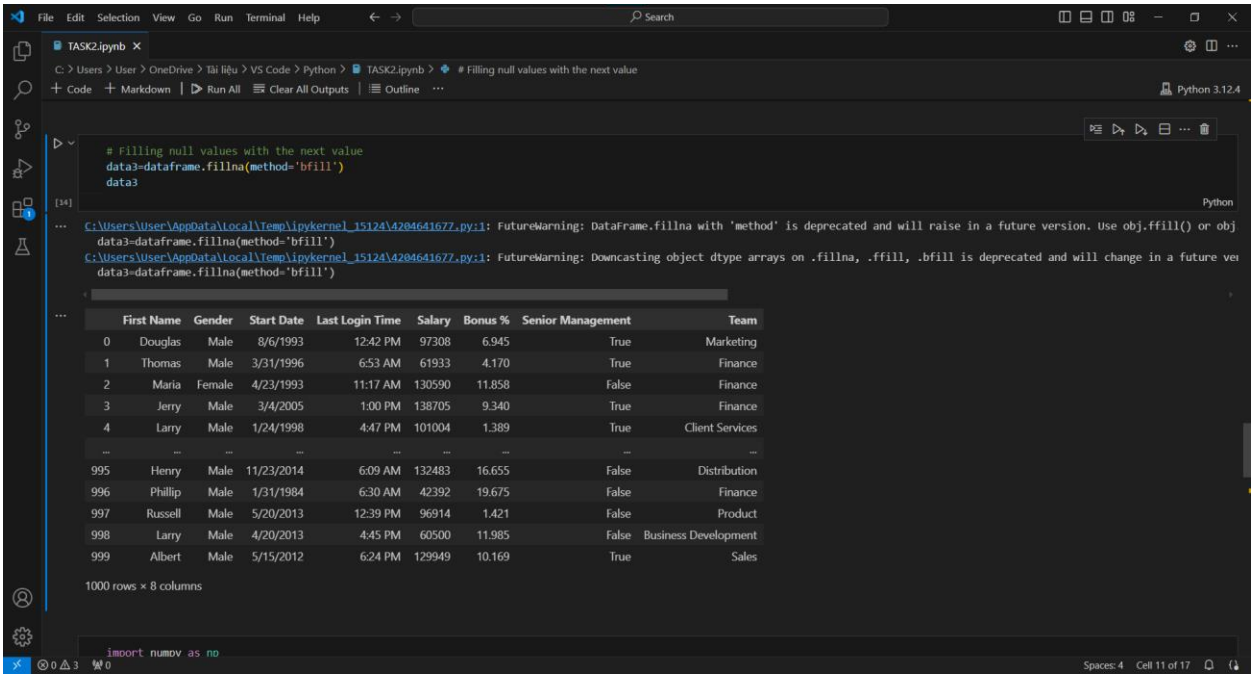

File Edit Selection View Go Run Terminal Help
Search
TASK2.ipynb
C: > Users > User > OneDrive > Tài liệu > VS Code > Python > TASK2.ipynb > dataframe.isnull()
Code Markdown Run All Clear All Outputs Outline
Python 3.12.4
995 Henry NaN 11/23/2014 6:09 AM 132483 16.655 False Distribution
996 Phillip Male 1/31/1984 6:30 AM 42392 19.675 False Finance
997 Russell Male 5/20/2013 12:39 PM 96914 1.421 False Product
998 Larry Male 4/20/2013 4:45 PM 60500 11.985 False Business Development
999 Albert Male 5/15/2012 6:24 PM 129949 10.169 True Sales
1000 rows × 8 columns
dataframe.isnull()
[6]
Python
First Name Gender Start Date Last Login Time Salary Bonus % Senior Management Team
0 False False False False False False False False
1 False False False False False False False True
2 False False False False False False False False
3 False False False False False False False False
4 False False False False False False False False
995 False True False False False False False False
996 False False False False False False False False
997 False False False False False False False False
998 False False False False False False False False
999 False False False False False False False False
1000 rows × 8 columns
Spaces: 4 Cell 5 of 17

