DATA SCIENCE MINOR PROJECT REPORT

CONTENTS OF THE REPORT

- Cover page
- Declaration
- Acknowledgement
- Table of Content

1. Introduction
2. Objectives/Scope of the Analysis
3. Source of dataset
4. ETL process
5. Analysis on dataset (for each analysis)
   i. Introduction
   ii. General Description
   iii. Specific Requirements, functions and formulas
   iv. Analysis results
   v. Visualization
6. List of Analysis with results
7. References
8. Bibliography

COVER PAGE

INTRODUCTION TO DATA MANAGEMENT PROJECT REPORT

(Project Semester August-December 2020)

NATIONAL NAMES

Submitted by

Pravallika

Registration No:11809836

Programme and Section B.tech CSE and KMO72

Course Code:INT 217

Under the Guidance of

Vasudha mam with U.Id:23036

Discipline of CSE/IT

Lovely School of Computer science and engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that Chidipudi.Pravallika bearing Registration no.11809836 has completed INT-217 project titled, **“National Names”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor
Designation of the Supervisor
School of Computer Science and engineering
Lovely Professional University
Phagwara, Punjab.

Date: 20-11-2020

DECLARATION

I, Chidipudi.Pravallika, student of Introduction to data management under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 20-11-2020 Signature

Registration No. 11809836 Pravallika

Acknowledgement

The success and final outcome of this project required a lot of guidance and assistance from my faculty. All that I have done is only due to such supervision and assistance and I would not forget to thank her.

I respect and thank my faculty, for providing me an opportunity to do the **Data management Project** and giving me all support and guidance which helped me to complete the project duty. Iam extremely thankful to her for providing such a good support and guidance.

Pravallika

INTRODUCTION:

I have undertaken a project titled “National Names”.

In this project I used various topics that I have learnt from this course.This product named “National Names” is about the names and their frequency and count of American children in different states of the country.

This project datasets also deals with most names in usage in U.S. between 1880 and present,with the count of each name given per year.Only names given to at least 5 babies in the same year are included in the datasets.The dataset consists of 5 columns and more than 1 lack rows.

General Description:

This project dataset consists of 5 columns namely:ID,Name,Year, Gender,State,Count. From that set I partitioned different sheets like: state wise count,Dashboard,Hyperlinking,Percentage of female and male,Top 20 repeated names,Year wise similar names,state names.

State wise count:It is about the number of children in each and every state.

Dashboard:It is a collection of all the graphs in state wise count,percentage of female and male,Top 20 repeated names,Year wise similar names.

Hyperlinking: It is a link that shows all the information of particular sheet of that repected link.

Percentage of female and male:It simply explains about the ratio of males and females in each and evry state.

Top 20 repeated names :It deals with the year-wise count of females and males .

Year wise similar names:It shows yearwise count of similar names of children.

Objectives/Scope of the Analysis:

The main objectives of my project are:

1. To know the names of U.S.children.
2. To know the replication value of same name.
3. Only names of 5 babies in the same year are counted.
4. Easy way to count the name for specific gender.

Source of dataset:

https://knowledge.domo.com/Training/Self-Service_Training/Onboarding_Resources/Fun_Sample_Datasets

ETL process:

What is ETL?

ETL is a process that extracts the data from different source systems, then transforms the data (like applying calculations, concatenations, etc.) and finally loads the data into the Data Warehouse system. Full form of ETL is Extract, Transform and Load.

It's tempting to think a creating a Data warehouse is simply extracting data from multiple sources and loading into database of a Data warehouse. This is far from the truth and requires a complex ETL process. The ETL process requires active inputs from various stakeholders including developers, analysts, testers, top executives and is technically challenging.

Step 1) Extraction

In this step, data is extracted from the source system into the staging area. Transformations if any are done in staging area so that performance of source system in not degraded. Also, if corrupted data is copied directly from the source into Data warehouse database, rollback will be a challenge. Staging area gives an opportunity to validate extracted data before it moves into the Data warehouse.

Step 2) Transformation

Data extracted from source server is raw and not usable in its original form. Therefore it needs to be cleansed, mapped and transformed. In fact, this is the key step where ETL process adds value and changes data such that insightful BI reports can be generated.

Step 3) Loading

Loading data into the target datawarehouse database is the last step of the ETL process. In a typical Data warehouse, huge volume of data needs to be loaded in a relatively short period (nights). Hence, load process should be optimized for performance.

