Python for Statistics and machine learning

by (Name)

The Name of the Class (Course)

Professor (Tutor)

The Name of the School (University)

The City and State where it is located

The Date

**Introduction**

In a world of fast moving technology and data, individuals and corporations are laying more emphasis on the quality of data that they are getting. There are several channels through which companies, organizations and corporations get their data. These channels range from social media platforms, enterprise resource planning applications (ERP), intelligent systems and customer portals.

The result of such massive amounts of data flowing into organization systems gives birth to big data. Big data is data is data that is fast moving, highly voluminous and comes in many forms. Unlike traditional data, big data is unstructured data that is not organized into any formal rows and columns. Rather, the data within the systems are organized into undefined columns or any meaningful columns.

Such data is stored in relational database systems such as Oracle, MySQL, and DB2 or PostgreSQL applications. These applications make it easier to query and get result sets from applications that they are currently running. On the other hand, non-relational database applications have other tools like Apache Hadoop and Kafka that they use to run analytics on big data applications.

However, the final objective and goal of every business application is to have the business gain visibility on other functions of the company. Such functions include Marketing, Sales and Distribution, I.T and H.R.

When it comes to sales, it is one thing to get a product moving in the market. Whereas the latter is true, it’s possible to have the business get visibility on its performance in future. This is made possible through sales and predictions. As crucial as it is for the business, there is always that undying need for the company to have clear visibility on its premade and rating in the near future. Such performance however are limited to the prevailing factors within the environment such as legal activities, pricing, cost of production, availability of raw materials and labor costs.

During predictive activities, it’s assumed that the business has a clear cut understanding of its current status and all the factors surrounding it. Some of these factors however are within the control of the business.

**Data collection:**

Is the initial step in the decision-making process. Top management relies on decision support tools to develop ready-to-use algorithms and tools that will necessitate the process of highlighting critical decisions to be made in various approaches to resolving company difficulties. Data points are another term for the sources of these data. Data points are critical entry points for data into storage devices.

Typically, the organization has servers within the apps on its network that are intended for information consumption and full use. These are some examples of data points. Customer relationship management systems, firm mobile apps, internet portals, points of sale, contact forms and links, call logs, email dialogues, and social media answers are all examples of customer relationship management systems. What happens if a corporation receives an excessive amount of data that it cannot handle? This is where big data comes in. The following questions will be addressed by big data;

1. What kind of information is this?

2. From where is it coming?

3. Who is submitting it?

4. Where should we keep it?

5. How will we examine it?

Big data has several characteristics, like being quick moving, enormous in volume, and containing a greater quantity of truth. The data storage is the next phase in the data processing. Also, various types of online data storage exist and leave data traces along the way. When you go shopping and swipe your card over the POS, card reader, or other machine, traces of your personal information are left all over the place.

Alternatively, when we go online to search for a specific product to buy, the Google search engine automatically and secretly monitors our key words, maps them to our locations, and then sends us recommendations from other sites based on the type of item and the desire to buy it, as well as how closely it is related to the items we are actually looking for online.

This is one of the reasons why, after looking for a product on, you may receive numerous recommendations from other websites, mobile apps, and social media pages. This type of targeted selling may appear to be illegal and uncalled for, although it is permissible in other ways. Perhaps we should revisit the laws and procedures in place to preserve the privacy of human data; such privileges should not be secured just by obscure terms and conditions that most consumers do not read or comprehend. When it comes to data storage, we can see that programs use three main sorts of formats to store data.

> Structured

> Unstructured

> No storage (web 3.0)

Structured database systems are traditional database systems that store data in the, metadata, and table information and then join different tables together, as we will see later in this discussion.

In contrast, a database management system (DBMS) is a tool for managing structured data.

MySQL, Oracle, DB2, and Postgres are examples of DBMS.

