

19AIE214 BIG DATA ANALYSIS

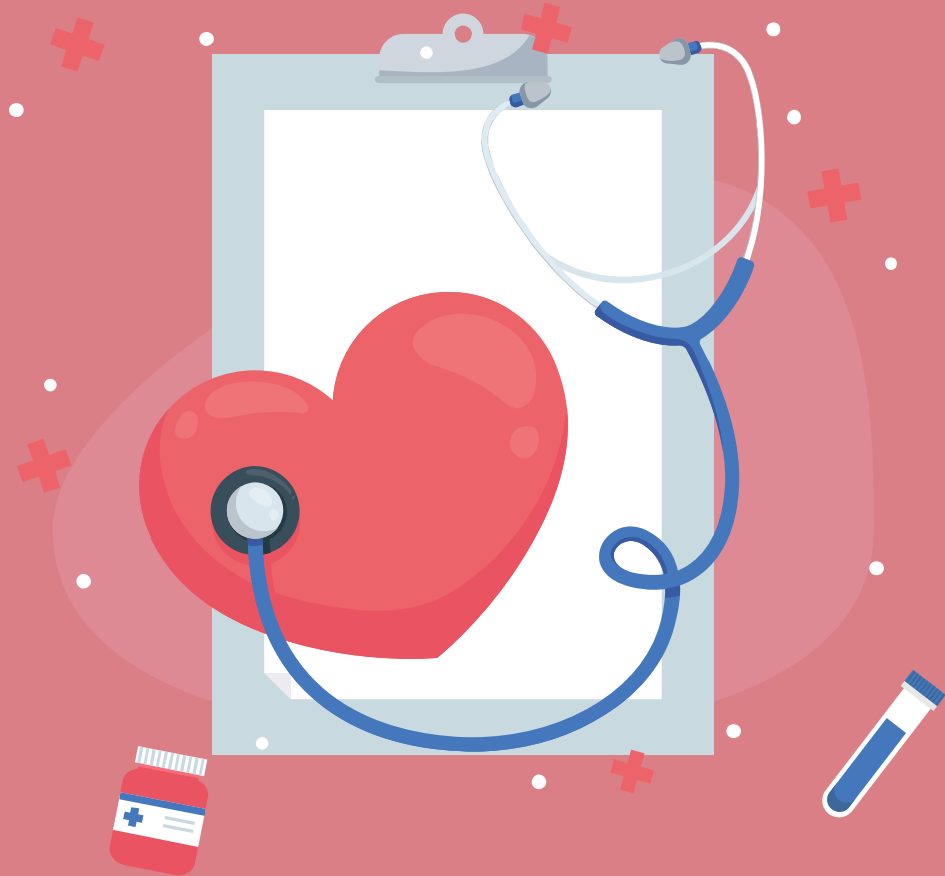


Table of contents

1

The Dataset

A small description of the dataset used.

2

Spark

Description of Spark.

3

Cluster

Creating a cluster.

4

Code

Brief explanation of the code.

5

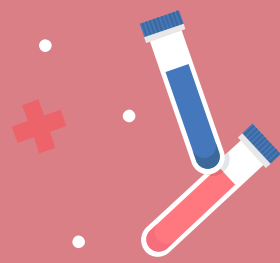
Analysis

Analysis done on the code with inference.

6

Conclusion

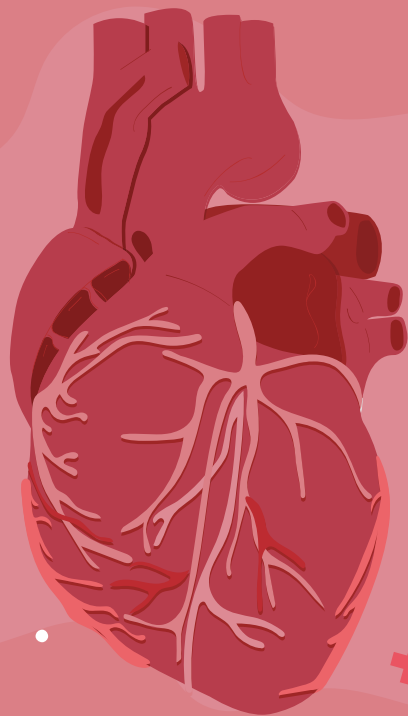
Ending.



01

ABOUT THE DATASET





Introduction to the dataset

The dataset used here is based on the series of heart diseases happened in the year 1988 covering the people of Cleveland, Hungary, Switzerland and Long Beach V.

Attribute Information

1. Age
2. Sex: 0 = Female; 1 = Male
3. Chest Pain (4 values)
4. Resting Blood Pressure
5. Serum Cholesterol in mg/dl
6. Fasting Blood Sugar > 120 mg/dl
7. Resting Electrocardiographic Results (values of 0,1,2)
8. Maximum Heart Rate Achieved
9. Exercise Induced Angina
10. OLDPEAK = ST depression induced by Exercise Relative to Rest
11. The slope of peak exercise ST segment
12. Number of Major Vessels (values 0,1,2,3) Colored By Fluoroscopy
13. Thal: 0 = normal; 1 = fixed defect; 2 = reversible defect
14. Target: 0 = no disease; 1 = disease confirmed

02 SPARK



ABOUT SPARK



It is a Data Management Framework and also big data framework

Hadoop and Spark differs mainly in the way they represent/ Abstract data

Basic data abstraction in spark is Distribute Sequence and Distributed file

We will be using spark with scala API

.



03

CLUSTER

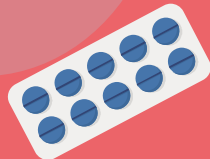




Starting Master

spark-class org.apache.spark.deploy.master.Master

```
C:\Users\Prasanth S N\spark-3.1.1\spark\bin>spark-class org.apache.spark.deploy.master.Master
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
21/05/16 18:51:41 INFO Master: Started daemon with process name: 14472@LAPTOP-8Q9QAENO
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/C:/Users/Prasanth%20S%20N/spark-3.1.1\spark\bin\spark-class.jar) of method java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
21/05/16 18:51:46 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java versions instead
21/05/16 18:51:46 INFO SecurityManager: Changing view acls to: Prasanth S N
21/05/16 18:51:46 INFO SecurityManager: Changing modify acls to: Prasanth S N
21/05/16 18:51:46 INFO SecurityManager: Changing view acls groups to:
21/05/16 18:51:46 INFO SecurityManager: Changing modify acls groups to:
21/05/16 18:51:46 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(); users with modify permissions: Set(Prasanth S N); groups with view permissions: Set(); groups with modify permissions: Set()
21/05/16 18:51:48 INFO Utils: Successfully started service 'sparkMaster' on port 7077.
21/05/16 18:51:48 INFO Master: Starting Spark master at spark://192.168.56.1:7077
21/05/16 18:51:48 INFO Master: Running Spark version 3.1.1
21/05/16 18:51:48 INFO Utils: Successfully started service 'MasterUI' on port 8080.
21/05/16 18:51:48 INFO MasterWebUI: Bound MasterWebUI to 0.0.0.0, and started at http://LAPTOP-8Q9QAENO:8080
21/05/16 18:51:48 INFO Master: I have been elected leader! New state: ALIVE
```