Analysis on dataset (for each analysis):

Introduction:

I have undertaken a project titled “National Names”.

In this project I used various topics that I have learnt from this course.This product named “National Names” is about the names and their frequency and count of American children in different states of the country.

This project datasets also deals with most names in usage in U.S. between 1880 and present,with the count of each name given per year.Only names given to at least 5 babies in the same year are included in the datasets.The dataset consists of 5 columns and more than 1 lack rows.

General Description:

This project dataset consists of 5 columns namely:ID,Name,Year, Gender,State,Count. From that set I partitioned different sheets like: state wise count,Dashboard,Hyperlinking,Percentage of female and male,Top 20 repeated names,Year wise similar names,state names.

State wise count:It is about the number of children in each and every state.

Dashboard:It is a collection of all the graphs in state wise count,percentage of female and male,Top 20 repeated names,Year wise similar names.

Hyperlinking: It is a link that shows all the information of particular sheet of that repected link.

Percentage of female and male:It simply explains about the ratio of males and females in each and evry state.

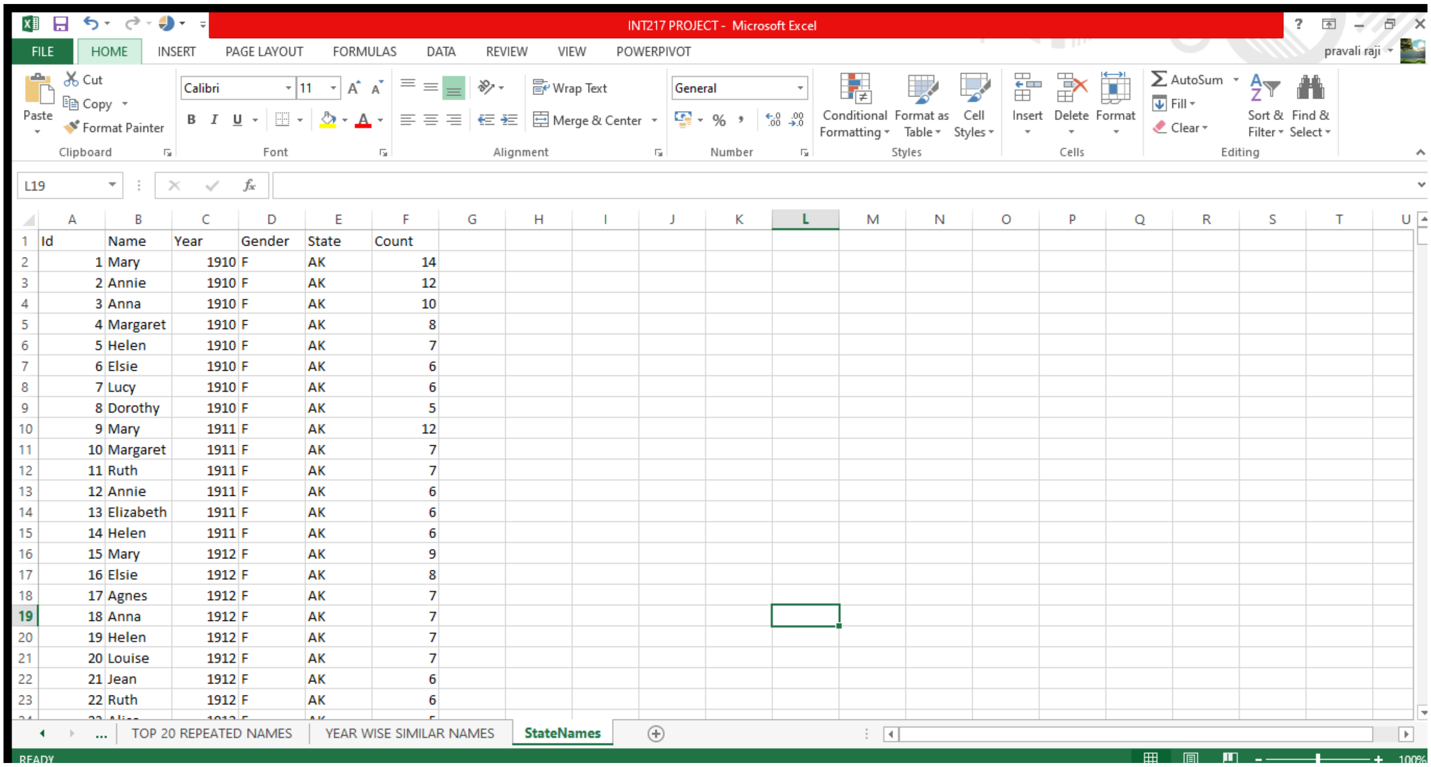
Top 20 repeated names :It deals with the year-wise count of females and males .

