CHAPTER-1

INTRODUCTION

1.1 Introduction

K-Means falls in the general category of clustering algorithms. Clustering is a form of unsupervised learning that tries to find structures in the data without using any labels or target values. Clustering partitions a set of observations into separate groupings such that an observation in a given group is more similar to another observation in the same group than to another observation in a different group. UKgas datset includes Quarterly UK gas consumption from 1960Q1 to 1986Q4, in millions of therms. UKgas dataset consists of 108 observations of 3 variables.

The columns in this dataset are:

- X1
- Time
- Value

X1: This field describes the row number.

Time: This field consists of year values from 1960 to 1986, and every year is divided into four parts merging three months into a part.

Value: This field describes the value of gas consumption in UK with respective to time.

1.2 Problem Statement

A dataset is usually a rectangular array of data with rows representing observations and columns representing variables. Different traditions have different names for the rows and columns of a dataset. Statisticians refer to them as observations and variables, database analyst’s call them records and fields, and those from the data mining/machine learning disciplines call them examples and attributes. UKgas dataset contains 108 observations on 3 variables.

1.3 Scope of the Project

R has a wide variety of objects for holding data, including scalars, vectors, matrices, arrays, data frames, and lists. They differ in terms of the type of data they can hold, how they’re created, their structural complexity, and the notation used to identify and access individual elements.

CHAPTER-2

ANALYSIS

2.1 Requirements Gathering

The analysis phase can be broken into two phases:

- Data gathering
- Data analysis.

2.2 Software Requirement Specification

2.2.1Purpose

The purpose of this document is to analyze the UKgas dataset using K-means Clustering Algorithm. Clustering is a form of unsupervised learning that tries to find structures in the data without using any labels or target values. Clustering partitions a set of observations into separate groupings such that an observation in a given group is more similar to another observation in the same group than to another observation in a different group.

2.2.2 Scope

R has a wide variety of objects for holding data, including scalars, vectors, matrices, arrays, data frames, and lists. They differ in terms of the type of data they can hold, how they're created, their structural complexity, and the notation used to identify and access individual elements**.**

2.2.3Functional Requirements

Functional Requirement defines a function of a system or its component, where a function is described as a specification of behavior between outputs and inputs. It may involve calculations, technical details, data manipulation and processing, and other specific functionality that define what a system is supposed to accomplish. Functional requirements are supported by non-functional requirements. Functional requirements lead to data requirements (i.e., the information needed to perform the desired functions). According to Martin, “A Functional Requirement is a functional-level capability or business rule which is necessary to solve a problem or an objective”.

2.2.4Non-Functional Requirements

A Non-Functional requirement is a requirement that specifies criteria that can be used to judge the operation of a system, rather than specific behavior. Some of the typical non-functional requirements are: Performance, Scalability, Reliability, Maintainability and Security.

2.2.4.1Performance

Performance is the amount of work accomplished by a computer system. Depending on the context, high computer performance may involve one or more of the following:

- Short response time
- High throughput
- Low utilization of computing resources
- High bandwidth
- Short data transmission time

2.2.4.2Scalability

Scalability is the capability of a system, network or process to handle a growing amount of work, or its potential to be enlarged to accommodate that growth. For example, a system is considered scalable if it is capable of increasing its total output under an increased load when resources are added.

2.2.4.3Reliability

Reliability is defined as the probability of success. For the system to be more reliable the frequency of failures must be reduced.

2.2.4.4Maintainability

Maintainability is the ease with which a product can be maintained in order to:

- Correct defects
- Repair or replace faulty or worn-out components without having to replace still working parts
- Prevent unexpected working condition
- Maximize a product’s life

2.2.4.5Security

Security is freedom from potential harm from external forces. Beneficiaries of security may be persons and social groups, objects and institutions, ecosystems, and any persons other entity or phenomenon vulnerable to unwanted change by its environment.

2.2.5Technical Specifications

In this the Software and Hardware requirements are specified.

