ABSTRACT

Due to the vast and phenomenal data increase undergoing in this fast-growing world of technology as a result of new inventions taking center stage, data is always increasing under a high rate. This makes data to become the crucial ingredient and byproduct of nearly the whole sphere of our modern world. Hence new ways of dealing with this kind of data are employed to be able to cope with day-to-day growing technology. Data science and analytics provides the skills and competencies to solve these vast amount of data to overcome any challenge that might be incurred in our current institutions and companies and to have future predictions through machine learning. Thus, as a technocrat, I decided to collect some data from companies to illustrate how to analyze data and generate appropriate presentations for a better understanding of data science and analysis.

ACKNOWLEDGEMENT

A wise man once said "If you want to go fast, go alone. If you want to go far, go together. Hence I would like to prolapse the department chairperson, ***Mr. James Mbao***, and the attachment planner***, Mr. Daniel Njuguna***, lecturers, and the school administration in whole for providing me with necessary tools to achieve this particular project. I would also like to recognize ***Oloolaiser water and supply company*** for their generosity of providing me with data sets for analysis. Lastly, I would like to thank my guardians for being to make sure that I have the tools needed to make me a reliable technocrat and be able to fulfil my needs.

Table of Contents

1. INTRODUCTION

1.1 Background

Most of the data collected tend to be dirty. This project entails the collection of dataset obtained from Oloolaiser water and supply company that need to be cleaned using appropriate tools and to make it easy to work on. This will be achieved with the help of tools like pandas, numpy, matplotlib, skilearn and machine learning to be able have a final analyzed data that will be able to be used in the company for future use.

1.2 Objective

The main objective of this project is to be able to come up with a dataset that will make it easy to be able to study machine learning in this particular domain. This will aid in predicting the ongoing and future trends and patterns in the water companies which will then result in the best way the company will be able to handle and deal with their customers to achieve optimal relationship in the business.

2. LITERATURE OVERVIEW

2.1 Overview of data analysis

Research has been done which shows the comprehensive overview of data science covers the analytics, programming, and business skills necessary to master discipline. *Field Cady. (2013).The Data Science Handbook.* Stated that computer science and software engineering offers an extensive coverage since they play the central role in daily work of data scientist. He also state that finding a good data scientist has been likened to hunting a unicorn: the required combination technical skills and flexibility of a person in a particular domain.

2.2 Methodological review

Niranjanamurthy. M. Hemant. (2012). Advances In Data Science and Analytics. Snapplify reader. Presenting the concepts and advances of data science and analytics. It majors on the practical applications that can be utilized across multiple disciplines and industries, for both the engineer and student focusing on machine learning, big data, business intelligence and analytics.

3. DATA COLLECTION

3.1 Data sources and description

Data can be collected from different reliable sources. This includes collecting data from companies that deals with vast amount of data. This includes companies like the banking industry, Water Company, learning institutions and many more companies that deals with a lot of data that are definitely not clean and need to be cleaned in order to be able to have future predictions and coping with fast dynamic changes occurring in this fast developing technology after analyzing them. It is the most convenient way of obtaining data if one has the urge of getting better understanding of data engineering and machine learning.

