INSTITUTE OF TECHNOLOGY
& MANAGEMENT
GWALIOR • MP • INDIA

Laboratory Manual

Data Analytics Lab
(CS-605)

For

Third Year Students
Department: Computer Science & Engineering

INSTITUTE OF TECHNOLOGY
& MANAGEMENT
GWALIOR • MP • INDIA

Department of Computer Science and Engineering

Vision of CSE Department:

The department envisions to nurture students to become technologically proficient, research competent and socially accountable for the welfare of the society.

Mission of the CSE Department:

I. To provide high quality education through effective teaching-learning process emphasizing active participation of students.

II. To build scientifically strong engineers to cater to the needs of industry, higher studies, research and startups.

III. To awaken young minds ingrained with ethical values and professional behaviors for the betterment of the society.

Program Educational Objectives:

Graduates will be able to

I. Our engineers will demonstrate application of comprehensive technical knowledge for innovation and entrepreneurship.

II. Our graduates will employ capabilities of solving complex engineering problems to succeed in research and/or higher studies.

III. Our graduates will exhibit team-work and leadership qualities to meet stakeholder business objectives in their careers.

IV. Our graduates will evolve in ethical and professional practices and enhance socioeconomic contributions to the society.

Program Outcomes (POs):

Engineering Graduates will be able to:

1. **Engineering knowledge**: Apply the knowledge of mathematics, science, engineering Fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem analysis**: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7. **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one’s own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
12. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

INSTITUTE OF TECHNOLOGY
& MANAGEMENT
GWALIOR • MP • INDIA

Course Outcomes

Data Analytics Lab (CS-605)

CO1:	Understand and apply the basic of data analytics concepts of statistics and probability.
CO2 :	Apply the data processing techniques on Data Frame using Python Libraries.
CO3 :	Implement and evaluate the data analytics techniques using MATLAB, R and Python tools.
CO4 :	Able to evaluate or assess models with the large volume of data with the help of modern tools
CO5 :	Define and explain to python for data cleaning and visualization as a data analytics tool.

Course	Course Outcomes	CO Attainment	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
CO1	Understand and apply the basic of data analytics concepts of statistics and probability.		1	1	1	1									2		
CO2	Apply the data processing techniques on Data Frame using Python Libraries.		1	1	1	1	1				1			1	1	2	
CO3	Implement and evaluate the data analytics techniques using MATLAB, R and Python tools.		1	1		1	1							1		2	2
CO4	Able to evaluate or assess models with the large volume of data with the help of modern tools			1		1	1				1	1			2		1
CO5	Define and explain to python for data cleaning and visualization as a data analytics tool.			1	1	2					1	1		2	2		2

ITM
INSTITUTE OF TECHNOLOGY
& MANAGEMENT
GWALIOR • MP • INDIA

List of Program

S. No.	List	Course Outcomes	Page No.
	Introduction to Python Introduction to Analytics		1-2
1.	Write a Python Program to Get Total Price of all FuelType from Toyota.csv file and show it using a line plot with the following Style properties. Generated line plot must include following Style properties: • Line Style dotted and Line-color should be red • Show legend at the lower right location. • X label name = Fuel Type • Y label name = Price • Add a circle marker. • Line marker color as red • Line width should be 3	CO1, CO5	3-4
2	Write a Python Program to Read 'Petrol' and 'CNG' FuelType sales data from Toyota.csv file and show it using the bar chart	CO1	5

3	Write a Python program to create and display a DataFrame from a specified dictionary data which has the index labels.
	exam_data = {'name': ['Dinesh', 'Suresh', 'Rahul', 'Ravi', 'Manoj', 'Hari', 'Yatharth', 'Saurabh', 'Kapil', 'Salini'],
	'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],
	'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],
	'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}
	labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']
		CO1, CO2	6