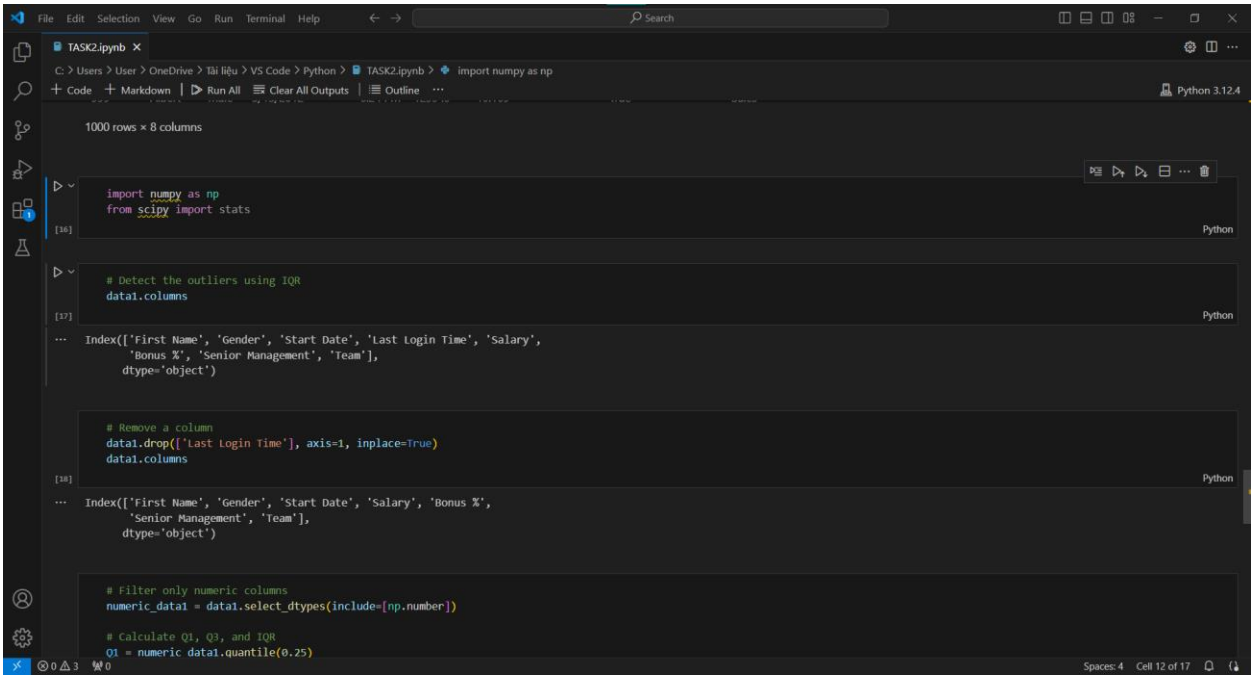

File Edit Selection View Go Run Terminal Help
Search
TASK2.ipynb
C: > Users > User > OneDrive > Tài liệu > VS Code > Python > TASK2.ipynb > dataframe.isnull().sum()
Code Markdown Run All Clear All Outputs Outline
Python 3.12.4
4 False False False False False False False False
995 False True False False False False False False
996 False False False False False False False False
997 False False False False False False False False
998 False False False False False False False False
999 False False False False False False False False
1000 rows × 8 columns
dataframe.isnull().sum()
Python
First Name 67
Gender 145
Start Date 0
Last Login Time 0
Salary 0
Bonus % 0
Senior Management 67
Team 43
dtype: int64
dataframe.notnull()
Python
First Name Gender Start Date Last Login Time Salary Bonus % Senior Management Team
0 True True True True True True True True
1 True True True True True True True False
Cell 6 of 17