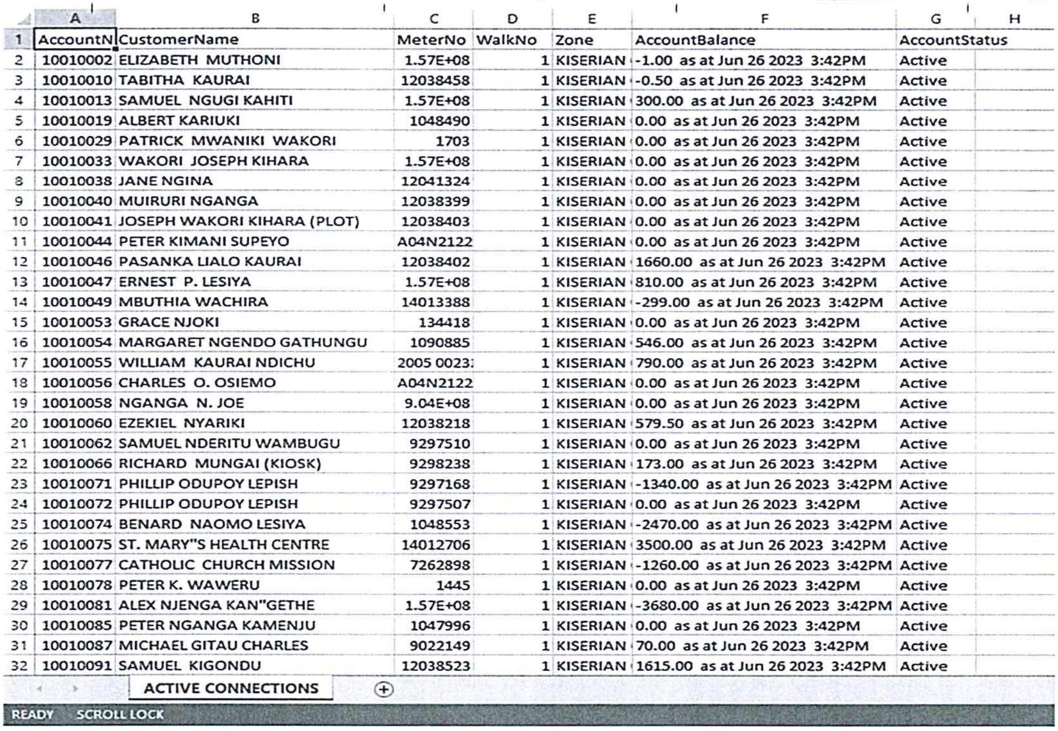
One can also have data by creating them from scratch using different reliable and already inbuilt algorithms. This is achieved by having the knowledge of programming languages which is essential for the data creation. Using this method may result in obtaining clean data but can also be used in data science and analysis.

Another way of obtaining data is by downloading datasets that has been uploaded in various website for the purpose for the purpose of data analysis. Obtaining datasets by using of this method depends on the sites that you will get the data. It tends to be the easiest but not more convenient as compared to data directly from an organization.

3.2 Data collection

Using the most convenient way of obtaining datasets for analysis, which means that one has obtain uncleaned data formally from a company or industry which deals with a lot of data, I decided to go for the water company. I collected my data from Oloolaiser water and supply company who provided me with dataset which probably was dirty to allow me to analyze it get them the feedback inform of presentation.

Here is a fracture of the data I collected from the company;



