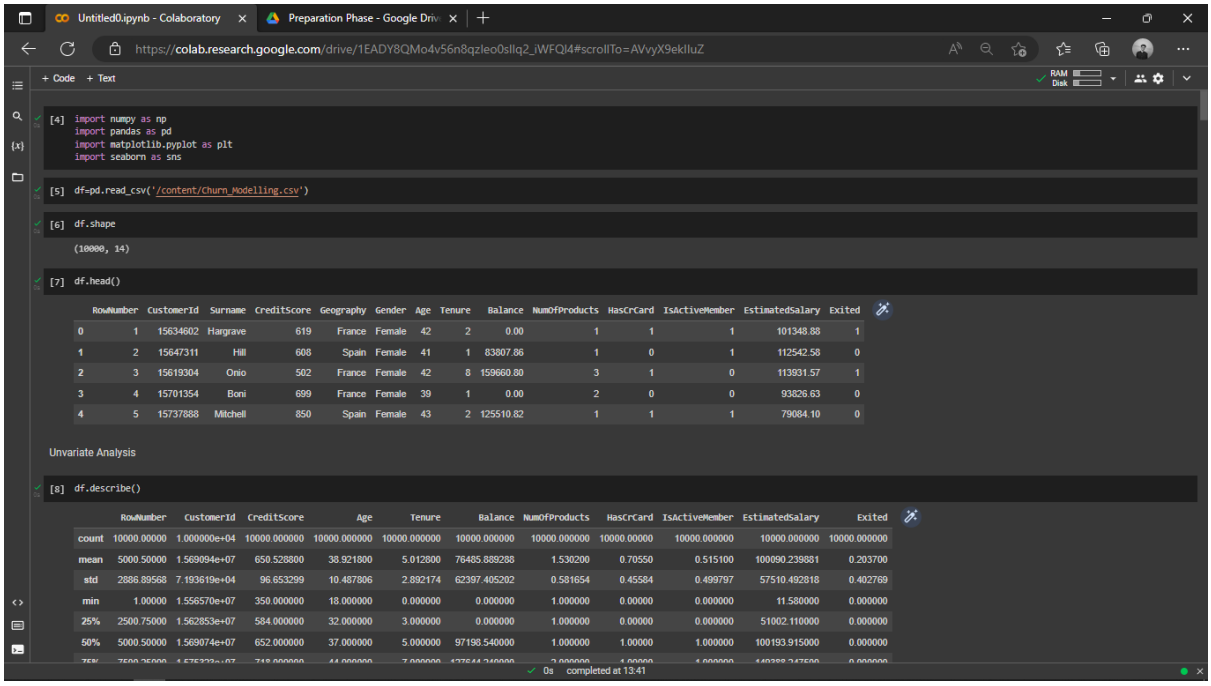
**Assignment -2**
**Python Programming**

| | |
|---|---|
| Assignment Date | 11 November 2022 |
| Student Name | Mr.S.Joshua |
| Student Roll Number | E1194020 |
| Maximum Marks | 2 Marks |



Untitled0.ipynb - Colaboratory | Preparation Phase - Google Driv

https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9eklluZ

+ Code + Text

```
[4] import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
```

```
[5] df=pd.read_csv('/content/Churn_Modelling.csv')
```

```
[6] df.shape
```

```
(10000, 14)
```

```
[7] df.head()
```

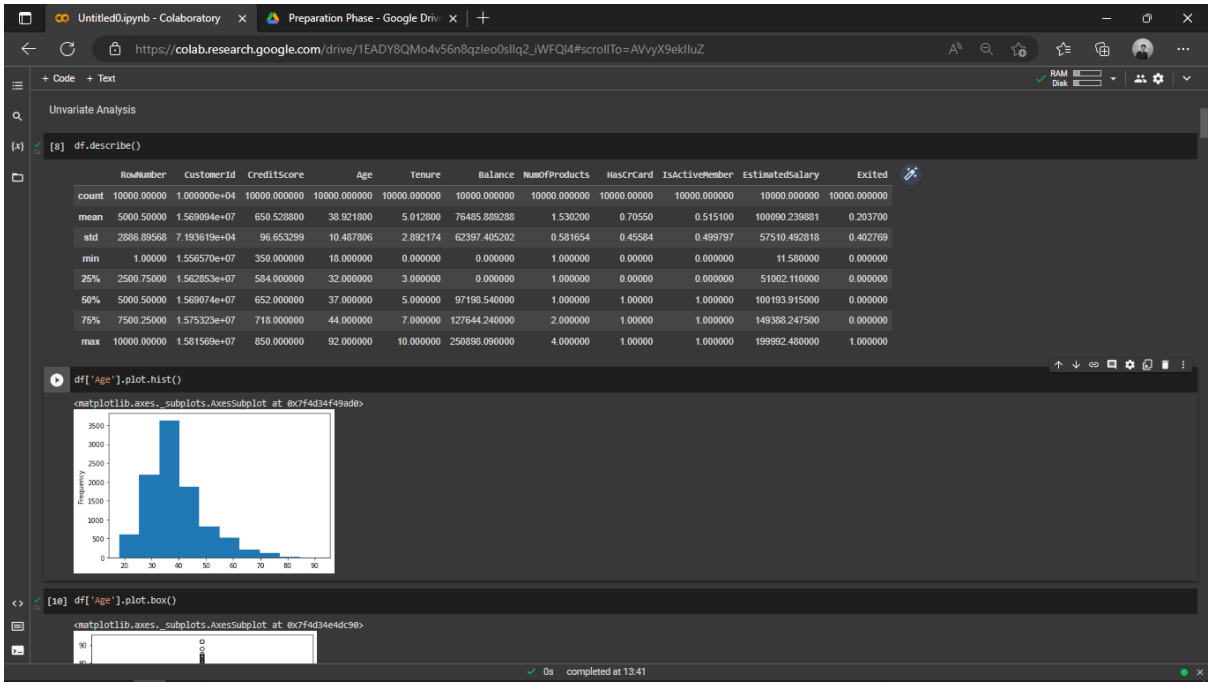
| | RowNumber | CustomerId | Surname | CreditScore | Geography | Gender | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 15634602 | Hargrave | 619 | France | Female | 42 | 2 | 0.00 | 1 | 1 | 1 | 101348.88 | 1 |
| 1 | 2 | 15647311 | Hill | 608 | Spain | Female | 41 | 1 | 83807.86 | 1 | 0 | 1 | 112542.58 | 0 |
| 2 | 3 | 15619304 | Onio | 502 | France | Female | 42 | 8 | 159660.80 | 3 | 1 | 0 | 113931.57 | 1 |
| 3 | 4 | 15701354 | Boni | 699 | France | Female | 39 | 1 | 0.00 | 2 | 0 | 0 | 93826.63 | 0 |
| 4 | 5 | 15737888 | Mitchell | 850 | Spain | Female | 43 | 2 | 125510.82 | 1 | 1 | 1 | 79084.10 | 0 |

Unvariate Analysis

```
[8] df.describe()
```

| | RowNumber | CustomerId | CreditScore | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited |
|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 10000.00000 | 1.000000e+04 | 10000.000000 | 10000.000000 | 10000.000000 | 10000.000000 | 10000.000000 | 10000.00000 | 10000.000000 | 10000.000000 | 10000.000000 |
| mean | 5000.50000 | 1.569094e+07 | 650.528800 | 38.921800 | 5.012800 | 76485.889288 | 1.530200 | 0.70550 | 0.515100 | 100090.239881 | 0.203700 |
| std | 2886.89568 | 7.193619e+04 | 96.653299 | 10.487806 | 2.892174 | 62397.405202 | 0.581654 | 0.45584 | 0.499797 | 57510.492818 | 0.402769 |
| min | 1.00000 | 1.556570e+07 | 350.000000 | 18.000000 | 0.000000 | 0.000000 | 1.000000 | 0.00000 | 0.000000 | 11.580000 | 0.000000 |
| 25% | 2500.75000 | 1.562853e+07 | 584.000000 | 32.000000 | 3.000000 | 0.000000 | 1.000000 | 0.00000 | 0.000000 | 51002.110000 | 0.000000 |
| 50% | 5000.50000 | 1.569074e+07 | 652.000000 | 37.000000 | 5.000000 | 97198.540000 | 1.000000 | 1.00000 | 1.000000 | 100193.915000 | 0.000000 |