Year wise similar names:It shows yearwise count of similar names of children .

Specific Requirements, functions and formulas:

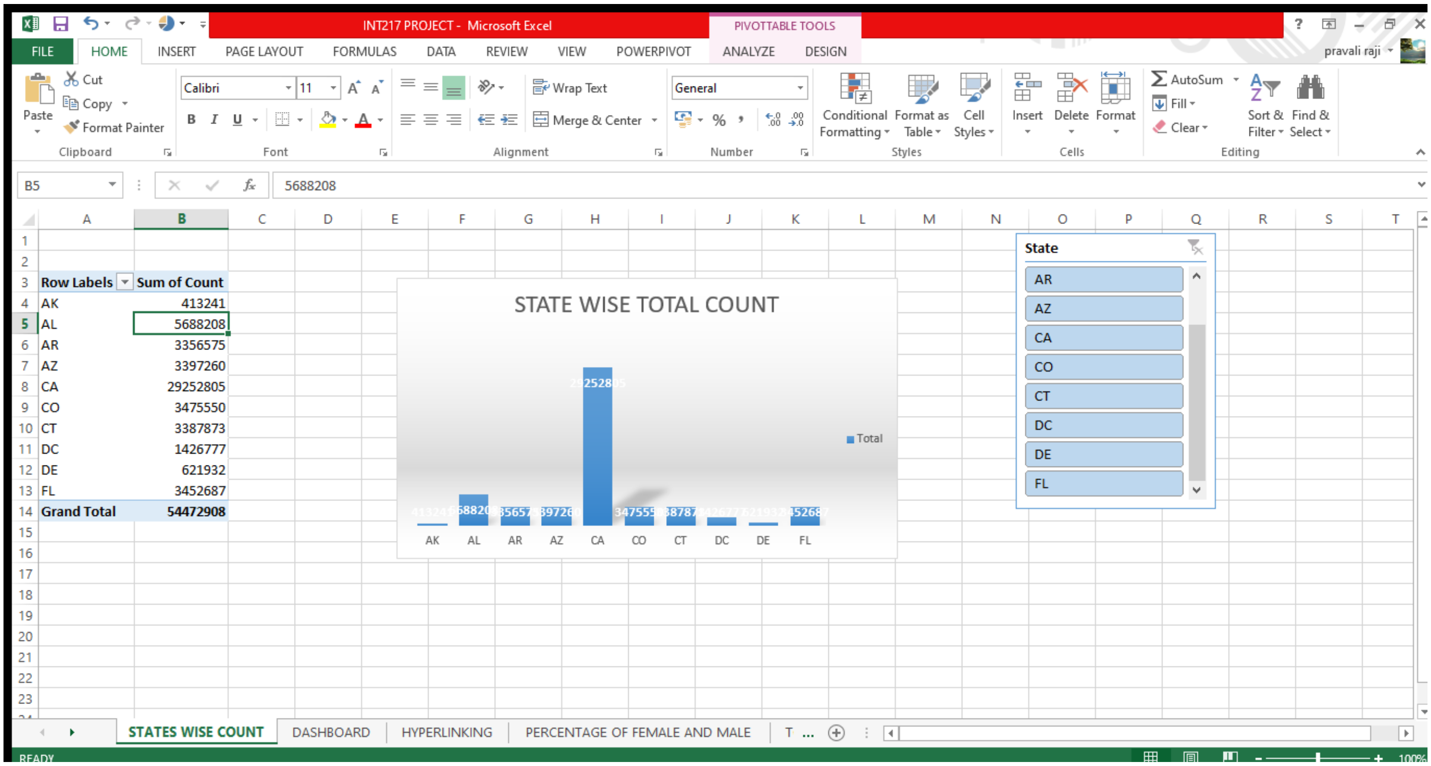
As there is no cleaning step in the data I selected I haven’t used Tableau in my project.I used different functions like filter,sort,pivot table,graphs and concepts like hyperlinking.slicer etc; for completion of my project.