Creating Worker

spark-class org.apache.spark.deploy.worker.Worker <Master URL>

```
C:\Users\Prasanth S N\spark-3.1.1\spark\bin>spark-class org.apache.spark.deploy.worker.Worker spark://192.168.56.1:7077
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
21/05/16 19:56:53 INFO Worker: Started daemon with process name: 18224@LAPTOP-8Q9QAENO
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/C:/Users/Prasanth S N\spark-3.1.1\spark\bin\spark-unsafe-3.1.1.jar)
or java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
21/05/16 19:56:58 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java class
21/05/16 19:56:58 INFO SecurityManager: Changing view acls to: Prasanth S N
21/05/16 19:56:58 INFO SecurityManager: Changing modify acls to: Prasanth S N
21/05/16 19:56:58 INFO SecurityManager: Changing view acls groups to:
21/05/16 19:56:58 INFO SecurityManager: Changing modify acls groups to:
21/05/16 19:56:58 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(); users with modify permissions: Set(Prasanth S N); groups with modify permissions: Set()
21/05/16 19:56:59 INFO Utils: Successfully started service 'sparkWorker' on port 50879.
21/05/16 19:56:59 INFO Worker: Worker decommissioning not enabled, SIGPWR will result in exiting.
21/05/16 19:56:59 INFO Worker: Starting Spark worker 192.168.56.1:50879 with 16 cores, 30.4 GiB RAM
21/05/16 19:56:59 INFO Worker: Running Spark version 3.1.1
21/05/16 19:56:59 INFO Worker: Spark home: C:\Users\Prasanth S N\spark-3.1.1\spark
21/05/16 19:56:59 INFO ResourceUtils: =====
21/05/16 19:56:59 INFO ResourceUtils: No custom resources configured for spark.worker.
21/05/16 19:56:59 INFO ResourceUtils: =====
21/05/16 19:56:59 INFO Utils: Successfully started service 'WorkerUI' on port 8081.
21/05/16 19:57:00 INFO WorkerWebUI: Bound WorkerWebUI to 0.0.0.0, and started at http://LAPTOP-8Q9QAENO:8081
21/05/16 19:57:00 INFO Worker: Connecting to master 192.168.56.1:7077...
21/05/16 19:57:00 INFO TransportClientFactory: Successfully created connection to /192.168.56.1:7077 after 35 ms (0 ms spent)
21/05/16 19:57:00 INFO Worker: Successfully registered with master spark://192.168.56.1:7077
```



MASTER UI



Spark Master at spark://192.168.56.1:7077

URL: spark://192.168.56.1:7077

Alive Workers: 1

Cores in use: 16 Total, 0 Used

Memory in use: 30.4 GiB Total, 0.0 B Used

Resources in use:

Applications: 0 [Running](#), 0 [Completed](#)

Drivers: 0 Running, 0 Completed

Status: ALIVE

▼ Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20210516195659-192.168.56.1-50879	192.168.56.1:50879	ALIVE	16 (0 Used)	30.4 GiB (0.0 B Used)	

▼ Running Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

▼ Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------



```

  _ _ _ _ _
 / V \ _ V _ _ \ / _ \
/_ _ \ _ _ \ _ _ \ / _ \
/_ _ \ _ _ \ _ _ \ / _ \
/_ _ \ _ _ \ _ _ \ / _ \
version 3.1.1

```

```
scala>
```



MASTER UI



Spark Master at spark://192.168.56.1:7077

URL: spark://192.168.56.1:7077

Alive Workers: 1

Cores in use: 16 Total, 16 Used

Memory in use: 30.4 GiB Total, 1024.0 MiB Used

Resources in use:

Applications: 1 [Running](#), 0 [Completed](#)

Drivers: 0 Running, 0 Completed

Status: ALIVE

▼ Workers (1)

Worker Id	Address	State	Cores	Memory	Resources
worker-20210516195659-192.168.56.1-50879	192.168.56.1:50879	ALIVE	16 (16 Used)	30.4 GiB (1024.0 MiB Used)	

▼ Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20210516200215-0000 (kill)	Spark shell	16	1024.0 MiB		2021/05/16 20:02:15	Prasanth S N	RUNNING	55 s

▼ Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
----------------	------	-------	---------------------	------------------------	----------------	------	-------	----------

WORKER



Spark Worker at 192.168.56.1:50879

ID: worker-20210516195659-192.168.56.1-50879

Master URL: spark://192.168.56.1:7077

Cores: 16 (16 Used)

Memory: 30.4 GiB (1024.0 MiB Used)

Resources:

[Back to Master](#)

▼ Running Executors (1)

ExecutorID	State	Cores	Memory	Resources	Job Details	Logs
0	RUNNING	16	1024.0 MiB		ID: app-20210516200215-0000 Name: Spark shell User: Prasanth S N	stdout stderr

- Note: After 1st Analysis

JOBS

Spark Jobs (?)

User: Prasanth S N

Total Uptime: 58 min

Scheduling Mode: FIFO

Completed Jobs: 5

▶ Event Timeline

▼ Completed Jobs (5)

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

Job Id ▼	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
4	show at <console>:26 show at <console>:26	2021/05/16 21:00:06	0.1 s	2/2 (2 skipped)	<div>3/3 (4 skipped)</div>
3	sortBy at <console>:25 sortBy at <console>:25	2021/05/16 21:00:05	0.5 s	3/3	<div>6/6</div>
2	sortBy at <console>:25 sortBy at <console>:25	2021/05/16 21:00:04	74 ms	1/1	<div>2/2</div>
1	show at <console>:26 show at <console>:26	2021/05/16 21:00:03	0.7 s	1/1	<div>1/1</div>
0	first at <console>:25 first at <console>:25	2021/05/16 20:59:59	0.8 s	1/1	<div>1/1</div>

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

STAGES

Stages for All Jobs

Completed Stages: 8

Skipped Stages: 2

▼ Completed Stages (8)

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

Stage Id ▾	Description		Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
9	show at <console>:26	+ details	2021/05/16 21:00:06	61 ms	1/1			853.0 B	
8	sortBy at <console>:25	+ details	2021/05/16 21:00:06	42 ms	2/2			1574.0 B	1692.0 B
5	sortBy at <console>:25	+ details	2021/05/16 21:00:06	52 ms	2/2			1574.0 B	
4	map at <console>:25	+ details	2021/05/16 21:00:05	0.3 s	2/2			40.4 KiB	1574.0 B
3	sortBy at <console>:25	+ details	2021/05/16 21:00:05	91 ms	2/2	55.8 KiB			40.4 KiB
2	sortBy at <console>:25	+ details	2021/05/16 21:00:04	66 ms	2/2	55.8 KiB			
1	show at <console>:26	+ details	2021/05/16 21:00:03	0.6 s	1/1	37.2 KiB			
0	first at <console>:25	+ details	2021/05/16 20:59:59	0.8 s	1/1	37.2 KiB			