0s completed at 13:41



Untitled0.ipynb - Colaboratory | Preparation Phase - Google Driv

https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9eklluZ

+ Code + Text

Unvariate Analysis

```
[8] df.describe()
```

| | RowNumber | CustomerId | CreditScore | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited |
|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 10000.00000 | 1.000000e+04 | 10000.000000 | 10000.000000 | 10000.000000 | 10000.000000 | 10000.000000 | 10000.00000 | 10000.000000 | 10000.000000 | 10000.000000 |
| mean | 5000.50000 | 1.569094e+07 | 650.528800 | 38.921800 | 5.012800 | 76485.889288 | 1.530200 | 0.70550 | 0.515100 | 100090.239881 | 0.203700 |
| std | 2886.89568 | 7.193619e+04 | 96.653299 | 10.487806 | 2.892174 | 62397.405202 | 0.581654 | 0.45584 | 0.499797 | 57510.492818 | 0.402769 |
| min | 1.00000 | 1.556570e+07 | 350.000000 | 18.000000 | 0.000000 | 0.000000 | 1.000000 | 0.00000 | 0.000000 | 11.580000 | 0.000000 |
| 25% | 2500.75000 | 1.562853e+07 | 584.000000 | 32.000000 | 3.000000 | 0.000000 | 1.000000 | 0.00000 | 0.000000 | 51002.110000 | 0.000000 |
| 50% | 5000.50000 | 1.569074e+07 | 652.000000 | 37.000000 | 5.000000 | 97198.540000 | 1.000000 | 1.00000 | 1.000000 | 100193.915000 | 0.000000 |
| 75% | 7500.25000 | 1.575323e+07 | 718.000000 | 44.000000 | 7.000000 | 127644.240000 | 2.000000 | 1.00000 | 1.000000 | 149388.247500 | 0.000000 |
| max | 10000.00000 | 1.581569e+07 | 850.000000 | 92.000000 | 10.000000 | 250898.090000 | 4.000000 | 1.00000 | 1.000000 | 199992.480000 | 1.000000 |

```
df['Age'].plot.hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4d34f49ad0>
```




```
[10] df['Age'].plot.box()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4d34e4dc90>
```



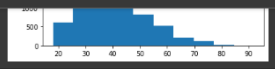

0s completed at 13:41

Untitled0.ipynb - Colaboratory | Preparation Phase - Google Driv

https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9ekIluZ

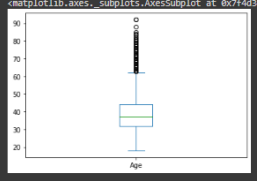
+ Code + Text

RAM Disk



```
[10] df['Age'].plot.box()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4d34e4dc90>
```

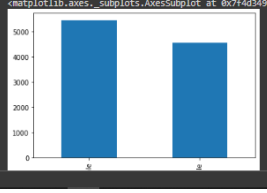



```
[11] df['Gender'].value_counts()
```

```
Male      5457
Female    4543
Name: Gender, dtype: int64
```

```
[12] df['Gender'].value_counts().plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4d34992550>
```



0s completed at 13:41

Untitled0.ipynb - Colaboratory | Preparation Phase - Google Driv

https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9ekIluZ

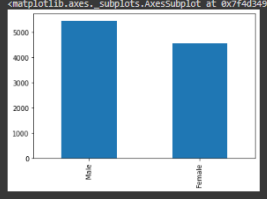
+ Code + Text

RAM Disk

```
Name: Gender, dtype: int64
```

```
[12] df['Gender'].value_counts().plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4d34992550>
```

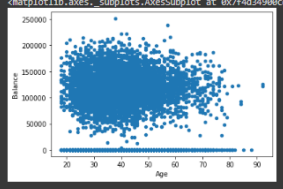



Bivarite Analysis

```
[13] import matplotlib.pyplot as plt
     %matplotlib inline
```

```
[14] df.plot.scatter('Age','Balance')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4d34900cd0>
```




0s completed at 13:41

Untitled0.ipynb - Colaboratory | Preparation Phase - Google Driv

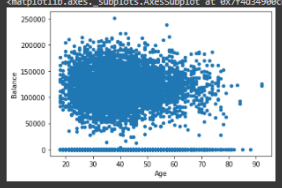
https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9eklluZ

+ Code + Text

RAM Disk

```
[14] df.plot.scatter('Age','Balance')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4d34900cd0>
```




```
[15] df.corr()
```

| | RowNumber | CustomerId | CreditScore | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited |
|---|---|---|---|---|---|---|---|---|---|---|---|
| RowNumber | 1.000000 | 0.004202 | 0.005840 | 0.000783 | -0.006495 | -0.009067 | 0.007246 | 0.000599 | 0.012044 | -0.005988 | -0.016571 |
| CustomerId | 0.004202 | 1.000000 | 0.005308 | 0.009497 | -0.014883 | -0.012419 | 0.016972 | -0.014025 | 0.001665 | 0.015271 | -0.006248 |
| CreditScore | 0.005840 | 0.005308 | 1.000000 | -0.003965 | 0.000842 | 0.006268 | 0.012238 | -0.005458 | 0.025651 | -0.001384 | -0.027094 |
| Age | 0.000783 | 0.009497 | -0.003965 | 1.000000 | -0.009997 | 0.028308 | -0.030680 | -0.011721 | 0.085472 | -0.007201 | 0.285323 |
| Tenure | -0.006495 | -0.014883 | 0.000842 | -0.009997 | 1.000000 | -0.012254 | 0.013444 | 0.022583 | -0.028362 | 0.007784 | -0.014001 |
| Balance | -0.009067 | -0.012419 | 0.006268 | 0.028308 | -0.012254 | 1.000000 | -0.304180 | -0.014858 | -0.010084 | 0.012797 | 0.118533 |
| NumOfProducts | 0.007246 | 0.016972 | 0.012238 | -0.030680 | 0.013444 | -0.304180 | 1.000000 | 0.003183 | 0.009612 | 0.014204 | -0.047820 |
| HasCrCard | 0.000599 | -0.014025 | -0.005458 | -0.011721 | 0.022583 | -0.014858 | 0.003183 | 1.000000 | -0.011866 | -0.009933 | -0.007138 |
| IsActiveMember | 0.012044 | 0.001665 | 0.025651 | 0.085472 | -0.028362 | -0.010084 | 0.009612 | -0.011866 | 1.000000 | -0.011421 | -0.156128 |
| EstimatedSalary | -0.005988 | 0.015271 | -0.001384 | -0.007201 | 0.007784 | 0.012797 | 0.014204 | -0.009933 | -0.011421 | 1.000000 | 0.012097 |
| Exited | -0.016571 | -0.006248 | -0.027094 | 0.285323 | -0.014001 | 0.118533 | -0.047820 | -0.007138 | -0.156128 | 0.012097 | 1.000000 |

```
[16] df['Age'].corr(df['Balance'])
```

```
0.02830836832749169
```

0s completed at 13:41

Untitled0.ipynb - Colaboratory | Preparation Phase - Google Driv

https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9eklluZ

+ Code + Text

RAM Disk

[16] df['Age'].corr(df['Balance'])

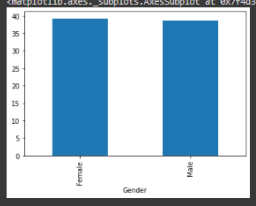
0.02830836832749169

```
[17] df.groupby('Gender')['Age'].mean()
```

```
Gender
Female    39.238389
Male      38.658237
Name: Age, dtype: float64
```

```
[18] df.groupby('Gender')['Age'].mean().plot.bar()
```

