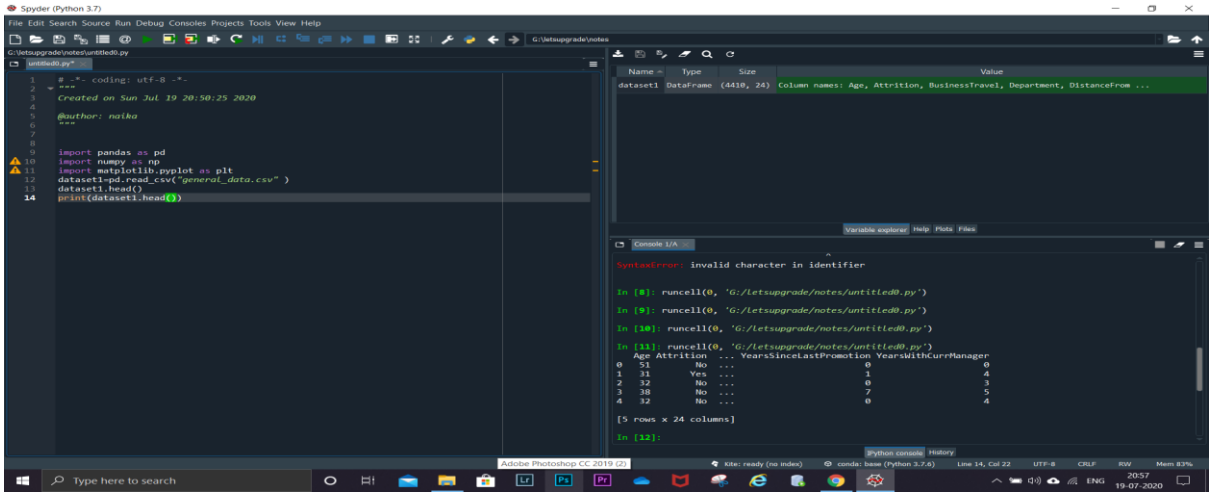
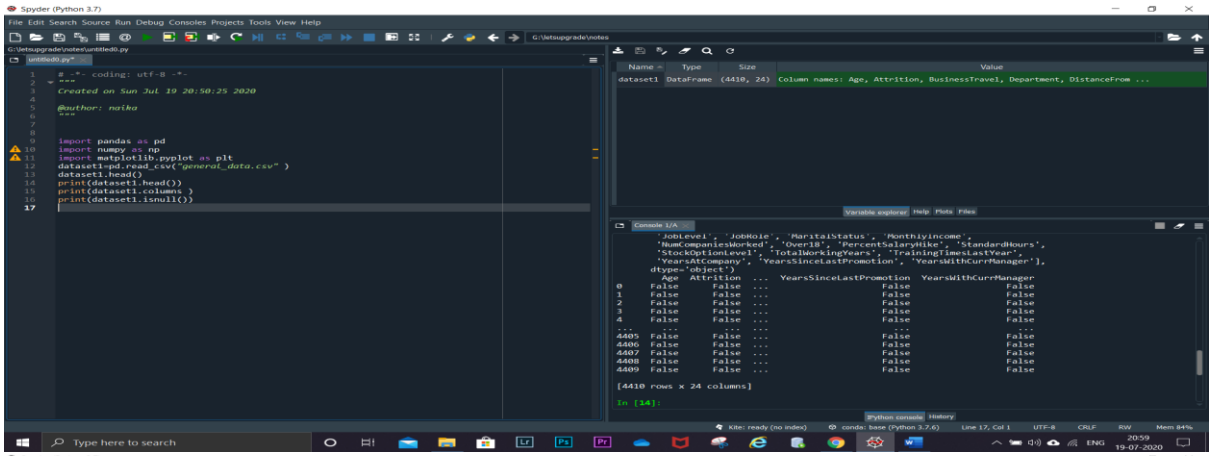
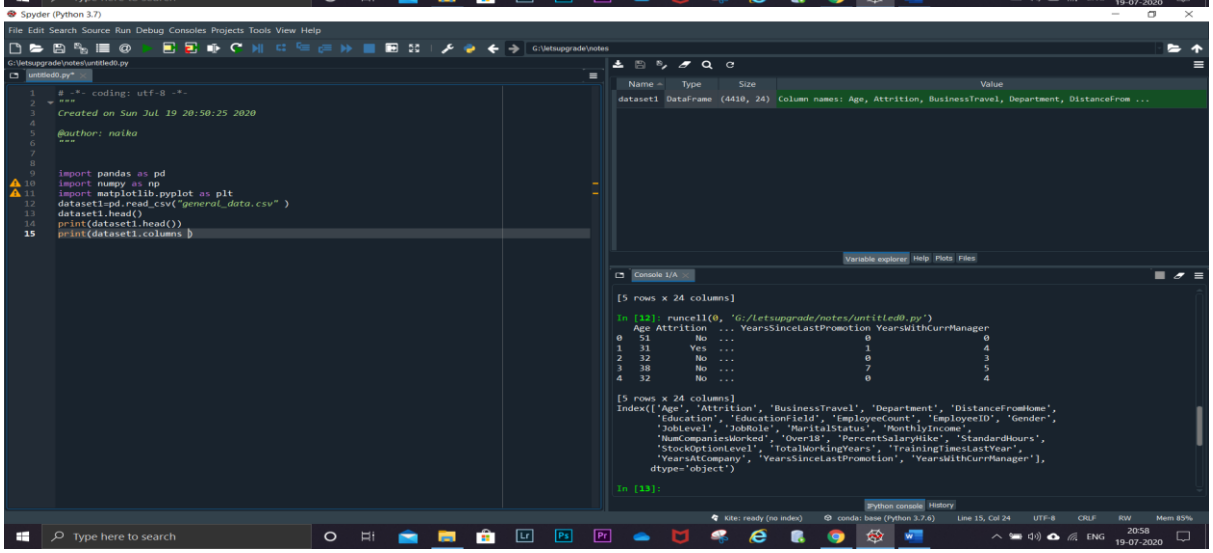
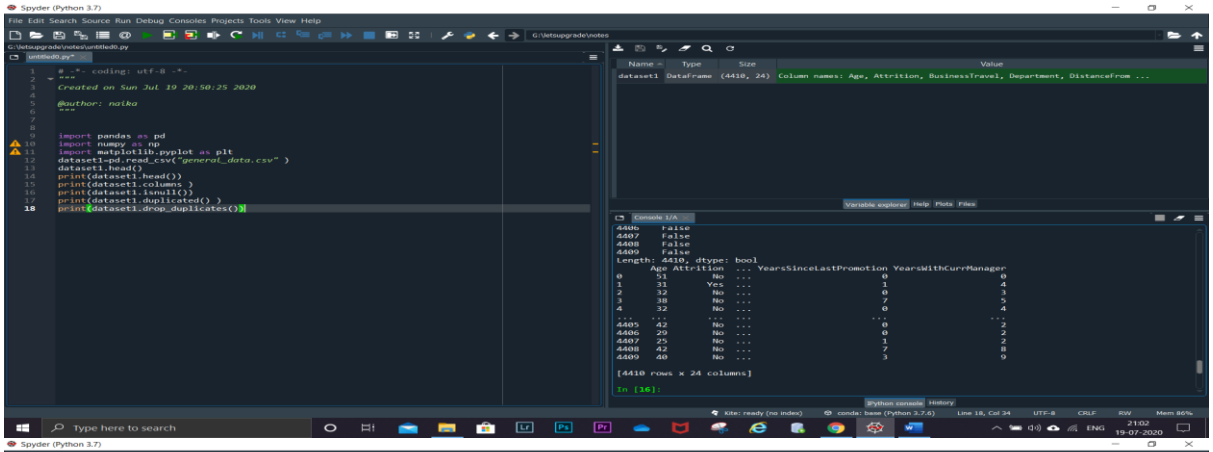
DAY 7 ASSIGNMENT