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

▼ Skipped Stages (2)

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

Stage Id ▾	Description		Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
7	map at <console>:25	+ details	Unknown	Unknown	0/2				
6	sortBy at <console>:25	+ details	Unknown	Unknown	0/2				

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

EXECUTORS

Executors

Show Additional Metrics

Summary

	▲ RDD Blocks	Storage Memory	On Heap Storage Memory	Off Heap Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(2)	0	54.2 KiB / 868.8 MiB	54.2 KiB / 868.8 MiB	0.0 B / 0.0 B	0.0 B	16	0	0	13	13	3 s (35.0 ms)	186.1 KiB	44.3 KiB	43.6 KiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B / 0.0 B	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(2)	0	54.2 KiB / 868.8 MiB	54.2 KiB / 868.8 MiB	0.0 B / 0.0 B	0.0 B	16	0	0	13	13	3 s (35.0 ms)	186.1 KiB	44.3 KiB	43.6 KiB	0

Executors

Show 20 entries

Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	On Heap Storage Memory	Off Heap Storage Memory	Peak JVM Memory OnHeap / OffHeap	Peak Execution Memory OnHeap / OffHeap	Peak Storage Memory OnHeap / OffHeap	Peak Pool Memory Direct / Mapped	Disk Used	Cores	Resources	Resource Profile Id	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs	Thread Dump
0	192.168.56.1:51126	Active	0	27.1 KiB / 434.4 MiB	27.1 KiB / 434.4 MiB	0.0 B / 0.0 B	0.0 B / 0.0 B	0.0 B / 0.0 B	0.0 B / 0.0 B	0.0 B / 0.0 B	0.0 B	16		0	0	0	13	13	3 s (35.0 ms)	186.1 KiB	44.3 KiB	43.6 KiB	stdout stderr	Thread Dump
driver	LAPTOP-8Q9QAENO:51055	Active	0	27.1 KiB / 434.4 MiB	27.1 KiB / 434.4 MiB	0.0 B / 0.0 B	222.4 MiB / 201.4 MiB	0.0 B / 0.0 B	265.6 KiB / 0.0 B	80.8 MiB / 0.0 B	0.0 B	0		0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B		Thread Dump

Showing 1 to 2 of 2 entries

Previous 1 Next

SQL



3.1.1

[Jobs](#)[Stages](#)[Storage](#)[Environment](#)[Executors](#)[SQL](#)

Spark shell application UI

SQL

Completed Queries: 2

▼ Completed Queries (2)

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

ID ▾	Description	Submitted	Duration	Job IDs
1	show at <console>:26 +details	2021/05/16 21:00:06	0.2 s	[4]
0	show at <console>:26 +details	2021/05/16 21:00:03	1 s	[1]

Page: 1

1 Pages. Jump to 1 . Show 100 items in a page. Go

04 CODE



Data Preprocessing

```
val df = sc.textFile("heart.csv");val header = df.first();
```

```
val data = df.filter(row => row!=header);
```

```
val sep = data.map{l=>
```

```
    val s0 = l.split(",")
```

```
    val
```

```
(age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal,target)=(s0(0).toInt,s0(1).toInt,  
s0(2).toInt,s0(3).toInt,s0(4).toInt,s0(5).toInt,s0(6).toInt,s0(7).toInt,s0(8).toInt,s0(9).toFloat,s0(10).toInt,s  
0(11).toInt,s0(12).toInt,s0(13).toInt)
```

```
(age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal,target))}
```

```
sep.toDF("age","sex","cp","trestbps","chol","fbs","restecg","thalach","exang","oldpeak","slope","ca","thal  
","target").show(false);
```



Age vs Target

```
val age_target =  
sep.sortBy(_._1).map{case (age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,ol  
dpeak,slope,ca,thal,target)=>(age,(target,1))};  
  
val age_target_sum = age_target.reduceByKey((x,y)=>(x._1+y._1,x._2+y._2));  
  
val age_target_avg =  
age_target_sum.map{case (a,b)=>(a,b._1/b._2.toFloat)}.sortBy(_._1);  
  
age_target_avg.toDF("Age","Affected Percent").show(false);
```





Sex vs Target

```
val sex_target =  
  sep.sortBy(_._1).map{case (age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,ol  
  dpeak,slope,ca,thal,target)=>(sex,(target,1))}  
  
val sex_target_sum = sex_target.reduceByKey((x,y)=>(x._1+y._1,x._2+y._2));  
  
val sex_target_avg =  
  sex_target_sum.map{case (a,b)=>(a,b._1/b._2.toFloat)}.sortBy(_._1);  
  
sex_target_avg.toDF("Sex","Affected Percent").show(false);
```





Max Heart Rate vs Target

```
val thal_target =  
  sep.sortBy(_._1).map{case (age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,ol  
  dpeak,slope,ca,thal,target)=>(thal,(target,1))};  
  
val thal_target_sum = thal_target.reduceByKey((x,y)=>(x._1+y._1,x._2+y._2));  
  
val thal_target_avg =  
  thal_target_sum.map{case (a,b)=>(a,b._1/b._2.toFloat)}.sortBy(_._1);  
  
thal_target_avg.toDF("Thal","Affected Percent").show(false);
```





Cholesterol vs Target

```
val chol_target =  
  sep.sortBy(_._1).map{case (age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,ol  
  dpeak,slope,ca,thal,target)=>(chol,(target,1))};  
  
val chol_target_sum = chol_target.reduceByKey((x,y)=>(x._1+y._1,x._2+y._2));  
  
val chol_target_avg =  
  chol_target_sum.map{case (a,b)=>(a,b._1/b._2.toFloat)}.sortBy(_._1);  
  
chol_target_avg.toDF("Cholesterol","Affected Percent").show(false);
```





Age and Sex vs Target

```
val ageSex_target = sep.sortBy(_._1).map(l=>((l._1,l._2),(l._14,1)))
```

```
val ageSex_target_reduced =  
ageSex_target.reduceByKey((x,y)=>((x._1+y._1),(x._2+y._2))).sortBy(_._1)
```

```
val ageSex_target_avg =  
ageSex_target_reduced.map{case(a,b)=>(a,b,((b._1).toFloat/(b._2).toFloat).toFloat)}
```

```
ageSex_target_avg.toDF("(Age,Sex)","(nAffected,nPersons)","Affected  
Percent").show(false)
```





Sex and Chestpain vs Target

```
val sexCP_target = sep.sortBy(_._1).map(l=>((l._2,l._3),(l._14,1)))
```

```
val sexCP_target_reduced =  
sexCP_target.reduceByKey((x,y)=>((x._1+y._1),(x._2+y._2))).sortBy(_._1)
```

```
val sexCP_target_avg =  
sexCP_target_reduced.map{case(a,b)=>(a,b,((b._1).toFloat/(b._2).toFloat).toFloat)}
```

```
sexCP_target_avg.toDF("(Sex,ChestPain)","(nAffected,nPersons)","Affected  
Percent").show(false)
```





