Big-Data

Part-2: Coursework

Name: Prakash Dahal

Student Id: 1828421

Date: 2019/19/04

**Group Members:**

* Prakash Dahal
* Rajan Sapkota (Sharma)

Table of Contents

[A. Report 3](#_Toc6591059)

[1. Introduction to Big Data 3](#_Toc6591060)

[2. Evaluation of the Tools and Techniques 4](#_Toc6591061)

[a. Oracle and Excel 4](#_Toc6591062)

[b. MongoDB 4](#_Toc6591063)

[c. MapReduce Framework 5](#_Toc6591064)

[d. Hadoop 5](#_Toc6591065)

[3. Matrix 6](#_Toc6591066)

[4. Conclusions and recommendations 6](#_Toc6591067)

[B. Investigation 7](#_Toc6591068)

[a. Cleaning 7](#_Toc6591069)

[b. Manipulation: 15](#_Toc6591070)

[c. Analysis of the data: 38](#_Toc6591071)

[d. Data Visualization: 39](#_Toc6591072)

[C. Contribution 40](#_Toc6591073)

[References 41](#_Toc6591074)

# Report

## Introduction to Big Data

The term Big-Data represents for large amount of data which can be analyzed and processed. The word ‘Big-data’ became popular since it was introduced by John Mashey in 1990s because data become so useful for analysis, to know the behavior or to get extract information out of it. Better operational efficiency, improving customer efficiency, intelligence for decision taking and many more are the best outcome of processing big data. Example; User data, sensor data, satellite data, etc. gives large data each second. Data also can be in structured or semi-structured or unstructured format.

Data is not regarded as big because of its volume but according to IBM, it is represented by three V’s dimensions. They are given below:

1. Volume:

Volume represents data in large size like hundreds of terabytes, petabytes, yottabytes or zettabytes.

1. Velocity:

Velocity represents for the speed of data change. At what speed data is being generated.

1. Variety:

Variety represents different types of data which may include text, audio, video, clicks streams, touch sensors and so on.

Other different V’s also represent big-data like Veracity, Value and so on (Sriramoju, 2017) (Chitresh Verma, 2016).

## Evaluation of the Tools and Techniques

Different tools and techniques are there for processing big-data. There are tools available for processing big data. Hadoop, NoSql, Hive, Mongo DB, etc. are some of the popular tools used for handling big-data (M. Sowmya, 2017). Some of these are explained below:

### Oracle and Excel

Excel is a spread sheet where data can be saved in the form of row and column. Data saved in excel can be directly imported in oracle and use the data. It’s easy to store data in excel and load data easily into database but it can store limited number of row and columns only. In other hand Excel is not good for storing various data like image, audio and so on.

Oracle is a relational database management system. It treats data as a unit which helps for the extraction of related data. Oracle is the first database designed for enterprise which is flexible and cost effective to maintain information. Roll back features make it more unique but still in terms of big data oracle is not good for handling large data and variety of data. But Oracle 12c has some updated features (Cyran, 2005) (Rick Greenwald, 2003).

### MongoDB

MongoDB is open-source and NoSQL database which means Not only Structured Query Language, so it is not restricted by structured query language. It is highly flexible and scalable which make it more popular. It supports for complex data structure and data index. It stores data in JSON format by automatic subdivision and geographical spatial index. In terms of big-data, it has document storage limitation, but it supports structured, semi structured and unstructured data. Convenient storage, functional diversity, reliability etc. makes mongo DB stronger (Chaokui Li, 2014) (Prabagaren, 2014).

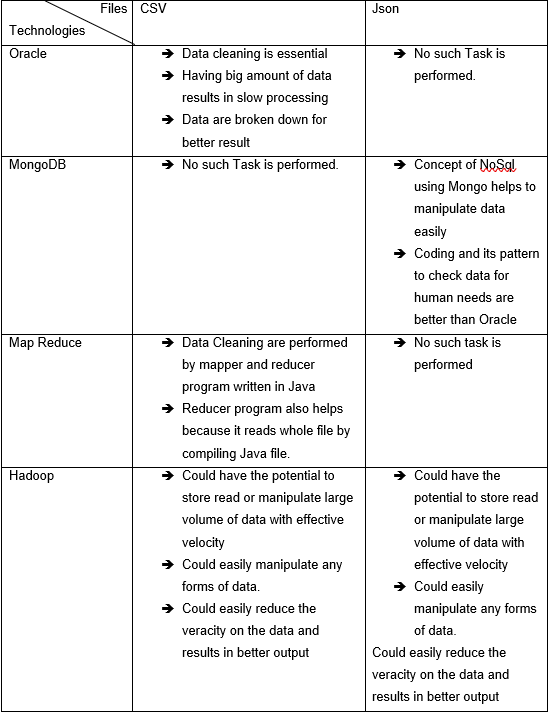
### MapReduce Framework

Map Reduce is a parallel data processing technique. It consists of two main phases mapper and reducer. Mapper reads the file and make mapping. It transforms input record to intermediate key-value pairs. After this reducer aggregate the intermediate result and gives output. Map reduce has merged with HDFS for better performance in HDFS. It does not support stream data processing also processing of complex data analytical is difficult. It performs on homogeneous data, so to solve this issue map join reduce was introduced.

Map reducer processes slowly for small file and not applicable for synchronized data (Ruchi Bhardwaj, 2014).

### Hadoop

The concept of Hadoop is officially established by Apache which is open source. The journey of Hadoop started from first version (i.e. Hadoop 1.0) has made lots of impact. The technique of cluster computing, grid computing, cloud computing, distributed file system handling etc. and security support and enhancement from Apache Rhino, Apache Ranger, and Apache Knox has made Hadoop stronger. It handles fault-tolerance. MapReduce of Hadoop is dependent on YARN for parallel processing which makes job scheduling and Cluster Resource management. Hadoop supports big-data characteristics like volume, velocity variety and veracity where each has its own role in the framework (Gurjit singh Bhathal, 2018).



## Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Volume** | **Velocity** | **Variety** | **Veracity** |
| **Oracle** | Increase in volume slowdowns the performance of the oracle. The processing time periods are a bit longer than Mongo DB and Hadoop. | Daily huge data are generated in an unstructured format where oracle falls back for this. | Oracle is based on relational format and supports structured data and not able to applicable for unstructured data. | Oracle must have cleaned and structured data format but in real world, impure data are generated daily. |
| **Mongo DB** | Sharding, fault-tolerance and replication are three major key factors on mongo DB to handle huge data. | It is based on NoSQL and has indexing feature which maintains changing data. | It supports structured to unstructured data like CSV, JSON and so on. | Accurate data can be achieved from Mongo DB since it handles impure data too. |
| **Hadoop** | It includes HDFS which work on distributed way and map reduce handles large data easily. | Highly changing data on Yahoo, IBM, Facebook are handled by Hadoop since it can run on multiple nodes and clusters. | It handles all types of data structured, semi structured and unstructured data. | Hadoop supports impure data and can work parallelly on the given data. |

## Conclusions and recommendations

Data are increasing day by day where comes the concept of Big-data. Data must be managed in the systematic way. Big data is represented by volume, velocity, variety or veracity. Oracle, MongoDB and Hadoop are highly used for handling large data.