4	Write a Python program to display a summary of the basic information like index, columns, non null values of each column, memory usage etc. about a specified DataFrame which is Toyota.csv.	CO1, CO2	7
5	WAP in Python to insert a column named “AGE_IN_MONTH” and then fill the values into inserted column on the basis of “AGE” column using user defined function where function return a value. AGE_IN_MONTH=AGE*12	CO3	8
6	Write a program that accepts a sequence of words as input and write the words in a comma-separated sequence file after sorting them	CO2	9

	alphabetically.		
7	WAP in Python to copy line-by-line contents of a file into another file.	CO2	10
8	WAP to enter characters one by one and then stored lower characters in the LOWER file and upper characters stored in the UPPER file and other characters stored in the OTHER file.	CO3	11
9	Explain and Understand the steps to solve the following system of linear equations using MATLAB: $2x-3y+4z=5$ $y+4z+x=10$ $-2z+3x+4y=0$	CO2, CO3	12-13
10	The dataset Toyota.csv may be found on the web page. This dataset contains Car information: Price, Age, KM, FuelType, KM, CC and Doors. Save this file and use read.table to import it into R. What are the means and standard deviations of the data variables (excluding Age)? Apply Data Analytic tool.	CO3, CO4	14

INTRODUCTION TO PYTHON

In Python programming section, the applications of Python are taken into account. Applications of Python which are taken into details according to the syllabus prescribed by RGPV Bhopal for this lab are:

a) Console Based Programming
b) OOPs Based Programming
c) GUI Based Programming
d) String
e) List
f) Tuple
g) Dictionary

LAB REQUIREMENTS

For Python Programming

- Python 3.7
- PyCharm
- Anaconda

This Interpreter has no special hardware requirements as such. Any System with a minimum 256 MB RAM and any normal processor can use for this lab.

INTRODUCTION DATA ANALYTICS

Data Analytics refers to the techniques used to analyze data to enhance productivity and business gain. Data is extracted from various sources and is cleaned and categorized to analyze various behavioral patterns. The techniques and the tools used vary according to the organization or individual. Data Analytics has a key role in improving your business as it is used to gather hidden insights, generate reports, perform market analysis, and improve business requirements.

Different role of Data Analytics:

- **Gather Hidden Insights –** Hidden insights from data are gathered and then analyzed with respect to business requirements.
- **Generate Reports –** Reports are generated from the data and are passed on to the respective teams and individuals to deal with further actions for a high rise in business.
- **Perform Market Analysis –** Market Analysis can be performed to understand the strengths and weaknesses of competitors.

- **Improve Business Requirement** – Analysis of Data allows improving Business to customer requirements and experience.

Different tools used in Data Analytics:

With the increasing demand for Data Analytics in the market, many tools have emerged with various functionalities for this purpose. Either open-source or user-friendly, the top tools in the data analytics market are as follows.

- **R programming** – This tool is the leading analytics tool used for statistics and data modeling. R compiles and runs on various platforms such as UNIX, Windows, and Mac OS. It also provides tools to automatically install all packages as per user-requirement.
- **Python** – Python is an open-source, object-oriented programming language that is easy to read, write, and maintain. It provides various machine learning and visualization libraries such as Scikit-learn, TensorFlow, Matplotlib, Pandas, Keras, etc. It also can be assembled on any platform like SQL server, a MongoDB database or JSON

- **Tableau Public** – This is a free software that connects to any data source such as Excel, corporate Data Warehouse, etc. It then creates visualizations, maps, dashboards etc with real- time updates on the web.
- **QlikView** – This tool offers in-memory data processing with the results delivered to the end-users quickly. It also offers data association and data visualization with data being compressed to almost 10% of its original size.
- **SAS** – A programming language and environment for data manipulation and analytics, this tool is easily accessible and can analyze data from different sources.
- **Microsoft Excel** – This tool is one of the most widely used tools for data analytics. Mostly used for clients' internal data, this tool analyzes the tasks that summarize the data with a preview of pivot tables.
- **RapidMiner** – A powerful, integrated platform that can integrate with any data source types such as Access, Excel, Microsoft SQL, Tera data, Oracle, Sybase etc. This tool is mostly used for predictive analytics, such as data mining, text analytics, machine learning.