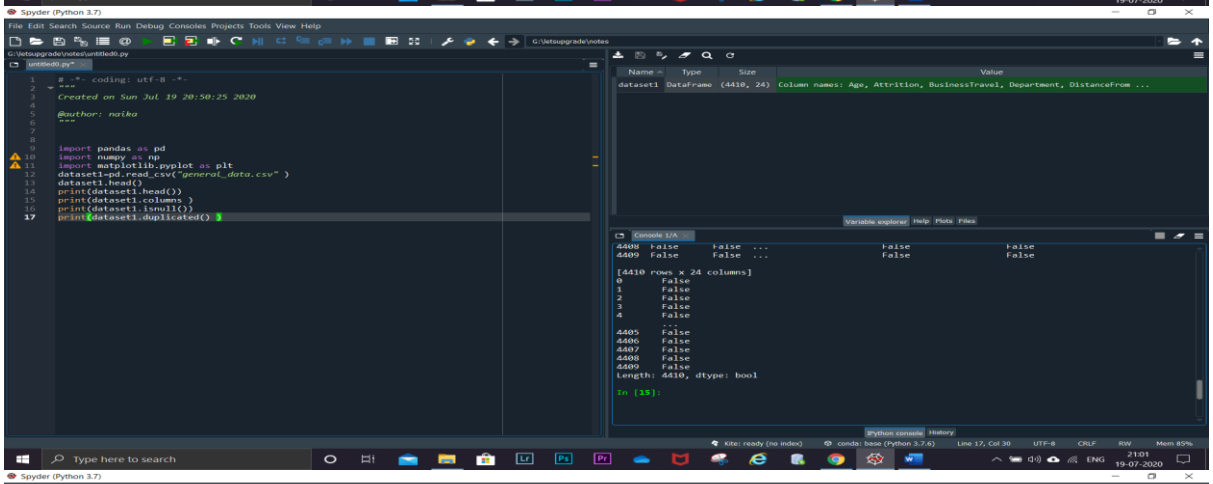
```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

dataset1=pd.read_csv("general_data.csv" )

dataset1.head()

print(dataset1.head())

print(dataset1.columns )

print(dataset1.isnull())

print(dataset1.duplicated() )

print(dataset1.drop_duplicates())

dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears',
'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()

print(dataset3 )

dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears',
'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].median()

print(dataset3 )

dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears',
'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].mode()

print(dataset3 )

dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears',
'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].var()

print(dataset3 )

dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears',
'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].skew()

print(dataset3 )

dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked', 'PercentSalaryHike','TotalWorkingYears',
'TrainingTimesLastYear', 'YearsAtCompany','YearsSinceLastPromotion', 'YearsWithCurrManager']].kurt()

print(dataset3 )

box_plot=dataset1.Age

plt.boxplot(box_plot)

box_plot=dataset1.MonthlyIncome

plt.boxplot(box_plot)

box_plot=dataset1.YearsAtCompany

plt.boxplot(box_plot)
```