Almost all of these databases operate in the same way; the differences may be minor syntax agreements here and there, the company name and the weather, or whether or not to accept capital SQL queries or simply ignore them. Postgress only accepts lower case queries. Any other query written in upper case will be converted to lower case automatically.

Consider the below table sample for organized /structured database system

|  |  |  |  |
| --- | --- | --- | --- |
| Student\_id | name | wing | grade |
| 0003 | Alex maercies | Blue | 4 |
| 773 | Jontahan Mario | Red | 3 |
| 6443 | Simon Trucey | Green | 1 |
| 5243 | Lucia Degraada | Yellow | 2 |
| 1109 | Kimberly Golden | Purple | 3 |

Structured database systems have been used for a long time and are the oldest in history due to their wide range of applications. It is the case of legacy systems and databases. Their applications range from financial systems to healthcare education systems, research and development studies, and financial stocks, among other developer environments.

Unstructured database systems, on the other hand, are the polar opposite of structured databases. Instead of data being stored in organized rows and columns, the data is tied in some unstructured format, making it impossible to tell which particular row or index value position the data is in. Unstructured data is the industry's second most recent type of database system. Even some developers and software engineers are unfamiliar with databases.Unfortunately, this is where most applicators generating large amounts of data will shift their attention and focus, which is ideal because UDBS is the only way current applications connect with big data.

Because UBD does not consider any unique rows and columns, the UDBS will store images, audio, characters, integers, and even documents that can only be accessed as an array.Consider the following example of an array calling the object:

for (i in siteInfo .users)

{

    for (j in siteInfo.users[i])

    {

        x = siteInfo.users[i][j];

        console.log(x);

    }

}

The unstructured database array above retrieves an object called site info and returns the associated array of items within the object. This is how a database like this works.

Google's Firebase is one of its products. This is a tool that grants developers and organizations access to a wide range of features, including Firebase data tools. This type of database does not store data in rows or columns. MongoDB is another type of database that is widely used for the same reasons as Firebase. Furthermore, the web 3.0 is the most recent introduction to data storage that is not even stored. You see, when the internet was invented, it all started with static sites, then moved on to dynamic systems, then to smart systems, and now the entire system is almost back to web 1,0, but in a smart and silent way.

Later on, developers would take advantage of the engineers' missed opportunity and leverage on the needs of getting calculations and arithmetic applications on their dataset. This is where statistical languages like R came into play. R has been a powerful child in attempting to gather quick summary statistics on the concerned measures of central tendencies, summaries on the values of comparison, and also by assisting in providing the necessary forecasting and data descriptive features and values of such datasets.

Furthermore, the language includes an inbuilt mode that allows the researcher to quickly obtain whatever kind of insights they require from such dataset. Among the models used in this language are: Linear regressions

* Logistic regressions
* ARIMA models
* KNN model
* Random forest and decision trees

By forecasting the models through a fit function, the models have been effective in attempting to assist scientists in discovering relationships between datasets while also being able to predict the future outcomes of such inputs. Python, despite being old and powerful, has been used to wrangle data throughout the entire data cleaning process, analysis, and even building models to predict the dataset. This application's data analysis and features include the ability to gather data from the internet using custom tools such as beautiful soup. Once his data collection and preparation is complete, Python's built-in libraries such as Numpy, SciePy, and pandas have been used as tools to aid in the data visualization process.

Nonetheless, an old but still widely used approach to data analysis is a tool that we are all familiar with: Microsoft Excel. This tool benefits almost every user and basic analyst, regardless of experience or age. Using Microsoft Office, as we will see in the following chapters, an individual can quickly clean their dataset, obtain measures of central tendencies, and then apply the relevant required functions on the data analyses tool pack to quickly obtain meaningful information from the dataset, as we will discuss in the final section of this research paper in the analyses section.