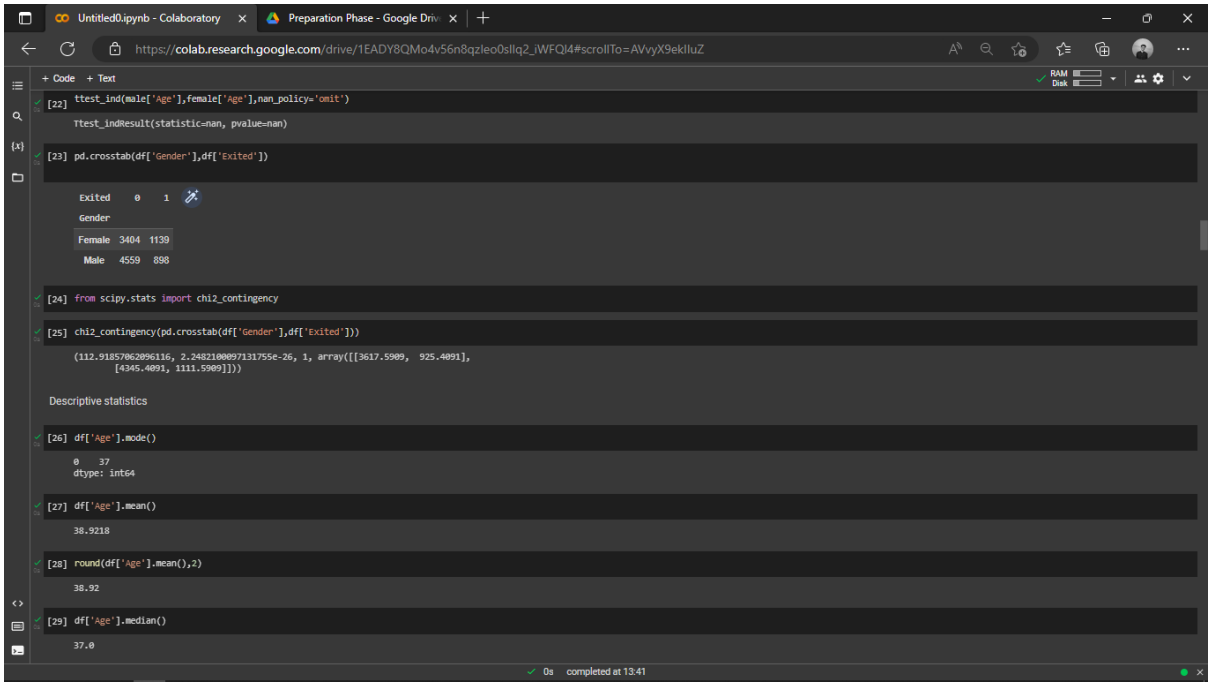
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4d348f0910>
```

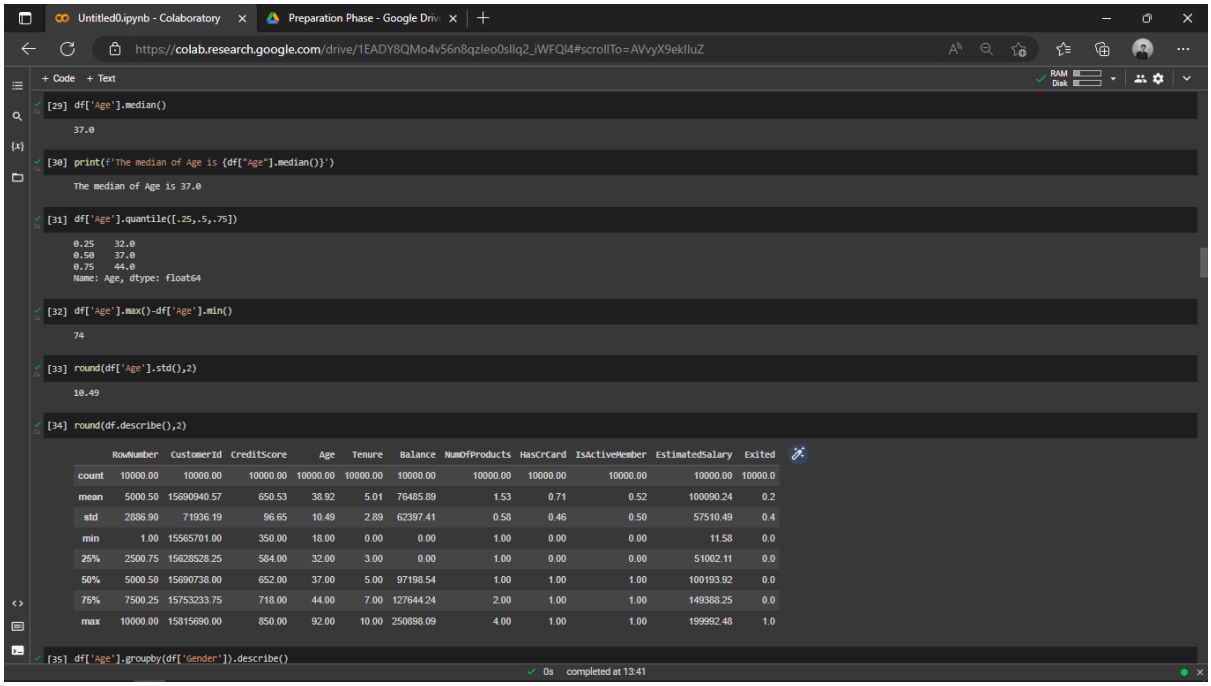



```
[19] from scipy.stats import ttest_ind

[20] male=df[df['Gender']=='male']

[21] female=df[df['Gender']=='female']

[22] ttest_ind(male['Age'],female['Age'],nan_policy='omit')
```

```
Ttest_indResult(statistic=nan, pvalue=nan)
```

0s completed at 13:41


Untitled0.ipynb - Colaboratory
Preparation Phase - Google Driv
https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9ekIluZ
+ Code + Text
RAM
Disk
[22] ttest_ind(male['Age'],female['Age'],nan_policy='omit')
Ttest_indResult(statistic=nan, pvalue=nan)
[23] pd.crosstab(df['Gender'],df['Exited'])
Exited 0 1
Gender
Female 3404 1139
Male 4559 898
[24] from scipy.stats import chi2_contingency
[25] chi2_contingency(pd.crosstab(df['Gender'],df['Exited']))
(112.91857062096116, 2.2482100971317555e-26, 1, array([[3617.5909, 925.4091],
[4345.4091, 1111.5909]]))
Descriptive statistics
[26] df['Age'].mode()
0 37
dtype: int64
[27] df['Age'].mean()
38.9218
[28] round(df['Age'].mean(),2)
38.92
[29] df['Age'].median()
37.0
0s completed at 13:41


Untitled0.ipynb - Colaboratory
Preparation Phase - Google Driv
https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9ekIluZ
+ Code + Text
RAM
Disk
[29] df['Age'].median()
37.0
[30] print(f'The median of Age is {df["Age"].median()}')
The median of Age is 37.0
[31] df['Age'].quantile([.25,.5,.75])
0.25 32.0
0.50 37.0
0.75 44.0
Name: Age, dtype: float64
[32] df['Age'].max()-df['Age'].min()
74
[33] round(df['Age'].std(),2)
10.49
[34] round(df.describe(),2)
RowNumber CustomerId CreditScore Age Tenure Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
count 10000.00 10000.00 10000.00 10000.00 10000.00 10000.00 10000.00 10000.00 10000.00 10000.00 10000.0
mean 5000.50 15690940.57 650.53 38.92 5.01 76485.89 1.53 0.71 0.52 100090.24 0.2
std 2886.90 71936.19 96.65 10.49 2.89 62397.41 0.58 0.46 0.50 57510.49 0.4
min 1.00 15565701.00 350.00 18.00 0.00 0.00 1.00 0.00 0.00 11.58 0.0
25% 2500.75 15628528.25 584.00 32.00 3.00 0.00 1.00 0.00 0.00 51002.11 0.0
50% 5000.50 15690738.00 652.00 37.00 5.00 97198.54 1.00 1.00 1.00 100193.92 0.0
75% 7500.25 15753233.75 718.00 44.00 7.00 127644.24 2.00 1.00 1.00 149388.25 0.0
max 10000.00 15815690.00 850.00 92.00 10.00 250898.09 4.00 1.00 1.00 199992.48 1.0
[35] df['Age'].groupby(df['Gender']).describe()
0s completed at 13:41

Untitled0.ipynb - Colaboratory | Preparation Phase - Google Driv

https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9ekIluZ

+ Code + Text

RAM Disk

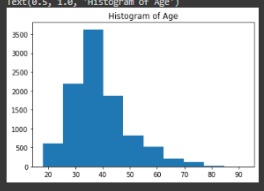
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| max | 10000.00 | 15815690.00 | 850.00 | 92.00 | 10.00 | 250898.09 | 4.00 | 1.00 | 1.00 | 199992.48 | 1.0 |

```
[35] df['Age'].groupby(df['Gender']).describe()
```

| Gender | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| Female | 4543.0 | 39.238389 | 10.588588 | 18.0 | 32.0 | 38.0 | 45.0 | 85.0 |
| Male | 5457.0 | 38.658237 | 10.396780 | 18.0 | 32.0 | 37.0 | 43.0 | 92.0 |

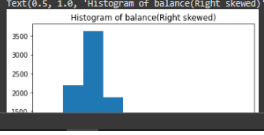
```
[36] plt.hist(df['Age'])
     plt.title('Histogram of Age')
```

```
Text(0.5, 1.0, 'Histogram of Age')
```




```
[37] plt.hist(df['Age'])
     plt.title('Histogram of balance(Right skewed)')
```

```
Text(0.5, 1.0, 'Histogram of balance(Right skewed)')
```




0s completed at 13:41

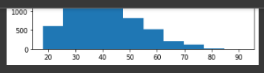
Untitled0.ipynb - Colaboratory | Preparation Phase - Google Driv

https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9ekIluZ

+ Code + Text

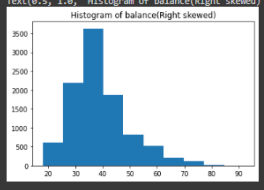
RAM Disk

[36]



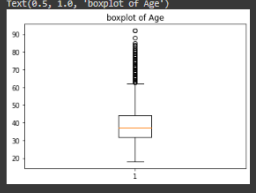
```
[37] plt.hist(df['Age'])
     plt.title('Histogram of balance(Right skewed)')
```

```
Text(0.5, 1.0, 'Histogram of balance(Right skewed)')
```




```
[38] plt.boxplot(df['Age'])
     plt.title('boxplot of Age')
```

```
Text(0.5, 1.0, 'boxplot of Age')
```




```
[39] import plotly.express as px
```

0s completed at 13:41

Untitled0.ipynb - Colaboratory | Preparation Phase - Google Driv

https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9eklluZ

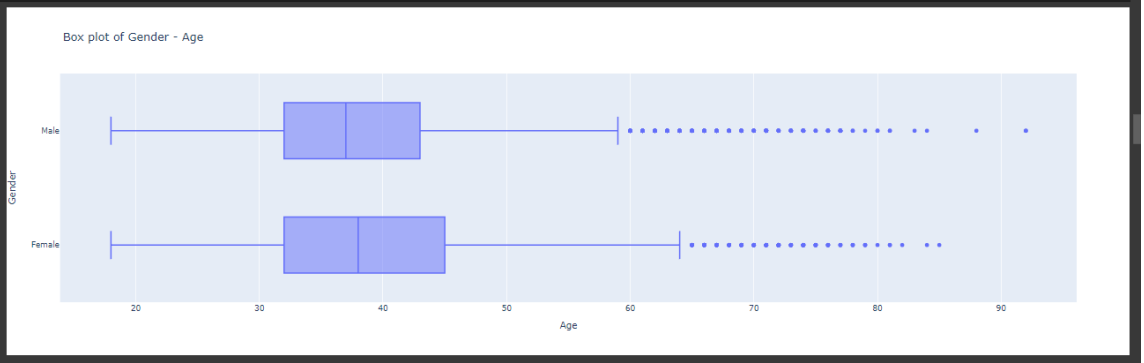
+ Code + Text

RAM Disk

[38]