Analysis results:



	A	B	C	D	E	F
1	Id	Name	Year	Gender	State	Count
2	1	Mary	1910	F	AK	14
3	2	Annie	1910	F	AK	12
4	3	Anna	1910	F	AK	10
5	4	Margaret	1910	F	AK	8
6	5	Helen	1910	F	AK	7
7	6	Elsie	1910	F	AK	6
8	7	Lucy	1910	F	AK	6
9	8	Dorothy	1910	F	AK	5
10	9	Mary	1911	F	AK	12
11	10	Margaret	1911	F	AK	7
12	11	Ruth	1911	F	AK	7
13	12	Annie	1911	F	AK	6
14	13	Elizabeth	1911	F	AK	6
15	14	Helen	1911	F	AK	6
16	15	Mary	1912	F	AK	9
17	16	Elsie	1912	F	AK	8
18	17	Agnes	1912	F	AK	7
19	18	Anna	1912	F	AK	7
20	19	Helen	1912	F	AK	7
21	20	Louise	1912	F	AK	7
22	21	Jean	1912	F	AK	6
23	22	Ruth	1912	F	AK	6

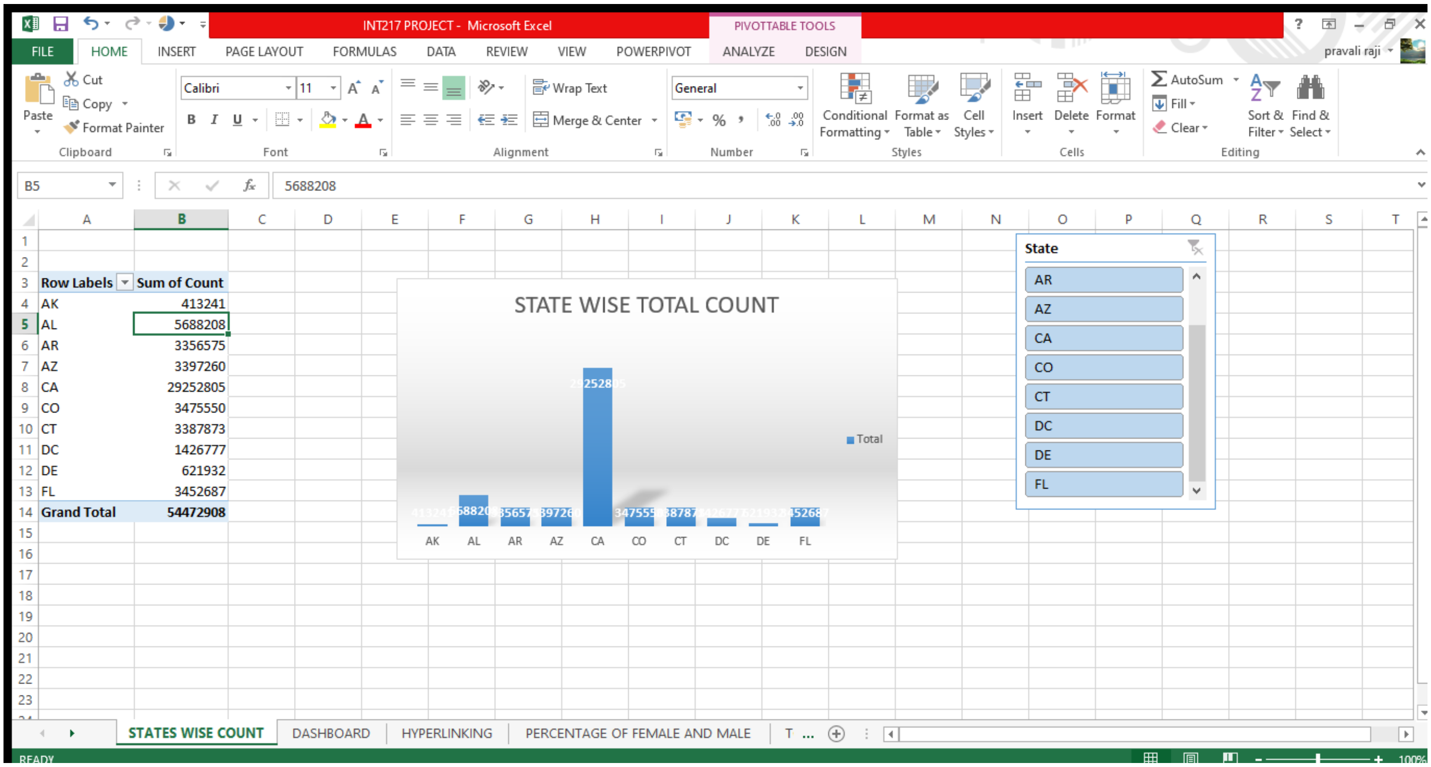
Fig-1.1)states dataset


INT217 PROJECT - Microsoft Excel
PIVOTTABLE TOOLS
FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW POWERPIVOT ANALYZE DESIGN
STATE WISE TOTAL COUNT
Total
AK AL AR AZ CA CO CT DC DE FL
State
AR
AZ
CA
CO
CT
DC
DE
FL
STATES WISE COUNT DASHBOARD HYPERLINKING PERCENTAGE OF FEMALE AND MALE
READY

B5 5688208

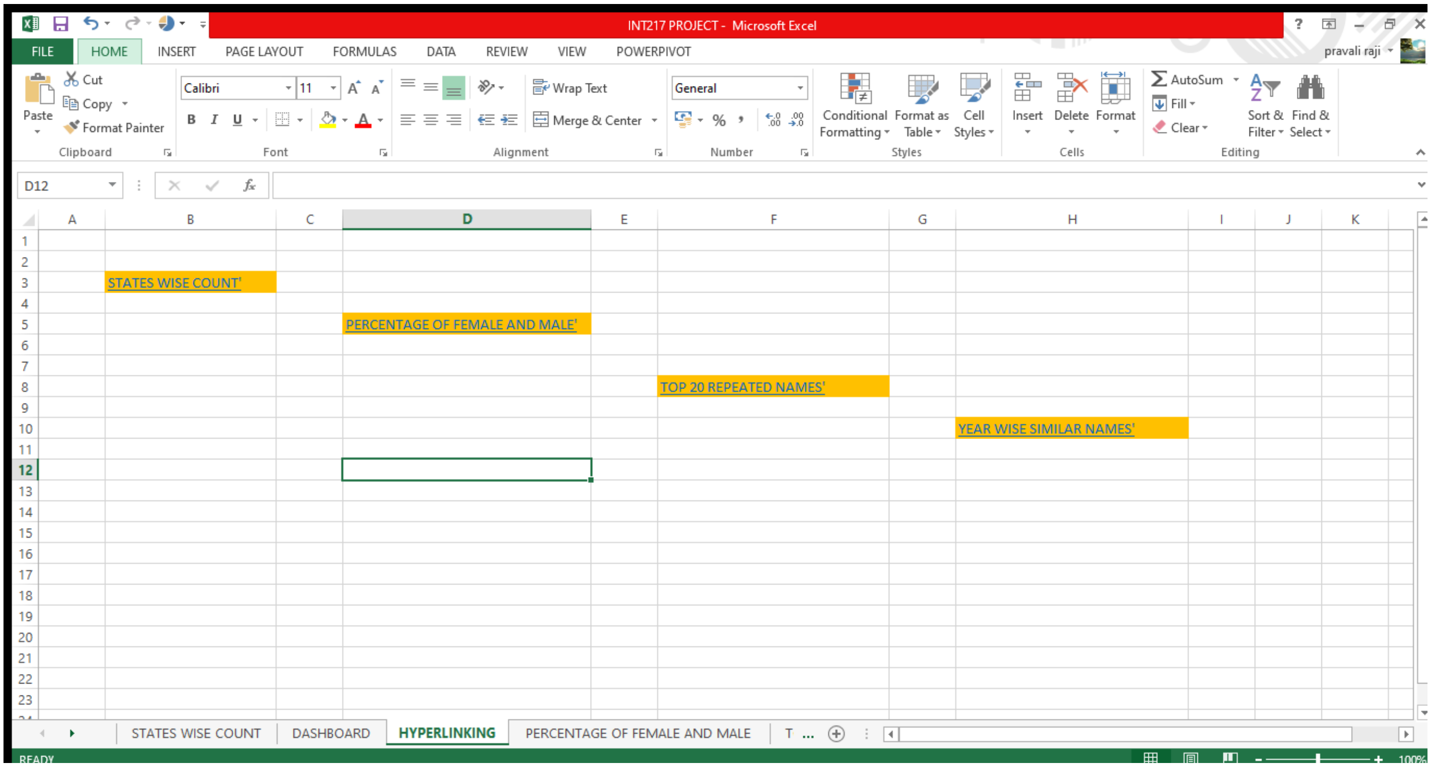
Row Labels	Sum of Count
AK	413241
AL	5688208
AR	3356575
AZ	3397260
CA	29252805
CO	3475550
CT	3387873
DC	1426777
DE	621932
FL	3452687
Grand Total	**54472908**