	A	B	C	D	E	F	G	H
1	AccountN	CustomerName	MeterNo	WalkNo	Zone	AccountBalance	AccountStatus	
2	10010002	ELIZABETH MUTHONI	1.57E+08	1	KISERIAN	-1.00 as at Jun 26 2023 3:42PM	Active	
3	10010010	TABITHA KAURAI	12038458	1	KISERIAN	-0.50 as at Jun 26 2023 3:42PM	Active	
4	10010013	SAMUEL NGUGI KAHITI	1.57E+08	1	KISERIAN	300.00 as at Jun 26 2023 3:42PM	Active	
5	10010019	ALBERT KARIUKI	1048490	1	KISERIAN	0.00 as at Jun 26 2023 3:42PM	Active	
6	10010029	PATRICK MWANIKI WAKORI	1703	1	KISERIAN	0.00 as at Jun 26 2023 3:42PM	Active	
7	10010033	WAKORI JOSEPH KIHARA	1.57E+08	1	KISERIAN	0.00 as at Jun 26 2023 3:42PM	Active	
8	10010038	JANE NGINA	12041324	1	KISERIAN	0.00 as at Jun 26 2023 3:42PM	Active	
9	10010040	MUIRURI NGANGA	12038399	1	KISERIAN	0.00 as at Jun 26 2023 3:42PM	Active	
10	10010041	JOSEPH WAKORI KIHARA (PLOT)	12038403	1	KISERIAN	0.00 as at Jun 26 2023 3:42PM	Active	
11	10010044	PETER KIMANI SUPEYO	A04N2122	1	KISERIAN	0.00 as at Jun 26 2023 3:42PM	Active	
12	10010046	PASANKA LIALO KAURAI	12038402	1	KISERIAN	1660.00 as at Jun 26 2023 3:42PM	Active	
13	10010047	ERNEST P. LESIYA	1.57E+08	1	KISERIAN	810.00 as at Jun 26 2023 3:42PM	Active	
14	10010049	MBUTHIA WACHIRA	14013388	1	KISERIAN	-299.00 as at Jun 26 2023 3:42PM	Active	
15	10010053	GRACE NJOKI	134418	1	KISERIAN	0.00 as at Jun 26 2023 3:42PM	Active	
16	10010054	MARGARET NGENDO GATHUNGU	1090885	1	KISERIAN	546.00 as at Jun 26 2023 3:42PM	Active	
17	10010055	WILLIAM KAURAI NDICHU	2005 0023:	1	KISERIAN	790.00 as at Jun 26 2023 3:42PM	Active	
18	10010056	CHARLES O. OSIEMO	A04N2122	1	KISERIAN	0.00 as at Jun 26 2023 3:42PM	Active	
19	10010058	NGANGA N. JOE	9.04E+08	1	KISERIAN	0.00 as at Jun 26 2023 3:42PM	Active	
20	10010060	EZEKIEL NYARIKI	12038218	1	KISERIAN	579.50 as at Jun 26 2023 3:42PM	Active	
21	10010062	SAMUEL NDERITU WAMBUGU	9297510	1	KISERIAN	0.00 as at Jun 26 2023 3:42PM	Active	
22	10010066	RICHARD MUNGAI (KIOSK)	9298238	1	KISERIAN	173.00 as at Jun 26 2023 3:42PM	Active	
23	10010071	PHILLIP ODUPOY LEPISH	9297168	1	KISERIAN	-1340.00 as at Jun 26 2023 3:42PM	Active	
24	10010072	PHILLIP ODUPOY LEPISH	9297507	1	KISERIAN	0.00 as at Jun 26 2023 3:42PM	Active	
25	10010074	BENARD NAOMO LESIYA	1048553	1	KISERIAN	-2470.00 as at Jun 26 2023 3:42PM	Active	
26	10010075	ST. MARY"S HEALTH CENTRE	14012706	1	KISERIAN	3500.00 as at Jun 26 2023 3:42PM	Active	
27	10010077	CATHOLIC CHURCH MISSION	7262898	1	KISERIAN	-1260.00 as at Jun 26 2023 3:42PM	Active	
28	10010078	PETER K. WAWERU	1445	1	KISERIAN	0.00 as at Jun 26 2023 3:42PM	Active	
29	10010081	ALEX NJENGA KAN"GETHE	1.57E+08	1	KISERIAN	-3680.00 as at Jun 26 2023 3:42PM	Active	
30	10010085	PETER NGANGA KAMENJU	1047996	1	KISERIAN	0.00 as at Jun 26 2023 3:42PM	Active	
31	10010087	MICHAEL GITAU CHARLES	9022149	1	KISERIAN	70.00 as at Jun 26 2023 3:42PM	Active	
32	10010091	SAMUEL KIGONDU	12038523	1	KISERIAN	1615.00 as at Jun 26 2023 3:42PM	Active	

Fig.3.1 Data obtained from the company which is dirty.

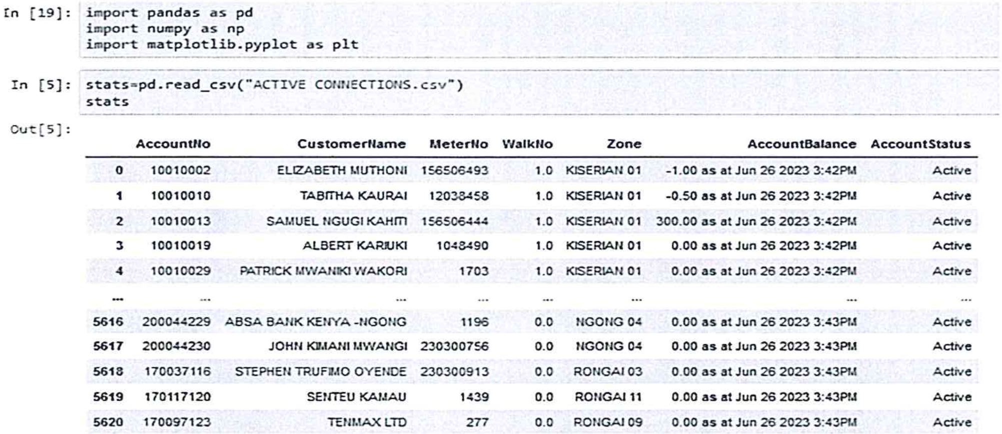
3.3 Data quality assessment

After obtaining this data I it checked out to meet my expectations since it was dirty and it is really a big data. Hence this will this will make it a good necessity for the analysis and thereafter for machine learning thus providing the best solution.