```
[39] import plotly.express as px
     import pandas as pd
```

```
[40] px.box(df,x='Age',y='Gender',title="Box plot of Gender - Age")
```




Handel the missing Values

```
[41] df.isnull()
```

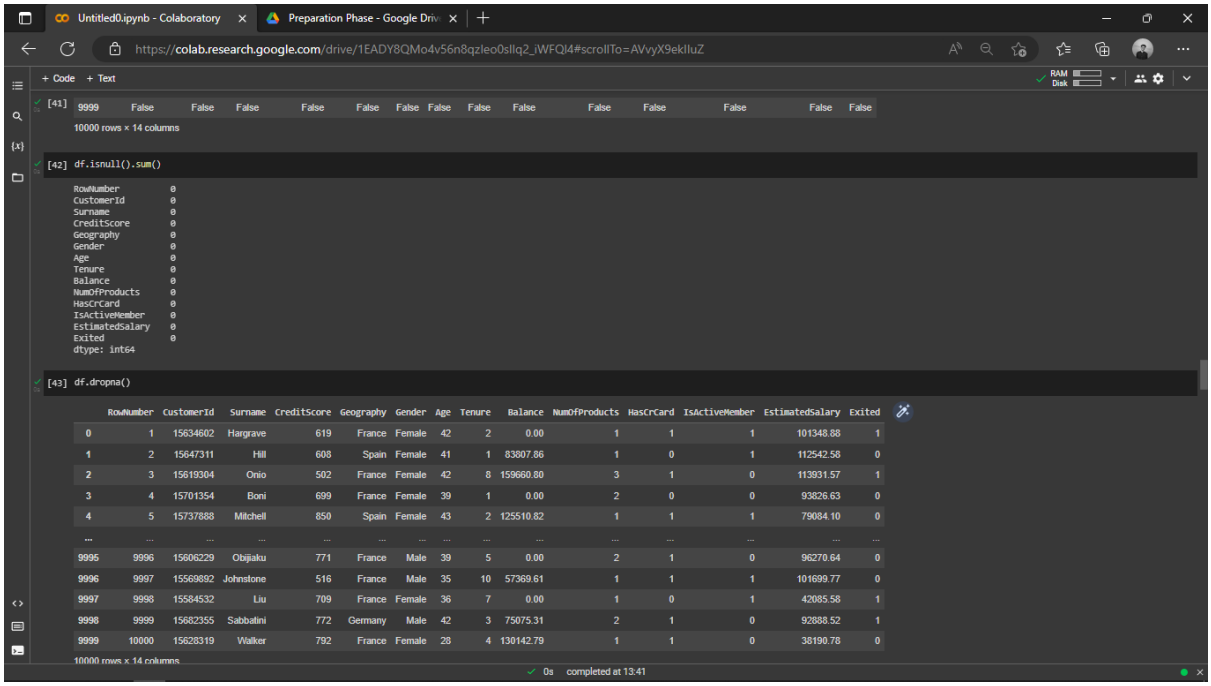
| | RowNumber | CustomerId | Surname | CreditScore | Geography | Gender | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 1 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 2 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 3 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 4 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 9995 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 9996 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 9997 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 9998 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |
| 9999 | False | False | False | False | False | False | False | False | False | False | False | False | False | False |

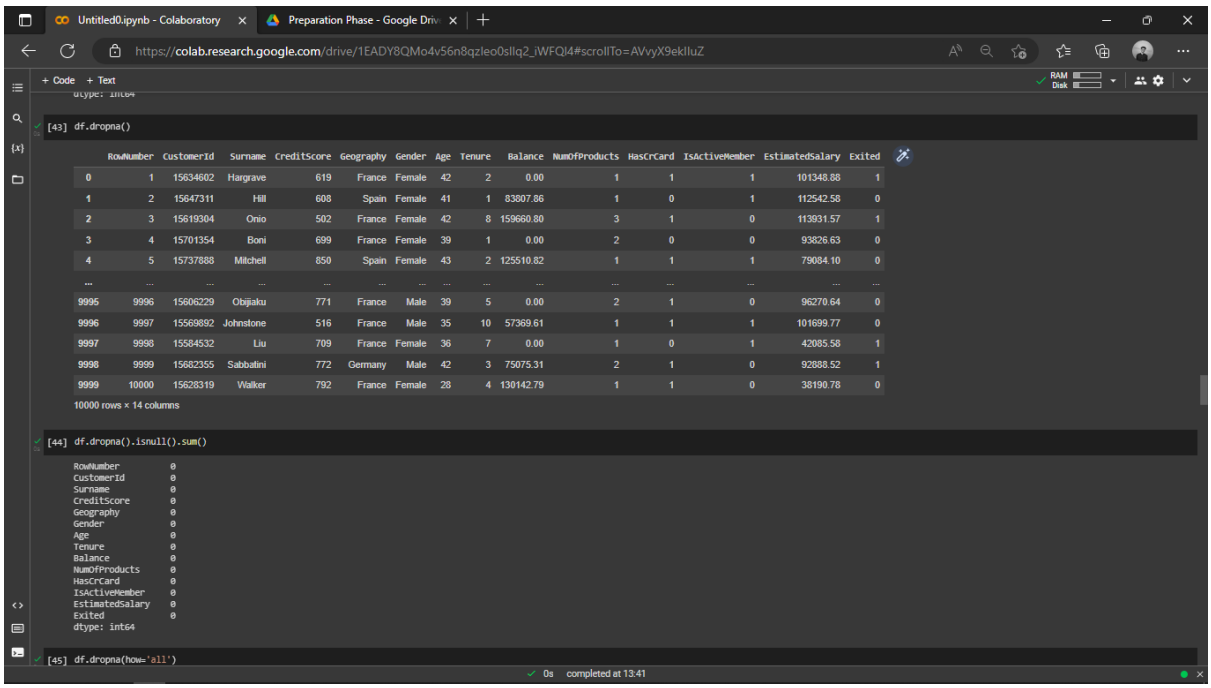
10000 rows × 14 columns

```
[42] df.isnull().sum()
```

```
RowNumber          0
CustomerId         0
Surname            0
CreditScore        0
Geography          0
Gender             0
Age                0
Tenure             0
Balance            0
NumOfProducts      0
HasCrCard          0
IsActiveMember     0
```

0s completed at 13:41


Untitled0.ipynb - Colaboratory
Preparation Phase - Google Driv
https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0sllq2_iWFQI4#scrollTo=AVvyX9eklluZ
+ Code + Text
RAM
Disk
[41] 9999 False False False False False False False False False False False False False
10000 rows × 14 columns
[42] df.isnull().sum()
RowNumber 0
CustomerId 0
Surname 0
CreditScore 0
Geography 0
Gender 0
Age 0
Tenure 0
Balance 0
NumOfProducts 0
HasCrCard 0
IsActiveMember 0
EstimatedSalary 0
Exited 0
dtype: int64
[43] df.dropna()
RowNumber CustomerId Surname CreditScore Geography Gender Age Tenure Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
0 1 15634602 Hargrave 619 France Female 42 2 0.00 1 1 1 101348.88 1
1 2 15647311 Hill 608 Spain Female 41 1 83807.86 1 0 1 112542.58 0
2 3 15619304 Onio 502 France Female 42 8 159660.80 3 1 0 113931.57 1
3 4 15701354 Boni 699 France Female 39 1 0.00 2 0 0 93826.63 0
4 5 15737888 Mitchell 850 Spain Female 43 2 125510.82 1 1 1 79084.10 0
... ... ... ... ... ... ... ... ... ... ... ... ... ...
9995 9996 15606229 Obijiaku 771 France Male 39 5 0.00 2 1 0 96270.64 0
9996 9997 15569892 Johnstone 516 France Male 35 10 57369.61 1 1 1 101699.77 0
9997 9998 15584532 Liu 709 France Female 36 7 0.00 1 0 1 42085.58 1
9998 9999 15682355 Sabbatini 772 Germany Male 42 3 75075.31 2 1 0 92888.52 1
9999 10000 15628319 Walker 792 France Female 28 4 130142.79 1 1 0 38190.78 0
10000 rows × 14 columns
0s completed at 13:41


Untitled0.ipynb - Colaboratory
Preparation Phase - Google Driv
https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0sllq2_iWFQI4#scrollTo=AVvyX9eklluZ
+ Code + Text
RAM
Disk
[43] df.dropna()
RowNumber CustomerId Surname CreditScore Geography Gender Age Tenure Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
0 1 15634602 Hargrave 619 France Female 42 2 0.00 1 1 1 101348.88 1
1 2 15647311 Hill 608 Spain Female 41 1 83807.86 1 0 1 112542.58 0
2 3 15619304 Onio 502 France Female 42 8 159660.80 3 1 0 113931.57 1
3 4 15701354 Boni 699 France Female 39 1 0.00 2 0 0 93826.63 0
4 5 15737888 Mitchell 850 Spain Female 43 2 125510.82 1 1 1 79084.10 0
... ... ... ... ... ... ... ... ... ... ... ... ... ...
9995 9996 15606229 Obijiaku 771 France Male 39 5 0.00 2 1 0 96270.64 0
9996 9997 15569892 Johnstone 516 France Male 35 10 57369.61 1 1 1 101699.77 0
9997 9998 15584532 Liu 709 France Female 36 7 0.00 1 0 1 42085.58 1
9998 9999 15682355 Sabbatini 772 Germany Male 42 3 75075.31 2 1 0 92888.52 1
9999 10000 15628319 Walker 792 France Female 28 4 130142.79 1 1 0 38190.78 0
10000 rows × 14 columns
[44] df.dropna().isnull().sum()
RowNumber 0
CustomerId 0
Surname 0
CreditScore 0
Geography 0
Gender 0
Age 0
Tenure 0
Balance 0
NumOfProducts 0
HasCrCard 0
IsActiveMember 0
EstimatedSalary 0
Exited 0
dtype: int64
[45] df.dropna(how='all')
0s completed at 13:41

Untitled0.ipynb - Colaboratory | Preparation Phase - Google Driv

https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9eklluZ

+ Code + Text

RAM Disk