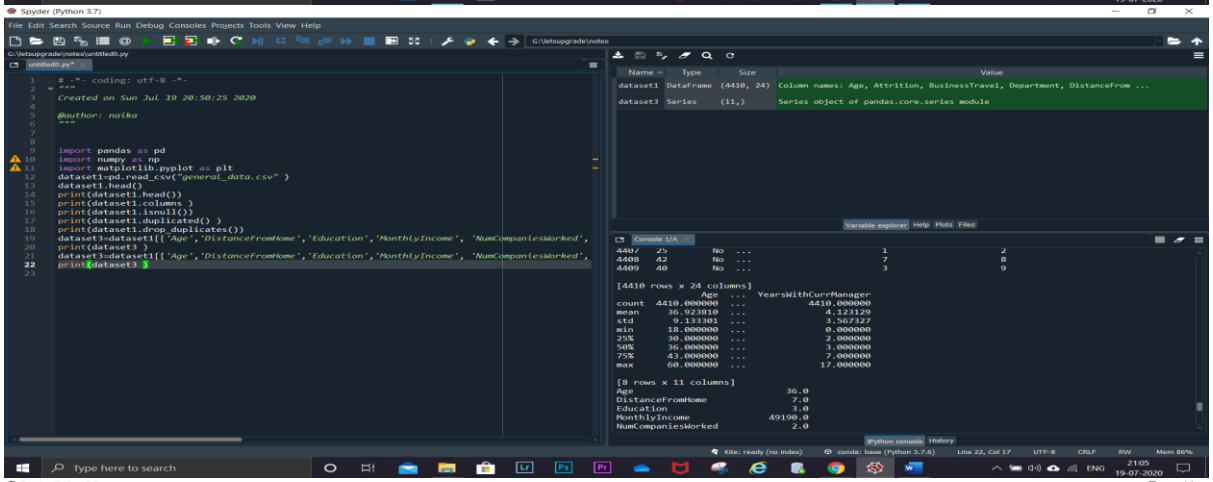

Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
G:\letsupgrade\notes
untitled0.py
-*- coding: utf-8 -*-
"""
Created on Sun Jul 19 20:50:25 2020
@author: naika
"""
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
dataset1=pd.read_csv("general_data.csv")
dataset1.head()
print(dataset1.head())
Name Type Size Value
dataset1 DataFrame (4410, 24) Column names: Age, Attrition, BusinessTravel, Department, DistanceFrom ...
Variable explorer Help Plots Files
Console 1/A
SyntaxError: invalid character in identifier
In [8]: runcell(0, 'G:/letsupgrade/notes/untitled0.py')
In [9]: runcell(0, 'G:/letsupgrade/notes/untitled0.py')
In [10]: runcell(0, 'G:/letsupgrade/notes/untitled0.py')
In [11]: runcell(0, 'G:/letsupgrade/notes/untitled0.py')
Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager
0 51 No ... 0 0
1 31 Yes ... 1 4
2 32 No ... 0 3
3 38 No ... 7 5
4 32 No ... 0 4
[5 rows x 24 columns]
In [12]:
Python console History
Kite: ready (no index)
conda: base (Python 3.7.6)
Line 14, Col 22
UTF-8
CRLF
RW
Mem 83%
Adobe Photoshop CC 2019 (2)
Type here to search
ENG
20:57
19-07-2020

