In [ ]:

```
#This analysis focuses on the behavior of bank customers who are more likely to leave t
he bank
#(i.e. close their bank account).
#I want to find out the most striking behaviors of customers through Exploratory Data A
nalysis and
#later on use some of the predictive analytics techniques to determine the customers wh
o are most likely to churn.
```

In [23]:

```
'''Descriptive Statistics is the building block of data science.
In simple terms, descriptive statistics can be defined as the measures that summarize a
given data,
and these measures can be broken down further into the measures of central tendency, me
asures of dispersion and Graphs.

Measures of central tendency include mean, median, and the mode, while the measures of
 variability include
standard deviation, variance, and the interquartile range. .

I will be explaining:

Mean
Median
Mode
Standard Deviation
Variance
Interquartile Range
Skewness'''
```

Out[23]:

```
'Descriptive Statistics is the building block of data science. \nIn simple
terms, descriptive statistics can be defined as the measures that summariz
e a given data,\nand these measures can be broken down further into the me
asures of central tendency, measures of dispersion and Graphs.\n\nMeasures
of central tendency include mean, median, and the mode, while the measures
of variability include\nstandard deviation, variance, and the interquartil
e range. .\n\nI will be explaining:\n\nMean\nMedian\nMode\nStandard Deviat
ion\nVariance\nInterquartile Range\nSkewness'
```

In [24]:

```
'''Data set:
CustomerId—contains random values and has no effect on customer leaving the bank.
CreditScore—can have an effect on customer churn, since a customer with a higher credit
score is less likely to leave the bank.
City—a customer's location can affect their decision to leave the bank.
Gender—it's interesting to explore whether gender plays a role in a customer leaving th
e bank.
Age—this is certainly relevant, since older customers are less likely to leave their ba
nk than younger ones.
BranchId - It is not relevant, all services of bank can be done from branch or online
Tenure—refers to the number of years that the customer has been a client of the bank. N
ormally, older clients are more loyal and less likely to leave a bank.
Balance—also a very good indicator of customer churn, as people with a higher balance i
n their accounts are less likely to leave the bank compared to those with lower balance
s.
NumOfProducts—refers to the number of products that a customer has purchased through th
e bank.
PrimaryAcHolder  - This is the person who is legally responsible for the debt and balan
ce along with the maintenance of the account.
HasOnlineService - Required for easy and 24/7 service
HasCrCard—denotes whether or not a customer has a credit card. This column is also rele
vant, since people with a credit card are less likely to leave the bank.
PrefContact - account holder contact details
IsActiveMember—active customers are less likely to leave the bank.
EstimatedSalary—as with balance, people with lower salaries are more likely to leave th
e bank compared to those with higher salaries.
Exited—whether or not the customer left the bank.'''
```

Out[24]:

```
'Data set:\nCustomerId—contains random values and has no effect on custome
r leaving the bank.\nCreditScore—can have an effect on customer churn, sin
ce a customer with a higher credit score is less likely to leave the ban
k.\nCity—a customer's location can affect their decision to leave the ban
k.\nGender—it's interesting to explore whether gender plays a role in a cu
stomer leaving the bank.\nAge—this is certainly relevant, since older cust
omers are less likely to leave their bank than younger ones.\nBranchId - I
t is not relevant, all services of bank can be done from branch or online
\nTenure—refers to the number of years that the customer has been a client
of the bank. Normally, older clients are more loyal and less likely to lea
ve a bank.\nBalance—also a very good indicator of customer churn, as peopl
e with a higher balance in their accounts are less likely to leave the ban
k compared to those with lower balances.\nNumOfProducts—refers to the numb
er of products that a customer has purchased through the bank.\nPrimaryAcH
older  - This is the person who is legally responsible for the debt and ba
lance along with the maintenance of the account. \nHasOnlineService - Requ
ired for easy and 24/7 service    \nHasCrCard—denotes whether or not a cus
tomer has a credit card. This column is also relevant, since people with a
credit card are less likely to leave the bank.\nPrefContact - account hold
er contact details         \nIsActiveMember—active customers are less like
ly to leave the bank.\nEstimatedSalary—as with balance, people with lower
salaries are more likely to leave the bank compared to those with higher s
alaries.\nExited—whether or not the customer left the bank.'
```