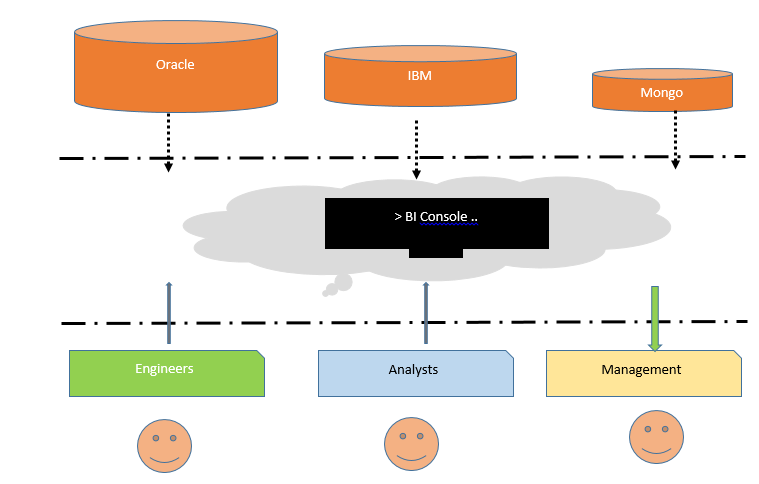
Software as a service (SAS) refers to web tools that are specifically designed to assist users in quickly navigating their analysis. With particular emphasis on the analyses of the products and outputs derived from the dataset that they intend to use. Furthermore, in this study, the SAS tools will form the foundation of our analysis, with key objectives and instructions sent on how to use the SAS software to convert a CSV dataset of our choice into the correct.bat file that we can use on our application to make meaningful and useful data derivations from.

We'll be able to quickly get the right analysis for this work by clicking and dragging.

All of this is necessary to assist the company in understanding and measuring the requirements of developing an effective business information system, with a particular focus on datasets, systems, and whether to consume a structured or unstructured database, the method and language of analysis, and the type of models to apply for each dataset.

Finally, the SAS online tool and the previously mentioned SQL language will be considered here for the development and management of the BI application to help management understand the features and capabilities of BI applications.

**Introducing the BI Application**



The BI is kind of a three phased program. The upper phase that contains the databases needed for data the application process. The lower phase also contains the analysis who add custom queries and models on the dataset, rewrite program logics and execute per ad hoc business requirements to top management. Finally, the managements role in the whole BI thing is to get real time, reports attached click of a button. These reports can be downloaded in the form of PDF documents, XML charts or Microsoft Excel sheets. Management are the top targets in any BI system. The tool should act as a decision support material for them.

**Significance of BI systems**

The value of BI systems are not fetched. It has been established that organisations are able to increase their incomes and revenue streams by close to 30% according to a research done by KPMG in 2018. The reason why this would be so is attributed to the fact that BI systems give the organisation a 360 degree view of what is happening in the company, managers from various departments can monitor data and statistical performances on sales, customer acquisition, geographical mappings of stores and clients and how these regions are doing sales wise.

This kind and level of information is important to any manager as it will help them understand and make key decision on how to manage the products across the various regions of interest and key decisions can now be made faster. Top managers do not have to wait for analysts to extract data from servers, spend several minutes and hours doing cleaning and analysing, but instead, the whole analysis process is managed by the smart BI tool, projects inns are done and where necessary, KPI indicators are monitored ahead of the period or week on due.

Minimal errors on reporting. Since the BI tool is managed and controlled by a large team of experienced engineers, it’s not easy to mess. Every engineer and analyst does the maker checker. Before one process is taken to the next, the period process checks for errors and alerts the staff working on that process, the process shall not proceed if the current process is not solved. This kind of scenario ensures that the final process only gives quality reporting. On the contrary though, this is not easier to achieve when doing manual reporting.

In this particular study, the researcher aims to establish the possibility of having a predictive model run for the sales and marketing so the business can tell its performance in the near future. Sample datasets are provided and graphical representations are illustrated. To have an efficient and effective predictive model, the following steps were followed at arriving to these models.