4. DATA PROCESSING

4.1 Data cleaning

This is done by use of data analyzing tools like the anaconda package which is purposely meant for this task. One just ought to upload the data to the environment and clean it using snippet commands like the *dropna()* and the *fillna()* found in the package. Eg,

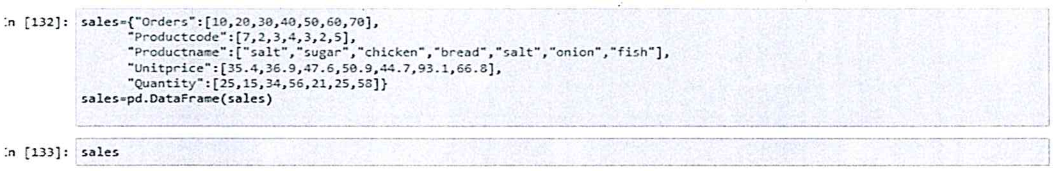


```
In [19]: import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt

In [5]:  stats=pd.read_csv("ACTIVE CONNECTIONS.csv")
         stats

Out[5]:
        AccountNo             CustomerName     MeterNo  WalkNo         Zone                      AccountBalance  AccountStatus
   0     10010002          ELIZABETH MUTHONI  156506493     1.0  KISERIAN 01  -1.00 as at Jun 26 2023 3:42PM           Active
   1     10010010              TABITHA KAURAI   12038458     1.0  KISERIAN 01  -0.50 as at Jun 26 2023 3:42PM           Active
   2     10010013        SAMUEL NGUGI KAHITI  156506444     1.0  KISERIAN 01  300.00 as at Jun 26 2023 3:42PM          Active
   3     10010019             ALBERT KARIUKI    1048490     1.0  KISERIAN 01   0.00 as at Jun 26 2023 3:42PM           Active
   4     10010029     PATRICK MWANIKI WAKORI       1703     1.0  KISERIAN 01   0.00 as at Jun 26 2023 3:42PM           Active
 ...          ...                        ...        ...     ...          ...                             ...              ...
5616    200044229  ABSA BANK KENYA -NGONG          1196     0.0     NGONG 04   0.00 as at Jun 26 2023 3:43PM           Active
5617    200044230        JOHN KIMANI MWANGI   230300756     0.0     NGONG 04   0.00 as at Jun 26 2023 3:43PM           Active
5618    170037116   STEPHEN TRUFIMO OYENDE   230300913     0.0    RONGAI 03   0.00 as at Jun 26 2023 3:43PM           Active
5619    170117120               SENTEU KAMAU       1439     0.0    RONGAI 11   0.00 as at Jun 26 2023 3:43PM           Active
5620    170097123                 TENMAX LTD        277     0.0    RONGAI 09   0.00 as at Jun 26 2023 3:43PM           Active
```

Fig.4.1 Data uploaded to the anaconda package for cleaning.

4.2 Data transformation

This is the act of making the availed data to compact with environment you are working on. For many cases one always have to convert the given data into data frame if you are working with anaconda package. Eg,



```
In [132]: sales={"Orders":[10,20,30,40,50,60,70],
                 "Productcode":[7,2,3,4,3,2,5],
                 "Productname":["salt","sugar","chicken","bread","salt","onion","fish"],
                 "Unitprice":[35.4,36.9,47.6,50.9,44.7,93.1,66.8],
                 "Quantity":[25,15,34,56,21,25,58]}
          sales=pd.DataFrame(sales)

In [133]: sales
```

Fig.4.2.1 An example of data program used to transform data.

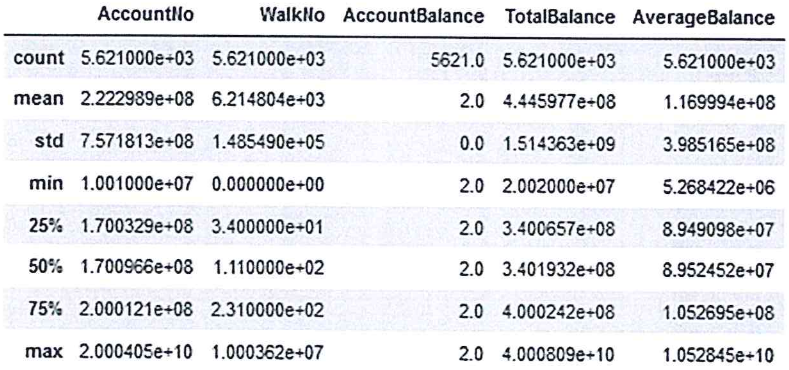
4.3 Handling missing data

This is done by using inbuilt commands to perform them. A good example is the fillna() and the dropna () which are used to fill in missing data and drop the unnecessary data respectively.