In [69]:

```
# Importing Libraries and Chanding directory

import os
import pandas as pd
import numpy as np
import statistics as st
import seaborn as sns
print(os.getcwd())
os.chdir("C:\\Leina\\Data_sets\\TD_assignment")
```

```
C:\Leina\Data_sets\TD_assignment
```

In [55]:

```
# Load the Data

df = pd.read_csv("hackathon_train_main.csv")
print(df.shape)
print(df.info())
```

```
(14646, 18)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14646 entries, 0 to 14645
Data columns (total 18 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   CustomerId       14646 non-null  int64  
 1   CreditScore      14646 non-null  float64
 2   City             14646 non-null  object 
 3   Gender           14646 non-null  object 
 4   Age              14646 non-null  int64  
 5   BranchId         14646 non-null  int64  
 6   Tenure           14646 non-null  int64  
 7   Balance          14646 non-null  float64
 8   CurrencyCode     14646 non-null  object 
 9   PrefLanguage     14646 non-null  object 
 10  NumOfProducts    14646 non-null  int64  
 11  PrimaryAcHolder  14646 non-null  int64  
 12  HasOnlineService 14646 non-null  int64  
 13  HasCrCard        14646 non-null  int64  
 14  PrefContact      14646 non-null  object 
 15  IsActiveMember   14646 non-null  int64  
 16  EstimatedSalary  14646 non-null  float64
 17  Exited           14646 non-null  int64  
dtypes: float64(3), int64(10), object(5)
memory usage: 2.0+ MB
None
```

In [27]:

```
#Five of the variables are categorical (labelled as 'object') while the remaining are n
umerical (labelled as 'int' or 'Float').
```

In [56]:

```
#Dropping some irrelavant features for Desriptive Analysis


df.drop(["CustomerId","City","BranchId","PrefLanguage","PrefContact"], axis = 'columns'
, inplace = True)
```

In [57]:

```
#Measures of Central Tendency
#Measures of central tendency describe the center of the data, and are represented by t
he mean, the median, and the mode.

df.mean()

round(df[["CreditScore", "Age", "Tenure", "Balance", "NumOfProducts", "PrimaryAcHolder"
, "HasCrCard"
         ,"IsActiveMember", "EstimatedSalary", "Exited"]].mean(),2)
```

Out[57]:

```
CreditScore           648.12
Age                    40.52
Tenure                  5.00
Balance             76542.07
NumOfProducts           1.57
PrimaryAcHolder         0.50
HasCrCard               0.70
IsActiveMember          0.55
EstimatedSalary    101700.60
Exited                  0.43
dtype: float64
```

In [31]:

```
#From the output, we can infer that the average age of the applicant is 40 years,
#the average balance 76542, average estimated salary is 101700 and the average tenure 5
years.
```

In [58]:

```
#Median
#Median represents the 50th percentile, or the middle value of the data, that separates
the distribution into two halves.
#The line of code below prints the median of the numerical variables in the data.
#The command df.median(axis = 0) will also give the same output.

round(df.median(),2)
```

Out[58]:

```
CreditScore            649.00
Age                     40.00
Tenure                   5.00
Balance              99681.60
NumOfProducts            2.00
PrimaryAcHolder          0.00
HasOnlineService         0.00
HasCrCard                1.00
IsActiveMember           1.00
EstimatedSalary     102443.05
Exited                   0.00
dtype: float64
```

In [ ]:

```
#From the output, we can infer that the median age of the applicants is 40 years,
#the median balance is 99681, esitimated salary is 102443 and the median tenure is 5 ye
ars.
#There is a difference between the mean and the median values of these variables, which
is because of the distribution of the data.
```

In [59]:

```
#Mode
#Mode represents the most frequent value of a variable in the data.
#This is the only central tendency measure that can be used with categorical variables,
#unlike the mean and the median which can be used only with quantitative data.

df.mode()
```

Out[59]:

	CreditScore	Gender	Age	Tenure	Balance	CurrencyCode	NumOfProducts	Primary.
0	850.0	Female	40.0	6.0	0.0	CAD	1.0	
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
...	...	...	...	...	...	...	...	
14641	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
14642	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
14643	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
14644	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
14645	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

14646 rows × 13 columns

In [60]:

```
df.loc[:,"Age"].mode()
```

Out[60]:

```
0    40
dtype: int64
```

In [ ]:

```
'''The interpretation of the mode is simple.
The output above shows that most of the applicants are female, as depicted by the 'Gend
er'.
Similar interpreation could be done for the other categorical variables like 'City' and
'PrefLanguage'.
For numerical variables, the mode value represents the value that occurs most frequentl
y.
For example, the mode value of 40 for the variable 'Age' means that the highest number
 (or frequency) of applicants are 40 years
old.'''
```

