Ashpak Sheikh

HIERARCY CLUSTERING

BATCH : DSWDMOD 020421

Business problem:- Perform clustering for the airlines data to obtain optimum number of clusters. Draw the inferences from the clusters obtained. Refer to EastWestAirlines.xlsx dataset.

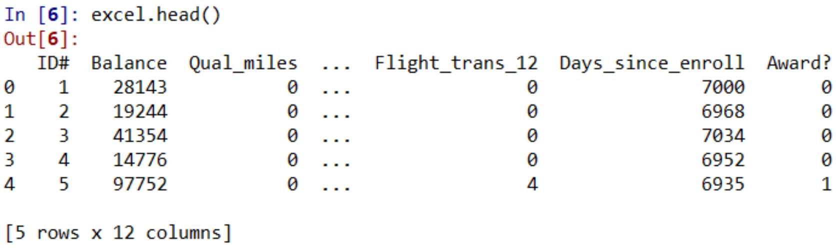
About Data: - We have been given data about EastWest Airline customers,their transaction, balance , bonus mils etc.

Analysis With Python: -

```
import pandas as pd

import numpy as np

excel = pd.read_excel("D:/DataScience/Class/assignment working/h_clustering/EastWestAirlines.xlsx",1)
```

looking at data types

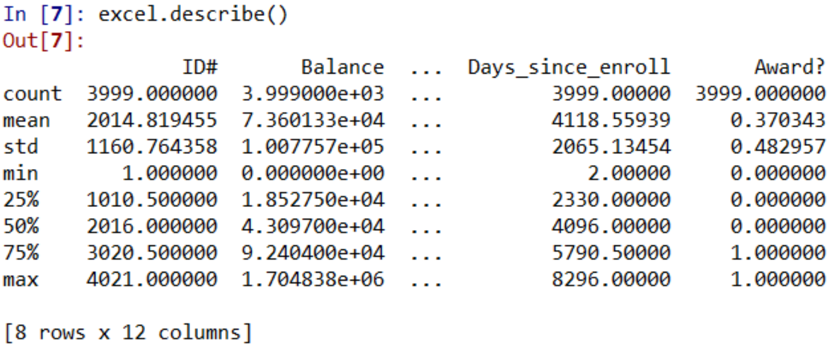
```
excel.head()
```



```
In [6]: excel.head()
Out[6]:
   ID#  Balance  Qual_miles  ...  Flight_trans_12  Days_since_enroll  Award?
0    1    28143           0  ...                0               7000       0
1    2    19244           0  ...                0               6968       0
2    3    41354           0  ...                0               7034       0
3    4    14776           0  ...                0               6952       0
4    5    97752           0  ...                4               6935       1

[5 rows x 12 columns]
```

checking EDA

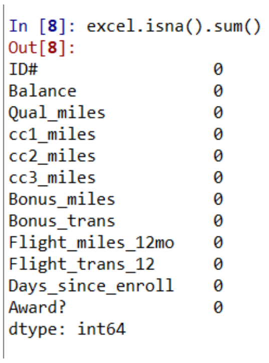
```
excel.describe()
```



```
In [7]: excel.describe()
Out[7]:
               ID#       Balance  ...  Days_since_enroll       Award?
count  3999.000000  3.999000e+03  ...         3999.00000  3999.000000
mean   2014.819455  7.360133e+04  ...         4118.55939     0.370343
std    1160.764358  1.007757e+05  ...         2065.13454     0.482957
min       1.000000  0.000000e+00  ...            2.00000     0.000000
25%    1010.500000  1.852750e+04  ...         2330.00000     0.000000
50%    2016.000000  4.309700e+04  ...         4096.00000     0.000000
75%    3020.500000  9.240400e+04  ...         5790.50000     1.000000
max    4021.000000  1.704838e+06  ...         8296.00000     1.000000

[8 rows x 12 columns]
```

checkinh null values

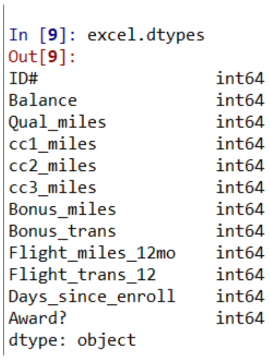
```
excel.isna().sum()
```



```
In [8]: excel.isna().sum()
Out[8]:
ID#                  0
Balance              0
Qual_miles           0
cc1_miles            0
cc2_miles            0
cc3_miles            0
Bonus_miles          0
Bonus_trans          0
Flight_miles_12mo    0
Flight_trans_12      0
Days_since_enroll    0
Award?               0
dtype: int64
```