5. DATA ANALYSIS AND METHODOLOGY

5.1 Descriptive analysis

After the data is cleaned one can smoothly commence the data analysis by starting with the simple statistical analysis. Here is an example of statistical analysis obtained from the dataset;



	AccountNo	WalkNo	AccountBalance	TotalBalance	AverageBalance
count	5.621000e+03	5.621000e+03	5621.0	5.621000e+03	5.621000e+03
mean	2.222989e+08	6.214804e+03	2.0	4.445977e+08	1.169994e+08
std	7.571813e+08	1.485490e+05	0.0	1.514363e+09	3.985165e+08
min	1.001000e+07	0.000000e+00	2.0	2.002000e+07	5.268422e+06
25%	1.700329e+08	3.400000e+01	2.0	3.400657e+08	8.949098e+07
50%	1.700966e+08	1.110000e+02	2.0	3.401932e+08	8.952452e+07
75%	2.000121e+08	2.310000e+02	2.0	4.000242e+08	1.052695e+08
max	2.000405e+10	1.000362e+07	2.0	4.000809e+10	1.052845e+10

Fig.5.1 Data snippet showing different relationships between the data.

5.2 Visualization techniques

This is the crucial and most important part of the data analysis. Since is the part which shows all what one has been doing and helps in visual understanding for the viewers and whoever is concerned with understanding the provided data without reviewing the data. This can be done with the help of bar graphs, histograms, scatter graphs, pie charts and many more. Eg,

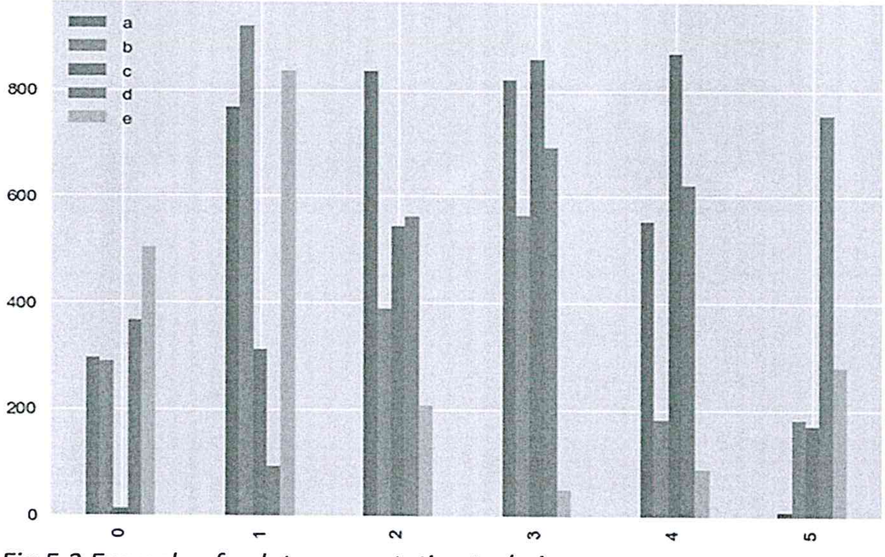


Fig.5.2 Example of a data presentation technique.

6. RESULTS

6.1 Summary of data analysis findings

With the data analysis done and the visual presentation generated ready to be presented and to be used in machine learning, I can summarize that the dataset of the Oloolaiser water and supply company show that there is positive correlation between various variable included in the data. Like the correlation between account balance and total amount payable has a positive relationship.