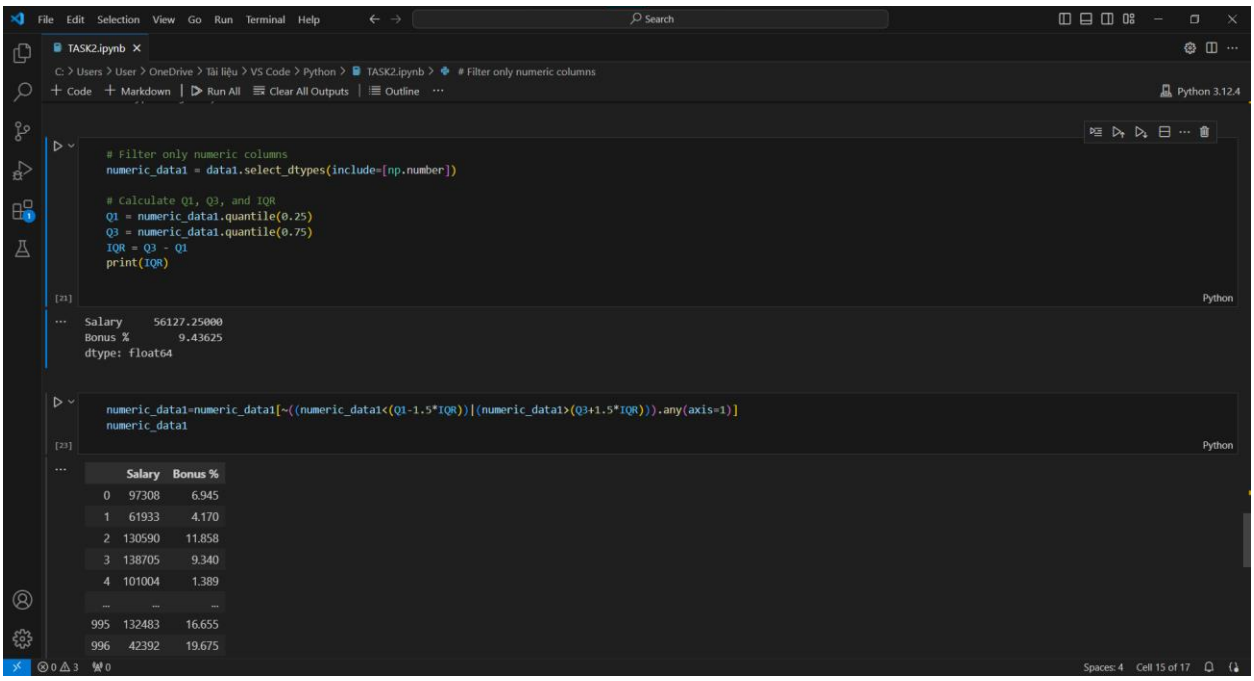

File Edit Selection View Go Run Terminal Help
Search
TASK2.ipynb
C: > Users > User > OneDrive > Tài liệu > VS Code > Python > TASK2.ipynb > dataframe.isnull().sum().sum()
Code Markdown Run All Clear All Outputs Outline
Python 3.12.4
dtype: int64
dataframe.notnull()
Python
First Name Gender Start Date Last Login Time Salary Bonus % Senior Management Team
0 True True True True True True True True
1 True True True True True True True False
2 True True True True True True True True
3 True True True True True True True True
4 True True True True True True True True
995 True False True True True True True True
996 True True True True True True True True
997 True True True True True True True True
998 True True True True True True True True
999 True True True True True True True True
1000 rows × 8 columns
dataframe.isnull().sum().sum()
Python
np.int64(322)
# Replace null values with 0
Spaces: 4 Cell 8 of 17