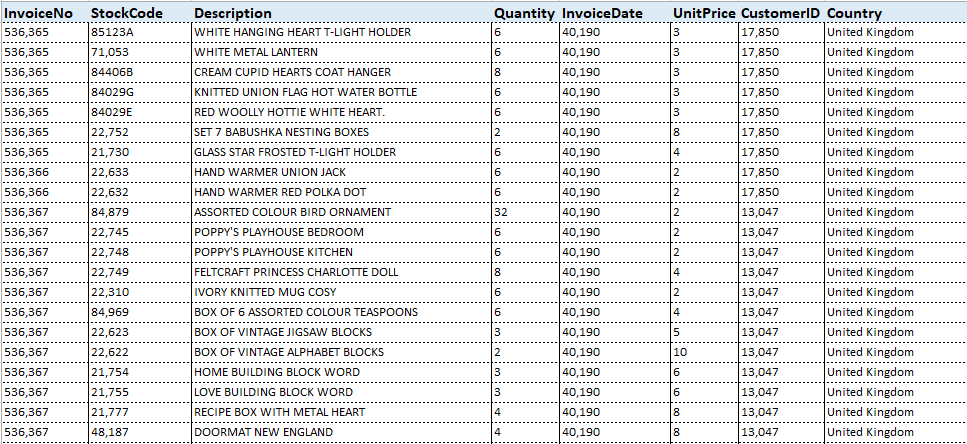
**Know your metrics:**

It is considered one of the key processes in establishing a successful business. This has something to do with where the business is coming from and where the business needs to go. Factors looked at here include profits, losses and costs of operations. Other factors looked at include the zones of operations and the distribution of the customers in these zones and locations.

Data collection and summary:

The dataset used for this particular part was obtained here:

<https://www.kaggle.com/vijayuv/onlineretail>



Then we move on with using Python to establish the monthly revenue:

# Get the necessary files

from datetime import datetime, timedelta

import pandas as pd

%matplotlib inline

import matplotlib.pyplot as plt

import numpy as np

import seaborn as sns

from \_\_future\_\_ import division

import plotly.plotly as py

import plotly.offline as pyoff

import plotly.graph\_objs as go

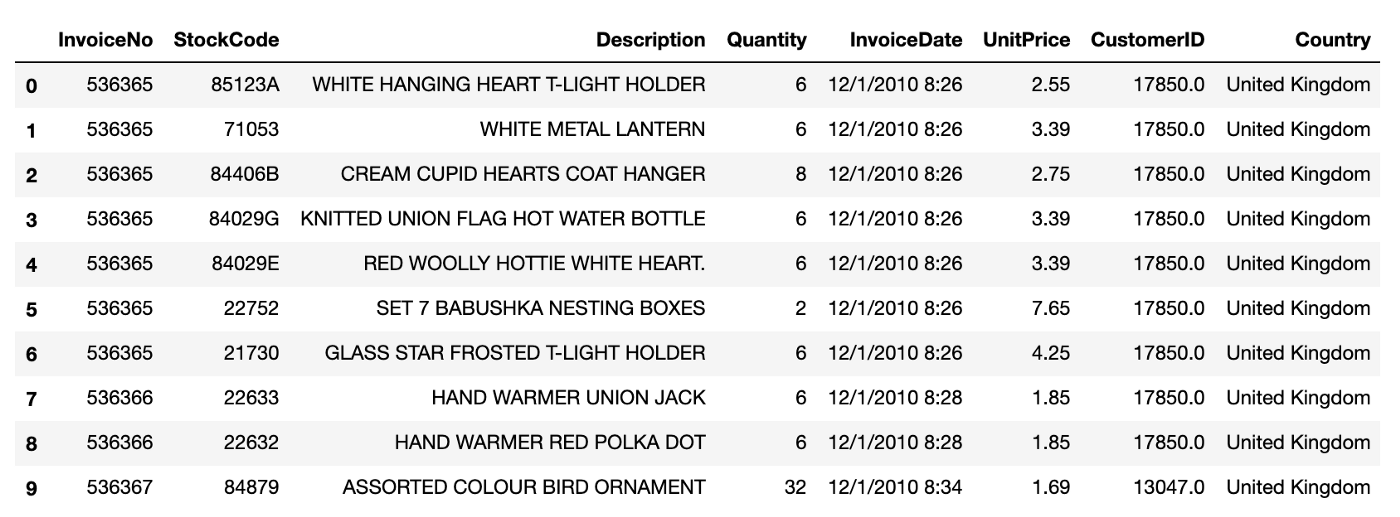
# visualize the dataset

pyoff.init\_notebook\_mode()

input\_data = pd.read\_csv('input.csv')

input\_data.head(10)

We run the code on Python Jupyter Notebook and get back the following feedback:



In order to get the monthly data, we write the Python code that will help us facilitate this:

input\_data['InvoiceDate'] = pd.to\_datetime(input\_data['InvoiceDate'])

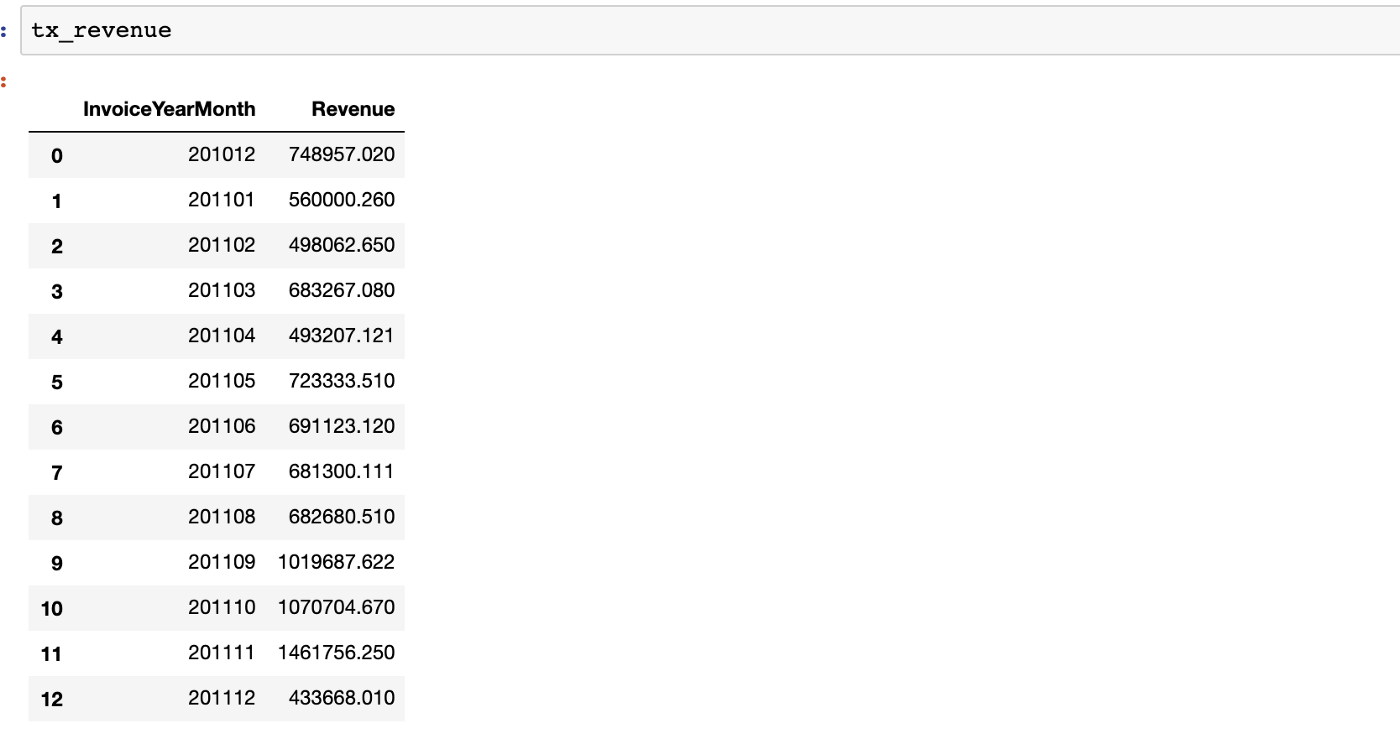
input\_data['InvoiceYearMonth'] = input\_data['InvoiceDate'].map(lambda date: 100\*date.year + date.month)