6.2 Visual presentation of key insights

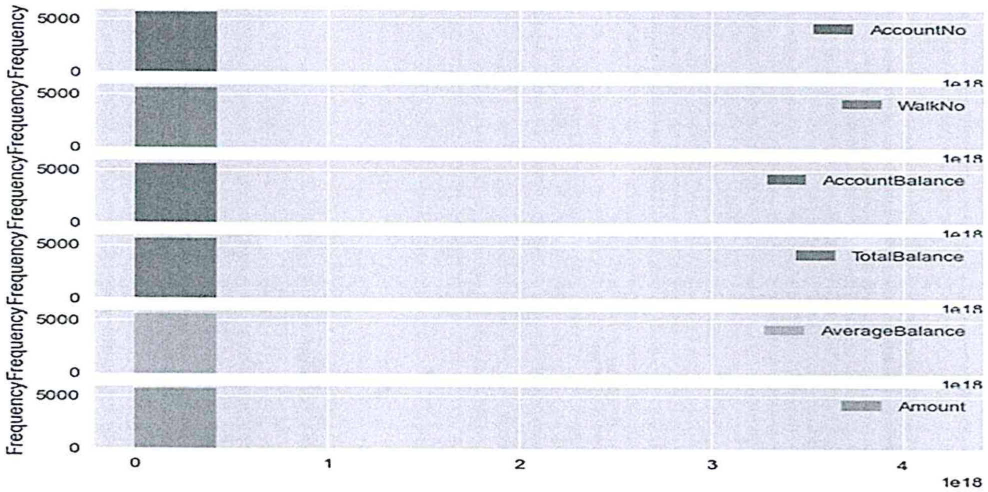


Fig.6.2 Visual representation of the key insight from the data.

6.3 Interpretation of results

The result can be represented in a single figure which will interpret the whole result for the dataset.

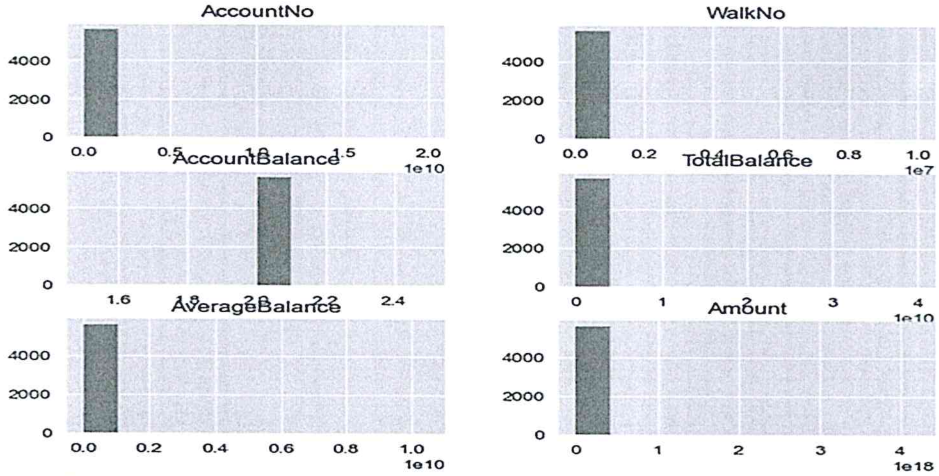


Fig.6.3 Interpretation of the result.

7 DISCUSSION

7.1 Discussion of results in relation to objectives

After carrying out the analysis of the dataset obtained from Oloolaiser water and Supply Company, I can comment that the necessary objectives has been achieved. First, the data has been cleaned to deal with out with unwanted data, then appropriate data summed up to the current data for swift and smooth analysis.

7.2 Limitations and assumptions

Some of the limitations incurred during the analysis project include;

- Collection of some data since some organizational data are sensitive.
- Having to erase some valuable data for accurate result.
- Some of essential tools are not available due to scarce resources.
- Hanging and malfunctioning of some device.

Hypothesis is an essential feature in project. Some of the assumptions made were;

- Any data can be analyzed.
- One can acquire data from any organization.
- Charges must be incurred during data collection.

8. CONCLUSION AND RECOMMENDATIONS

As a technocrat and data scientist and analyst it has come to my conclusion that, data is the backbone of future technology since it covers a span of about all the universe in each and every domain. This is because it is essential in every prospect. Hence, this can be achieved by being able to have the skills and flexibility to deal with data with the aid of machine learning in order to be able to cope with future problems or be able to hinder them from happening early in advance.

Recommendations

- Have the necessaryskills of data science.
- Embrace the use of data analysis and engineering to be able to predict future trends and patterns.
- Educating masses the benefits of data science and analytics.
- Having the reliable tools to achieve data analysis.
- Storing data in a structured way to make it easy in analyzing it.

With all this, there will be guaranteed the best converging relation between service providers and the end users which will nurture an increase in production in our companies and industries.

9. REFERENCES

Field Cady. (2013). *The Data Science Handbook.*

Niranjanamurthy. M. Hemant. (2012). *Advances In Data Science and Analytics.* Snapplify reader.

Vanderplas .T. Jacob. (2016). *Python Data Science Handbook*. Google books

https://www.google.com/search?q=data+science+books&oq=data+science+book&aqs=chrome.1.69i57j0i512l5j46i512j0i512l3.50473j0j1&sourceid=chrome&ie=UTF-8.

Here is the link to access to the portfolio;

https://drive.google.com/drive/folders/1BXxcyfg45n3d_F_cOeXj3U6cVHgMhXrR