```
Exited          0
dtype: int64
```

```
[45] df.dropna(how='all')
```

| | RowNumber | CustomerId | Surname | CreditScore | Geography | Gender | Age | Tenure | Balance | NumOfProducts | HasCrCard | IsActiveMember | EstimatedSalary | Exited |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 15634602 | Hargrave | 619 | France | Female | 42 | 2 | 0.00 | 1 | 1 | 1 | 101348.88 | 1 |
| 1 | 2 | 15647311 | Hill | 608 | Spain | Female | 41 | 1 | 83807.86 | 1 | 0 | 1 | 112542.58 | 0 |
| 2 | 3 | 15619304 | Onio | 502 | France | Female | 42 | 8 | 159660.80 | 3 | 1 | 0 | 113931.57 | 1 |
| 3 | 4 | 15701354 | Boni | 699 | France | Female | 39 | 1 | 0.00 | 2 | 0 | 0 | 93826.63 | 0 |
| 4 | 5 | 15737888 | Mitchell | 850 | Spain | Female | 43 | 2 | 125510.82 | 1 | 1 | 1 | 79084.10 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 9995 | 9996 | 15606229 | Obijiaku | 771 | France | Male | 39 | 5 | 0.00 | 2 | 1 | 0 | 96270.64 | 0 |
| 9996 | 9997 | 15569892 | Johnstone | 516 | France | Male | 35 | 10 | 57369.61 | 1 | 1 | 1 | 101699.77 | 0 |
| 9997 | 9998 | 15584532 | Liu | 709 | France | Female | 36 | 7 | 0.00 | 1 | 0 | 1 | 42085.58 | 1 |
| 9998 | 9999 | 15682355 | Sabbatini | 772 | Germany | Male | 42 | 3 | 75075.31 | 2 | 1 | 0 | 92888.52 | 1 |
| 9999 | 10000 | 15628319 | Walker | 792 | France | Female | 28 | 4 | 130142.79 | 1 | 1 | 0 | 38190.78 | 0 |

10000 rows × 14 columns

```
[46] df.dropna(how='all').shape
```

```
(10000, 14)
```

```
[47] df.dropna(axis=1).shape
```

```
(10000, 14)
```

```
[48] df.dropna(axis=1,how='all').shape
```

```
(10000, 14)
```

Outliers

0s completed at 13:41

Untitled0.ipynb - Colaboratory | Preparation Phase - Google Driv

https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9eklluZ

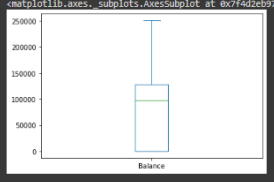
+ Code + Text

RAM Disk

Outliers

```
[49] df['Balance'].plot.box()
```

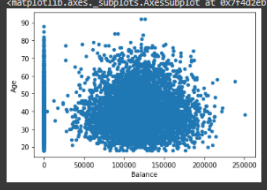
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4d2eb97750>
```




Bivatate Outliers Detection

```
[50] df.plot.scatter('Balance','Age')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4d2eb7fd0>
```




Removing Outliers from the Dataset

0s completed at 13:41

Untitled0.ipynb - Colaboratory | Preparation Phase - Google Driv

https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0sllq2_iWFQl4#scrollTo=AVvyX9eklluZ

+ Code + Text

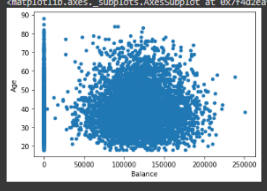
RAM Disk

Removing Outliers from the Dataset

```
[51] df=df[df['Age']<90]
```

```
[52] df.plot.scatter('Balance','Age')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4d2eaf8f50>
```

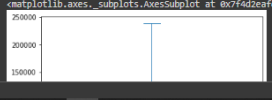



```
[53] df.loc[df['Balance']>250000,'Balance']=np.mean(df['Balance'])
```

```
/usr/local/lib/python3.7/dist-packages/pandas/core/indexing.py:1817: SettingWithCopyWarning:


A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
```

```
[54] df['Balance'].plot.box()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f4d2eafdb50>
```




0s completed at 13:41

Untitled0.ipynb - Colaboratory | Preparation Phase - Google Driv

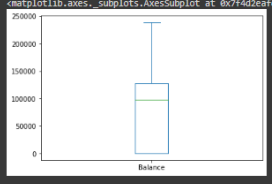
https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0sllq2_iWFQl4#scrollTo=AVvyX9eklluZ

+ Code + Text

RAM Disk

[54] df['Balance'].plot.box()

<matplotlib.axes._subplots.AxesSubplot at 0x7f4d2eafdb50>




```
[55] data_tips=pd.get_dummies('df')
     data_tips
```

| | df |
|---|---|
| 0 | 1 |

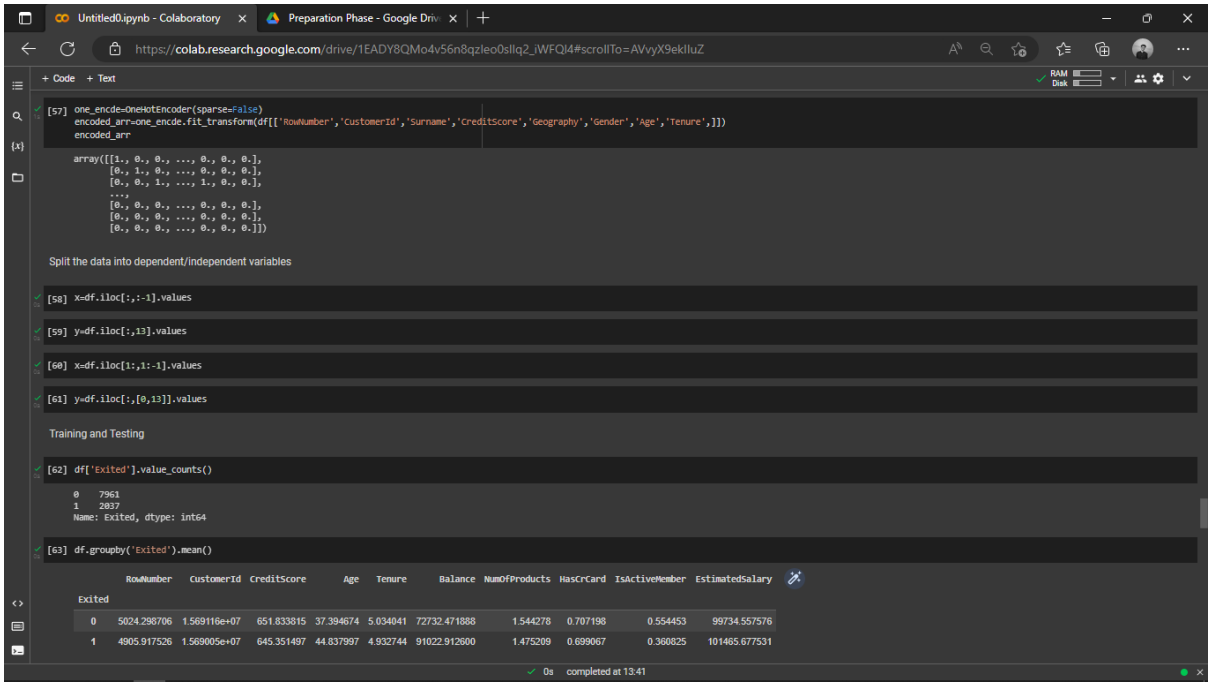
```
[56] from sklearn.preprocessing import OneHotEncoder
     df.keys()
     data_tips.keys()
```