2.2.5.1 Hardware Configuration

• Processor	-	Pentium IV or above
• Speed	-	1.5 GHz
• RAM	-	512MB or above
• Hard Disk	-	10GB or above

2.2.5.2 Software Configuration

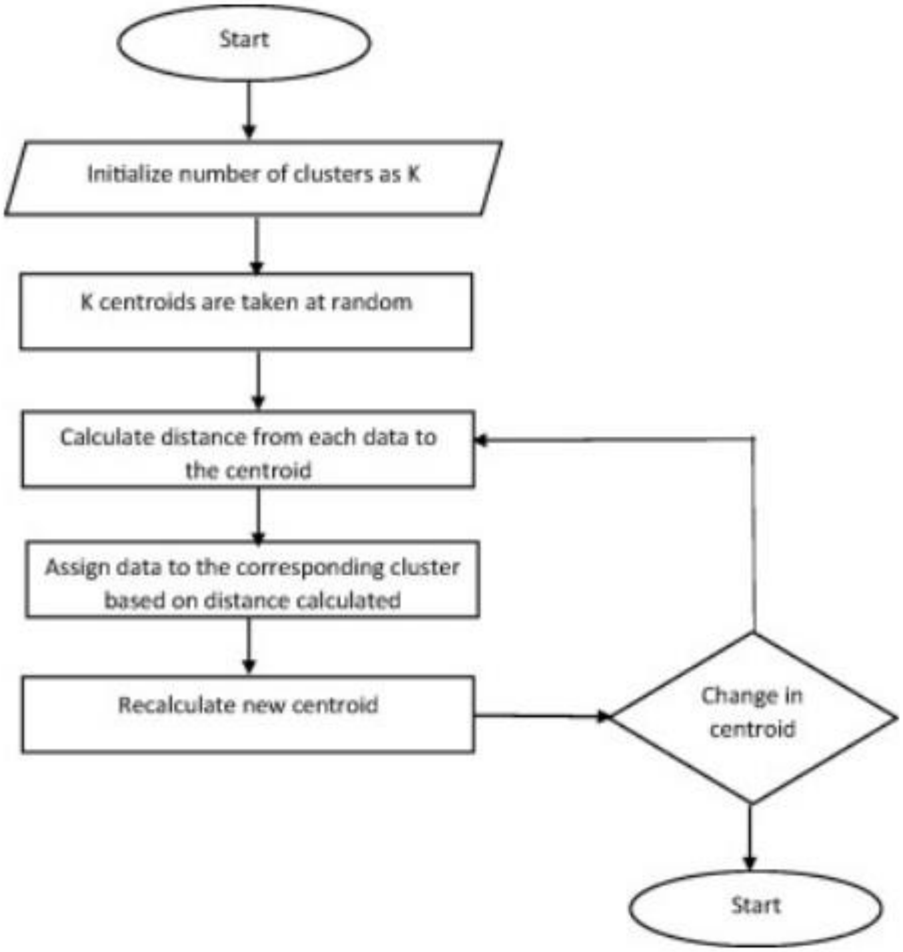
- Operating System - Windows 7
- Language - R language
- Software - R studio

CHAPTER-3

DESIGN

3.1 Data Flow Diagram

A flow diagram represents a set of dynamic relationships. It usually captures the physical or metaphorical flow of people, materials, communications, or objects through a set of nodes in a network.


Start
Initialize number of clusters as K
K centroids are taken at random
Calculate distance from each data to the centroid
Assign data to the corresponding cluster based on distance calculated
Recalculate new centroid
Change in centroid
Start

CHAPTER-4

IMPLEMENTATION

4.1. R Software Installation

To use R, first we need to install the R program on your computer.

4.1.1 How to check if R is installed on a Windows PC

- Before we install R on our computer, the first thing to do is to check whether R is already installed on your computer (for example, by a previous user).
- These instructions will focus on installing R on a Windows PC.
- If we are using a Windows PC, there are two ways you can check whether R is already installed on your computer:

1. Check if there is an "R" icon on the desktop of the computer that you are using. If so, double-click on the "R" icon to start R. If you cannot find an "R" icon, try step 2 instead.
2. Click on the "Start" menu at the bottom left of your Windows desktop, and then move your mouse over "All Programs" in the menu that pops up. See if "R" appears in the list of programs that pops up. If it does, it means that R is already installed on your computer, and you can start R by selecting "R" from the list.

If either (1) or (2) above does succeed in starting R, it means that R is already installed on the computer that you are using. (If neither succeeds, R is not installed yet). If there is an old version of R installed on the Windows PC that you are using, it is worth installing the latest version of R, to make sure that you have all the latest R functions available to you to use.

4.1.2 Finding out what is the latest version of R

To find out what is the latest version of R, you can look at the CRAN (Comprehensive R Network) website, http://cran.r-project.org/.

Beside “The latest release” (about half way down the page), it will say something like “R-X.X.X.tar.gz” (eg: “R-2.12.1.tar.gz”). This means that the latest release of R is X.X.X (for example, 2.12.1).

New releases of R are made very regular (approximately once a month), as R is actively being improved all the time. It is worthwhile installing new versions of R regularly, to make sure that you have a recent version of R (to ensure compatibility with all the latest versions of the R packages that you have downloaded).