Fig:1.2)State-wise count


B5 5688208

Row Labels	Sum of Count
AK	413241
AL	5688208
AR	3356575
AZ	3397260
CA	29252805
CO	3475550
CT	3387873
DC	1426777
DE	621932
FL	3452687
Grand Total	**54472908**

Fig:1.3)Dashboard



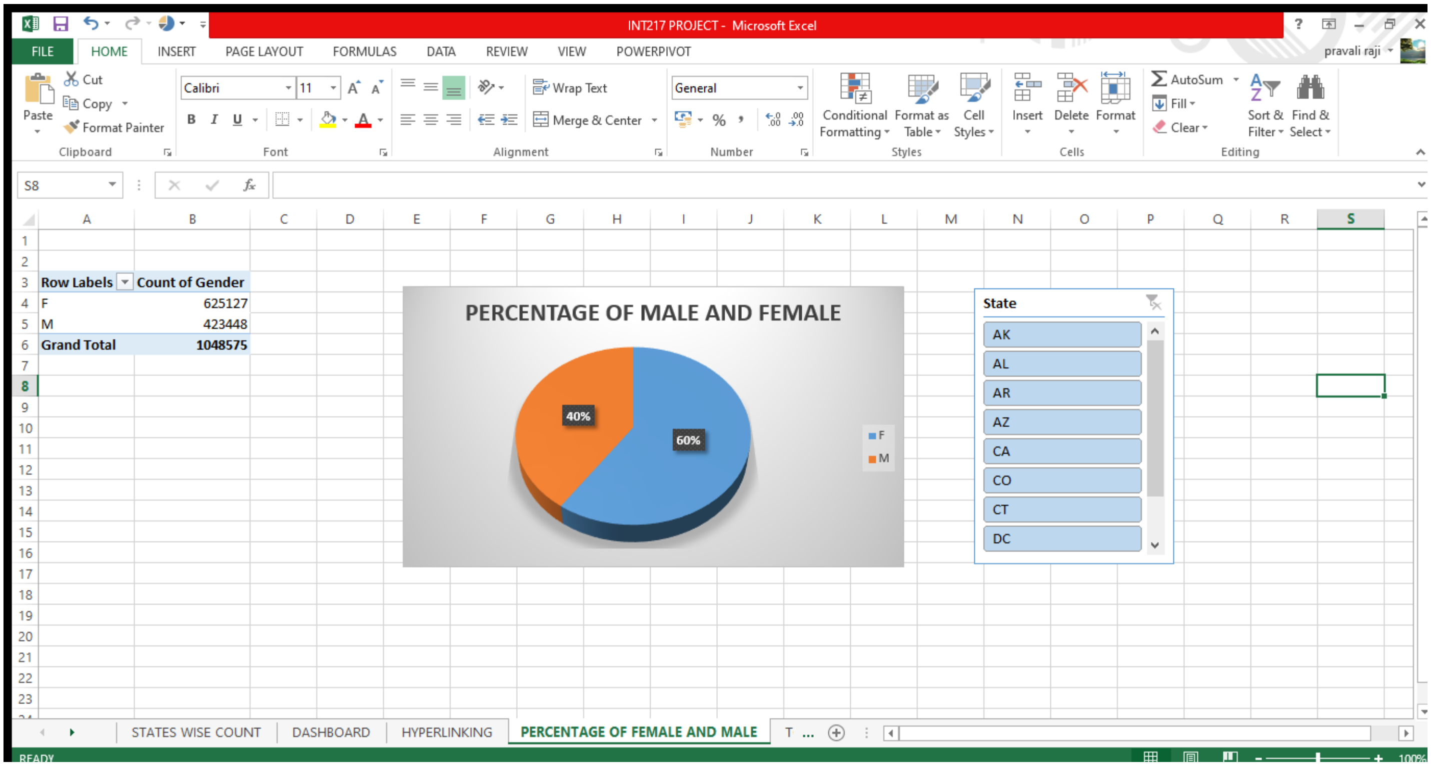
B3: STATES WISE COUNT'

D5: PERCENTAGE OF FEMALE AND MALE'

F8: TOP 20 REPEATED NAMES'