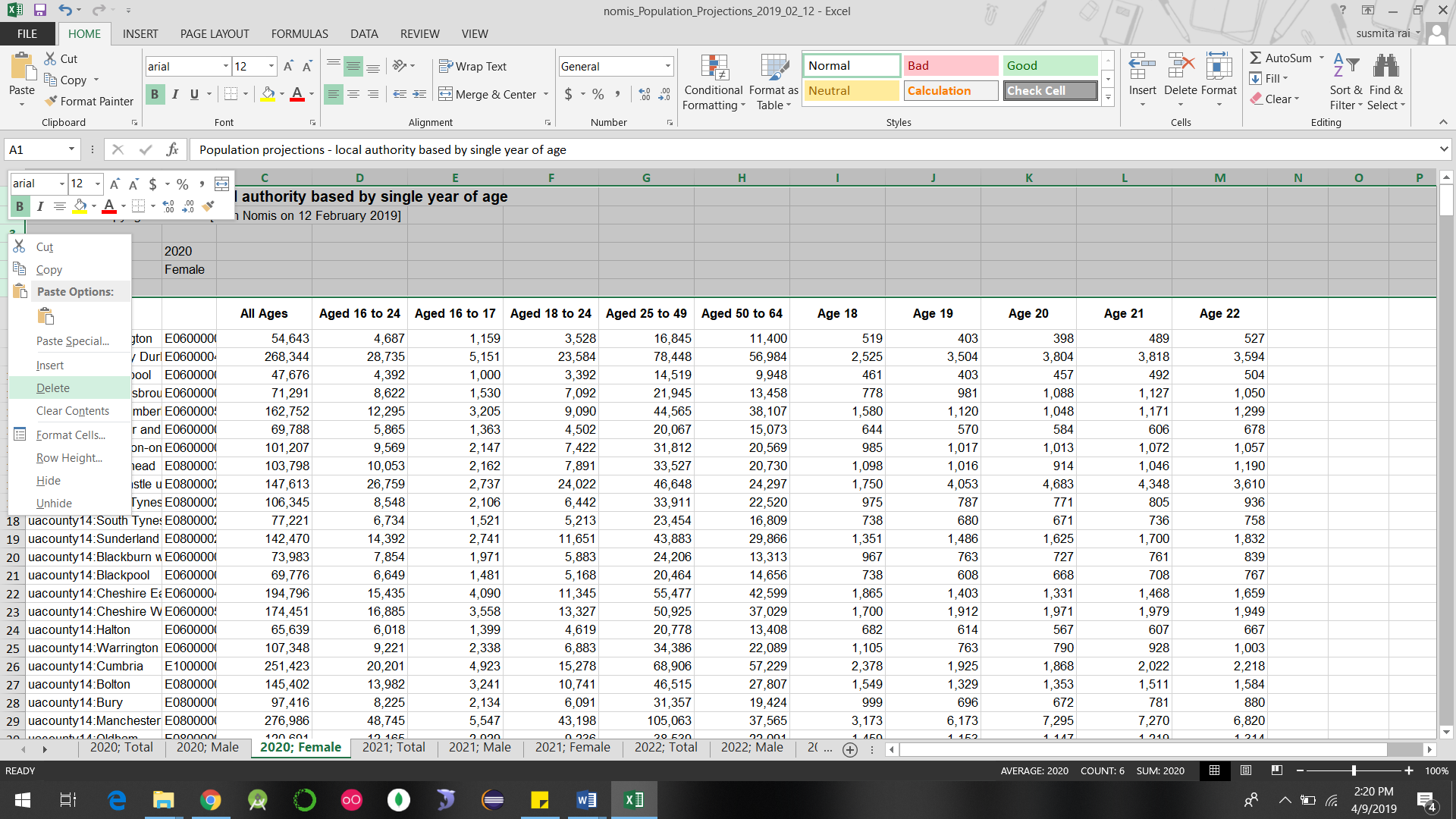
All tools have its own merits and demerits. Oracle and mongo DB are suitable for small data where large volume data can be handled by Hadoop since it used map reduce technique. Mongo also supports for semi and unstructured data where oracle lags.