4.1.3 R Installation Steps

To install R on your Windows computer, follow these steps:

1. Go to http://ftp.heanet.ie/mirrors/cran.r-project.org.
2. Under “Download and Install R”, click on the “Windows” link.
3. Under “Subdirectories”, click on the “base” link.
4. On the next page, you should see a link saying something like “Download R 2.10.1 for Windows” (or R X.X.X, where X.X.X gives the version of R, eg. R 2.11.1). Click on this link.
5. You may be asked if you want to save or run a file “R-2.10.1-win32.exe”. Choose “Save” and save the file on the Desktop. Then double-click on the icon for the file to run it.
6. You will be asked what language to install it in - choose English.
7. The R Setup Wizard will appear in a window. Click “Next” at the bottom of the R Setup wizard window.
8. The next page says “Information” at the top. Click “Next” again.
9. The next page says “Information” at the top. Click “Next” again.
10. The next page says “Select Destination Location” at the top. By default, it will suggest to install R in “C:\Program Files” on your computer.
11. Click “Next” at the bottom of the R Setup wizard window.
12. The next page says “Select components” at the top. Click “Next” again.
13. The next page says “Startup options” at the top. Click “Next” again.
14. The next page says “Select start menu folder” at the top. Click “Next” again.
15. The next page says “Select additional tasks” at the top. Click “Next” again.

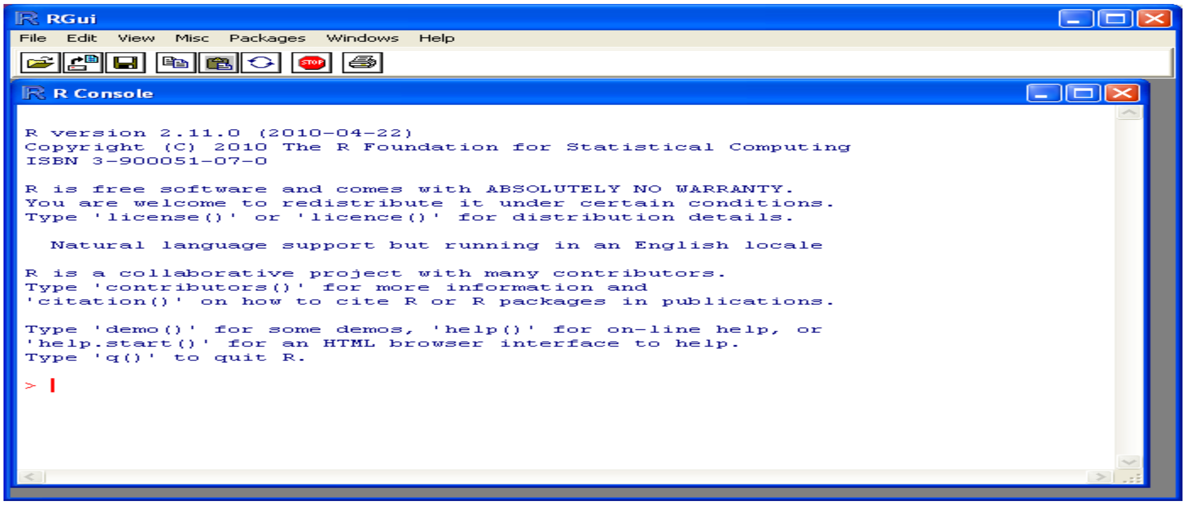
16. R should now be installed. This will take about a minute. When R has finished, you will see "Completing the R for Windows Setup Wizard" appear. Click "Finish".

17. To start R, you can either follow step 18, or 19:

18. Check if there is an "R" icon on the desktop of the computer that you are using. If so, double-click on the "R" icon to start R. If you cannot find an "R" icon, try step 19 instead.

19. Click on the "Start" button at the bottom left of your computer screen, and then choose "All programs", and start R by selecting "R" (or R X.X.X, where X.X.X gives the version of R, eg. R 2.10.0) from the menu of programs.