Age and Chest Pain vs Target

```
val ageCP_target = sep.sortBy(_._1).map(l=>((l._1,l._3),(l._14,1)))
```

```
val ageCP_target_reduced =  
ageCP_target.reduceByKey((x,y)=>((x._1+y._1),(x._2+y._2))).sortBy(_._1)
```

```
val ageCP_target_avg =  
ageCP_target_reduced.map{case(a,b)=>(a,b,((b._1).toFloat/(b._2).toFloat).toFloat)}
```

```
ageCP_target_avg.toDF("(Age,ChestPain)","(nAffected,nPersons)","Affected  
Percent").show(false)
```



05

ANALYSIS



Gender VS Target



Gender	Affected Percentage	Not-Affected
0	0.724359	0.275641
1	0.42075735	0.57924265

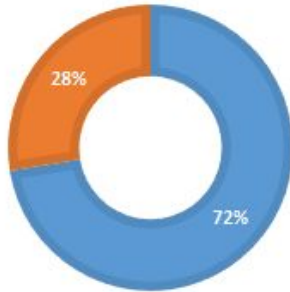


Analysis Done

OUTPUT

WOMEN

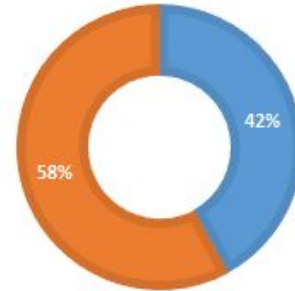
■ Affected ■ Not-Affected



OUTPUT

MEN

■ Affected ■ Not-Affected



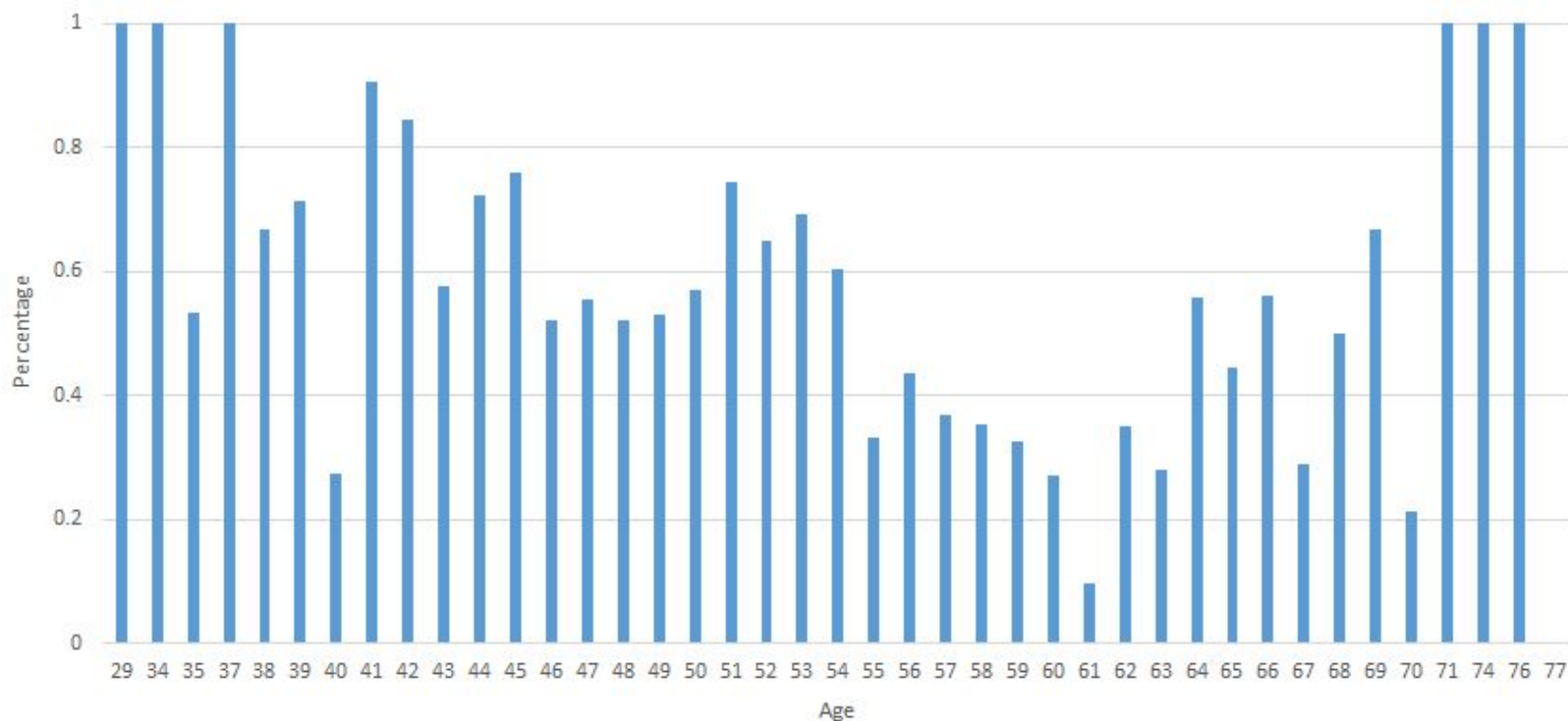
Age VS Target

Age	Percentage
29	1
34	1
35	0.533333
37	1
38	0.666667
39	0.714286
40	0.272727
41	0.90625
42	0.846154