Program 1

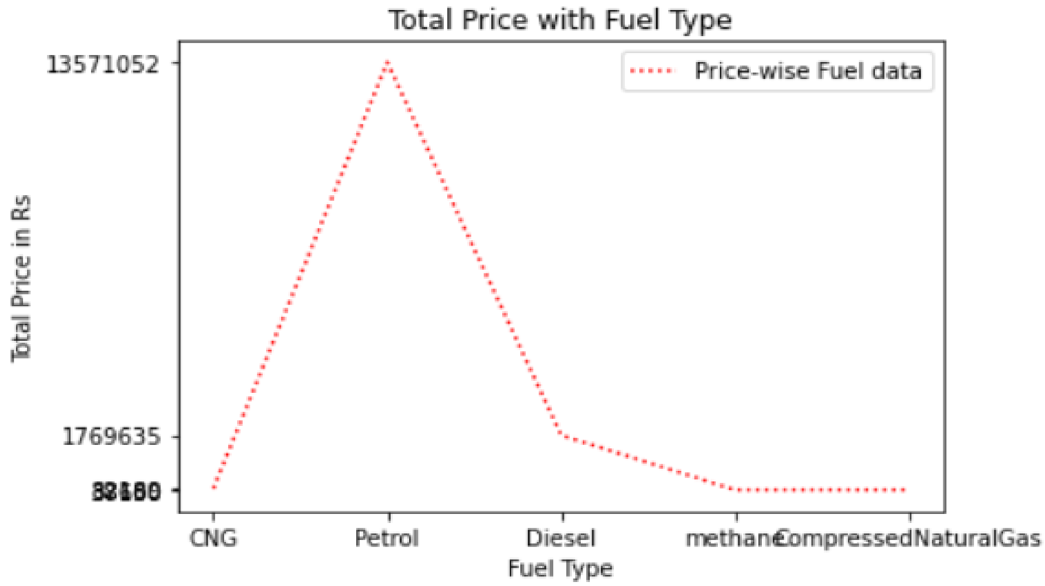
Write a Python Program to Get Total Price of all FuelType from Toyota.csv file and show it using a line plot with the following Style properties. Generated line plot must include following Style properties

- Line Style dotted and Line-color should be red
- Show legend at the lower right location.
- X label name = Fuel Type
- Y label name = Price

Procedure:

```
price1=price2=price3=price4=price5=0
for x in df.index:
    if df.loc[x,'FuelType']=='CNG':
        price1=price1+(df.loc[x,'Price'])
        fuel1='CNG'
    elif (df.loc[x,'FuelType']=='Petrol') | (df.loc[x,'FuelType']=='petrol'):
        price2=price2+(df.loc[x,'Price'])
        fuel2='Petrol'
    elif (df.loc[x,'FuelType']=='Diesel') | (df.loc[x,'FuelType']=='diesel'):
        price3=price3+(df.loc[x,'Price'])
        fuel3='Diesel'
    elif df.loc[x,'FuelType']=='methane':
        price4=price4+(df.loc[x,'Price'])
        fuel4='methane'
    elif df.loc[x,'FuelType']=='CompressedNaturalGas':
        price5=price5+(df.loc[x,'Price'])
        fuel5='CompressedNaturalGas'
pricelist=[price1,price2,price3,price4,price5]
#print(l6)
fuelList  = [fuel1,fuel2,fuel3,fuel4,fuel5]
plt.plot(fuelList, pricelist, label = 'Price-wise Fuel data',linestyle='dotted',color='red')
plt.xlabel('Fuel Type')
plt.ylabel('Total Price in Rs')
plt.xticks(fuelList)
plt.title('Total Price with Fuel Type')
plt.yticks([82189, 13571052, 1769635, 37600, 38150])
plt.legend()
plt.show()
```

Result:


Total Price with Fuel Type
Price-wise Fuel data
Total Price in Rs
13571052
1769635
CNG
Petrol
Diesel
methane
CompressedNaturalGas
Fuel Type