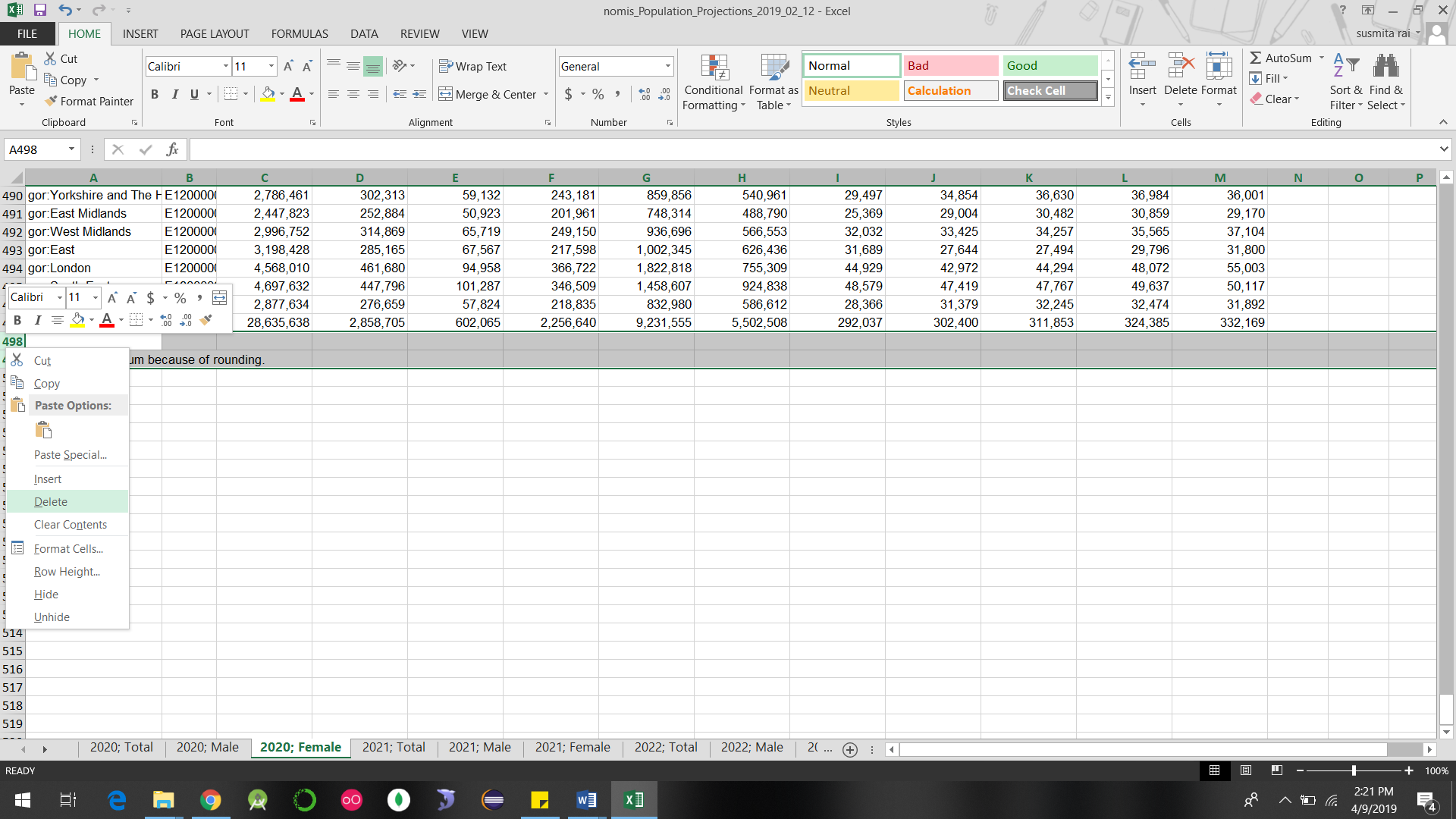
# Investigation

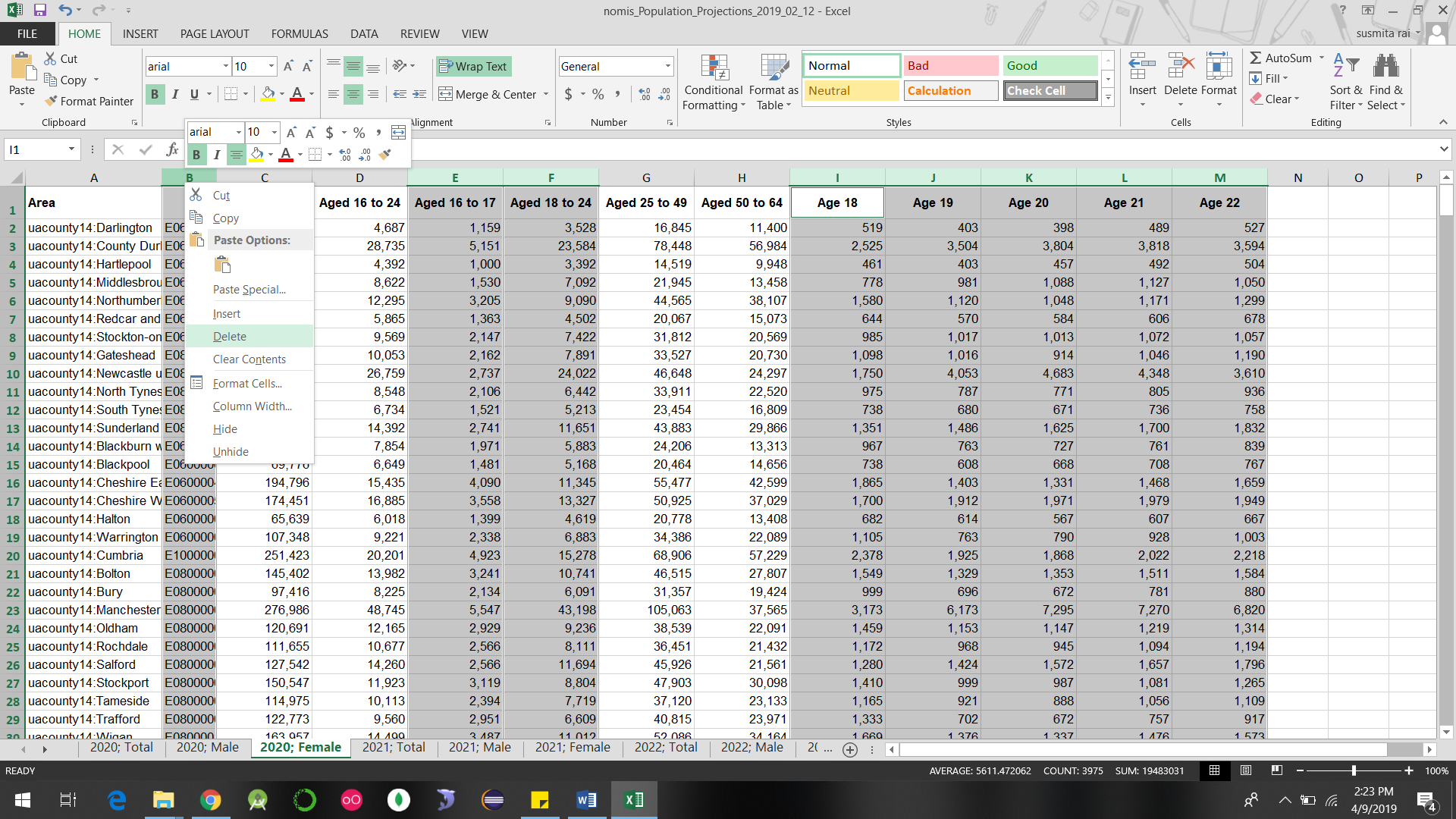
## Cleaning

Data cleaning is necessary for investigation so excel is used for cleaning. Cleaning Female data 2020:

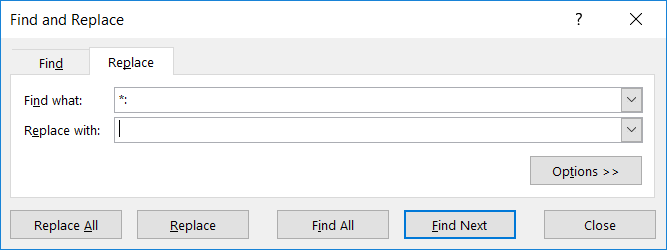
Deleting unrequired data

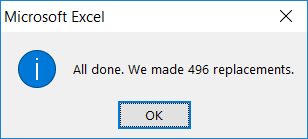


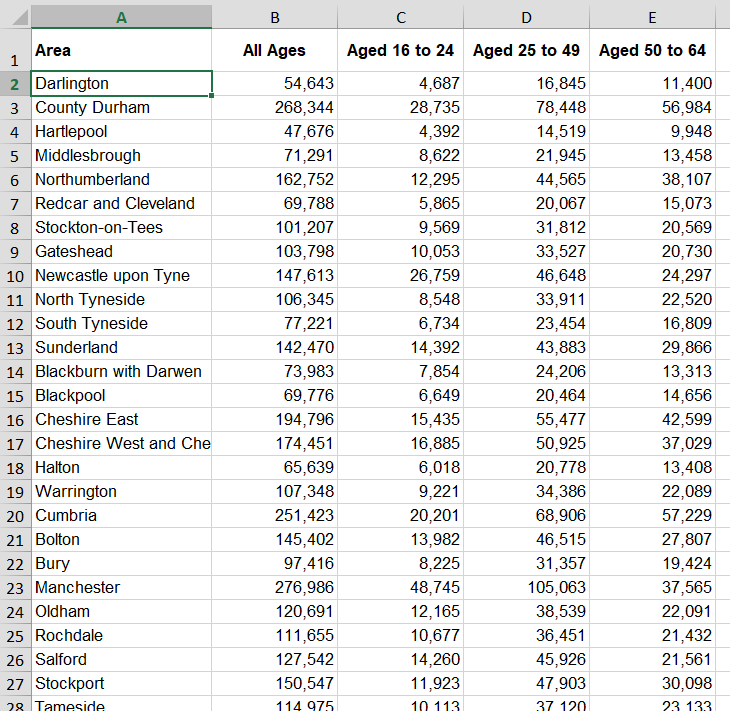
­­­­



Replacing value before ‘**:**’