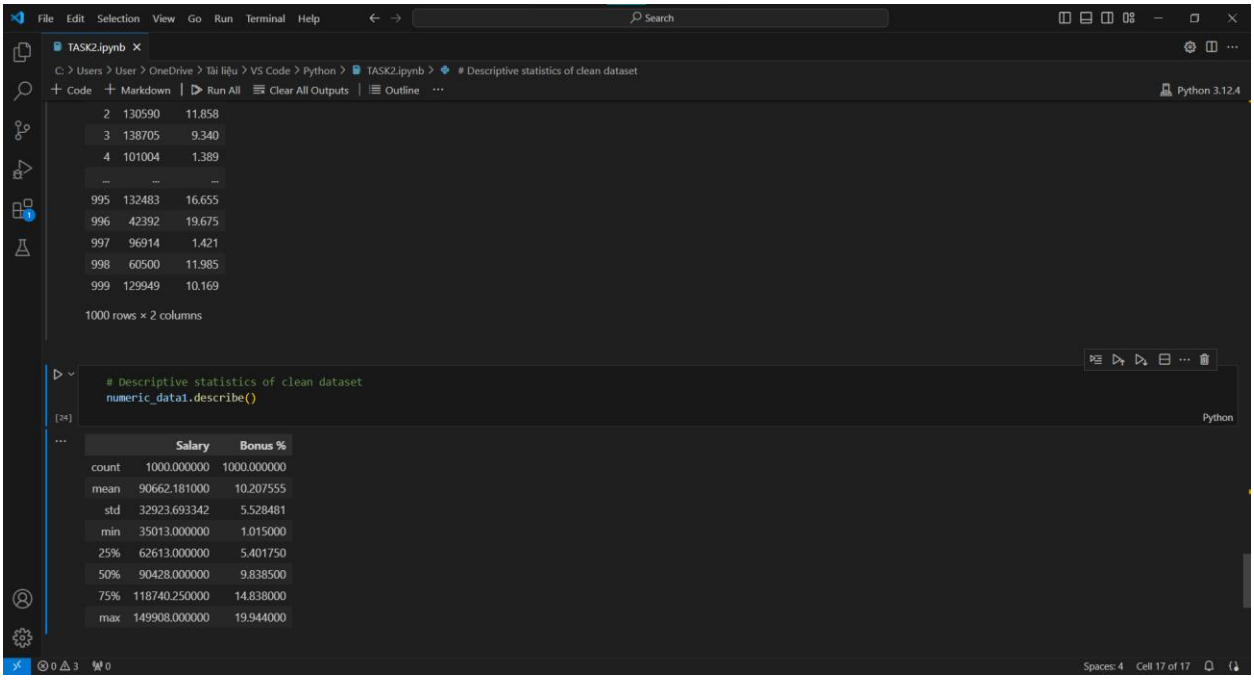

File Edit Selection View Go Run Terminal Help
Search
TASK2.ipynb
C: > Users > User > OneDrive > Tài liệu > VS Code > Python > TASK2.ipynb > # Replace null values with 0
+ Code + Markdown | Run All Clear All Outputs | Outline ...
Python 3.12.4
dataframe.isnull().sum().sum()
[9]
Python
np.int64(322)
# Replace null values with 0
data1=dataframe.fillna(value=0)
data1
[12]
Python
First Name Gender Start Date Last Login Time Salary Bonus % Senior Management Team
0 Douglas Male 8/6/1993 12:42 PM 97308 6.945 True Marketing
1 Thomas Male 3/31/1996 6:53 AM 61933 4.170 True 0
2 Maria Female 4/23/1993 11:17 AM 130590 11.858 False Finance
3 Jerry Male 3/4/2005 1:00 PM 138705 9.340 True Finance
4 Larry Male 1/24/1998 4:47 PM 101004 1.389 True Client Services
995 Henry 0 11/23/2014 6:09 AM 132483 16.655 False Distribution
996 Phillip Male 1/31/1984 6:30 AM 42392 19.675 False Finance
997 Russell Male 5/20/2013 12:39 PM 96914 1.421 False Product
998 Larry Male 4/20/2013 4:45 PM 60500 11.985 False Business Development
999 Albert Male 5/15/2012 6:24 PM 129949 10.169 True Sales
1000 rows × 8 columns
# Filling null values with the previous value
data2=dataframe.fillna(method='pad')
data2
Spaces: 4 Cell 9 of 17


File Edit Selection View Go Run Terminal Help
Search
TASK2.ipynb
C: > Users > User > OneDrive > Tài liệu > VS Code > Python > TASK2.ipynb > # Filling null values with the previous value
+ Code + Markdown | Run All Clear All Outputs | Outline ...
Python 3.12.4
# Filling null values with the previous value
data2=dataframe.fillna(method='pad')
data2
[13]
Python
C:\Users\User\AppData\Local\Temp\ipykernel_15124\3312497195.py:1: FutureWarning: DataFrame.fillna with 'method' is deprecated and will raise in a future version. Use obj.ffill() or obj
data2=dataframe.fillna(method='pad')
C:\Users\User\AppData\Local\Temp\ipykernel_15124\3312497195.py:1: FutureWarning: Downcasting object dtype arrays on .fillna, .ffill, .bfill is deprecated and will change in a future ver
data2=dataframe.fillna(method='pad')
First Name Gender Start Date Last Login Time Salary Bonus % Senior Management Team
0 Douglas Male 8/6/1993 12:42 PM 97308 6.945 True Marketing
1 Thomas Male 3/31/1996 6:53 AM 61933 4.170 True Marketing
2 Maria Female 4/23/1993 11:17 AM 130590 11.858 False Finance
3 Jerry Male 3/4/2005 1:00 PM 138705 9.340 True Finance
4 Larry Male 1/24/1998 4:47 PM 101004 1.389 True Client Services
995 Henry Male 11/23/2014 6:09 AM 132483 16.655 False Distribution
996 Phillip Male 1/31/1984 6:30 AM 42392 19.675 False Finance
997 Russell Male 5/20/2013 12:39 PM 96914 1.421 False Product
998 Larry Male 4/20/2013 4:45 PM 60500 11.985 False Business Development
999 Albert Male 5/15/2012 6:24 PM 129949 10.169 True Sales
1000 rows × 8 columns
Filling null values with the next value
Spaces: 4 Cell 10 of 17