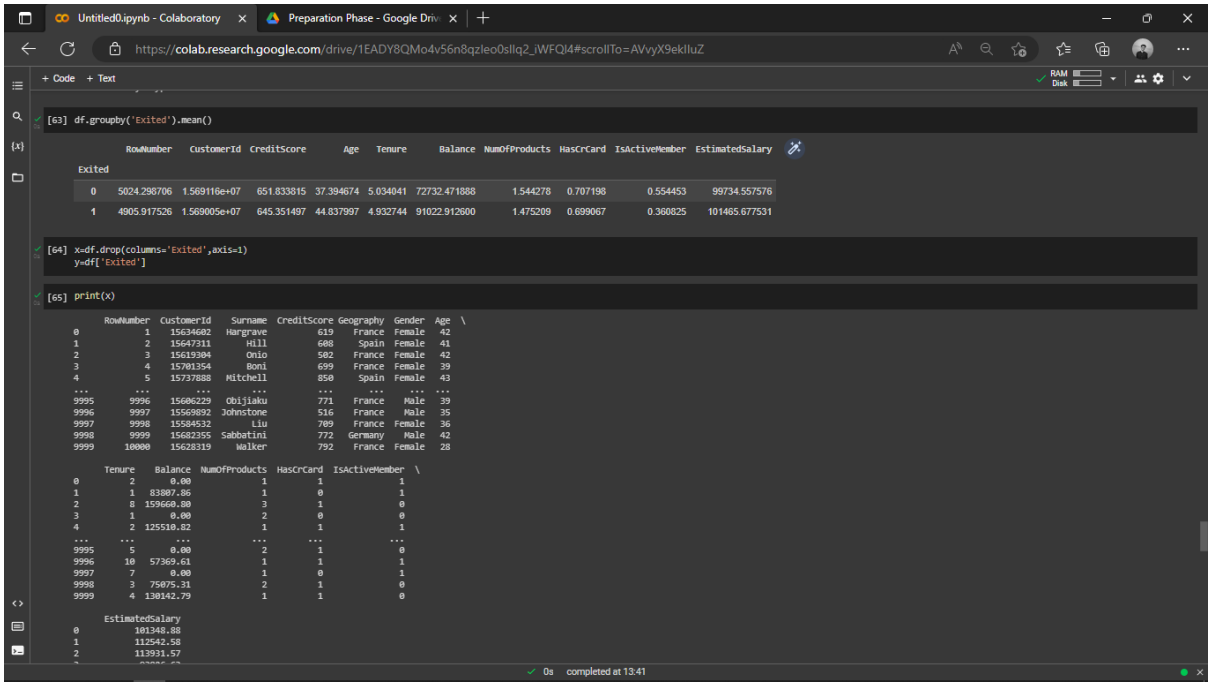
```
Index(['df'], dtype='object')
```

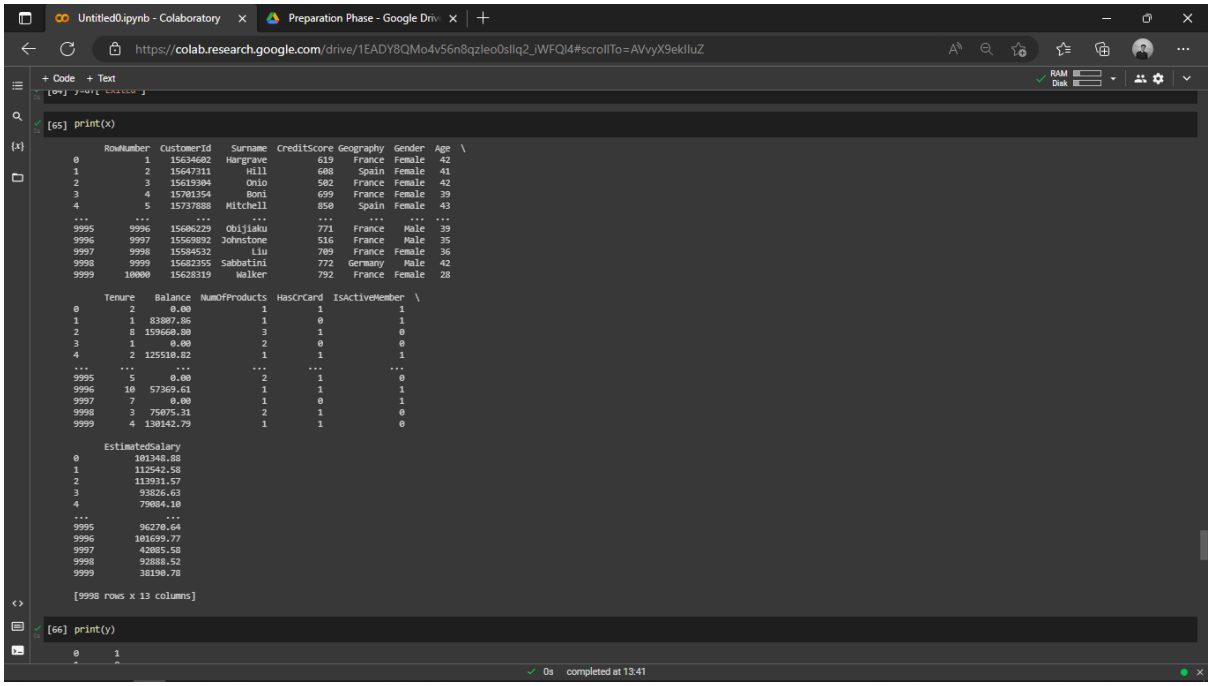
```
[57] one_encde=OneHotEncoder(sparse=False)
     encoded_arr=one_encde.fit_transform(df[['RowNumber','CustomerId','Surname','CreditScore','Geography','Gender','Age','Tenure',]])
     encoded_arr
```

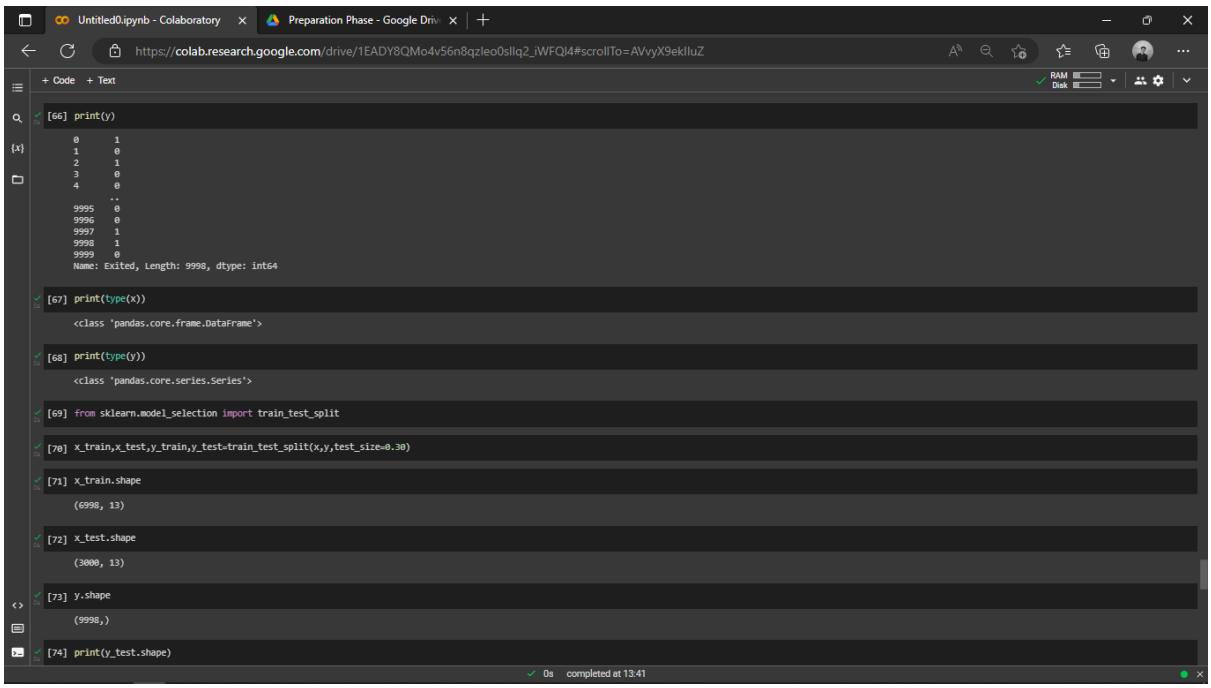
```
array([[1., 0., 0., ..., 0., 0.],
       [0., 1., 0., ..., 0., 0.],
       [0., 0., 1., ..., 1., 0.],
       ...,
       [0., 0., 0., ..., 0., 0.],
       [0., 0., 0., ..., 0., 0.],
       [0., 0., 0., ..., 0., 0.]])
```