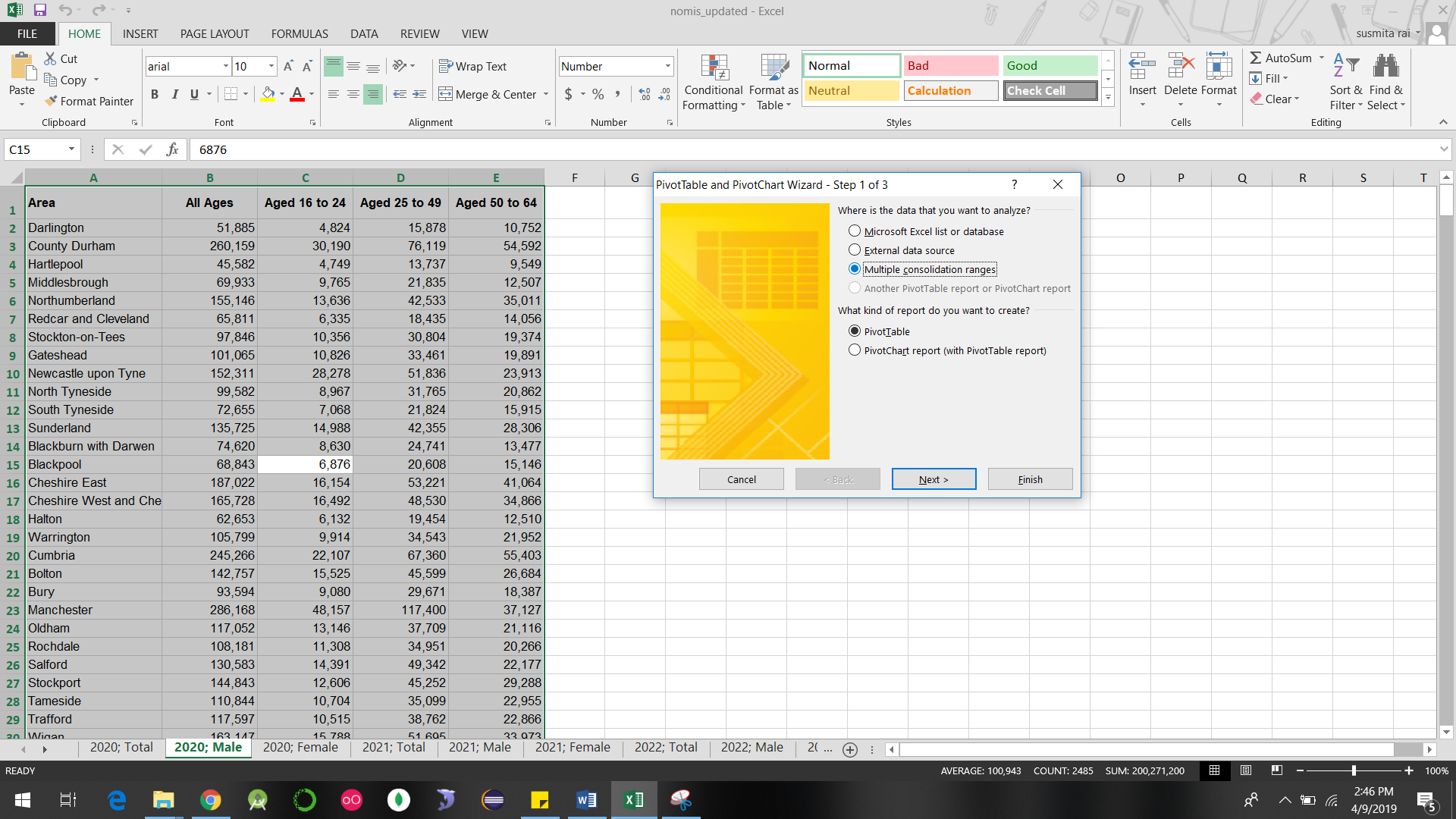


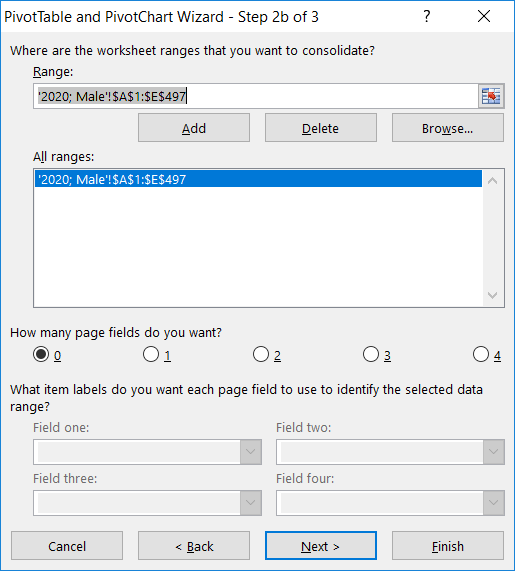


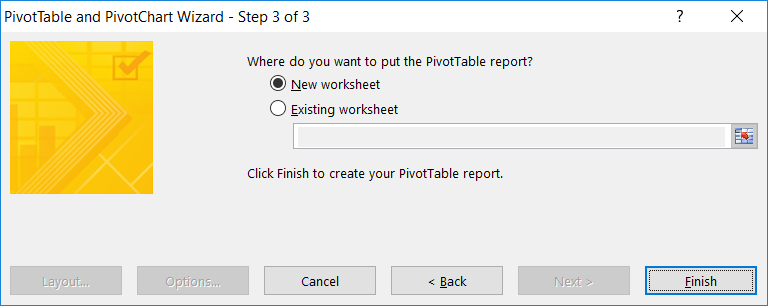
Similarly, other data were cleaned.

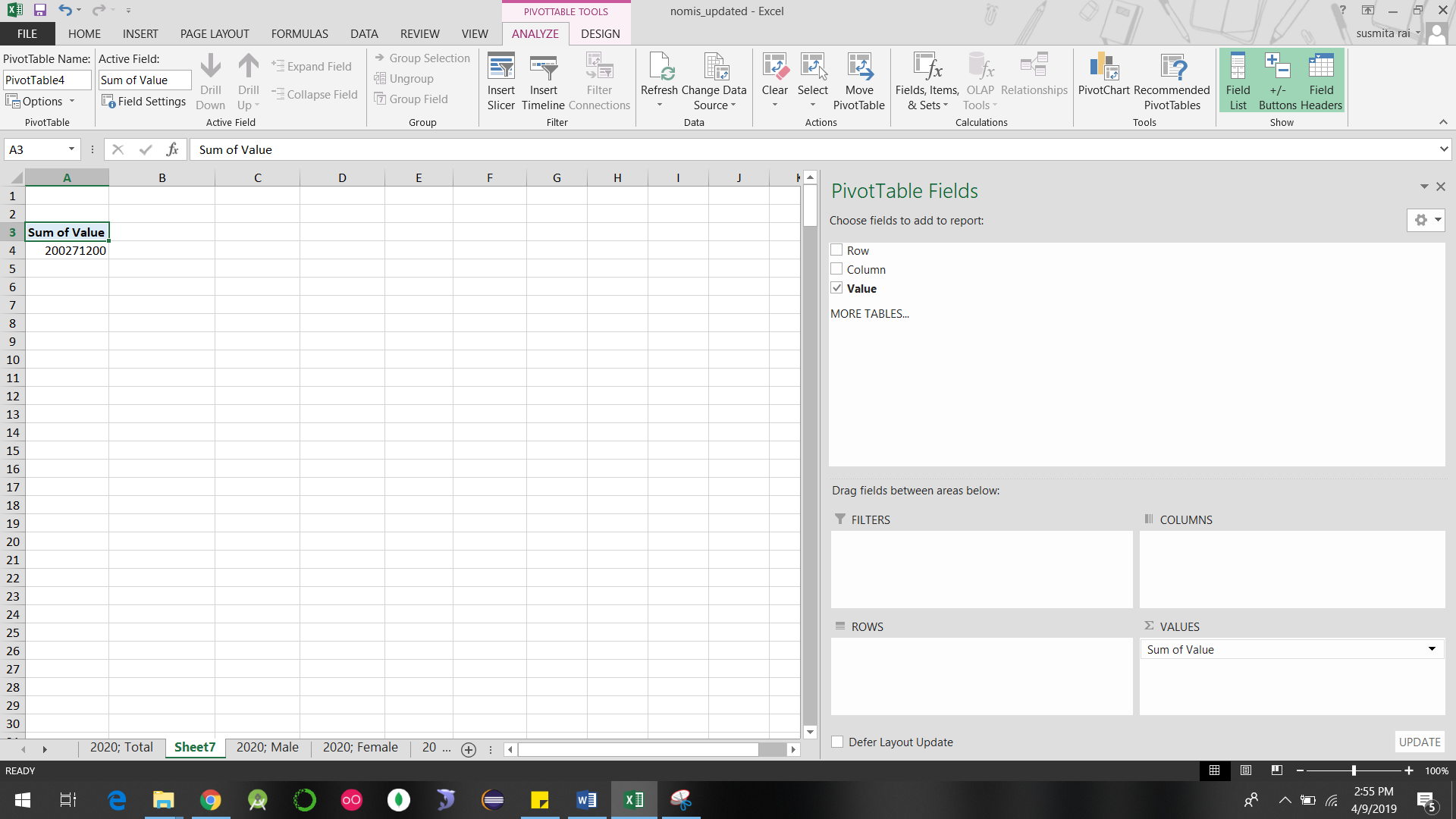
Manipulation:

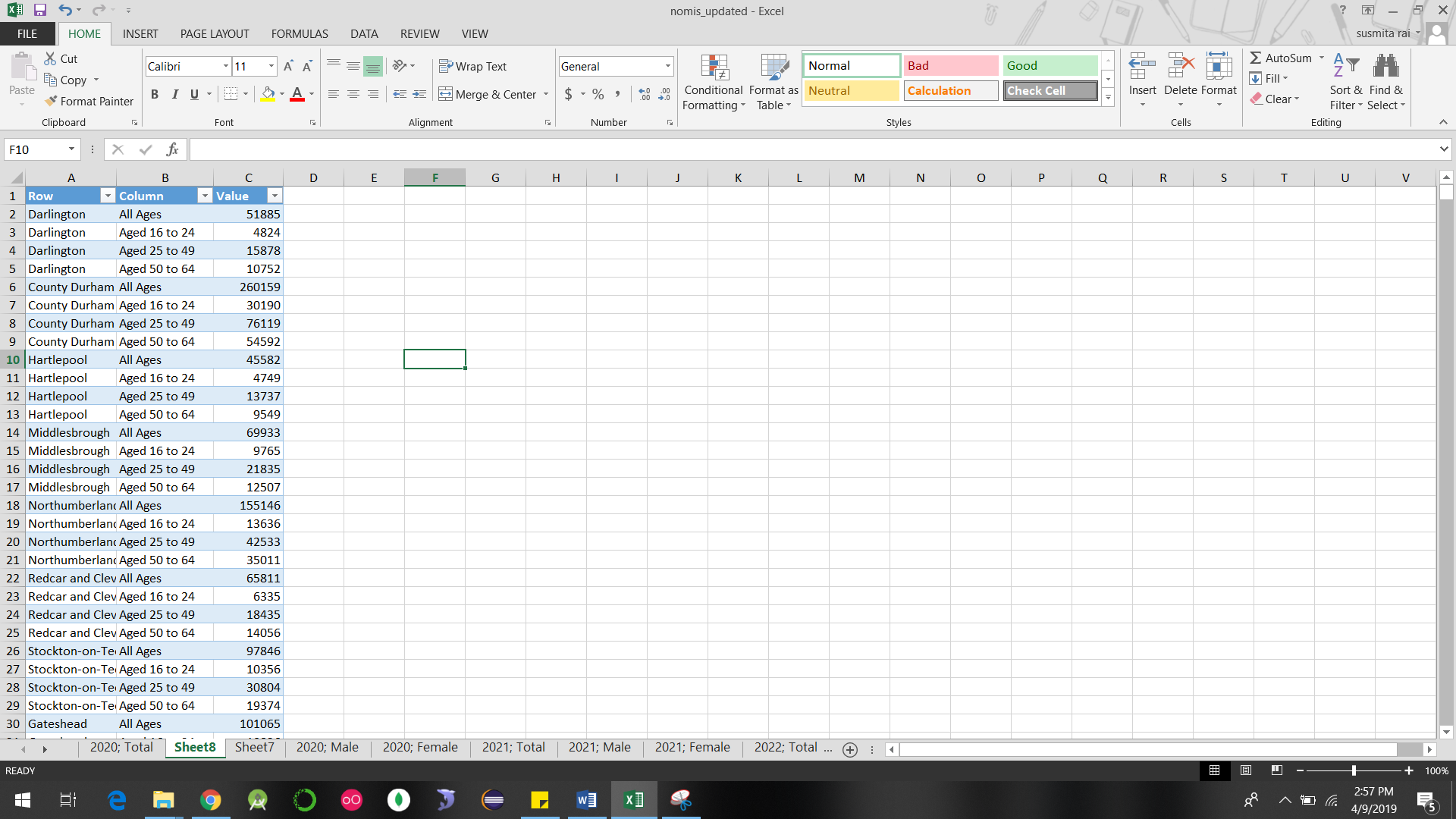
Data need to be converted into required format changing it into pivot.



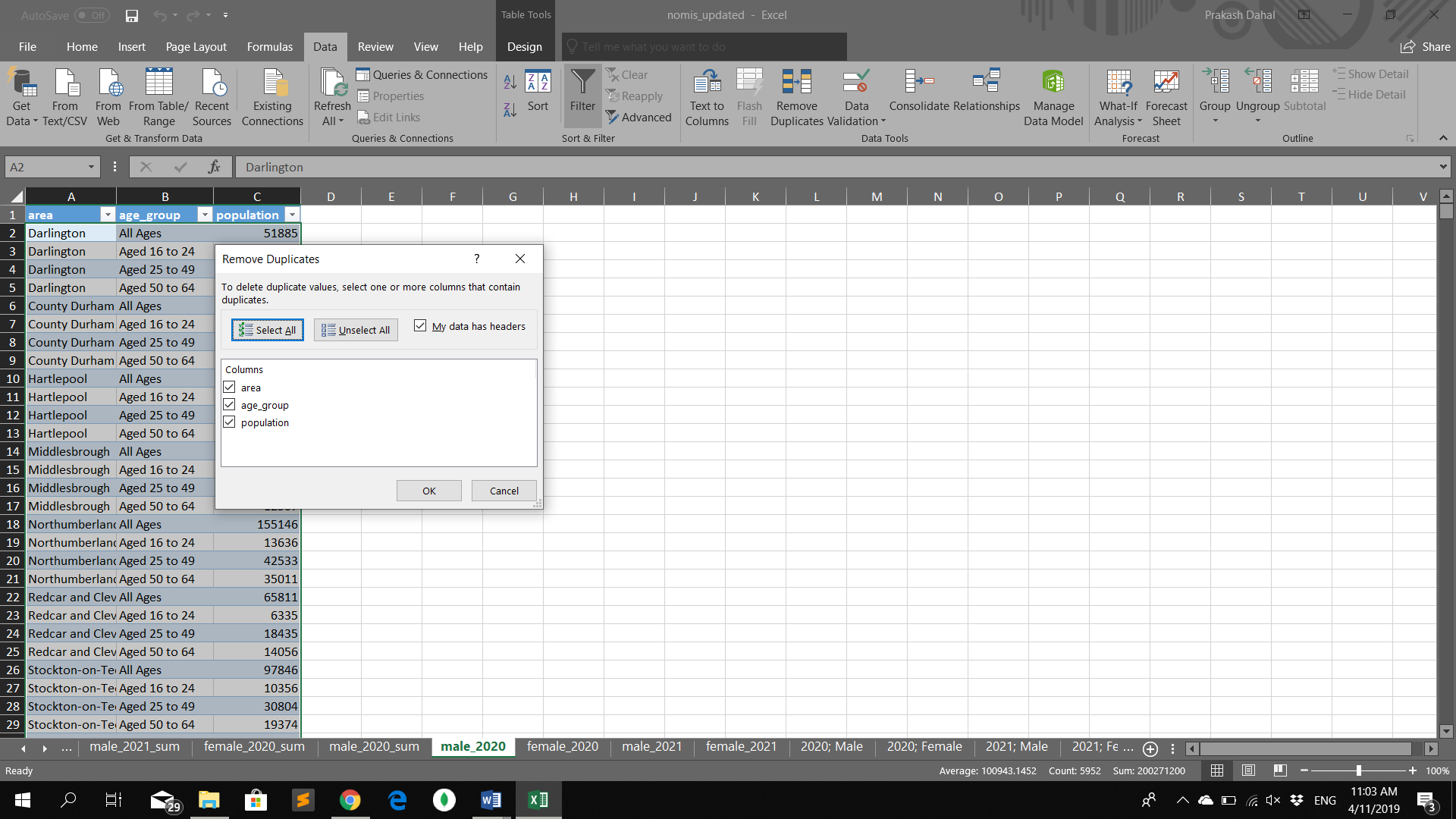


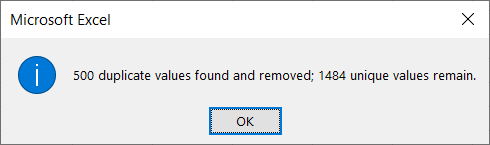






After analyzing these data, redundant rows are cleaned again.