20. The R console (a rectangle) should pop up:



```
R version 2.11.0 (2010-04-22)
Copyright (C) 2010 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

4.2 R Studio Installation

- Go to www.rstudio.com and click on the "Download RStudio" button.
- Click on "Download RStudio Desktop".
- Click on the version recommended for your system, or the latest Windows version, and save the executable file. Run the .exe file and follow the installation instructions.

4.2.1 SDS Foundations Package Installation

- Download SDSFoundations to your desktop (make sure it has the ".zip" extension).
- Open RStudio.
- Click on the Packages tab in the bottom right window.

- Click "Install."
- Select install from "Package Archive File."
- Select the SDSFoundations package file from your desktop.
- Click install. You are done! You can now delete the SDSpackage file from your desktop.

4.2.2 How to work with R :

- First, we need to write our code regarding UKgas in the R console.
- We can use many packages and methods for executing our code related to UKgas.
- After writing our code we need to save our file as filename with the extension .r as we are using R software, we need to use .r as the extension to save our file in R studio.
- After saving our file successfully, we need to run our file by keeping the cursor at the starting of the line and press Ctrl + Enter in order to execute our code.
- So, after the execution of our code, we can see the output related to our UKgas dataset in the right bottom corner.
- The outputs are obtained in the different forms of graphs by using bar graphs, dotted graphs, box plots and so on.
- By using these outputs, we can have a clear analysis of the data about our selected dataset using this R software.

4.2.3 Code:

```
help("read.csv")