checking data types

excel.dtypes



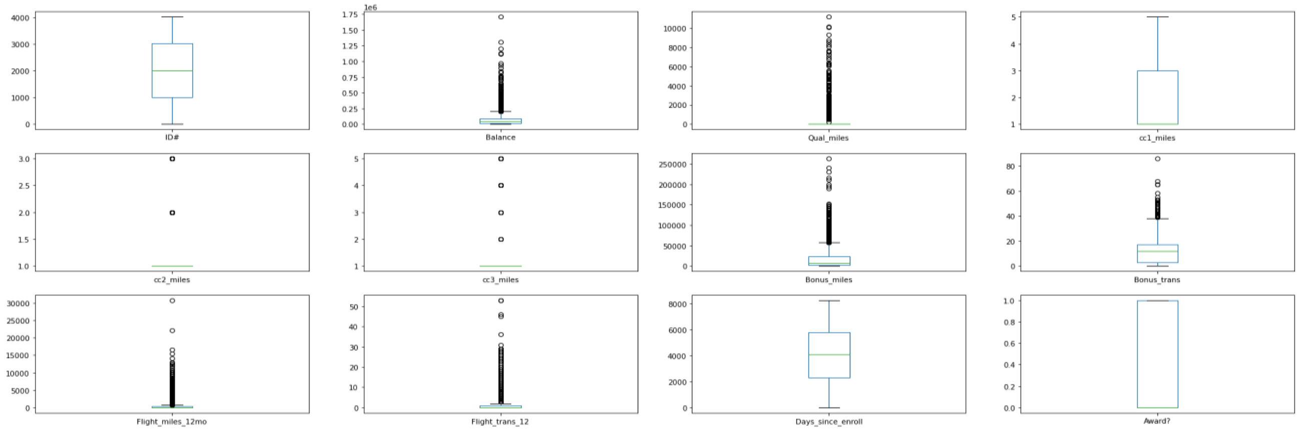
```
In [9]: excel.dtypes
Out[9]:
ID#                  int64
Balance              int64
Qual_miles           int64
cc1_miles            int64
cc2_miles            int64
cc3_miles            int64
Bonus_miles          int64
Bonus_trans          int64
Flight_miles_12mo    int64
Flight_trans_12      int64
Days_since_enroll    int64
Award?               int64
dtype: object
```

checking Duplicates

excel.duplicated().sum()

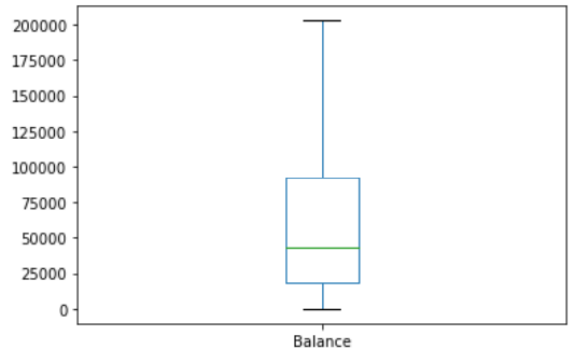
#Checking outliers

excel.plot(kind="box",subplots=True,layout=(4,4),figsize=(30,15))

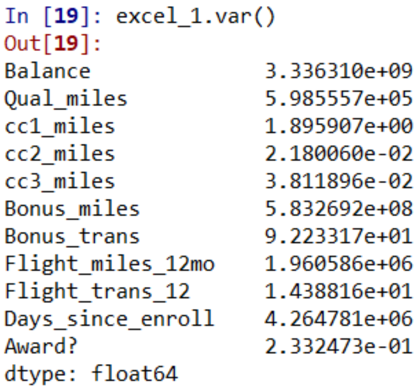


outliers treatment

```
q1=excel["Balance"].quantile(0.25)

q3=excel["Balance"].quantile(0.75)

H_limit=q3+1.5*(q3-q1)

win_quant=excel.Balance.quantile(0.93)

excel['Balance']=np.where(excel["Balance"]>H_limit,win_quant,excel["Balance"])

excel["Balance"].plot(kind="box")
```



```
excel["Qual_miles"].describe()

#droping ID

excel_1=excel.drop(["ID#"],axis=1)


excel_1.var()
```