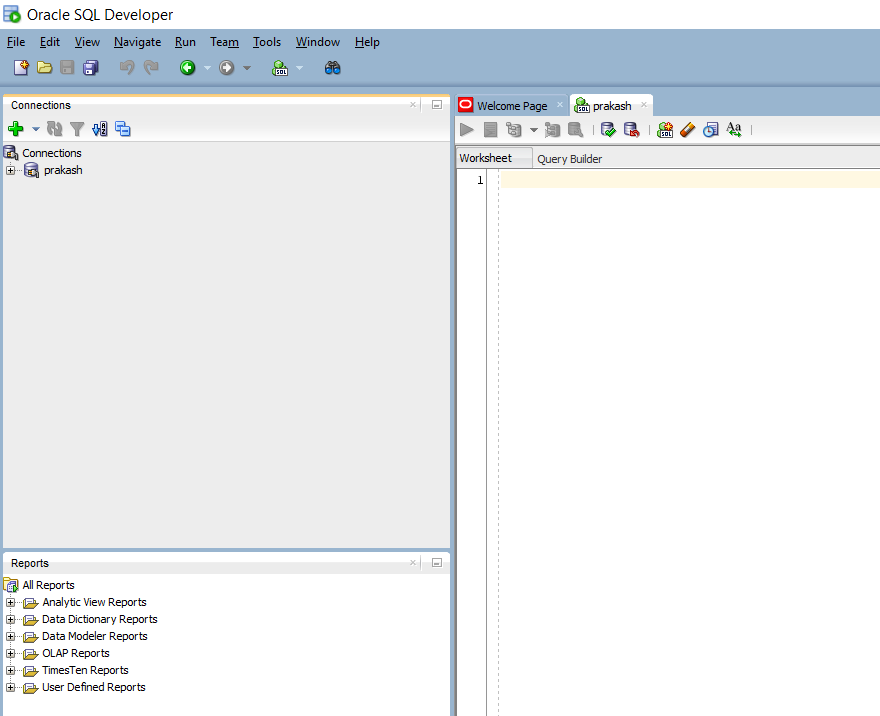


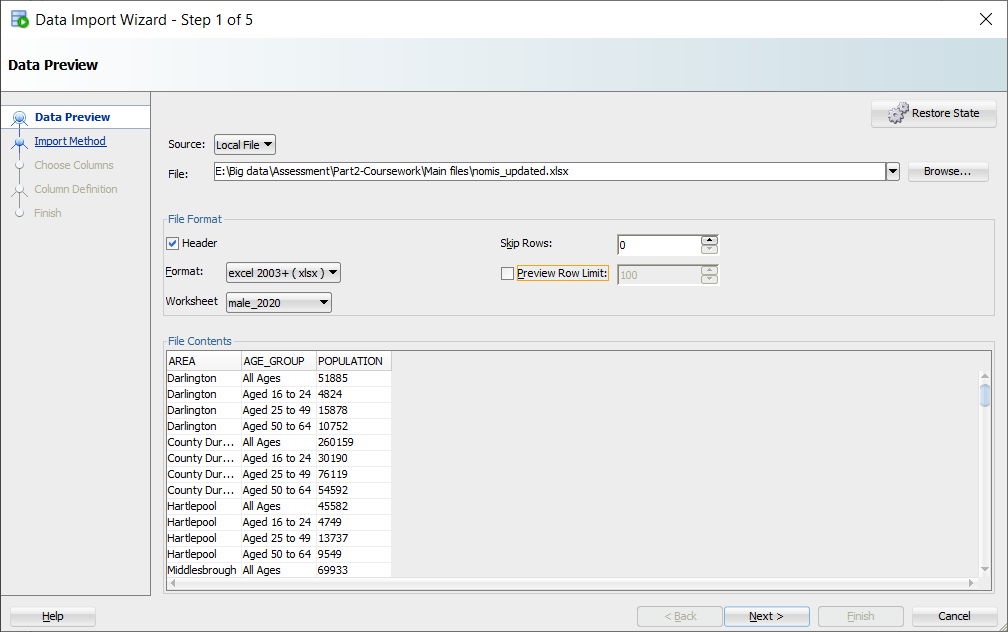
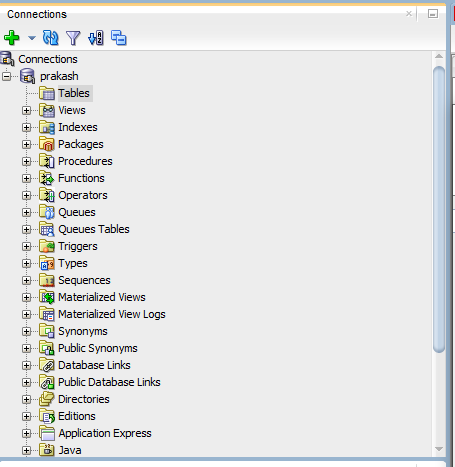
Similarly, other files are changed in and made pivot.

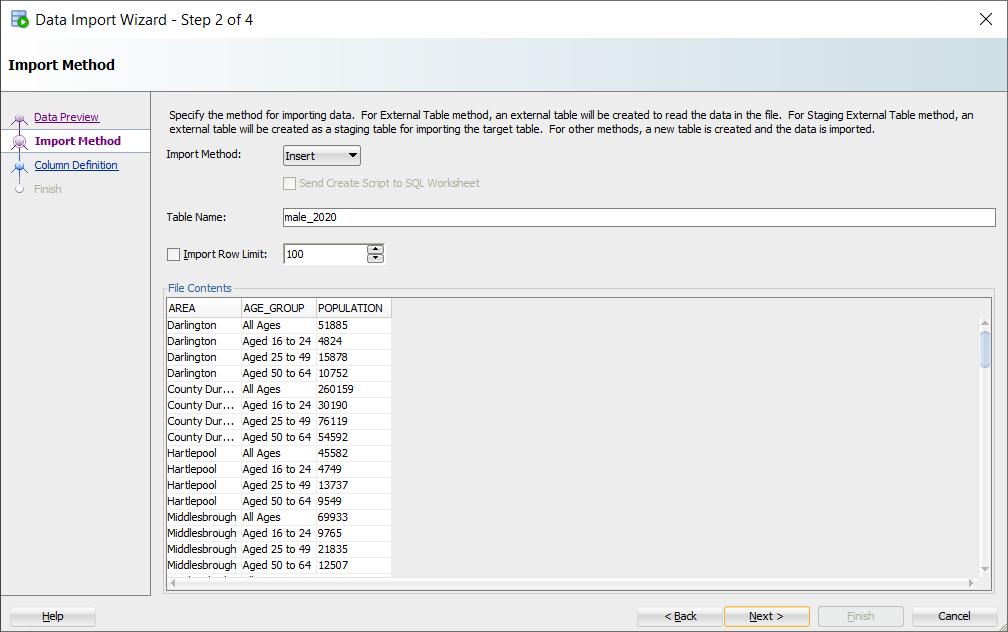
## Manipulation:

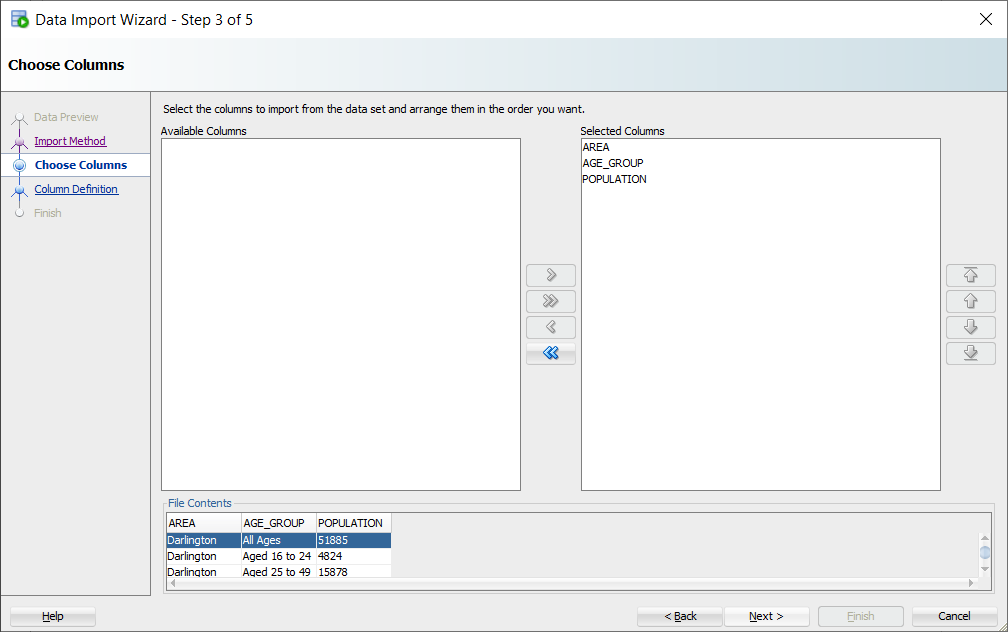
Importing in Oracle:

Importing required data in **prakash** databases

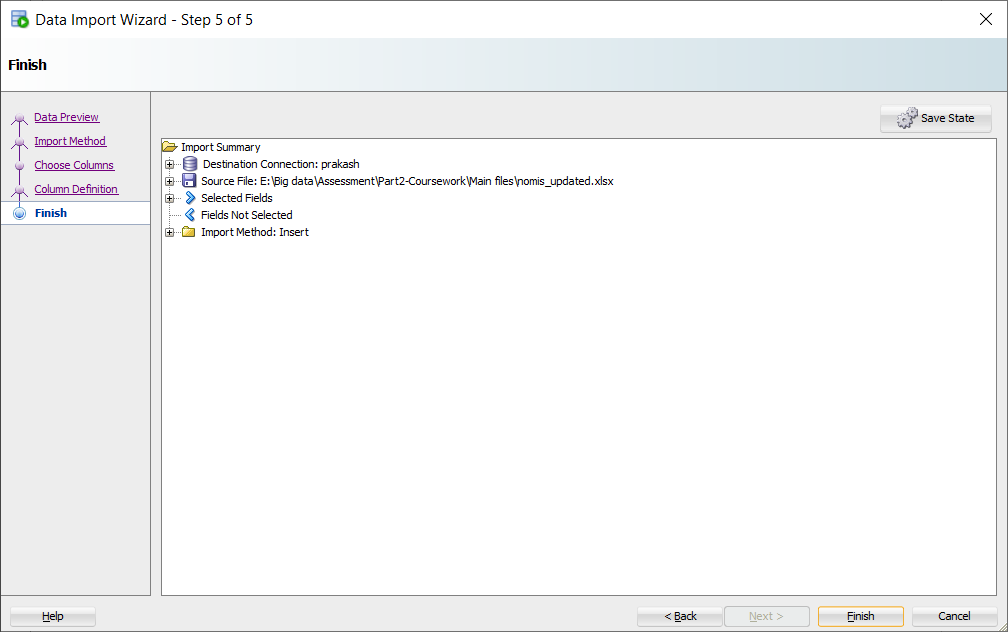
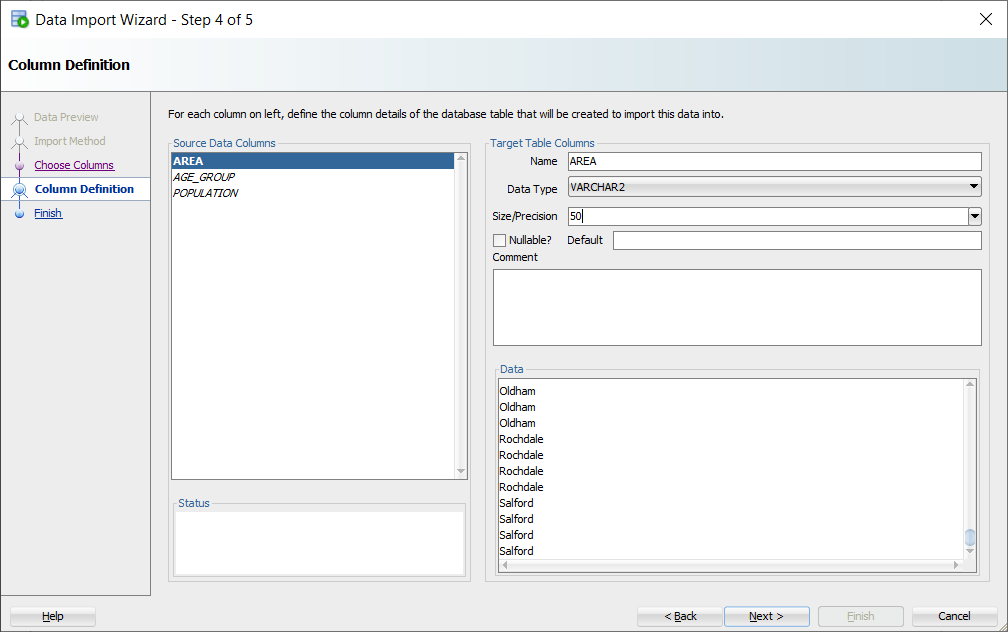


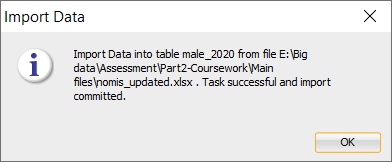


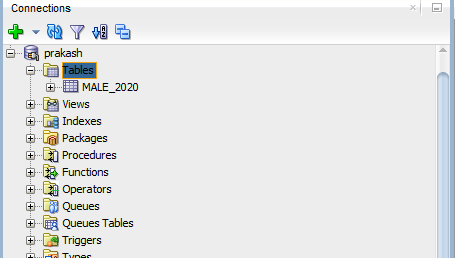




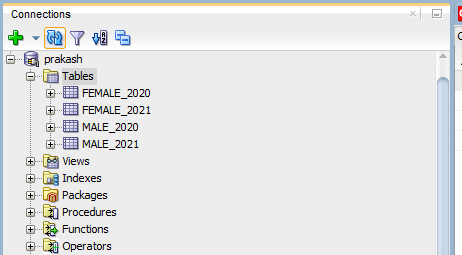
Changing size to 50 so that the area and age\_group field could be longer and making nullable. But leaving population as it is.



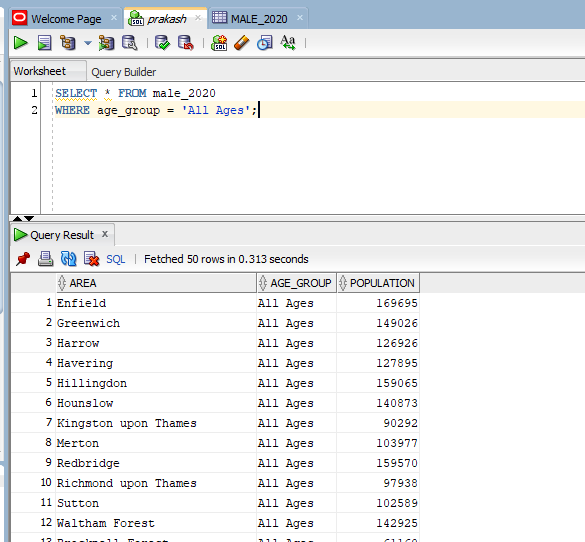




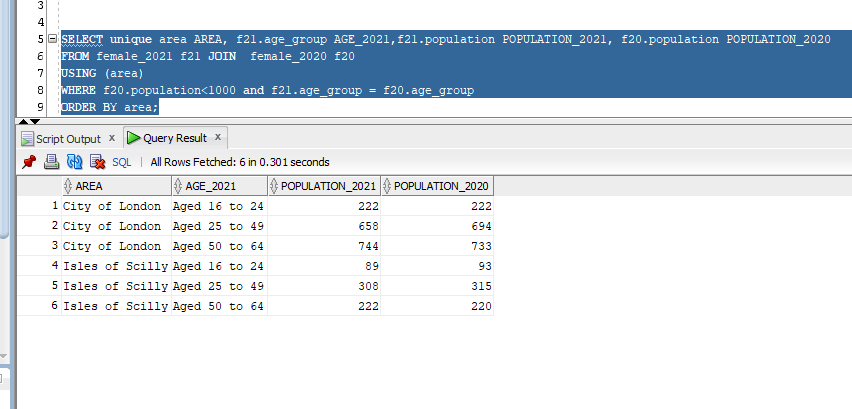
Similarly importing all required sheet.



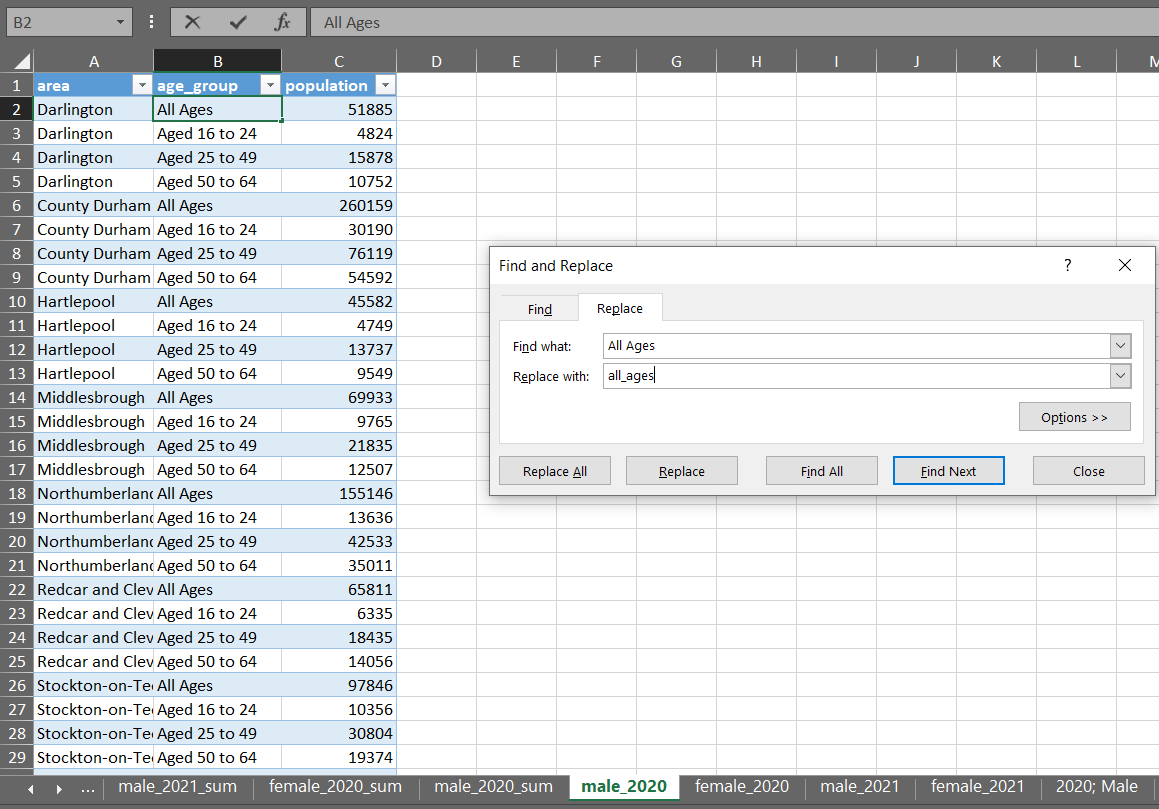
**SQL Query:**

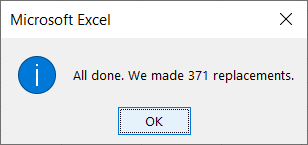


Query represents total age of each city of male2020.

