Week 8 Report

Team Member's details

Group Name: The Potent self

Name: Abdulrahman Afifi

Email: abdulrahmanafifi33@gmail.com

Country: Turkey

College/Company: Bahcesehir University

Specialization: Data Analyst

Problem Description

XYZ firm is gathering client data using Google Forms/Survey Monkey and has published a variety of n forms on the web.

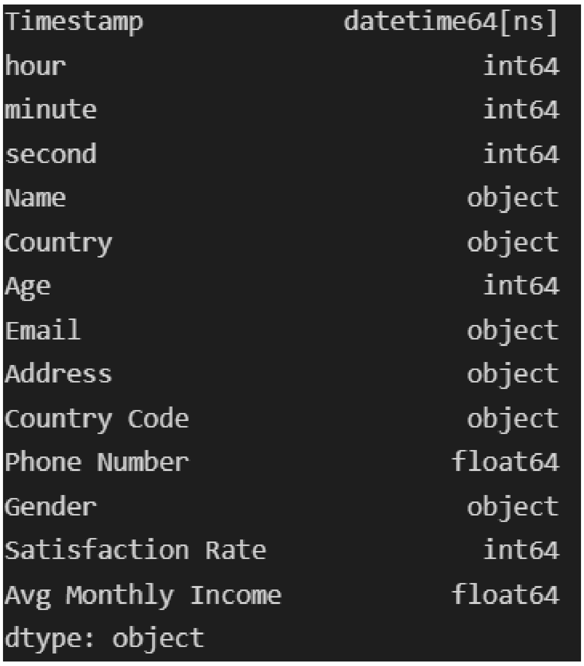
The company want to build a pipeline that would collect all of the data from these Google forms/survey surveys and visualise it on the dashboard.

The company needs clean data, and if there are any data issues in the data, they should be addressed via this process (duplicate data or junk data). A dedup check should be run on the customer's email address.

Data understanding

Total number of observations	3
Total number of files	1
Total number of features	13(including 3 derived features)
Base format of the file	xls
Size of the data	6KB

What type of data you have got for analysis?

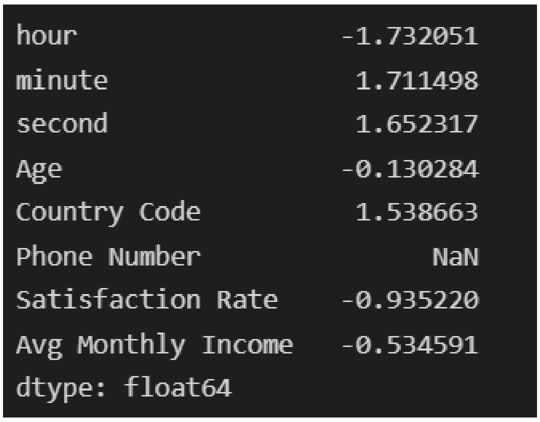


```
Timestamp              datetime64[ns]
hour                            int64
minute                          int64
second                          int64
Name                           object
Country                        object
Age                             int64
Email                          object
Address                        object
Country Code                   object
Phone Number                  float64
Gender                         object
Satisfaction Rate               int64
Avg Monthly Income            float64
dtype: object
```

We have 5 int64 features, 2 float64 features, and 6 object features

What are the problems in the data?

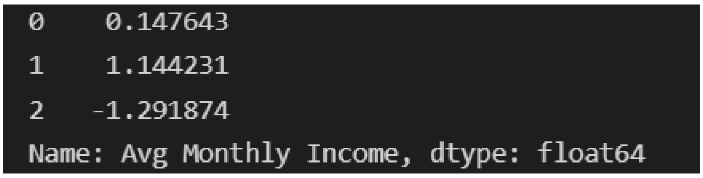
Due to few number of responses retrieved. I can say that there is no problem in the data. No NA values detected. Only one record has dropped below minimum threshold or outlier($13000) of avg monthly income. According to data skewness, most of data appear to be negatively/left skewed except the country code, minute, and second are positively/right skewed. The phone number appears to be symmetric since it has nan skewness ratio.



```
hour                  -1.732051
minute                 1.711498
second                 1.652317
Age                   -0.130284
Country Code           1.538663
Phone Number                NaN
Satisfaction Rate     -0.935220
Avg Monthly Income    -0.534591
dtype: float64
```

What approaches you are trying to apply on your data set to overcome problems like NA value, outlier etc and why?

No NA values detected. If there is I can simply use drop_duplicates function() in pandas. After applying the zscore test it has been found there is no outlier since all zscore values do not exceed value 3 . To treat any outlier, we simply can set this value to mean of the avg monthly income of all customers who responded to this form. Zscore test can be found below!



```
0    0.147643
1    1.144231
2   -1.291874
Name: Avg Monthly Income, dtype: float64
```

GitHub Repo link:

https://github.com/AfifiGhost2000/DataCollectionPipelineProject.git