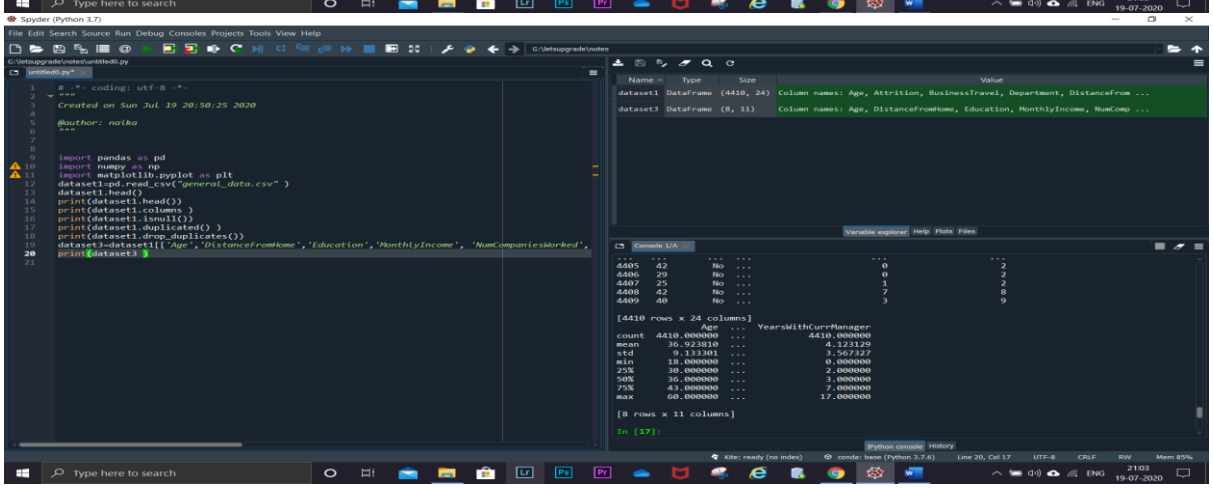

Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
G:\letsupgrade\notes
untitled0.py
-*- coding: utf-8 -*-
"""
Created on Sun Jul 19 20:50:25 2020
@author: naika
"""
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
dataset1=pd.read_csv("general_data.csv")
dataset1.head()
print(dataset1.head())
print(dataset1.columns)
print(dataset1.isnull())
Name Type Size Value
dataset1 DataFrame (4410, 24) Column names: Age, Attrition, BusinessTravel, Department, DistanceFrom ...
Variable explorer Help Plots Files
Console 1/A
'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',
'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',
'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager'],
dtype='object')
Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager
0 False False ... False False
1 False False ... False False
2 False False ... False False
3 False False ... False False
4 False False ... False False
... ...
4405 False False ... False False
4406 False False ... False False
4407 False False ... False False
4408 False False ... False False
4409 False False ... False False
[4410 rows x 24 columns]
In [14]:
Python console History
Kite: ready (no index)
conda: base (Python 3.7.6)
Line 17, Col 1
UTF-8
CRLF
RW
Mem 84%
Type here to search
ENG
20:59
19-07-2020

