In [1]:
```
from sqlalchemy import create_engine
import pymysql
import pandas as pd

connection = pymysql.connect(host = 'data-analytics-2018.cbrosir2cswx.us-east-1.rds.am
user = 'deepAnalytics',
password = 'Sqltask1234!',
database = 'Credit',
charset = 'utf8mb4',
cursorclass = pymysql.cursors.DictCursor)

df = pd.read_sql('SELECT * FROM credit', con = connection)
```

```
D:\Purdue_Data_Analytics_Cert\Andaconda\envs\DataScience\lib\site-packages\pandas\io
\sql.py:762: UserWarning: pandas only support SQLAlchemy connectable(engine/connectio
n) ordatabase string URI or sqlite3 DBAPI2 connectionother DBAPI2 objects are not tes
ted, please consider using SQLAlchemy
  warnings.warn(
```

In [2]:
```
#Preview the First lines of the Data
df.head()
```

Out[2]:

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	...	X15
0	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	...	BILL_AMT4
1	20000	female	university	1	24	2	2	-1	-1	-2	...	0
2	120000	female	university	2	26	-1	2	0	0	0	...	3272
3	90000	female	university	2	34	0	0	0	0	0	...	14331
4	50000	female	university	1	37	0	0	0	0	0	...	28314

5 rows × 24 columns

In [3]:
```
#Givese Statistical breakdown of the Data
df.describe()
```

Out[3]:

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	...	X15	X16	X17	X
count	3670	3670	3670	3670	3670	3670	3670	3670	3670	3670	...	3670	3670	3670	3
unique	63	3	5	5	53	10	11	11	10	9	...	2009	1984	1948	1
top	50000	female	university	2	29	0	0	0	0	0	...	0	0	0	
freq	453	2130	1644	2045	214	1741	1901	1875	1995	1996	...	424	460	532	

4 rows × 24 columns

In [4]:
```
#Gives info on Data type of the variables
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3670 entries, 0 to 3669
Data columns (total 24 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   X1      3670 non-null   object
 1   X2      3670 non-null   object
 2   X3      3670 non-null   object
 3   X4      3670 non-null   object
 4   X5      3670 non-null   object
 5   X6      3670 non-null   object
 6   X7      3670 non-null   object
 7   X8      3670 non-null   object
 8   X9      3670 non-null   object
 9   X10     3670 non-null   object
 10  X11     3670 non-null   object
 11  X12     3670 non-null   object
 12  X13     3670 non-null   object
 13  X14     3670 non-null   object
 14  X15     3670 non-null   object
 15  X16     3670 non-null   object
 16  X17     3670 non-null   object
 17  X18     3670 non-null   object
 18  X19     3670 non-null   object
 19  X20     3670 non-null   object
 20  X21     3670 non-null   object
 21  X22     3670 non-null   object
 22  X23     3670 non-null   object
 23  Y       3670 non-null   object
dtypes: object(24)
memory usage: 688.2+ KB
```

In [5]:
```
#Checking to see if there is any missing Data
print(df.isnull().sum())
```

```
X1     0
X2     0
X3     0
X4     0
X5     0
X6     0
X7     0
X8     0
X9     0
X10    0
X11    0
X12    0
X13    0
X14    0
X15    0
X16    0
X17    0
X18    0
X19    0
X20    0
X21    0
X22    0
X23    0
Y      0
dtype: int64
```

In [6]:
```
#Gives info on all of data types, "Checking to see if all are correct"
df.dtypes
```

Out[6]:
```
X1     object
X2     object
X3     object
X4     object
X5     object
X6     object
X7     object
X8     object
X9     object
X10    object
X11    object
X12    object
X13    object
X14    object
X15    object
X16    object
X17    object
X18    object
X19    object
X20    object
X21    object
X22    object
X23    object
Y      object
dtype: object
```

In [7]:
```
#Removing all the Duplicate Data
df = df.drop_duplicates()

#Getting Info after removing Duplicates
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2397 entries, 0 to 2397
Data columns (total 24 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   X1      2397 non-null   object
 1   X2      2397 non-null   object
 2   X3      2397 non-null   object
 3   X4      2397 non-null   object
 4   X5      2397 non-null   object
 5   X6      2397 non-null   object
 6   X7      2397 non-null   object
 7   X8      2397 non-null   object
 8   X9      2397 non-null   object
 9   X10     2397 non-null   object
 10  X11     2397 non-null   object
 11  X12     2397 non-null   object
 12  X13     2397 non-null   object
 13  X14     2397 non-null   object
 14  X15     2397 non-null   object
 15  X16     2397 non-null   object
 16  X17     2397 non-null   object
 17  X18     2397 non-null   object
 18  X19     2397 non-null   object
 19  X20     2397 non-null   object
 20  X21     2397 non-null   object
 21  X22     2397 non-null   object
 22  X23     2397 non-null   object
 23  Y       2397 non-null   object
dtypes: object(24)
memory usage: 468.2+ KB
```

In [8]:
```
#Removing the First Row of data
df = df.drop(labels = 0, axis = 0)
```

In [ ]:
```

```