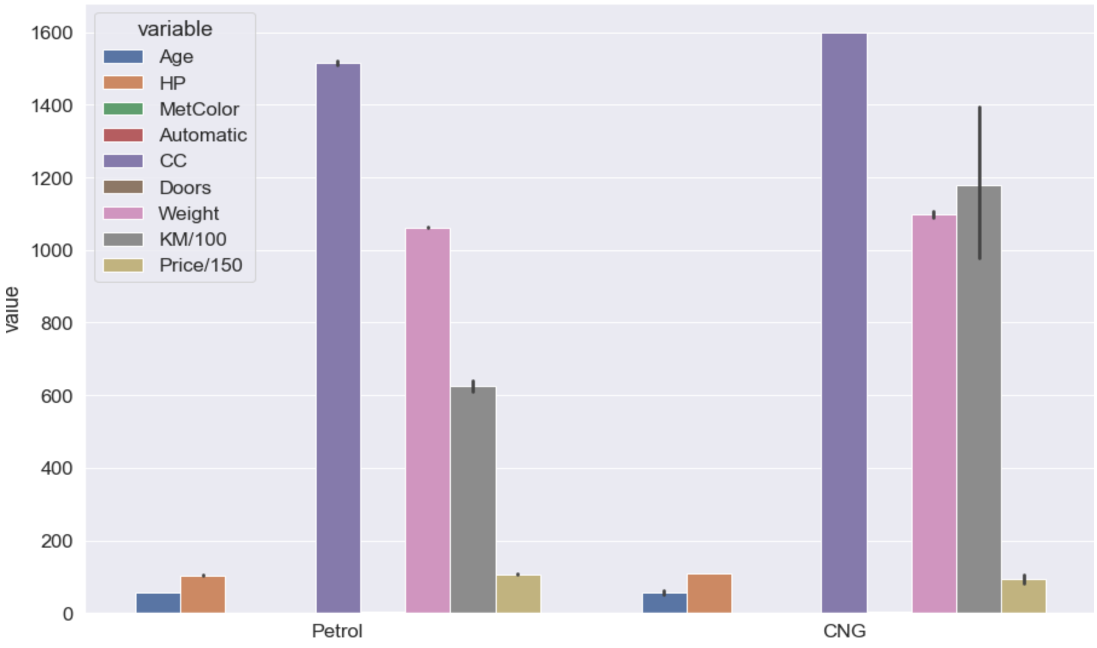
Program-2

Write a Python Program to Read 'Petrol' and 'CNG' FuelType sales data from Toyota.csv file and show it using the bar chart.

Procedure:

```
toyota_copy= toyota.copy()
toyota_copy['KM/100'] = toyota_copy['KM'].apply(lambda x: x/100)
toyota_copy['Price/150'] = toyota_copy['Price'].apply(lambda x: x/100)
toyota_copy.drop(['KM','Price'],axis=1,inplace=True)
data = toyota_copy.melt(id_vars='FuelType')
sns.barplot(x='FuelType', y='value', hue='variable', data=data[data['FuelType']!='Diesel'])
plt.gcf().set_size_inches(15, 10)
```

Output:



Program 3

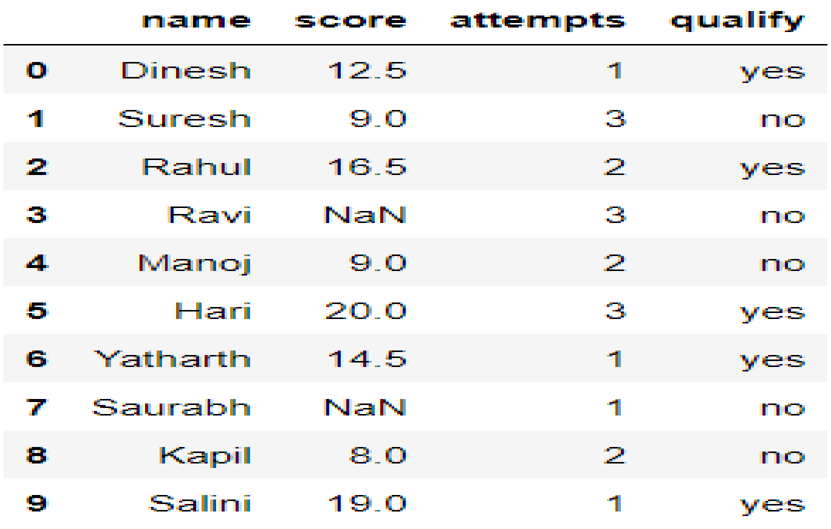
Write a Python program to create and display a DataFrame from a specified dictionary data which has the index labels.

```
exam_data = {'name': ['Dinesh', 'Suresh', 'Rahul', 'Ravi', 'Manoj', 'Hari', 'Yatharth', 'Saurabh', 'Kapil',
'Salini'],
            'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],
     'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],
     'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}
     labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']
```

Procedure:

```
import numpy as np
exam_data = {'name': ['Dinesh', 'Suresh', 'Rahul', 'Ravi', 'Manoj', 'Hari', 'Yatharth', 'Saurabh', 'Kapil',
'Salini'],
'score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],
'attempts': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],
'qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}
labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']
data=pd.DataFrame(exam_data)
print(data)
```