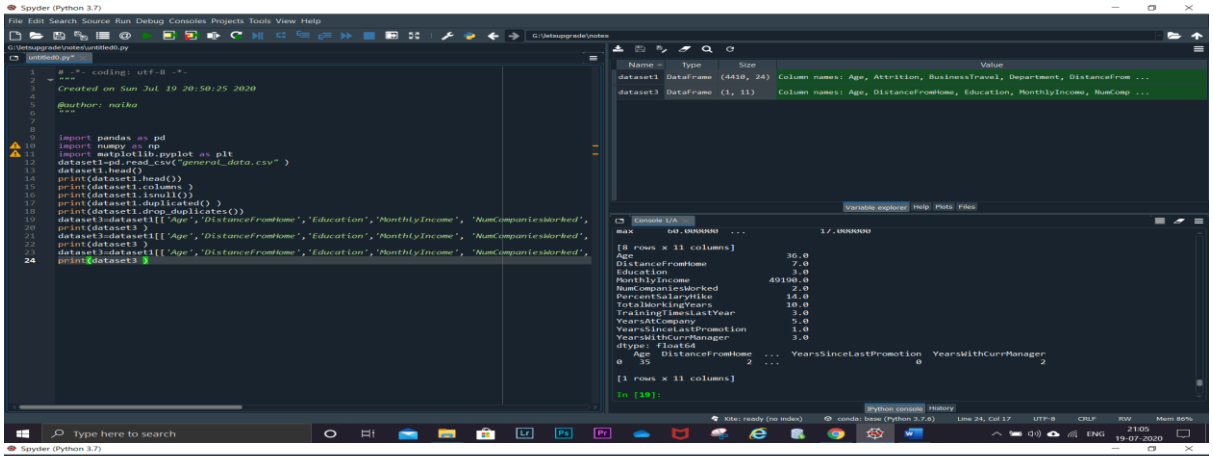

Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
G:\letsupgrade\notes
untitled0.py
-*- coding: utf-8 -*-
"""
Created on Sun Jul 19 20:50:25 2020
@author: naika
"""
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
dataset1=pd.read_csv("general_data.csv")
dataset1.head()
print(dataset1.head())
print(dataset1.columns)
Name Type Size Value
dataset1 DataFrame (4410, 24) Column names: Age, Attrition, BusinessTravel, Department, DistanceFrom ...
Variable explorer Help Plots Files
Console 1/A
[5 rows x 24 columns]
In [12]: runcell(0, 'G:/letsupgrade/notes/untitled0.py')
Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager
0 51 No ... 0 0
1 31 Yes ... 1 4
2 32 No ... 0 3
3 38 No ... 7 5
4 32 No ... 0 4
[5 rows x 24 columns]
Index(['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',
'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender',
'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',
'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',
'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager'],
dtype='object')
In [13]:
Python console History
Kite: ready (no index)
conda: base (Python 3.7.6)
Line 15, Col 24
UTF-8
CRLF
RW
Mem 85%
Type here to search
ENG
20:58
19-07-2020