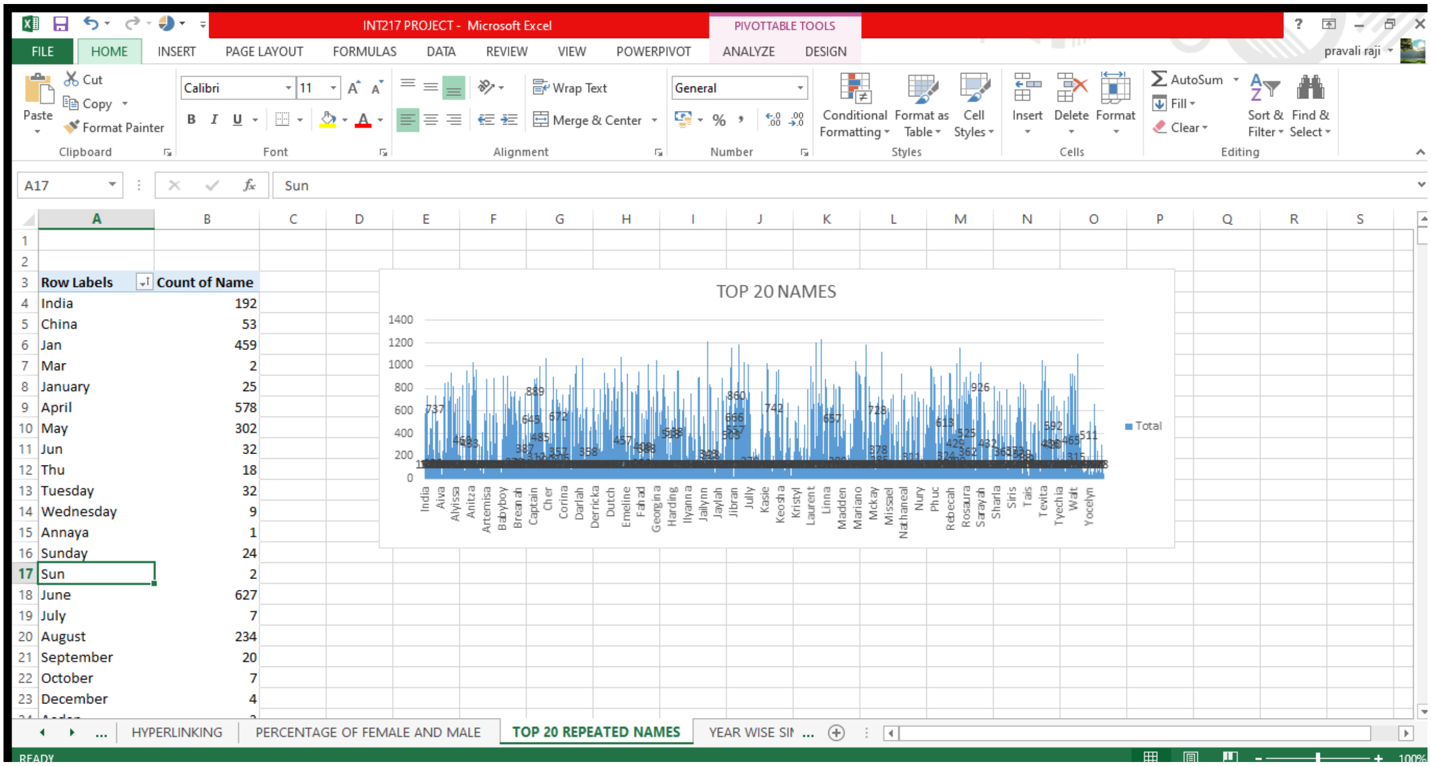
H10: YEAR WISE SIMILAR NAMES'

Fig:1.4)Hyperlinking


INT217 PROJECT - Microsoft Excel
FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW POWERPIVOT
S8
PERCENTAGE OF MALE AND FEMALE
40%
60%
F
M
State
AK
AL
AR
AZ
CA
CO
CT
DC
STATES WISE COUNT
DASHBOARD
HYPERLINKING
PERCENTAGE OF FEMALE AND MALE
READY