Output:



	name	score	attempts	qualify
0	Dinesh	12.5	1	yes
1	Suresh	9.0	3	no
2	Rahul	16.5	2	yes
3	Ravi	NaN	3	no
4	Manoj	9.0	2	no
5	Hari	20.0	3	yes
6	Yatharth	14.5	1	yes
7	Saurabh	NaN	1	no
8	Kapil	8.0	2	no
9	Salini	19.0	1	yes

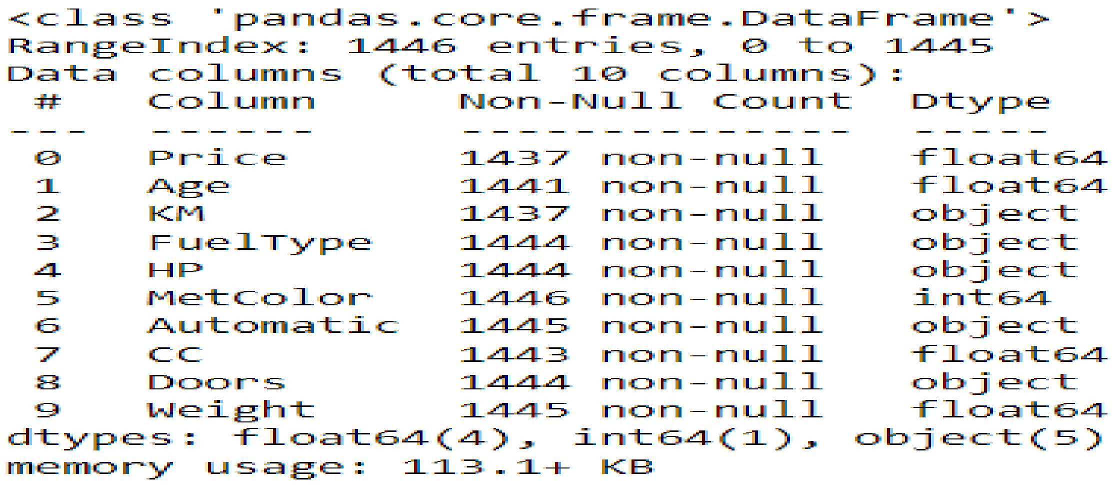
Program-4

Write a Python program to display a summary of the basic information like index, columns, non-null values of each column, memory usage etc. about a specified DataFrame which is Toyota.csv.

Procedure:

```
import pandas as pd
data=pd.read_csv('Toyota1.csv',na_values=['####','????'])
data.info()
```

Output:



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1446 entries, 0 to 1445
Data columns (total 10 columns):
 #    Column      Non-Null Count  Dtype
---   ------      --------------  -----
 0    Price       1437 non-null   float64
 1    Age         1441 non-null   float64
 2    KM          1437 non-null   object
 3    FuelType    1444 non-null   object
 4    HP          1444 non-null   object
 5    MetColor    1446 non-null   int64
 6    Automatic   1445 non-null   object
 7    CC          1443 non-null   float64
 8    Doors       1444 non-null   object
 9    Weight      1445 non-null   float64
dtypes: float64(4), int64(1), object(5)
memory usage: 113.1+ KB
```

‘

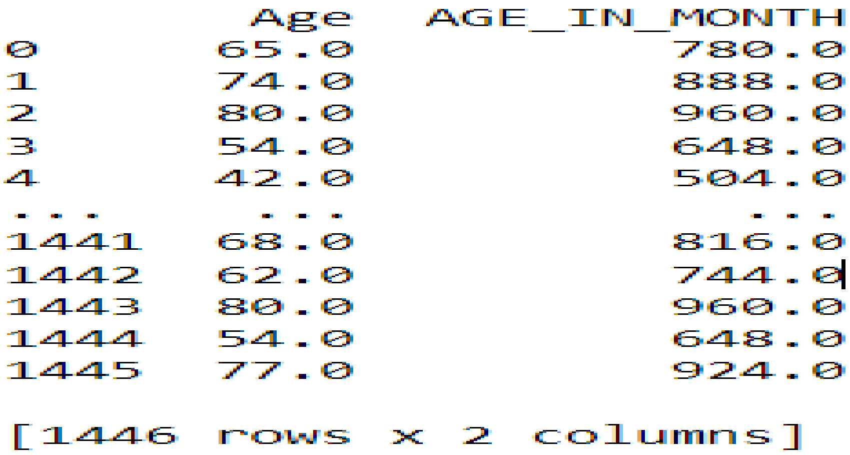
Program 5

WAP in Python to insert a column named “AGE_IN_MONTH” and then fill the values into inserted column on the basis of “AGE” ” column using user defined function where function return a value.
AGE_IN_MONTH=AGE*12

Procedure:

```
import pandas as pd
data=pd.read_csv('Toyota1.csv',na_values=['####','????'])
data.insert(2,'AGE_IN_MONTH',0)
def age_update(val):
    age_in_month=val*12
      return age_in_month
data['AGE_IN_MONTH']=age_update(data['Age'])
print(data[['Age','AGE_IN_MONTH']])
```

Output:



```
          Age  AGE_IN_MONTH
0        65.0         780.0
1        74.0         888.0
2        80.0         960.0
3        54.0         648.0
4        42.0         504.0
...       ...           ...
1441     68.0         816.0
1442     62.0         744.0
1443     80.0         960.0
1444     54.0         648.0
1445     77.0         924.0

[1446 rows x 2 columns]
```

Program 6

Write a program that accepts a sequence of words as input and write the words in a comma-separated sequence file after sorting them alphabetically.

Procedure:

```
wrd = input("Input words: ")
wrd_list = wrd.split(",")
wrd_list.sort()
print((', ').join(wrd_list))
```

Output:

```
Input words: ITM Gwalior
ITM Gwalior
```

Program-7

WAP in Python to copy line-by-line contents of a file into another file.

Procedure:

```
# open both files
with open('f1.txt','r') as f1file, open('f2.txt','a') as f2file:
        # read content from first file
        for line in f1file:
        # append content to second file
                F2file.write(line)
```

Program 8

WAP to enter characters one by one and then stored lower characters in the LOWER file and upper characters stored in the UPPER file and other characters stored in the OTHER file.

Procedure:

```
f1 = open(‘LOWER.txt’, ‘w’)
f2 = open(‘UPPER.txt’, ‘w’)
f3 = open(‘OTHERS.txt’, ‘w’)

c = True
while c:

    c = input(‘Enter a character to write or False to terminate the program : ‘)
        if c is False:
        break

    elif c.islower(): # checks for lower character
        f1.write(c)
        elif c.isupper(): # checks for upper character
        f2.write(c)

    else:
        f3.write(c)
```

Program 9

Explain and understand the steps to solve the following system of linear equations using MATLAB:

$3x-3y+2z=6$

$2x+2y+3z=12$

$-z+2x+2y=5$

Procedure:

```
% declaring the matrices based on the equations
A = [3 3 2; 2 2 3; -1 2 2]
b = [6; 12; 5]
% creating augmented matrix
Ab = [A b]
% checking the ranks
if rank(A) == rank(Ab)
   display("Unique solution exists")
else
   display("Unique solution does not exist")
end

%Now we can find the solution to this system of equations by using 3 methods:

%Conventional way : inv(A) * b
%Using mid-divide routine : A \ b
%Using linsolve routine : linsolve(A, b)

% conventional way of finding solution
x_inv = inv(A) * b
% using mid-divide routine of MATLAB
x_bslash = A \ b

% using linsolve routine of MATLAB
x_linsolve = linsolve(A, b)
```

Output:

```
x_inv =

   2.0000e+00

   8.8818e-16

  -3.0000e+00
```

```
x_bslash =
   2.0000e+00
   9.6892e-16
  -3.0000e+00
x_linsolve =
   2.0000e+00
   9.6892e-16
  -3.0000e+00
```

Program 10

The dataset Toyota.csv may be found on the web page. This dataset contains Car information: Price, Age, KM, FuelType, KM, CC and Doors. Save this file and use read.table to import it into R. What are the means and standard deviations of the data variables (excluding Age)? Apply Data Analytic tool.

Procedure:

```
#convert .csv file into dataframe
Data = read.table("data.csv", header= T, sep=";")
#To calculate the mean, standard deviation (Exclude Age)
Print('Price Mean:',mean(data$Price))
Print('Price Standard Deviation:',sd(data$Price))
Print('CC Mean:',mean(data$CC))
Print('CC Standard Deviation:',sd(data$CC))
```

Output:

```
Price Mean:10729.272094641614
Price Standard Deviation:3624.7192359440314
CC Mean:1566.988911988912
CC Standard Deviation:187.74178090826746
```