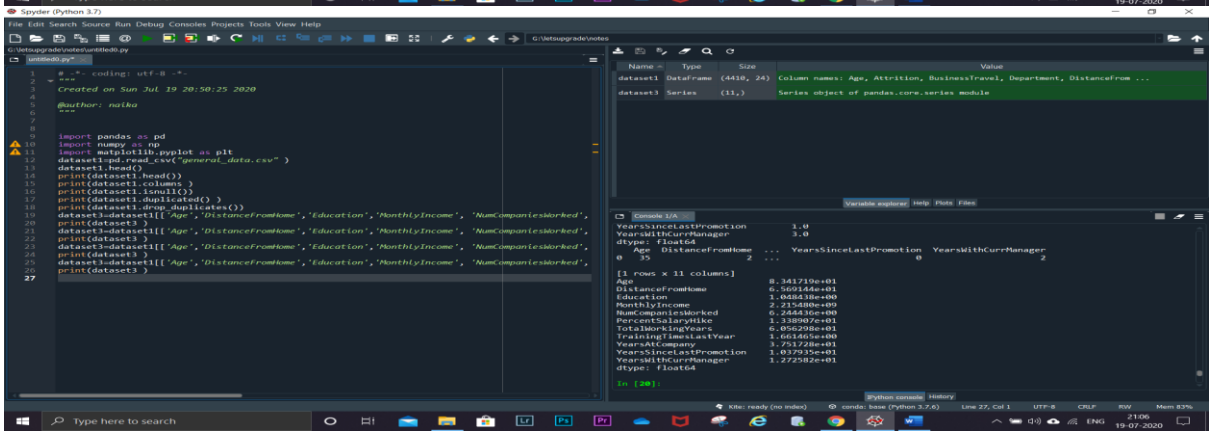

Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
G:\letsupgrade\notes
untitled0.py
-*- coding: utf-8 -*-
"""
Created on Sun Jul 19 20:50:25 2020
@author: naika
"""
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
dataset1=pd.read_csv("general_data.csv")
dataset1.head()
print(dataset1.head())
print(dataset1.columns)
print(dataset1.isnull())
print(dataset1.duplicated())
print(dataset1.drop_duplicates())
Name Type Size Value
dataset1 DataFrame (4410, 24) Column names: Age, Attrition, BusinessTravel, Department, DistanceFrom ...
Variable explorer Help Plots Files
Console 1/A
4406 False
4407 False
4408 False
4409 False
Length: 4410, dtype: bool
Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager
0 51 No ... 0 0
1 31 Yes ... 1 4
2 32 No ... 0 3
3 38 No ... 7 5
4 32 No ... 0 4
... ...
4405 42 No ... 0 2
4406 29 No ... 0 2
4407 25 No ... 1 2
4408 42 No ... 7 8
4409 40 No ... 3 9
[4410 rows x 24 columns]
In [16]:
Python console History
Kite: ready (no index)
conda: base (Python 3.7.6)
Line 18, Col 34
UTF-8
CRLF
RW
Mem 86%
Type here to search
ENG
21:02
19-07-2020

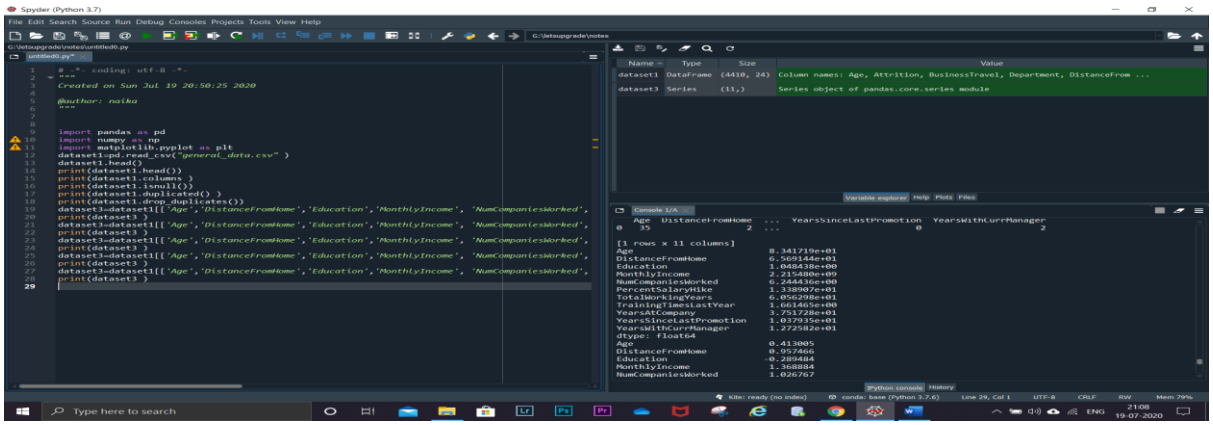

Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
G:\letsupgrade\notes
untitled0.py*
-*- coding: utf-8 -*-
"""
Created on Sun Jul 19 20:50:25 2020
@author: naika
"""
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
dataset1=pd.read_csv("general_data.csv")
dataset1.head()
print(dataset1.head())
print(dataset1.columns)
print(dataset1.isnull())
print(dataset1.duplicated())
Name Type Size Value
dataset1 DataFrame (4410, 24) Column names: Age, Attrition, BusinessTravel, Department, DistanceFrom ...
Variable explorer Help Plots Files
Console 1/A
4408 False False ... False False
4409 False False ... False False
[4410 rows x 24 columns]
0 False
1 False
2 False
3 False
4 False
...
4405 False
4406 False
4407 False
4408 False
4409 False
Length: 4410, dtype: bool
In [15]:
Python console History
Kite: ready (no index)
conda: base (Python 3.7.6)
Line 17, Col 30
UTF-8
CRLF
RW
Mem 85%
Type here to search
ENG
21:01
19-07-2020


Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
G:\letsupgrade\notes
untitled0.py*
-*- coding: utf-8 -*-
"""
Created on Sun Jul 19 20:50:25 2020
@author: naika
"""
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
dataset1=pd.read_csv("general_data.csv")
dataset1.head()
print(dataset1.head())
print(dataset1.columns)
print(dataset1.isnull())
print(dataset1.duplicated())
print(dataset1.drop_duplicates())
dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
print(dataset3)
Name Type Size Value
dataset1 DataFrame (4410, 24) Column names: Age, Attrition, BusinessTravel, Department, DistanceFrom ...
dataset3 Series (11,) Series object of pandas.core.series module
Variable explorer Help Plots Files
Console 1/A
4407 25 No ... 1 2
4408 42 No ... 7 8
4409 40 No ... 3 9
[4410 rows x 24 columns]
Age ... YearsWithCurrManager
count 4410.000000 ... 4410.000000
mean 36.923810 ... 4.123129
std 9.133301 ... 3.567327
min 18.000000 ... 0.000000
25% 30.000000 ... 2.000000
50% 36.000000 ... 3.000000
75% 43.000000 ... 7.000000
max 60.000000 ... 17.000000
[8 rows x 11 columns]
Age 36.0
DistanceFromHome 7.0
Education 3.0
MonthlyIncome 49190.0
NumCompaniesWorked 2.0
Python console History
Kite: ready (no index)
conda: base (Python 3.7.6)
Line 22, Col 17
UTF-8
CRLF
RW
Mem 86%
Type here to search
ENG
21:05
19-07-2020