File Edit Selection View Go Run Terminal Help
Search
TASK2.ipynb
C: > Users > User > OneDrive > Tài liệu > VS Code > Python > TASK2.ipynb > # Filling null values with the next value
Code Markdown Run All Clear All Outputs Outline
Python 3.12.4
# Filling null values with the next value
data3=dataframe.fillna(method='bfill')
data3
[14]
Python
C:\Users\User\AppData\Local\Temp\ipykernel_15124\4204641677.py:1: FutureWarning: DataFrame.fillna with 'method' is deprecated and will raise in a future version. Use obj.ffill() or obj
data3=dataframe.fillna(method='bfill')
C:\Users\User\AppData\Local\Temp\ipykernel_15124\4204641677.py:1: FutureWarning: Downcasting object dtype arrays on .fillna, .ffill, .bfill is deprecated and will change in a future ver
data3=dataframe.fillna(method='bfill')
First Name Gender Start Date Last Login Time Salary Bonus % Senior Management Team
0 Douglas Male 8/6/1993 12:42 PM 97308 6.945 True Marketing
1 Thomas Male 3/31/1996 6:53 AM 61933 4.170 True Finance
2 Maria Female 4/23/1993 11:17 AM 130590 11.858 False Finance
3 Jerry Male 3/4/2005 1:00 PM 138705 9.340 True Finance
4 Larry Male 1/24/1998 4:47 PM 101004 1.389 True Client Services
995 Henry Male 11/23/2014 6:09 AM 132483 16.655 False Distribution
996 Phillip Male 1/31/1984 6:30 AM 42392 19.675 False Finance
997 Russell Male 5/20/2013 12:39 PM 96914 1.421 False Product
998 Larry Male 4/20/2013 4:45 PM 60500 11.985 False Business Development
999 Albert Male 5/15/2012 6:24 PM 129949 10.169 True Sales
1000 rows × 8 columns
import numpy as np
Spaces: 4 Cell 11 of 17


File Edit Selection View Go Run Terminal Help
Search
TASK2.ipynb
C: > Users > User > OneDrive > Tài liệu > VS Code > Python > TASK2.ipynb > import numpy as np
Code Markdown Run All Clear All Outputs Outline
Python 3.12.4
1000 rows × 8 columns
import numpy as np
from scipy import stats
[16]
Python
Detect the outliers using IQR
data1.columns
[17]
Python
Index(['First Name', 'Gender', 'Start Date', 'Last Login Time', 'Salary',
'Bonus %', 'Senior Management', 'Team'],
dtype='object')
# Remove a column
data1.drop(['Last Login Time'], axis=1, inplace=True)
data1.columns
[18]
Python
Index(['First Name', 'Gender', 'Start Date', 'Salary', 'Bonus %',
'Senior Management', 'Team'],
dtype='object')
# Filter only numeric columns
numeric_data1 = data1.select_dtypes(include=[np.number])
# Calculate Q1, Q3, and IQR
Q1 = numeric_data1.quantile(0.25)
Spaces: 4 Cell 12 of 17


File Edit Selection View Go Run Terminal Help
Search
TASK2.ipynb
C: > Users > User > OneDrive > Tài liệu > VS Code > Python > TASK2.ipynb > # Filter only numeric columns
Code Markdown Run All Clear All Outputs Outline
Python 3.12.4
# Filter only numeric columns
numeric_data1 = data1.select_dtypes(include=[np.number])
# Calculate Q1, Q3, and IQR
Q1 = numeric_data1.quantile(0.25)
Q3 = numeric_data1.quantile(0.75)
IQR = Q3 - Q1
print(IQR)
[21]
Python
Salary 56127.25000
Bonus % 9.43625
dtype: float64
numeric_data1=numeric_data1[~((numeric_data1<(Q1-1.5*IQR))|(numeric_data1>(Q3+1.5*IQR))).any(axis=1)]
numeric_data1
[23]
Python
Salary Bonus %
0 97308 6.945
1 61933 4.170
2 130590 11.858
3 138705 9.340
4 101004 1.389
... ...
995 132483 16.655
996 42392 19.675
Spaces: 4 Cell 15 of 17


File Edit Selection View Go Run Terminal Help
Search
TASK2.ipynb
C: > Users > User > OneDrive > Tài liệu > VS Code > Python > TASK2.ipynb > # Descriptive statistics of clean dataset
Code Markdown Run All Clear All Outputs Outline
Python 3.12.4
2 130590 11.858
3 138705 9.340
4 101004 1.389
... ...
995 132483 16.655
996 42392 19.675
997 96914 1.421
998 60500 11.985
999 129949 10.169
1000 rows × 2 columns
# Descriptive statistics of clean dataset
numeric_data1.describe()
[24]
Python
Salary Bonus %
count 1000.000000 1000.000000
mean 90662.181000 10.207555
std 32923.693342 5.528481
min 35013.000000 1.015000
25% 62613.000000 5.401750
50% 90428.000000 9.838500
75% 118740.250000 14.838000
max 149908.000000 19.944000
Spaces: 4 Cell 17 of 17