input\_data['Revenue'] = input\_data['UnitPrice'] \* input\_data['Quantity']

tx\_revenue = tx\_data.groupby(['InvoiceYearMonth'])['Revenue'].sum().reset\_index()

tx\_revenue

Once this is done, we get the below result:



In order to visualize this data, we use some Python code to visualize the variables:

graphly = [

    go.Scatter(

        x=tx\_revenue['MonthYear'],

        y=tx\_revenue['Total Revenue'],

    )

]

graph\_layout = go.Layout(

        xaxis={"type": "category"},

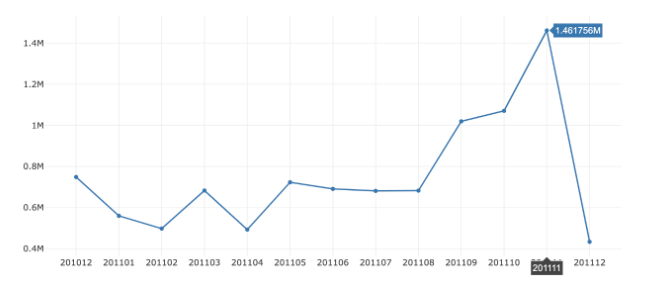
        title='Montly Revenue'

    )

figure = go.Figure(data=graphly, layout=plot\_layout)

pyoff.iplot(figure)

Once this is done, the result looks as below:

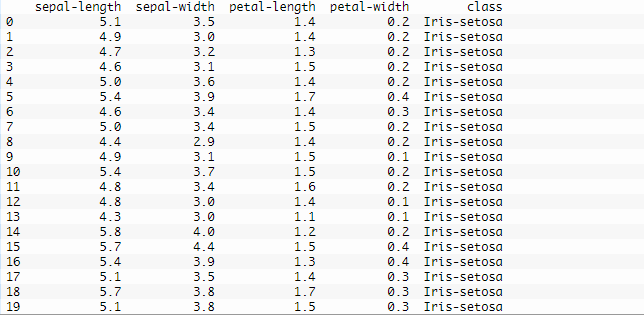


From the graph, it can be seen that most of the sales were made after 2011 October before the sales dropped to 0.4 million in December same year.

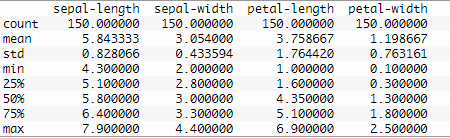
**Supervised learning:**

In this section, an analysis and forecasting is done based on the iris dataset that is publicly available on git hub. Then we shall try to use a model to classify the iris flower datasets.

A quick heads on the dataset reveals the below:



A statistical summary of the dataset reveals the below:



from pandas import read\_csv

url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/iris.csv"

names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']

dataset = read\_csv(url, names=names)

print(dataset.shape)