Analysis Done

Percentage of affected based on Age

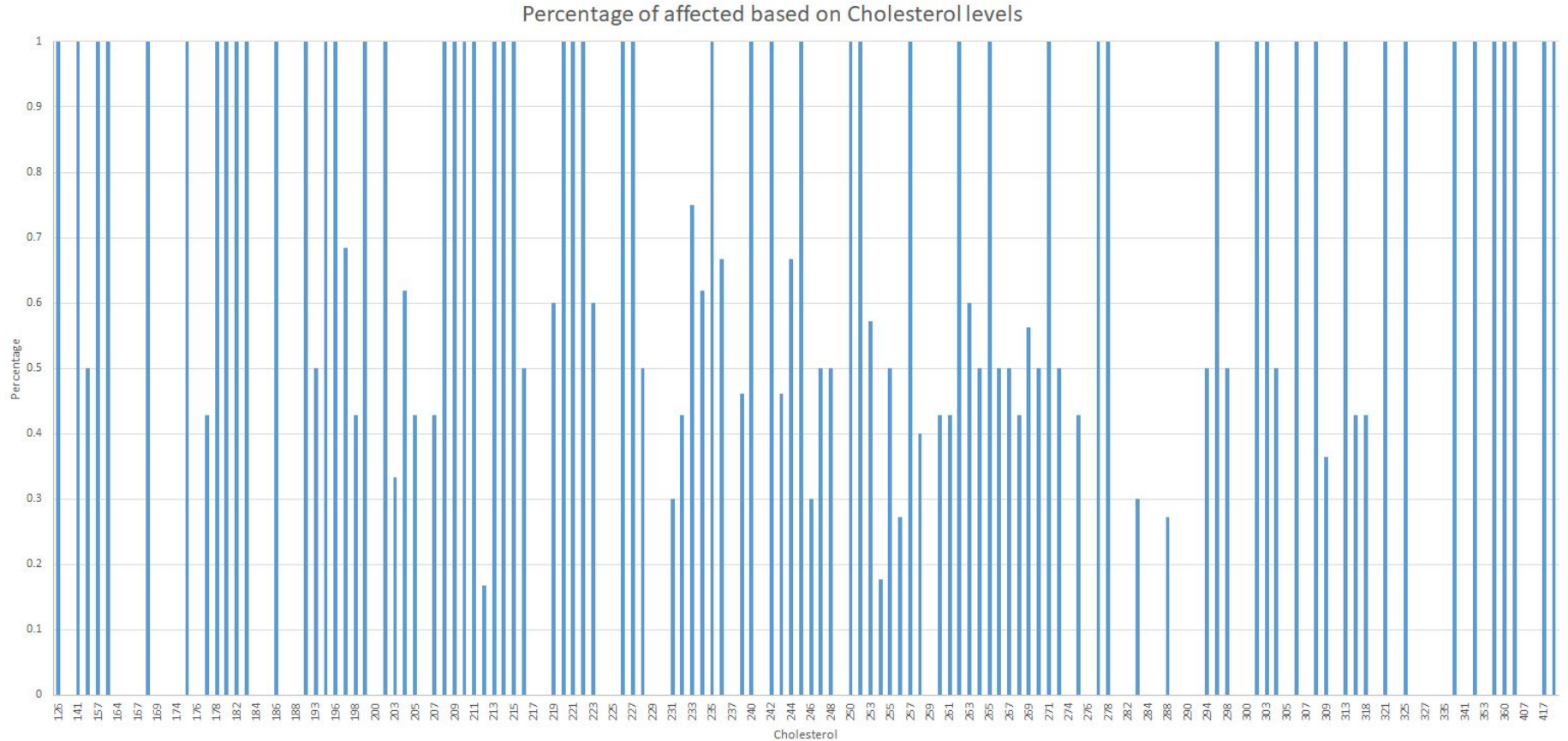


Cholesterol VS Target⁺

Cholesterol	Percentage
126	1
131	0
141	1
149	0.5
157	1
160	1
164	0
166	0
167	0



Analysis Done



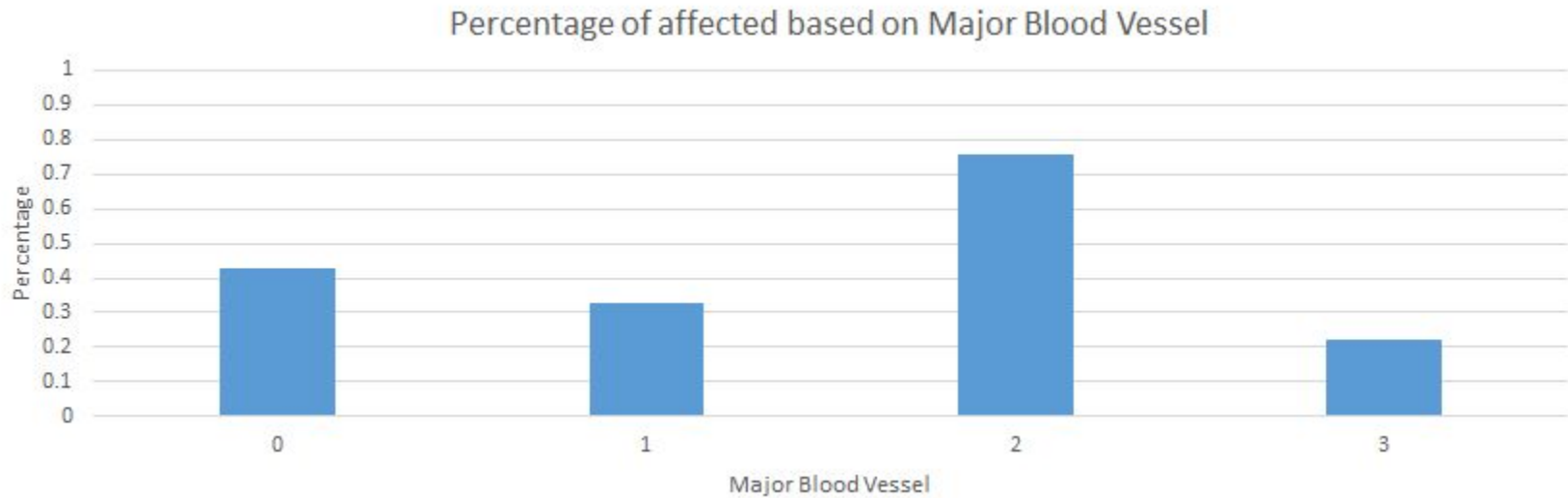
Major Blood Vessels VS Target



Major Vessels	Percentage
0	0.42857143
1	0.328125
2	0.75735295
3	0.2195122



Analysis Done



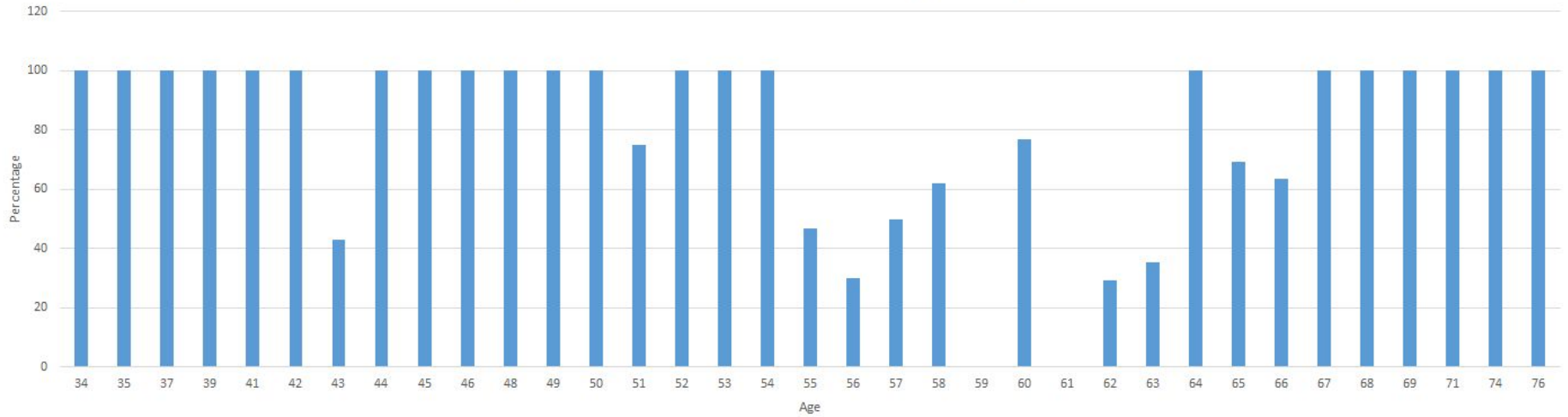
Age, Gender vs Target

Age	Gender	Affected	Total	Affected/total	Percentage
34	0	3	3	1	100
35	0	4	4	1	100
37	0	3	3	1	100
39	0	7	7	1	100
41	0	12	12	1	100
42	0	6	6	1	100
43	0	3	7	0.42857143	42.85714286
44	0	6	6	1	100
45	0	10	10	1	100