In [61]:

```
#Measures of Dispersion
#We have seen in the data, the values of central tendency measures differ for many vari
ables.
#This is because of the extent to which a distribution is stretched or squeezed.
#In statistics, this is measured by dispersion which is also referred to as variabilit
y, scatter, or spread.
#The most popular measures of dispersion are standard deviation, variance, and the inte
rquartile range.

#Standard Deviation: it is a measure that is used to quantify the amount of variation o
f a set of data values from its mean.
#A low standard deviation for a variable indicates that the data points tend to be clos
e to its mean, and vice versa.
#The line of code below prints the standard deviation of all the numerical variables in
the data.

df.std()
```

Out[61]:

```
CreditScore            86.326143
Age                     9.510517
Tenure                  2.586318
Balance             62165.262069
NumOfProducts           0.608287
PrimaryAcHolder         0.500004
HasOnlineService        0.499999
HasCrCard               0.456194
IsActiveMember          0.497655
EstimatedSalary     56721.249335
Exited                  0.494831
dtype: float64
```

In [36]:

```
#While interpreting standard deviation values, it is important to understand them in co
njunction with the mean.
#For example, in the above output, the standard deviation of the variable 'Balance' is
 much higher than that of the
#variable 'CreditScore'. However, the unit of these two variables is different and, the
refore,
#comparing the dispersion of these two variables on the basis of standard deviation alo
ne will be incorrect.
#This needs to be kept in mind.

print(df.loc[:,'Age'].std())
print(df.loc[:,'Balance'].std())

#calculate the standard deviation of the first five rows
df.std(axis = 1)[0:3]
```

```
9.5105171301812
62165.26206857463
```

Out[36]:

```
0    4.329047e+06
1    4.322886e+06
2    4.346901e+06
dtype: float64
```

In [62]:

```
#Variance
#Variance is another measure of dispersion.
#It is the square of the standard deviation and the covariance of the random variable w
ith itself.

df.var()
```

Out[62]:

```
CreditScore        7.452203e+03
Age                9.044994e+01
Tenure             6.689041e+00
Balance            3.864520e+09
NumOfProducts      3.700128e-01
PrimaryAcHolder    2.500045e-01
HasOnlineService   2.499986e-01
HasCrCard          2.081131e-01
IsActiveMember     2.476602e-01
EstimatedSalary    3.217300e+09
Exited             2.448574e-01
dtype: float64
```

In [63]:

```
#Interquartile Range (IQR)
#The Interquartile Range (IQR) is a measure of statistical dispersion,
#and is calculated as the difference between the upper quartile (75th percentile) and t
he lower quartile (25th percentile).
#The IQR is also a very important measure for identifying outliers and could be visuali
zed using a boxplot.

#IQR can be calculated using the iqr() function.

from scipy.stats import iqr
iqr(df['Age'])
```

Out[63]:

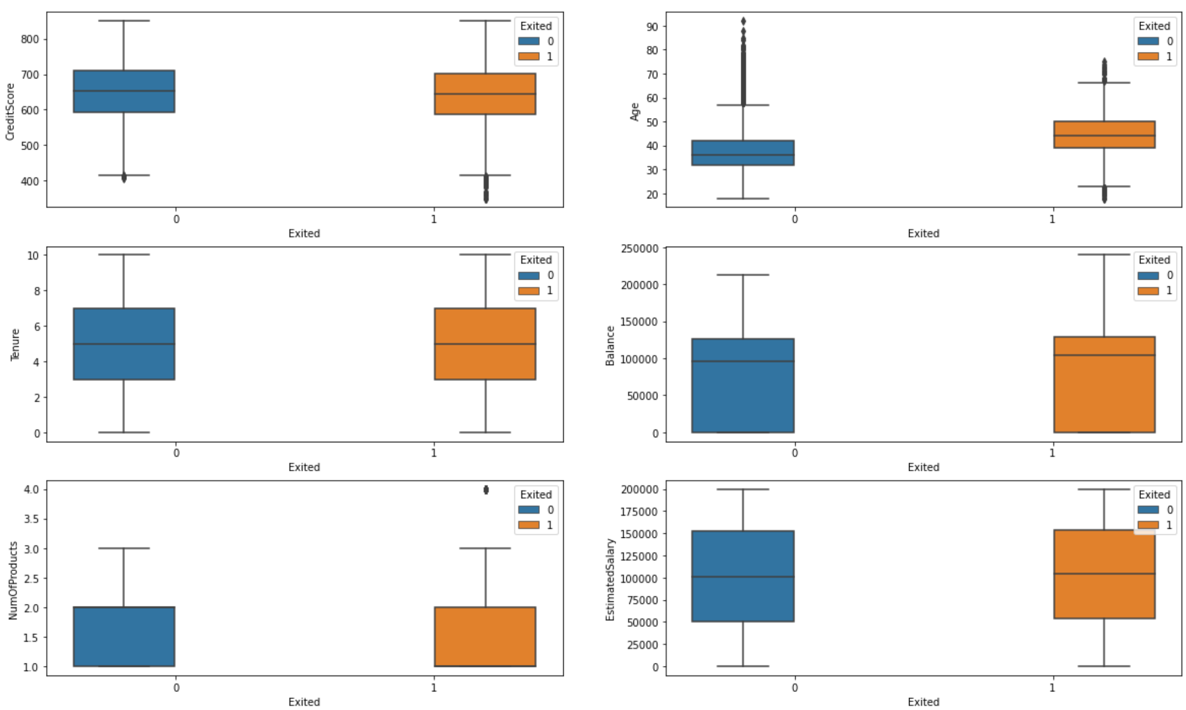
```
12.0
```

In [64]:

```
import seaborn as sns
import matplotlib.pyplot as plt
fig, axarr = plt.subplots(3, 2, figsize=(20, 12))
sns.boxplot(y='CreditScore',x = 'Exited', hue = 'Exited',data = df, ax=axarr[0][0])
sns.boxplot(y='Age',x = 'Exited', hue = 'Exited',data = df , ax=axarr[0][1])
sns.boxplot(y='Tenure',x = 'Exited', hue = 'Exited',data = df, ax=axarr[1][0])
sns.boxplot(y='Balance',x = 'Exited', hue = 'Exited',data = df, ax=axarr[1][1])
sns.boxplot(y='NumOfProducts',x = 'Exited', hue = 'Exited',data = df, ax=axarr[2][0])
sns.boxplot(y='EstimatedSalary',x = 'Exited', hue = 'Exited',data = df, ax=axarr[2][1])
```

Out[64]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x20d26d68850>
```



In [65]:

```
#Skewness
#It is the measure of the symmetry, or lack of it,
#The skewness value can be positive, negative, or undefined.
#In a perfectly symmetrical distribution, the mean, the median, and the mode will all h
ave the same value.
#However, the variables in our data are not symmetrical, resulting in different values
 of the central tendency.

print(df.skew())
```

```
CreditScore        -0.068362
Age                 0.626668
Tenure              0.016474
Balance            -0.170069
NumOfProducts       0.764637
PrimaryAcHolder     0.014204
HasOnlineService    0.017208
HasCrCard          -0.897537
IsActiveMember     -0.195124
EstimatedSalary    -0.026830
Exited              0.290355
dtype: float64
```

In [66]:

```
#Putting Everything Together
#We have learned the measures of central tendency and dispersion.
#It is important to analyse these individually, however, because there are certain usef
ul functions in python
#that can be called upon to find these values.
#One such important function is the .describe() function that prints the summary statis
tic of the numerical variables.