Row Labels	Count of Gender
F	625127
M	423448
Grand Total	1048575

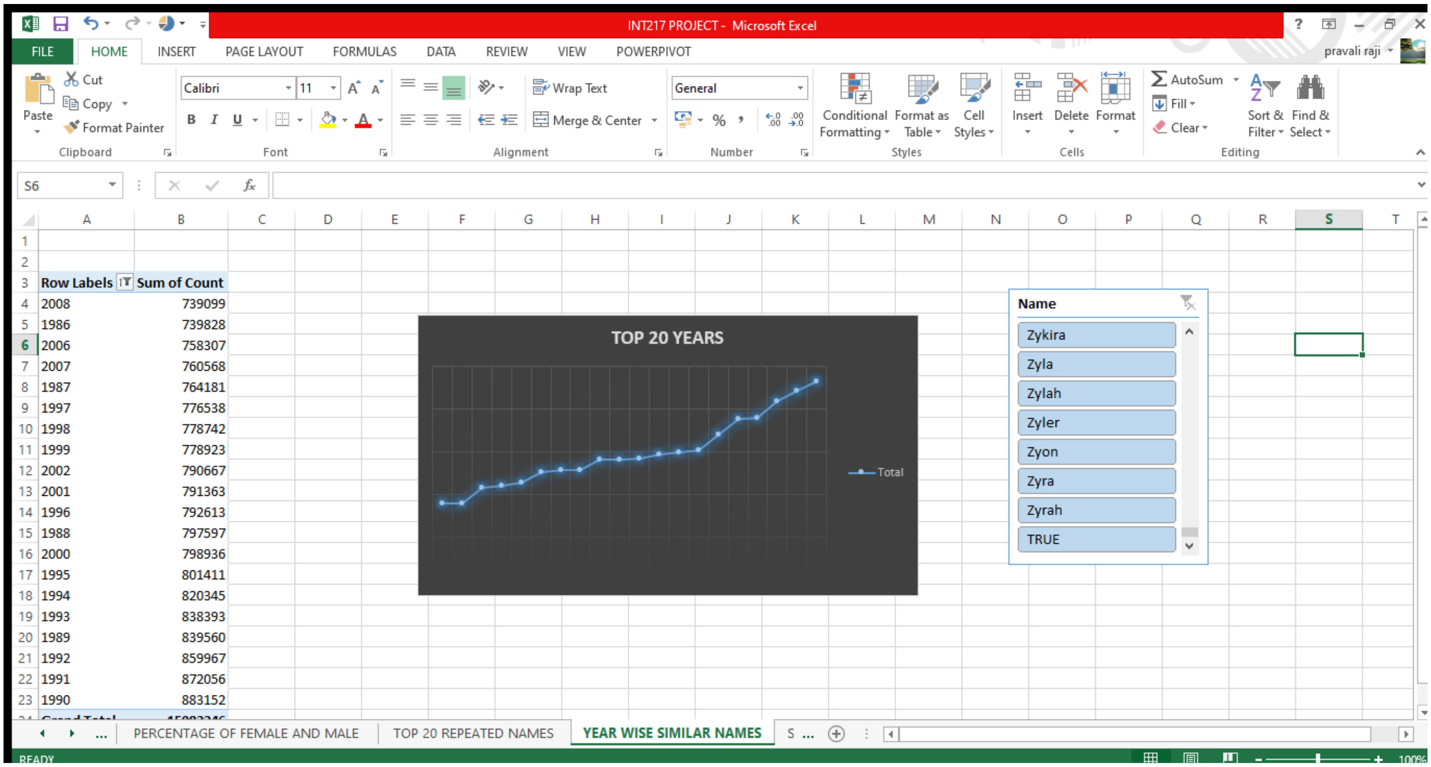
Fig:1.5)Percentage of female and male



Formula bar (A17): Sun

Row Labels	Count of Name
India	192
China	53
Jan	459
Mar	2
January	25
April	578
May	302
Jun	32
Thu	18
Tuesday	32
Wednesday	9
Annaya	1
Sunday	24
Sun	2
June	627
July	7
August	234
September	20
October	7
December	4

Fig-1.6)Top 20 Repeated Names



Row Labels	Sum of Count
2008	739099
1986	739828
2006	758307
2007	760568
1987	764181
1997	776538
1998	778742
1999	778923
2002	790667
2001	791363
1996	792613
1988	797597
2000	798936
1995	801411
1994	820345
1993	838393
1989	839560
1992	859967
1991	872056
1990	883152

Name
- Zykira
- Zyla
- Zylah
- Zyler
- Zyon
- Zyra
- Zyrah
- TRUE

Fig:1.7)Year wise similar names

References:

https://knowledge.domo.com/Training/Self-Service_Training/Onboarding_Resources/Fun_Sample_Datasets