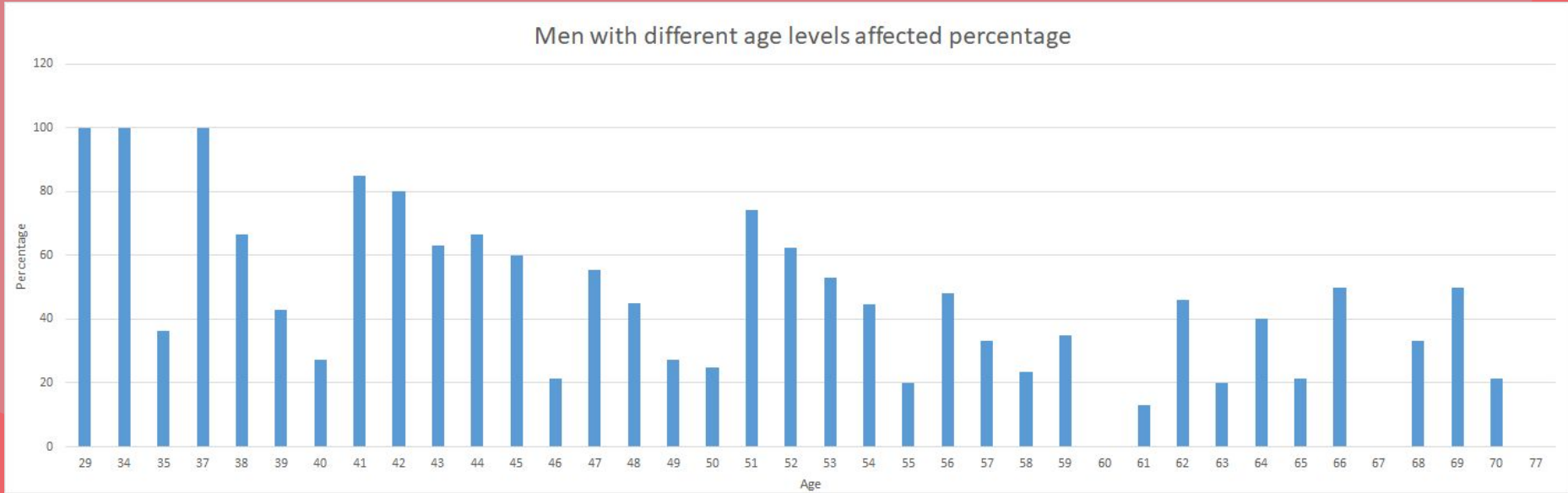


Analysis Done

Women in different age levels affected percentage



Analysis Done



Analysis Done

TOTAL	1025
Affected	526
Not Affected	499

Affected vs Not Affected



■ Affected ■ Not Affected



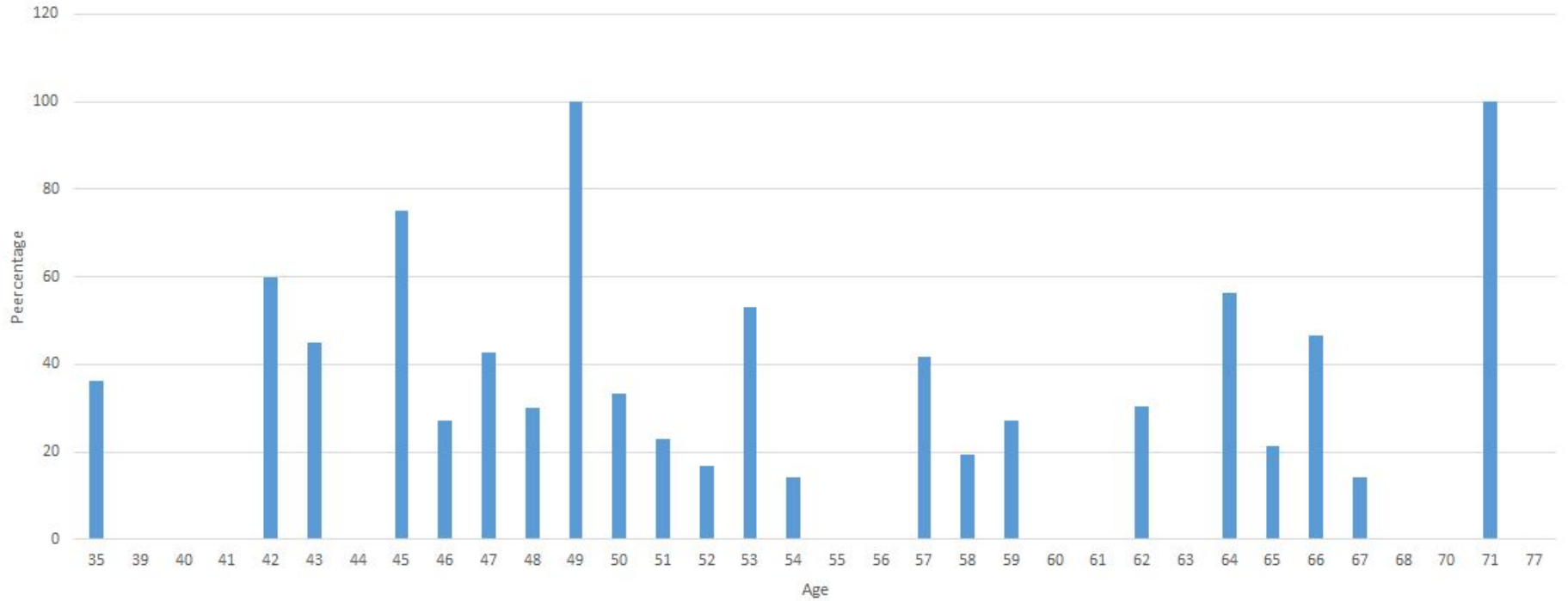
Age, Gender vs Target

Age	Chest Pain	Affected	Total	Affected/total	Percentage
35	0	4	11	0.36363637	36.36363636
39	0	0	4	0	0
40	0	0	8	0	0
41	0	0	3	0	0
42	0	6	10	0.6	60
43	0	9	20	0.45	45
44	0	0	10	0	0
45	0	9	12	0.75	75
46	0	3	11	0.27272728	27.27272727