?read.csv

UKgas<-read.csv(file.choose(),header=T)
```

```
getwd()
dim(UKgas)
names(UKgas)
head(UKgas)
table(UKgas$time)
UKgas1<-UKgas
head(UKgas1)
UKgas1$time=NULL
head(UKgas1)
res<-kmeans(UKgas1,6)
res
plot(UKgas1,col=(res$cluster))
plot(UKgas1[c("value")],col=res$cluster)
plot(UKgas1[c("value")],col=UKgas1$time)
```

CHAPTER-5

TESTING

5.1 Purpose

The purpose of testing is to execute or evaluate programs or systems to measure the results against the requirements, to document the difference between the expected and the actual result and to assist in resolving those differences by providing the proper debug aids.

Testing purpose examples

- Uncovering defects and finding important problems
- Assessing quality and risk
- Certifying to standards
- Minimizing safety-related risks
- Minimizing technical support costs
- Maximizing efficiency
- Verifying correctness
- Assessing conformance to specifications or regulations

5.2 Black Box Testing

Black Box Testing, also known as Behavioral Testing, is a software testing method in which the internal structure or design or implementation of the item being tested is not known to the tester.

Black box test design technique is a procedure to derive or select testcases based on an analysis of the specification, either functional or non-functional, of a component or system without reference to its internal structure. This method is named so because the software program, in the eyes of the tester, is like a black box, inside which one cannot see.

5.2.1Testcase Design

Test Case ID	TC_OA01	**Test Case Description**	Number of Clusters for UKgas Dataset		
Created By	Phanitha	**Reviewed By**	Radha	Version	1.0
QA Tester's Log	Review comments from Quality Assurance Group				
Tester's Name	Mahesh	**Date Tested**	Feb 11,2019	Test Case (Pass/Fail/Not Executed)	Pass

S #	**Prerequisites:**
1	Various attributes to be Clustered

S #	Test Data
1	Clusters of data based on Value attribute

Test Scenario	K-means Clustering on UKgas Dataset

Step #	**Step Details**	**Expected Results**	Actual Results	Pass / Fail / Not executed / Suspended
1	Load and View the dataset	Viewing the dataset rows and attributes	As Expected	Pass
2	Preprocessing the dataset	Removing the null values in the dataset	As Expected	Pass
3	Applying K-means Clustering	Generating Clusters for the attributes	As Expected	Pass

5.3 White Box Testing

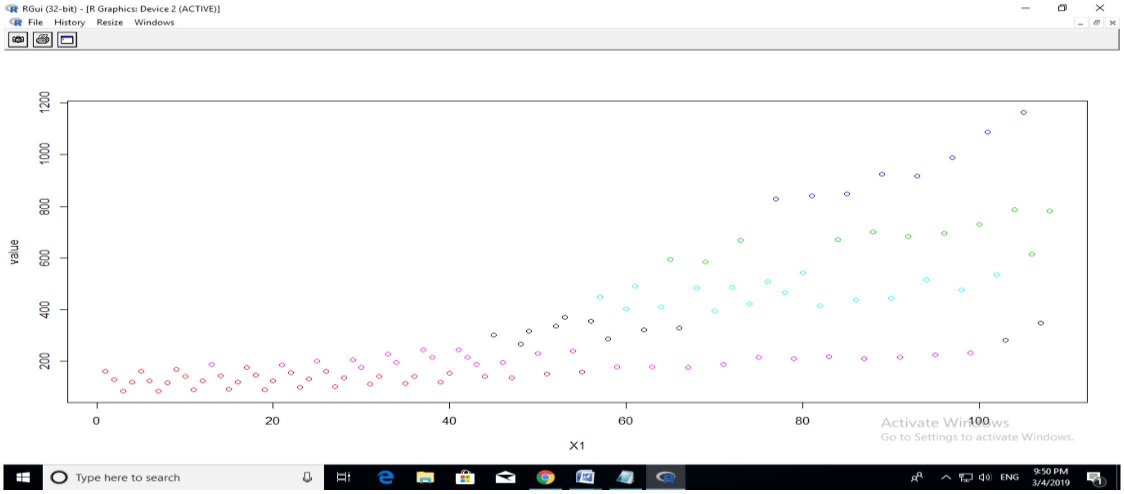
White-box testing is also known as clear box testing or transparent box testing or structural testing. It is a method of testing software that tests internal structures or workings of an application, as opposed to its functionality. White-box testing can be applied at the unit, integration and system levels of the software testing process. It can test paths within a unit, paths between units during integration, and between subsystem during a system-level test.

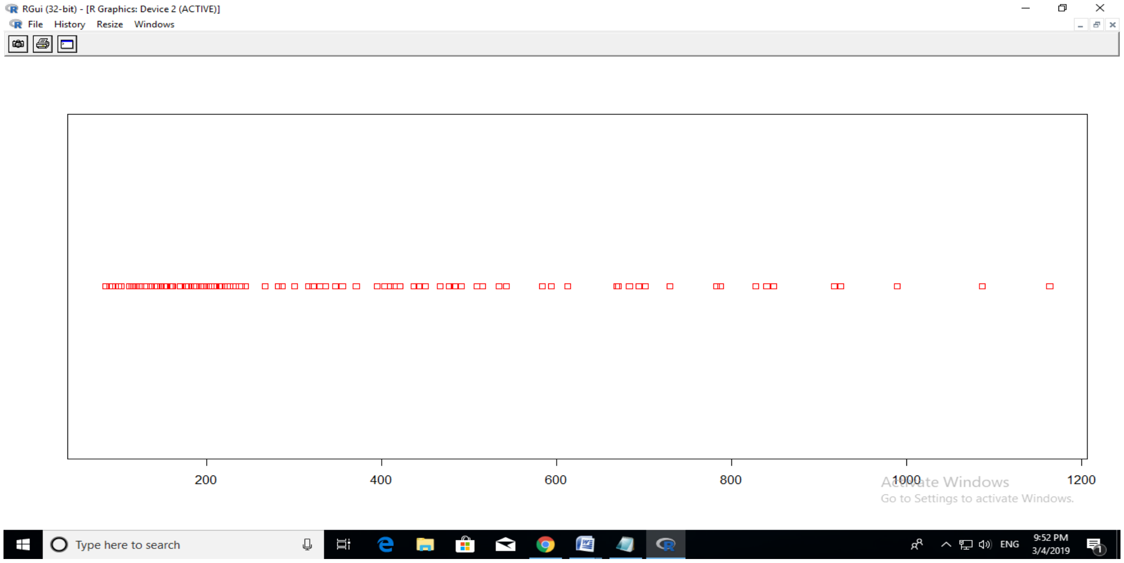
White-box test design techniques include the following code coverage criteria:

- Control flow testing
- Data flow testing
- Branch testing
- Statement coverage
- Decision coverage
- Path testing.

CHAPTER-6

SCREEN SHOTS:


RGui (32-bit) - [R Graphics: Device 2 (ACTIVE)]
File History Resize Windows
value
200
400
600
800
1000
1200
0
20
40
60
80
100
X1
Activate Windows
Go to Settings to activate Windows.
Type here to search
ENG
9:50 PM
3/4/2019


RGui (32-bit) - [R Graphics: Device 2 (ACTIVE)]
File History Resize Windows
200
400
600
800
1000
1200
Activate Windows
Go to Settings to activate Windows.
Type here to search
ENG
9:52 PM
3/4/2019

CHAPTER-7

CONCLUSION AND FUTURE SCOPE

7.1 Conclusion

We can conclude that the median of UKgas consumption is about 200.Most people would consume gas around 100 to 200. Our graph is skewed to the left that means UKgas consumption is less than its median. We can clearly conclude that people consume gas more and more as the time is more approaching to 21st century because the technology is improving year by year, people heavily rely on the transportation such as cars.

7.2 Future Scope

The results shows that the consumption of gas is directly proportional to the time. People consumed gas more and more as time approached to 21st century because of improved technology. This project helps government to analyze how much gas consumed in every year , in which year people consumed more gas and in which year people consumed less gas and how many resources are required to provide gas to people during particular years.

.

CHAPTER-8

BIBILOGRAPHY

Reference: [1] http://rpubs.com/K-means_UKgas

[2] http://www.ssfpack.com/dkbook/

[3] https://www.rstudio.com/