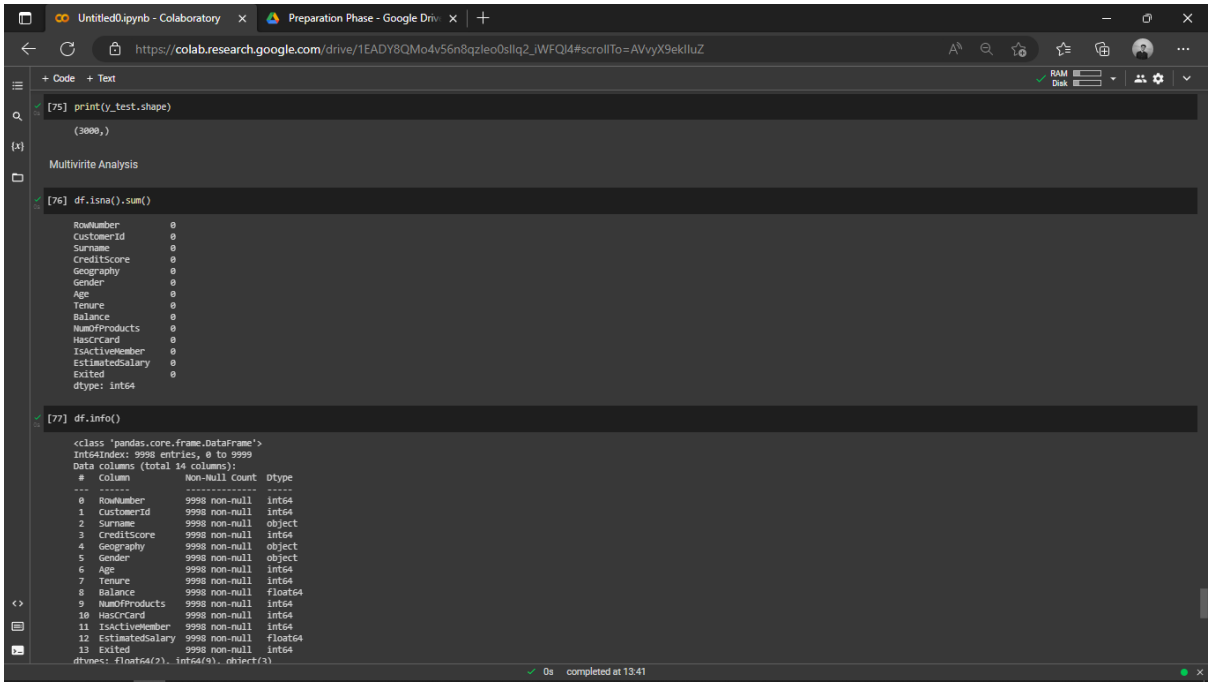
0s completed at 13:41

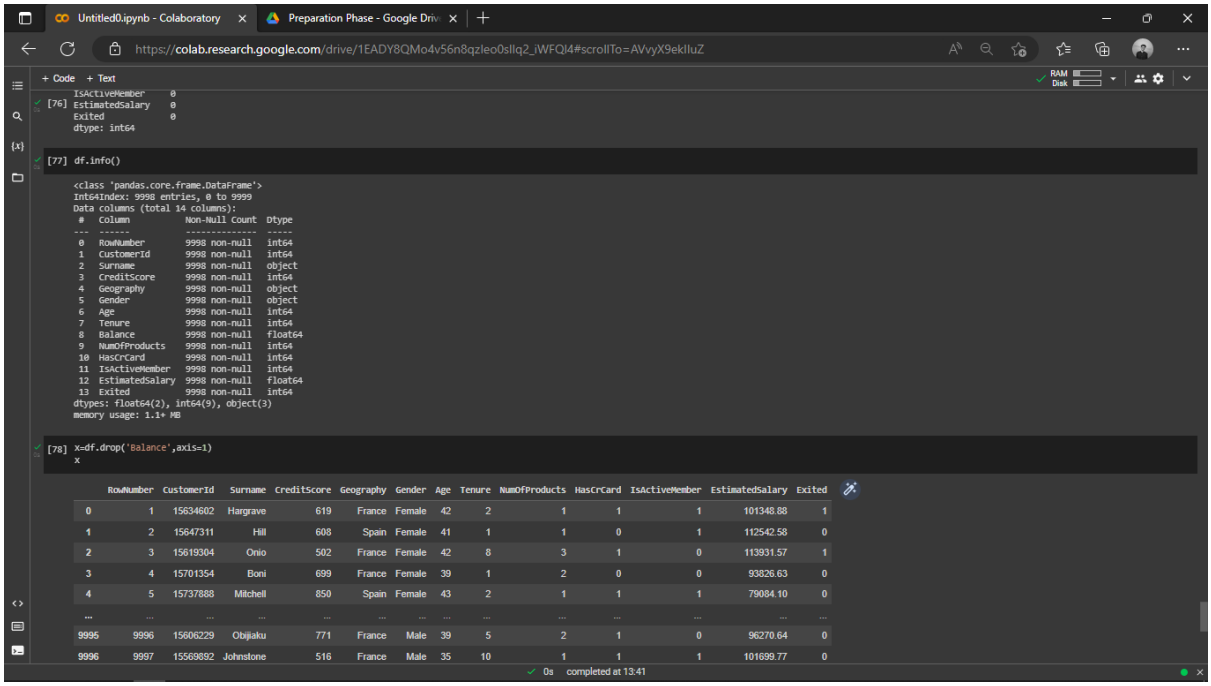

Untitled0.ipynb - Colaboratory
Preparation Phase - Google Driv
https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9eklluZ
+ Code + Text
RAM
Disk
[57] one_encde=OneHotEncoder(sparse=False)
encoded_arr=one_encde.fit_transform(df[['RowNumber','CustomerId','Surname','CreditScore','Geography','Gender','Age','Tenure',]])
encoded_arr
array([[1., 0., 0., ..., 0., 0.],
[0., 1., 0., ..., 0., 0.],
[0., 0., 1., ..., 1., 0.],
...,
[0., 0., 0., ..., 0., 0.],
[0., 0., 0., ..., 0., 0.],
[0., 0., 0., ..., 0., 0.]])
Split the data into dependent/independent variables
[58] X=df.iloc[:,:-1].values
[59] y=df.iloc[:,13].values
[60] X=df.iloc[1:,1:-1].values
[61] y=df.iloc[:,[0,13]].values
Training and Testing
[62] df['Exited'].value_counts()
0 7961
1 2037
Name: Exited, dtype: int64
[63] df.groupby('Exited').mean()
RowNumber CustomerId CreditScore Age Tenure Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary
Exited
0 5024.298706 1.569116e+07 651.833815 37.394674 5.034041 72732.471888 1.544278 0.707198 0.554453 99734.557576
1 4905.917526 1.569005e+07 645.351497 44.837997 4.932744 91022.912600 1.475209 0.699067 0.360825 101465.677531
0s completed at 13:41