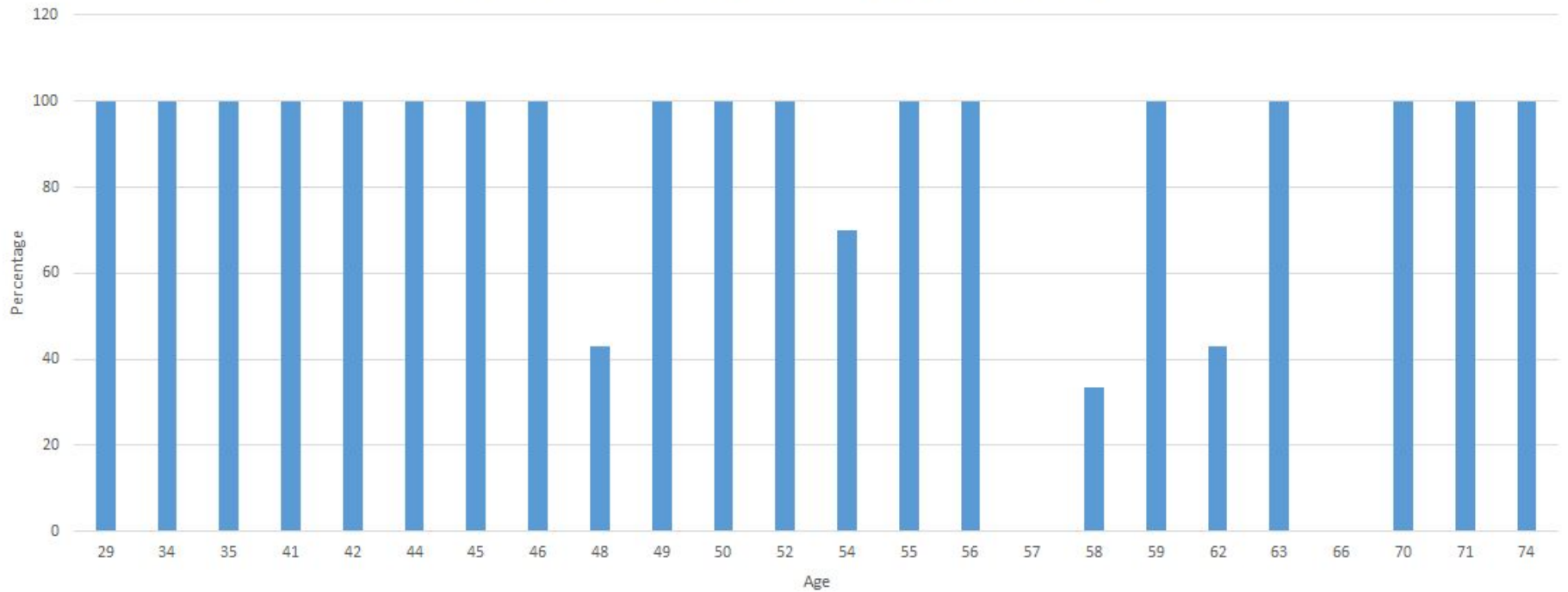
Analysis Done

Chest Pain when is it marked 0 affected people



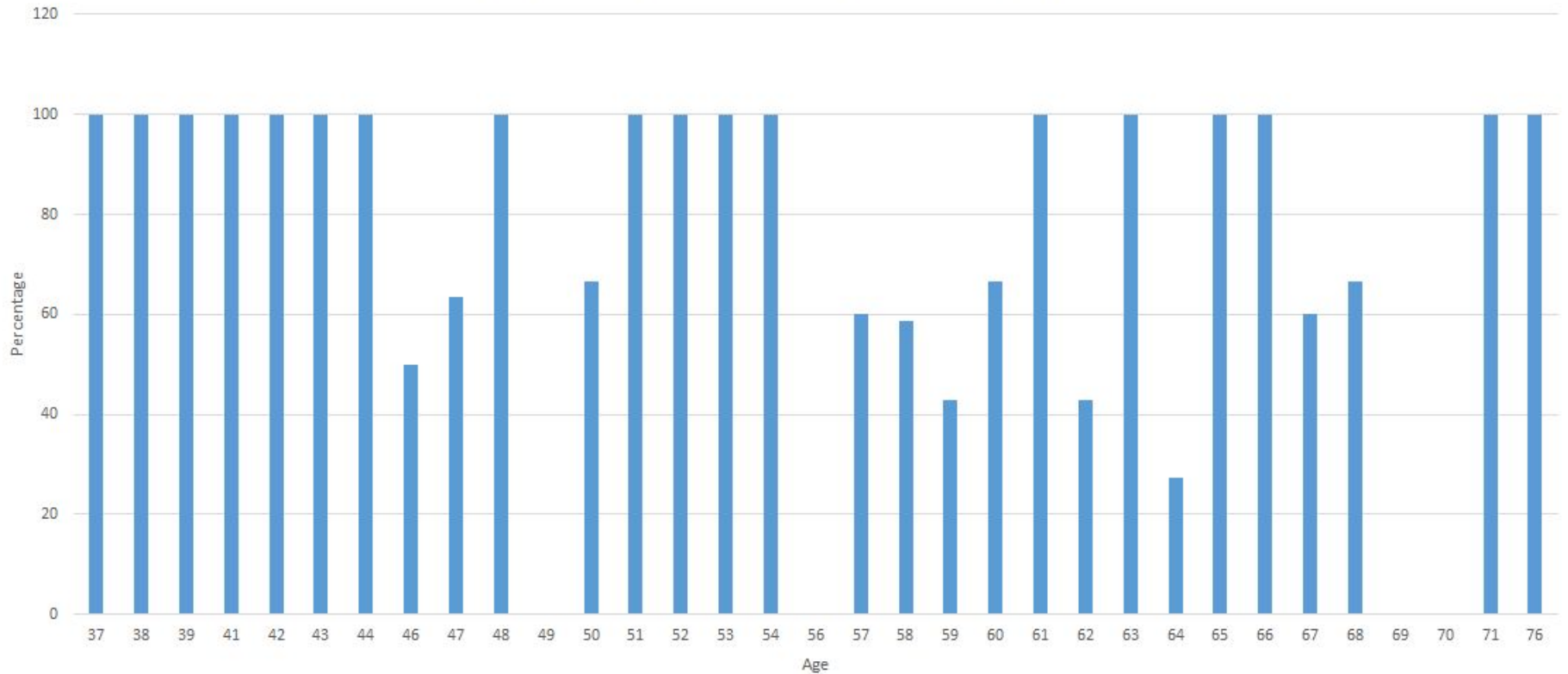
Analysis Done

Chest Pain when is it marked 1 affected people



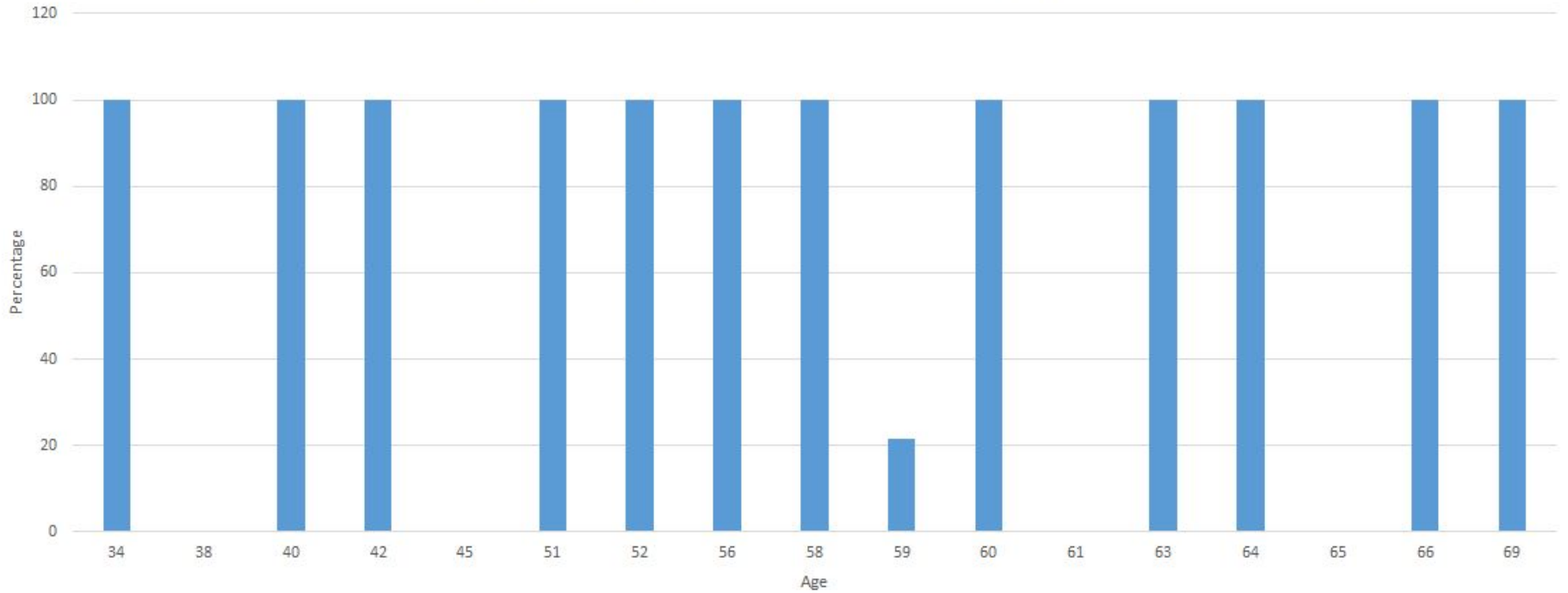
Analysis Done

Chest Pain when is it marked 2 affected people



Analysis Done

Chest Pain when is it marked 3 affected people



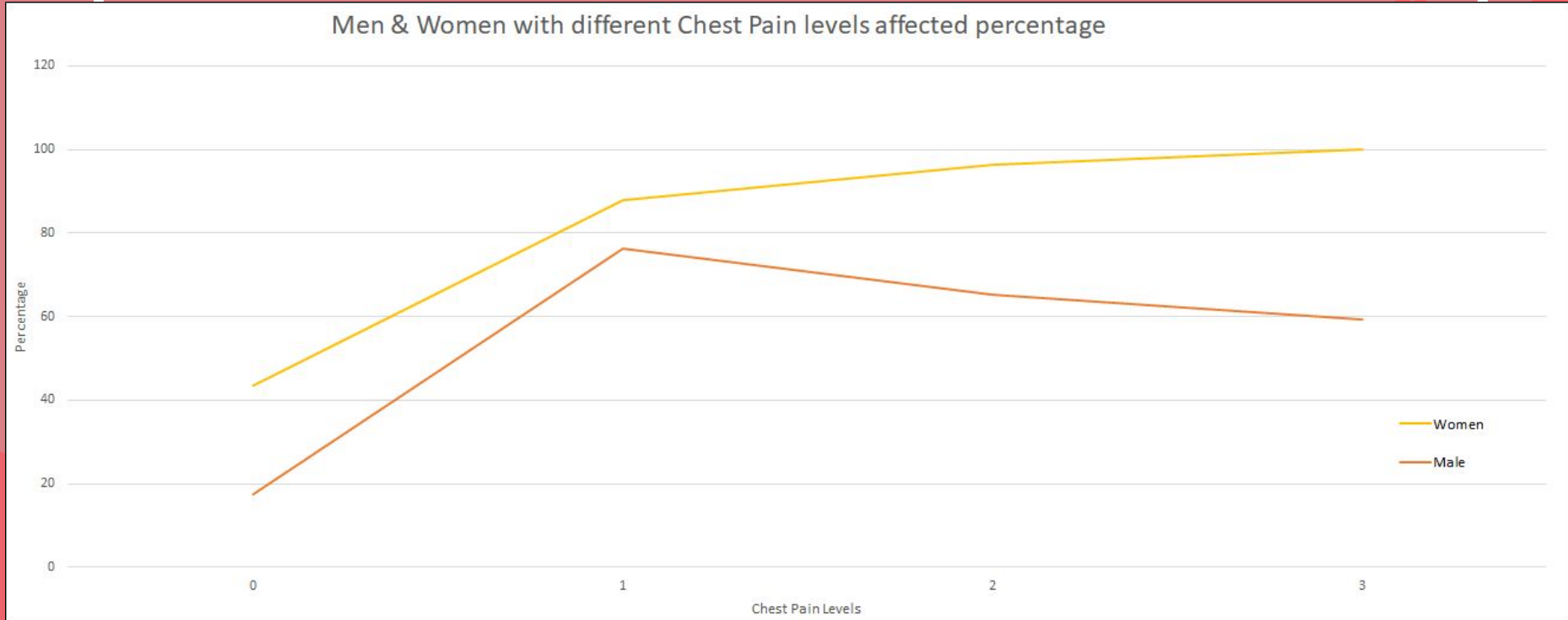


Gender, Chest Pain vs Target

Gender	Chest Pain	Affected	Total	Affected/ Total	Percentage
0	0	58	133	0.43609023	43.60902256
0	1	50	57	0.877193	87.71929825
0	2	105	109	0.96330273	96.33027523
0	3	13	13	1	100
1	0	64	364	0.17582418	17.58241758
1	1	84	110	0.76363635	76.36363636
1	2	114	175	0.6514286	65.14285714
1	3	38	64	0.59375	59.375



Analysis Done



06

Conclusion



Inference

1. Woman had a higher rate of getting heart diseases than men.
2. People with chest pain of '0' are least prone to heart disease than people with chest pain of other types.
3. People with Major Blood Vessel of '2' are highly prone to heart disease than people with Major Blood Vessel of other types.



THANKS!