df.describe(include = "all")
```

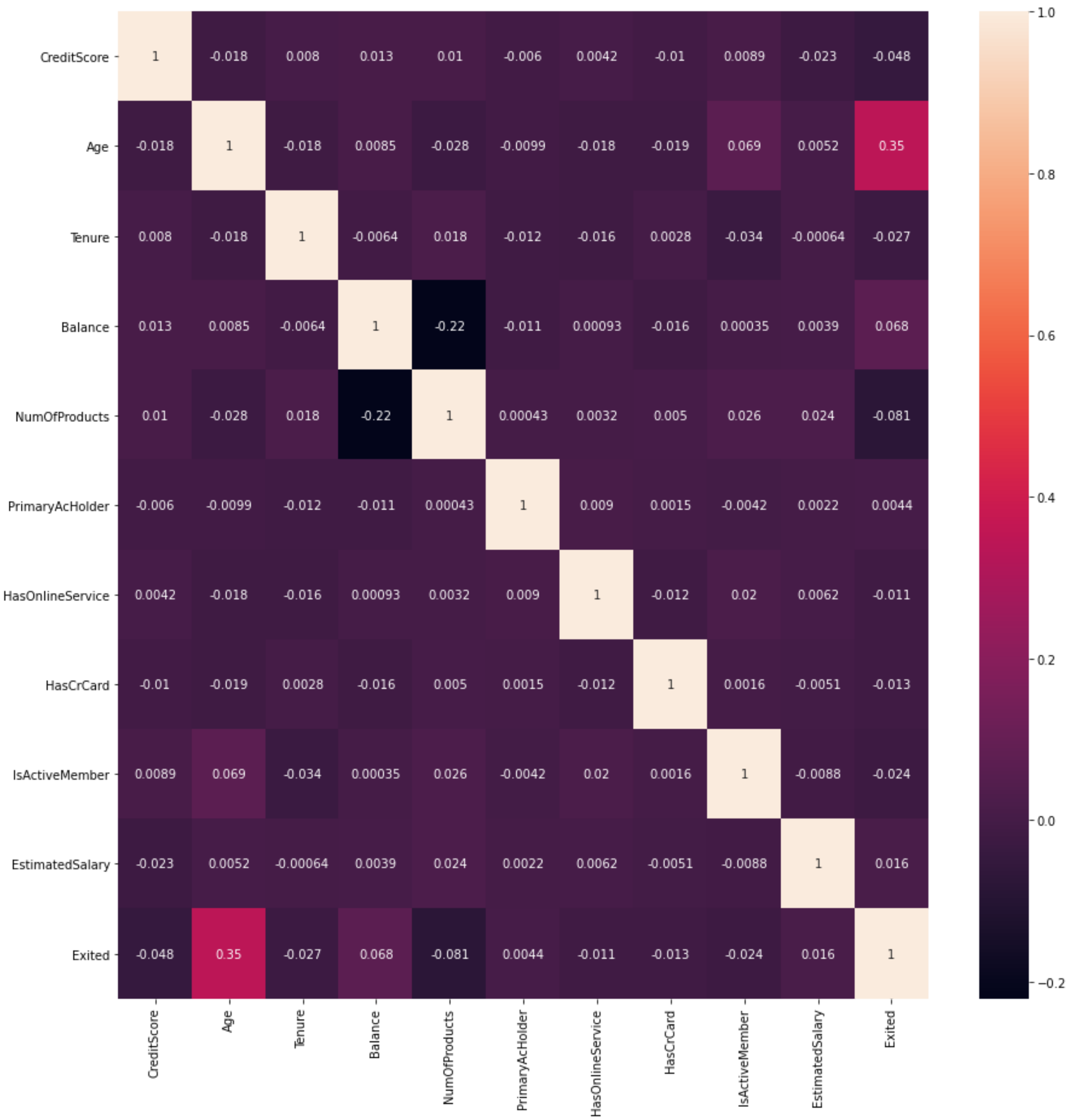
Out[66]:

	CreditScore	Gender	Age	Tenure	Balance	CurrencyCode	Nu
count	14646.000000	14646	14646.000000	14646.000000	14646.000000	14646	
unique	NaN	2	NaN	NaN	NaN	2	
top	NaN	Female	NaN	NaN	NaN	CAD	
freq	NaN	8418	NaN	NaN	NaN	11715	
mean	648.119085	NaN	40.520552	4.999795	76542.069687	NaN	
std	86.326143	NaN	9.510517	2.586318	62165.262069	NaN	
min	350.000000	NaN	18.000000	0.000000	0.000000	NaN	
25%	590.000000	NaN	34.000000	3.000000	0.000000	NaN	
50%	649.000000	NaN	40.000000	5.000000	99681.604705	NaN	
75%	707.000000	NaN	46.000000	7.000000	127261.310425	NaN	
max	850.000000	NaN	92.000000	10.000000	239583.326600	NaN	

In [72]:

```
#Correlation Between Features
#Correlation is a statistical term which in common usage refers to how close two variab
les are to having a
#linear relationship with each other.
#eatures with high correlation are more linearly dependent and hence have almost the sa
me effect on the dependent variable.
#So, when two features have high correlation, we can drop one of the two features.

plt.figure(figsize=(15, 15
                   ))
corrMatrix = df.corr()
sns.heatmap(corrMatrix, annot=True)
plt.show()
```



In [ ]:

```

```