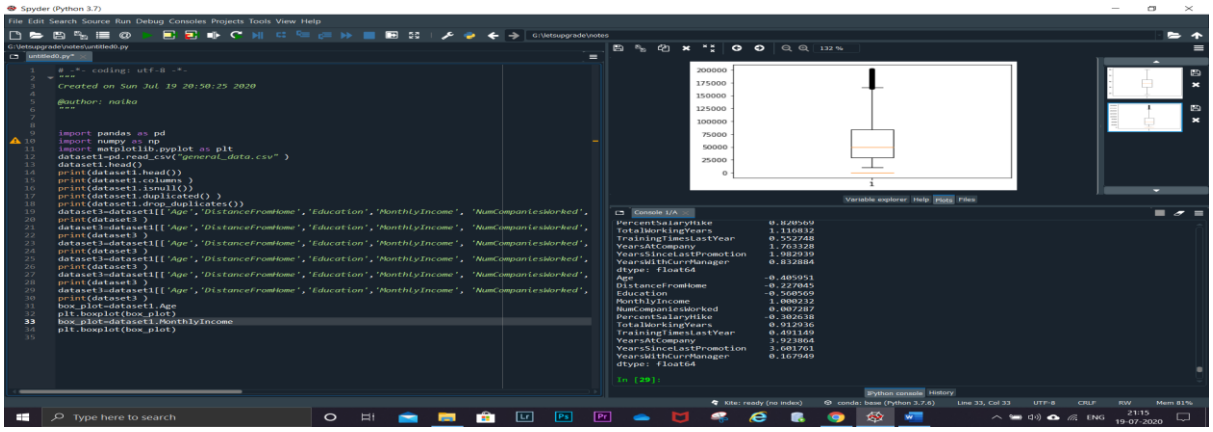

Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
G:\letsupgrade\notes
untitled0.py*
-*- coding: utf-8 -*-
"""
Created on Sun Jul 19 20:50:25 2020
@author: naika
"""
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
dataset1=pd.read_csv("general_data.csv")
dataset1.head()
print(dataset1.head())
print(dataset1.columns)
print(dataset1.isnull())
print(dataset1.duplicated())
print(dataset1.drop_duplicates())
dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
print(dataset3)
Name Type Size Value
dataset1 DataFrame (4410, 24) Column names: Age, Attrition, BusinessTravel, Department, DistanceFrom ...
dataset3 DataFrame (8, 11) Column names: Age, DistanceFromHome, Education, MonthlyIncome, NumComp ...
Variable explorer Help Plots Files
Console 1/A
4405 42 No ... 0 2
4406 29 No ... 0 2
4407 25 No ... 1 2
4408 42 No ... 7 8
4409 40 No ... 3 9
[4410 rows x 24 columns]
Age ... YearsWithCurrManager
count 4410.000000 ... 4410.000000
mean 36.923810 ... 4.123129
std 9.133301 ... 3.567327
min 18.000000 ... 0.000000
25% 30.000000 ... 2.000000
50% 36.000000 ... 3.000000
75% 43.000000 ... 7.000000
max 60.000000 ... 17.000000
[8 rows x 11 columns]
In [17]:
Python console History
Kite: ready (no index)
conda: base (Python 3.7.6)
Line 20, Col 17
UTF-8
CRLF
RW
Mem 85%
Type here to search
ENG
21:03
19-07-2020


Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
G:\letsupgrade\notes
untitled0.py
-*- coding: utf-8 -*-
"""
Created on Sun Jul 19 20:50:25 2020
@author: naika
"""
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
dataset1=pd.read_csv("general_data.csv")
dataset1.head()
print(dataset1.head())
print(dataset1.columns)
print(dataset1.isnull())
print(dataset1.duplicated())
print(dataset1.drop_duplicates())
dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
print(dataset3)
dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
print(dataset3)
dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
print(dataset3)
Name Type Size Value
dataset1 DataFrame (4410, 24) Column names: Age, Attrition, BusinessTravel, Department, DistanceFrom ...
dataset3 DataFrame (1, 11) Column names: Age, DistanceFromHome, Education, MonthlyIncome, NumComp ...
Variable explorer Help Plots Files
Console 1/A
max 60.000000 ... 17.000000
[8 rows x 11 columns]
Age 36.0
DistanceFromHome 7.0
Education 3.0
MonthlyIncome 49190.0
NumCompaniesWorked 2.0
PercentSalaryHike 14.0
TotalWorkingYears 10.0
TrainingTimesLastYear 3.0
YearsAtCompany 5.0
YearsSinceLastPromotion 1.0
YearsWithCurrManager 3.0
dtype: float64
Age DistanceFromHome ... YearsSinceLastPromotion YearsWithCurrManager
0 35 2 ... 0 2
[1 rows x 11 columns]
In [19]:
Python console History
Kite: ready (no index)
conda: base (Python 3.7.6)
Line 24, Col 17
UTF-8
CRLF
RW
Mem 86%
Type here to search
ENG
21:05
19-07-2020

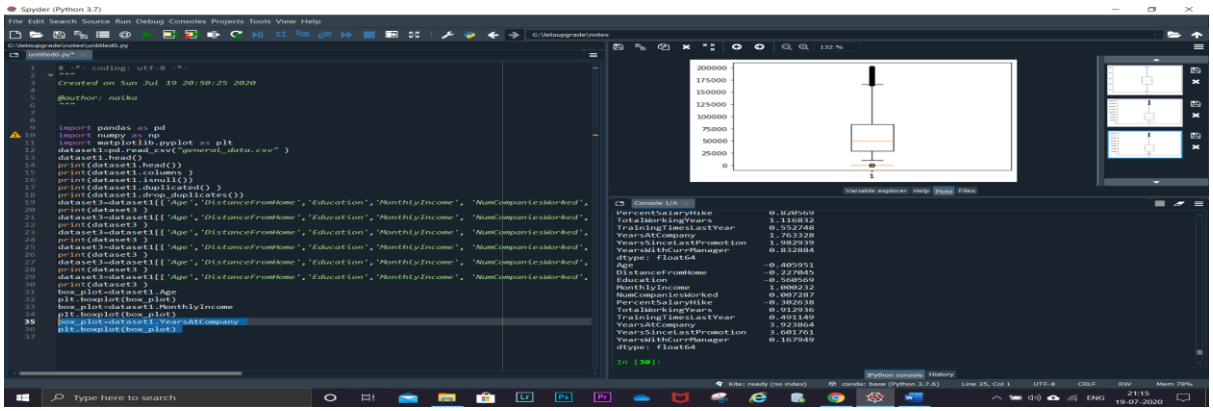

Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
G:\letsupgrade\notes
untitled0.py
-*- coding: utf-8 -*-
"""
Created on Sun Jul 19 20:50:25 2020
@author: naika
"""
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
dataset1=pd.read_csv("general_data.csv")
dataset1.head()
print(dataset1.head())
print(dataset1.columns)
print(dataset1.isnull())
print(dataset1.duplicated())
print(dataset1.drop_duplicates())
dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
print(dataset3)
dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
print(dataset3)
dataset3=dataset1[['Age','DistanceFromHome','Education','MonthlyIncome', 'NumCompaniesWorked',
print(dataset3)
Name Type Size Value
dataset1 DataFrame (4410, 24) Column names: Age, Attrition, BusinessTravel, Department, DistanceFrom ...
dataset3 Series (11,) Series object of pandas.core.series module
Variable explorer Help Plots Files
Console 1/A
YearsSinceLastPromotion 1.0
YearsWithCurrManager 3.0
dtype: float64
Age DistanceFromHome ... YearsSinceLastPromotion YearsWithCurrManager
0 35 2 ... 0 2
[1 rows x 11 columns]
Age 8.341719e+01
DistanceFromHome 6.569144e+01
Education 1.048418e+00
MonthlyIncome 2.215480e+09
NumCompaniesWorked 6.244436e+00
PercentSalaryHike 1.338907e+01
TotalWorkingYears 6.056298e+01
TrainingTimesLastYear 1.661465e+00
YearsAtCompany 3.751728e+01
YearsSinceLastPromotion 1.037935e+01
YearsWithCurrManager 1.272582e+01
dtype: float64
In [20]:
Python console History
Kite: ready (no index)
conda: base (Python 3.7.6)
Line 27, Col 1
UTF-8
CRLF
RW
Mem 83%
Type here to search
ENG
21:06
19-07-2020



Age is normally distributed without any outliers



Monthly Income is Right skewed with several outliers



Years at company is also Right Skewed with several outliers observed.