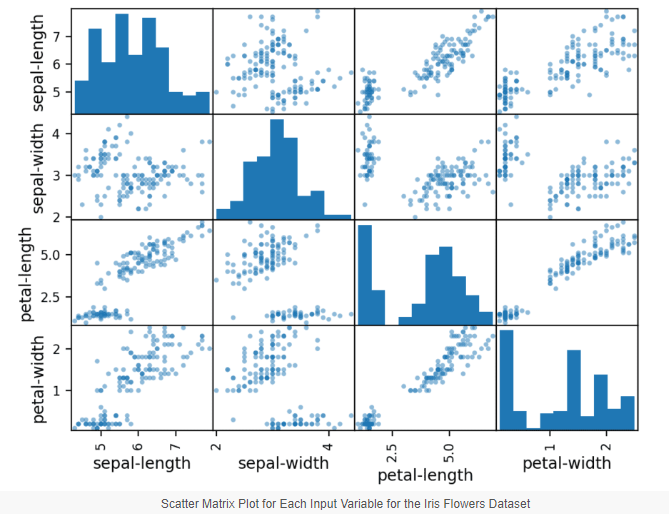
print(dataset.head(20))

print(dataset.describe())

print(dataset.groupby('class').size())

dataset.plot(kind='box', subplots=True, layout=(2,2), sharex=False, sharey=False)

pyplot.show()



#

from pandas import read\_csv

from matplotlib import pyplot

from sklearn.model\_selection import train\_test\_split

from sklearn.model\_selection import cross\_val\_score

from sklearn.model\_selection import StratifiedKFold

from sklearn.linear\_model import LogisticRegression

from sklearn.tree import DecisionTreeClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.discriminant\_analysis import LinearDiscriminantAnalysis

from sklearn.naive\_bayes import GaussianNB

from sklearn.svm import SVC

#

url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/iris.csv"

names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']

dataset = read\_csv(url, names=names)

#

array = dataset.values

X = array[:,0:4]

y = array[:,4]

X\_train, X\_validation, Y\_train, Y\_validation = train\_test\_split(X, y, test\_size=0.20, random\_state=1, shuffle=True)

#

models = []

models.append(('LR', LogisticRegression(solver='liblinear', multi\_class='ovr')))

models.append(('LDA', LinearDiscriminantAnalysis()))

models.append(('KNN', KNeighborsClassifier()))

models.append(('CART', DecisionTreeClassifier()))

models.append(('NB', GaussianNB()))

models.append(('SVM', SVC(gamma='auto')))

#

results = []

names = []

for name, model in models:

    kfold = StratifiedKFold(n\_splits=10, random\_state=1, shuffle=True)

    cv\_results = cross\_val\_score(model, X\_train, Y\_train, cv=kfold, scoring='accuracy')

    results.append(cv\_results)

    names.append(name)

    print('%s: %f (%f)' % (name, cv\_results.mean(), cv\_results.std()))

# Compare Algorithms

pyplot.boxplot(results, labels=names)

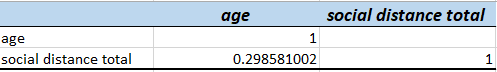
pyplot.title('Algorithm Comparison')

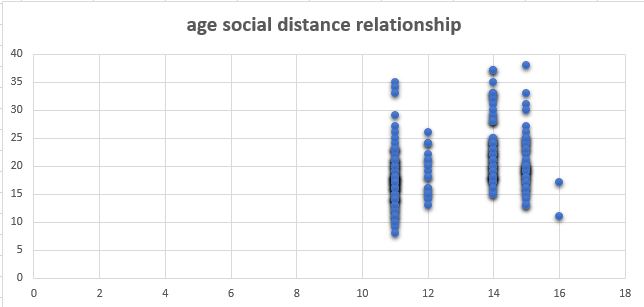
pyplot.show()

**Correlational analysis**

In this study, the researcher sought to establish the relationship between age, gender and social distance. Analysis of the initial dataset was established as below;