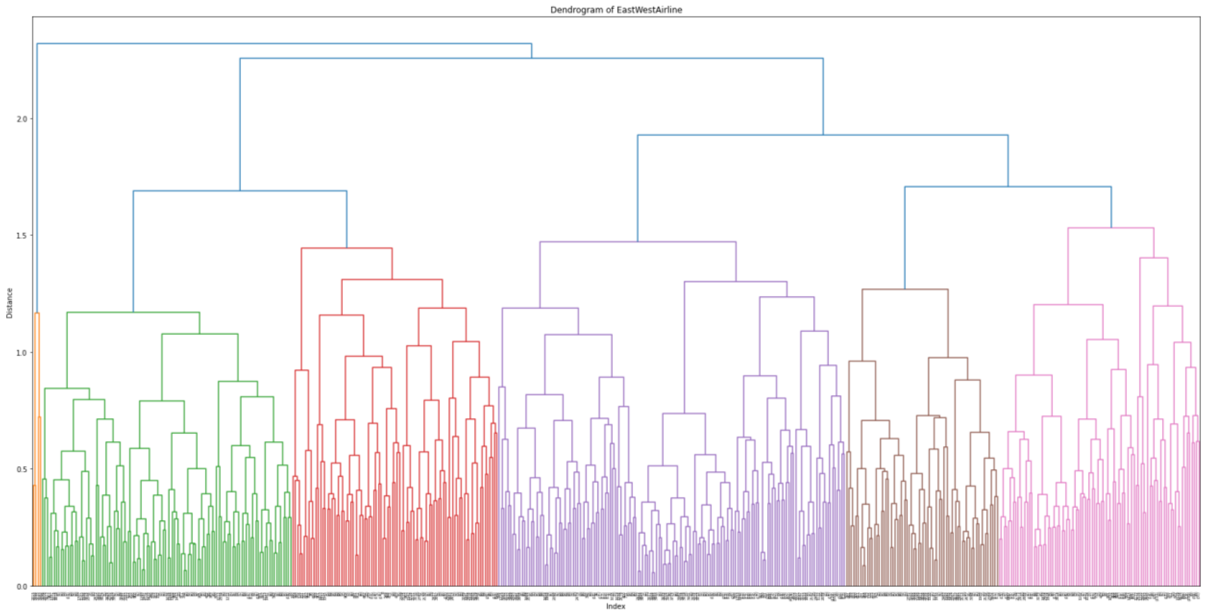
```
In [19]: excel_1.var()
Out[19]:
Balance              3.336310e+09
Qual_miles           5.985557e+05
cc1_miles            1.895907e+00
cc2_miles            2.180060e-02
cc3_miles            3.811896e-02
Bonus_miles          5.832692e+08
Bonus_trans          9.223317e+01
Flight_miles_12mo    1.960586e+06
Flight_trans_12      1.438816e+01
Days_since_enroll    4.264781e+06
Award?               2.332473e-01
dtype: float64
```

cc2_miles and cc3_miles have near zero variance so it wont help in model lerning hence removing them

```
excel_1=excel_1.drop(["cc2_miles","cc3_miles"],axis=1)
```

normalizing data

```
def norm(x):
    z=(x-x.min())/(x.max()-x.min())
    return z
norm_data=norm(excel_1)
from scipy.cluster.hierarchy import linkage
import scipy.cluster.hierarchy as sch
import matplotlib.pyplot as plt


link=linkage(norm_data,method="complete",metric="euclidean")
plt.figure(figsize=(30,15));plt.title("Dendrogram of EastWestAirline");plt.xlabel("Index");plt.ylabel("Distance")
sch.dendrogram(link)
plt.show()
```



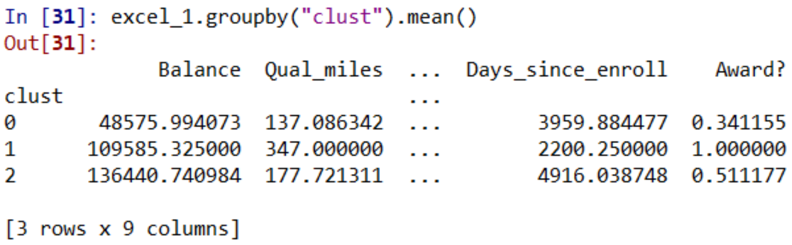
Ashpak Sheikh

HIERARCY CLUSTERING

BATCH : DSWDMOD 020421

Now applying AgglomerativeClustering

```
from sklearn.cluster import AgglomerativeClustering

h_complete = AgglomerativeClustering(n_clusters = 3, linkage = 'complete', affinity = "euclidean").fit(norm_data)

h_complete.labels_



excel_1["clust"] = h_complete.labels_ # creating a new column and assigning it to new column



excel_1.head()


# Aggregate mean of each cluster

excel_1.groupby("clust").mean()
```



```
In [31]: excel_1.groupby("clust").mean()
Out[31]:
             Balance  Qual_miles  ...  Days_since_enroll    Award?
clust                             ...
0       48575.994073  137.086342  ...        3959.884477  0.341155
1      109585.325000  347.000000  ...        2200.250000  1.000000
2      136440.740984  177.721311  ...        4916.038748  0.511177

[3 rows x 9 columns]
```