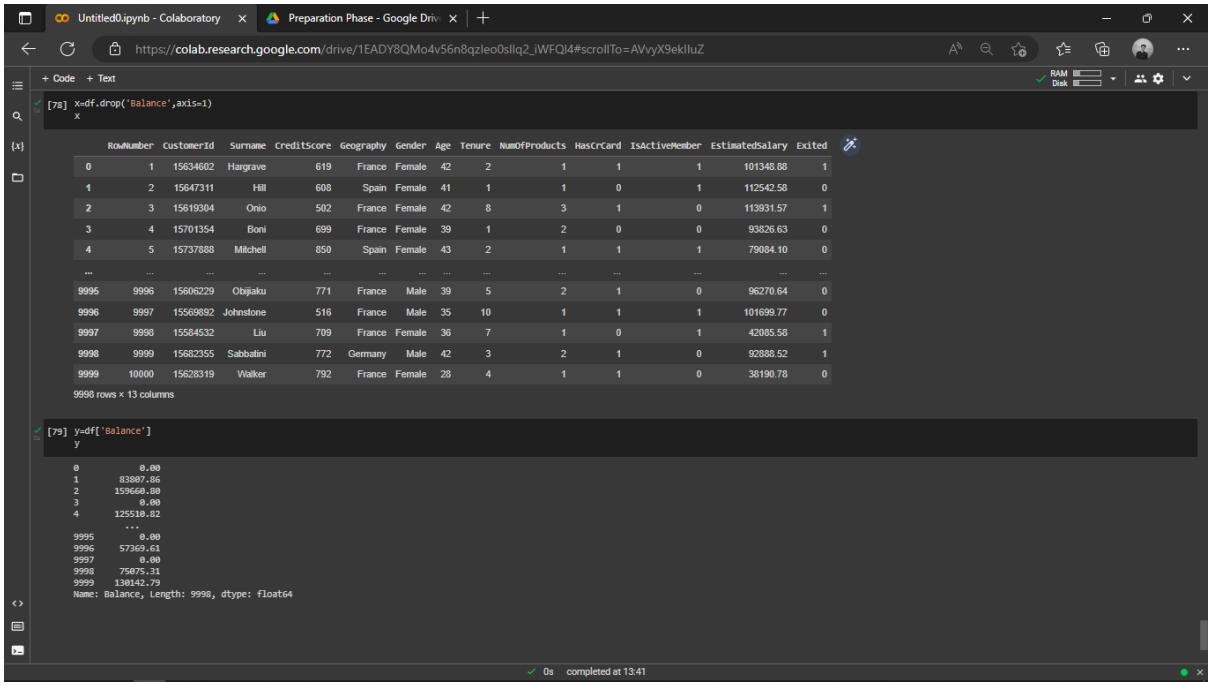

Untitled0.ipynb - Colaboratory
Preparation Phase - Google Driv
https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9eklluZ
+ Code + Text
RAM
Disk
[63] df.groupby('Exited').mean()
RowNumber CustomerId CreditScore Age Tenure Balance NumOfProducts HasCrCard IsActiveMember EstimatedSalary
Exited
0 5024.298706 1.569116e+07 651.833815 37.394674 5.034041 72732.471888 1.544278 0.707198 0.554453 99734.557576
1 4905.917526 1.569005e+07 645.351497 44.837997 4.932744 91022.912600 1.475209 0.699067 0.360825 101465.677531
[64] X=df.drop(columns='Exited',axis=1)
y=df['Exited']
[65] print(x)
RowNumber CustomerId Surname CreditScore Geography Gender Age \
0 1 15634602 Hargrave 619 France Female 42
1 2 15647311 Hill 608 Spain Female 41
2 3 15619304 Onio 502 France Female 42
3 4 15701354 Boni 699 France Female 39
4 5 15737888 Mitchell 850 Spain Female 43
... ... ... ... ... ... ... ...
9995 9996 15606229 Obijiaku 771 France Male 39
9996 9997 15569892 Johnstone 516 France Male 35
9997 9998 15584532 Liu 709 France Female 36
9998 9999 15682355 Sabbatini 772 Germany Male 42
9999 10000 15628319 Walker 792 France Female 28
Tenure Balance NumOfProducts HasCrCard IsActiveMember \
0 2 0.00 1 1 1
1 1 83807.86 1 0 1
2 8 159660.80 3 1 0
3 1 0.00 2 0 0
4 2 125510.82 1 1 1
... ... ... ... ... ...
9995 5 0.00 2 1 0
9996 10 57369.61 1 1 1
9997 7 0.00 1 0 1
9998 3 75075.31 2 1 0
9999 4 130142.79 1 1 0
EstimatedSalary
0 101348.88
1 112542.58
2 113931.57
0s completed at 13:41


Untitled0.ipynb - Colaboratory
Preparation Phase - Google Driv
https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9ekIluZ
+ Code + Text
RAM
Disk
[65] print(x)
RowNumber CustomerId Surname CreditScore Geography Gender Age \
0 1 15634602 Hargrave 619 France Female 42
1 2 15647311 Hill 608 Spain Female 41
2 3 15619304 Onio 502 France Female 42
3 4 15701354 Boni 699 France Female 39
4 5 15737888 Mitchell 850 Spain Female 43
... ... ... ... ... ... ... ...
9995 9996 15606229 Obijiaku 771 France Male 39
9996 9997 15569892 Johnstone 516 France Male 35
9997 9998 15584532 Liu 709 France Female 36
9998 9999 15682355 Sabbatini 772 Germany Male 42
9999 10000 15628319 Walker 792 France Female 28
Tenure Balance NumOfProducts HasCrCard IsActiveMember \
0 2 0.00 1 1 1
1 1 83807.86 1 0 1
2 8 159660.80 3 1 0
3 1 0.00 2 0 0
4 2 125510.82 1 1 1
... ... ... ... ... ...
9995 5 0.00 2 1 0
9996 10 57369.61 1 1 1
9997 7 0.00 1 0 1
9998 3 75075.31 2 1 0
9999 4 130142.79 1 1 0
EstimatedSalary
0 101348.88
1 112542.58
2 113931.57
3 93826.63
4 79084.10
... ...
9995 96270.64
9996 101699.77
9997 42085.58
9998 92888.52
9999 38190.78
[9998 rows x 13 columns]
[66] print(y)
0 1
0s completed at 13:41


Untitled0.ipynb - Colaboratory
Preparation Phase - Google Driv
https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9ekIluZ
+ Code + Text
RAM
Disk
[66] print(y)
0 1
1 0
2 1
3 0
4 0
..
9995 0
9996 0
9997 1
9998 1
9999 0
Name: Exited, Length: 9998, dtype: int64
[67] print(type(x))
<class 'pandas.core.frame.DataFrame'>
[68] print(type(y))
<class 'pandas.core.series.Series'>
[69] from sklearn.model_selection import train_test_split
[70] X_train,X_test,y_train,y_test=train_test_split(x,y,test_size=0.30)
[71] X_train.shape
(6998, 13)
[72] X_test.shape
(3000, 13)
[73] y.shape
(9998,)
[74] print(y_test.shape)
0s completed at 13:41


Untitled0.ipynb - Colaboratory
Preparation Phase - Google Driv
https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9eklluZ
+ Code + Text
RAM
Disk
[75] print(y_test.shape)
(3000,)
Multivirite Analysis
[76] df.isna().sum()
RowNumber 0
CustomerId 0
Surname 0
CreditScore 0
Geography 0
Gender 0
Age 0
Tenure 0
Balance 0
NumOfProducts 0
HasCrCard 0
IsActiveMember 0
EstimatedSalary 0
Exited 0
dtype: int64
[77] df.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9998 entries, 0 to 9999
Data columns (total 14 columns):
# Column Non-Null Count Dtype
0 RowNumber 9998 non-null int64
1 CustomerId 9998 non-null int64
2 Surname 9998 non-null object
3 CreditScore 9998 non-null int64
4 Geography 9998 non-null object
5 Gender 9998 non-null object
6 Age 9998 non-null int64
7 Tenure 9998 non-null int64
8 Balance 9998 non-null float64
9 NumOfProducts 9998 non-null int64
10 HasCrCard 9998 non-null int64
11 IsActiveMember 9998 non-null int64
12 EstimatedSalary 9998 non-null float64
13 Exited 9998 non-null int64
dtypes: float64(2), int64(9), object(3)
0s completed at 13:41


Untitled0.ipynb - Colaboratory
Preparation Phase - Google Driv
https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9eklluZ
+ Code + Text
RAM
Disk
IsActiveMember 0
[76] EstimatedSalary 0
Exited 0
dtype: int64
[77] df.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9998 entries, 0 to 9999
Data columns (total 14 columns):
# Column Non-Null Count Dtype
0 RowNumber 9998 non-null int64
1 CustomerId 9998 non-null int64
2 Surname 9998 non-null object
3 CreditScore 9998 non-null int64
4 Geography 9998 non-null object
5 Gender 9998 non-null object
6 Age 9998 non-null int64
7 Tenure 9998 non-null int64
8 Balance 9998 non-null float64
9 NumOfProducts 9998 non-null int64
10 HasCrCard 9998 non-null int64
11 IsActiveMember 9998 non-null int64
12 EstimatedSalary 9998 non-null float64
13 Exited 9998 non-null int64
dtypes: float64(2), int64(9), object(3)
memory usage: 1.1+ MB
[78] X=df.drop('Balance',axis=1)
X
RowNumber CustomerId Surname CreditScore Geography Gender Age Tenure NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
0 1 15634602 Hargrave 619 France Female 42 2 1 1 1 101348.88 1
1 2 15647311 Hill 608 Spain Female 41 1 1 0 1 112542.58 0
2 3 15619304 Onio 502 France Female 42 8 3 1 0 113931.57 1
3 4 15701354 Boni 699 France Female 39 1 2 0 0 93826.63 0
4 5 15737888 Mitchell 850 Spain Female 43 2 1 1 1 79084.10 0
... ... ... ... ... ... ... ... ... ... ... ... ...
9995 9996 15606229 Obijiaku 771 France Male 39 5 2 1 0 96270.64 0
9996 9997 15569892 Johnstone 516 France Male 35 10 1 1 1 101699.77 0
0s completed at 13:41


Untitled0.ipynb - Colaboratory
Preparation Phase - Google Driv
https://colab.research.google.com/drive/1EADY8QMo4v56n8qzleo0slIq2_iWFQI4#scrollTo=AVvyX9ekIluZ
+ Code + Text
RAM
Disk
[78] X=df.drop('Balance',axis=1)
x
RowNumber CustomerId Surname CreditScore Geography Gender Age Tenure NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
0 1 15634602 Hargrave 619 France Female 42 2 1 1 1 101348.88 1
1 2 15647311 Hill 608 Spain Female 41 1 1 0 1 112542.58 0
2 3 15619304 Onio 502 France Female 42 8 3 1 0 113931.57 1
3 4 15701354 Boni 699 France Female 39 1 2 0 0 93826.63 0
4 5 15737888 Mitchell 850 Spain Female 43 2 1 1 1 79084.10 0
... ... ... ... ... ... ... ... ... ... ... ... ...
9995 9996 15606229 Obijiaku 771 France Male 39 5 2 1 0 96270.64 0
9996 9997 15569892 Johnstone 516 France Male 35 10 1 1 1 101699.77 0
9997 9998 15584532 Liu 709 France Female 36 7 1 0 1 42085.58 1
9998 9999 15682355 Sabbatini 772 Germany Male 42 3 2 1 0 92888.52 1
9999 10000 15628319 Walker 792 France Female 28 4 1 1 0 38190.78 0
9998 rows × 13 columns
[79] y=df['Balance']
y
0 0.00
1 83807.86
2 159660.80
3 0.00
4 125510.82
...
9995 0.00
9996 57369.61
9997 0.00
9998 75075.31
9999 130142.79
Name: Balance, Length: 9998, dtype: float64
0s completed at 13:41