**Correlation between age and social distance results**



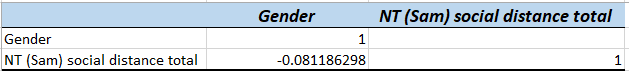


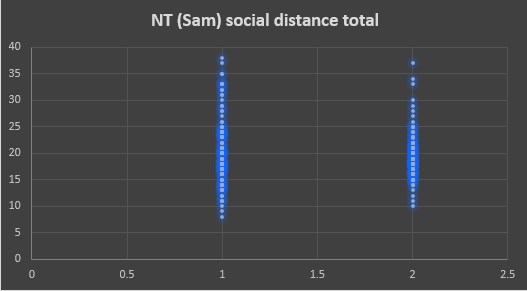
**Discussion:**

From the results of the analysis above, it can be seen that there is weak relationship between the age of a participant and their social distance. In this case, the independent variable selected as was social distance does seem to change significantly as the age of the participant increases. And the vice versa is true that social distance does not seem to change valiantly as the age reduces. The result is some kind of a scattered distribution, actually, if a line of best it is established on the graph, the linear regression determination becomes null since no pattern exist. Further, the correlation matrix produces a value of 0.29858 which is very low and far away from 1, which clearly indicates that the two variables do not depend on each other in any way. Age and social distance of the participant are not related.

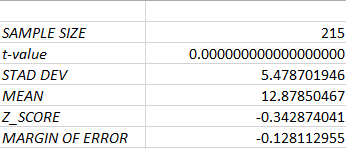
**Gender and social distance**

The results from gender and social distance are summarised below based on the analysis done;





**Summary of the correlation model**



**Discussion**

From the graph obtained above between gender and the social distance of the participants, it can be seen that since we are measuring two variables that are almost Boolean in nature, the relationship is inferentially skewed, even though it can be most observed that chances of males dominating higher social distance is higher because they are the majority in the population. However, it terms of linear correlation, the two datasets are not related, in fact, the correlation coefficient value gives –0.081, which is weaker from 1; this can be interpreted to mean that there is no relationship between gender and social distance. I.e. if gender changes, this does not affect the social distance of the participant. Additionally, if we establish a line of best fit from the data, it establishes no uniform movement of the data or objective being established, so we can conclude that gender does not affect the social distance of the participant.

Further analysis done based on as a sample dataset distribution of 215 entities in the population compared age and the social distance between the participants. The relationship established relationship between values as follows;

T value of less than *0.0000.* Showing a greater difference from the objected null hypotheses earlier discussed in the chapter, this is just to highlight that the two variables against confirm greater none linear relation among each other.

The standard deviation output from this sample of 215 indicated that a value of 5.478 was established; if we extrapolate this against an inferential value of 1.0, this further establishes that that their data point relationship among these two are widely spread out and none, is a causative agent of the other. Further Z score value derived as -0.34 indicates a reflection further away from the mean score value derived from the same dataset. Inferentially from sample mean of 12.88, the margin of error scored at -0.12, what this means is that the margin of error boundary is lightly significant to the researcher, revealing a near to less correlation between the given two datasets, i.e. considering age and the social distance.

**BIBLIOGRAPHY**

Andersen, L.B., Eriksson, T., Kristensen, N. and Pedersen, L.H., 2012. Attracting public service motivated employees: How to design compensation packages. International Review of Administrative Sciences, 78(4), pp.615-641.

Chaudhry, M.S., Sabir, H.M., Rafi, N. and Kalyar, M.N., 2011. Exploring the relationship between salary satisfaction and job satisfaction: A comparison of public and private sector organizations. The Journal of Commerce, 3(4), p.1.

D’Ambrosio, C., Jäntti, M. and Lepinteur, A., 2020. Money and happiness: Income, wealth and subjective well-being. Social Indicators Research, 148(1), pp.47-66.

Henderson, Andrew D., and James W. Fredrickson. "Top management team coordination needs and the CEO pay gap: A competitive test of economic and behavioral views." *Academy of Management Journal* 44.1 (2001): 96-117.

Singh, S.K., Gupta, S., Busso, D. and Kamboj, S., 2021. Top management knowledge value, knowledge sharing practices, open innovation and organizational performance. Journal of Business Research, 128, pp.788-798.