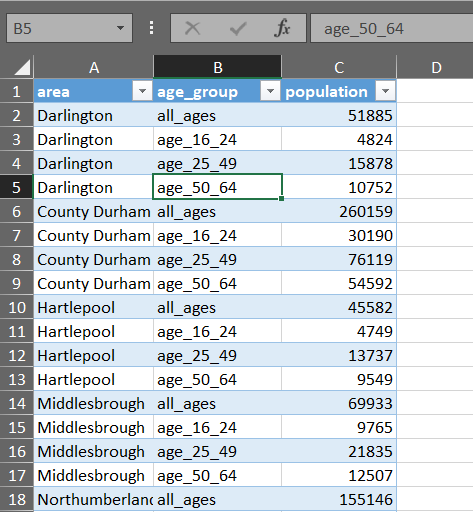


While analyzing the cleaned pivot csv file, the name given is not suitable so changing the name for mongo DB and Hadoop.

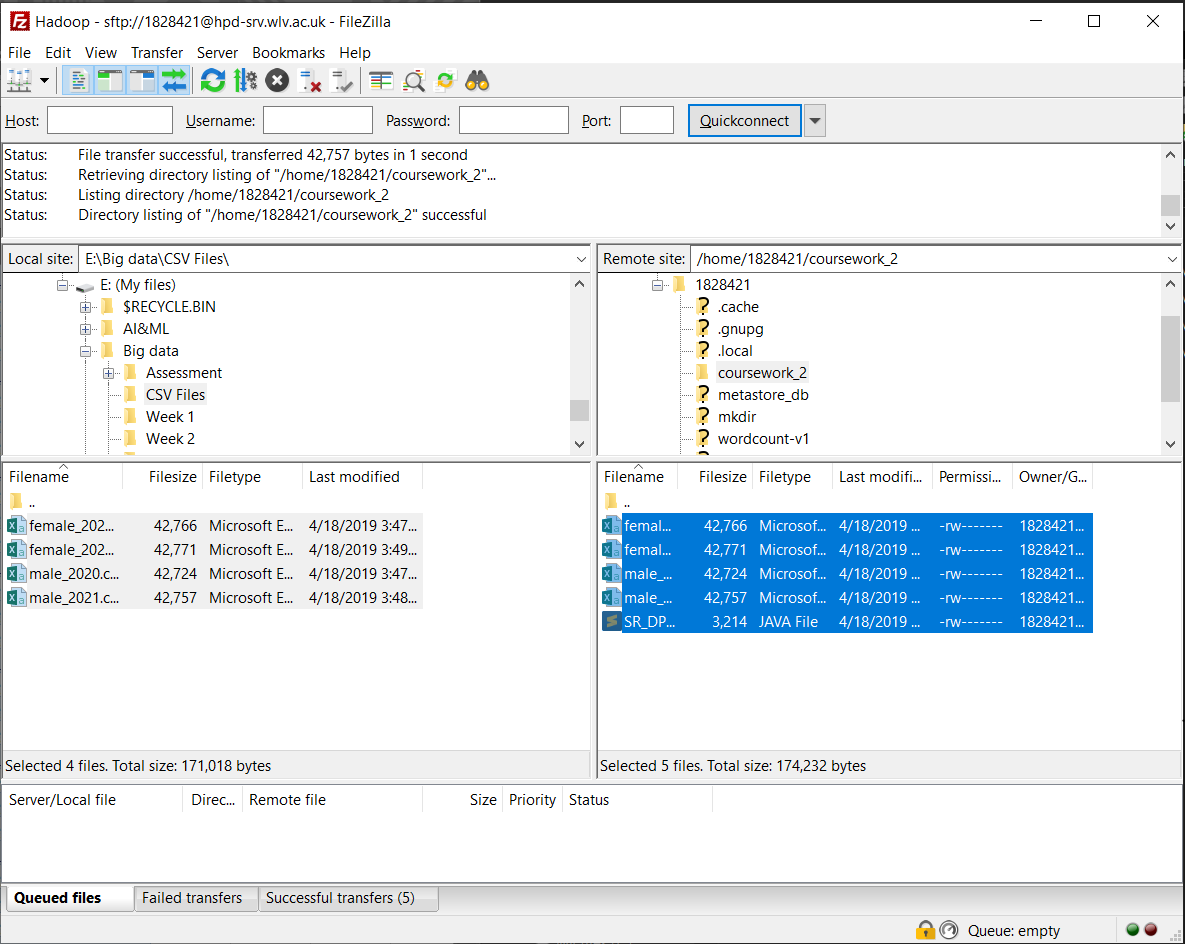




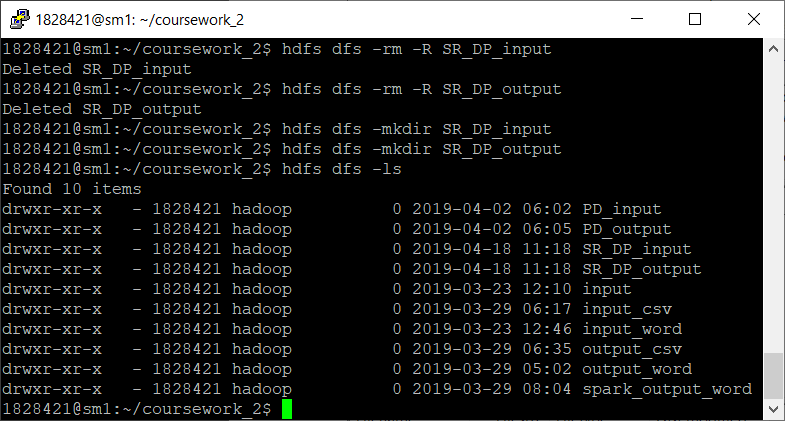
Similarly, giving suitable name for all on each sheet

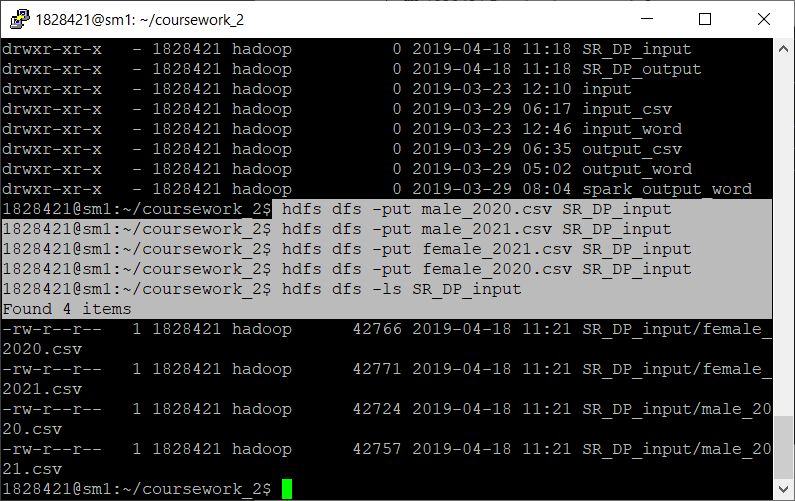


Hadoop:

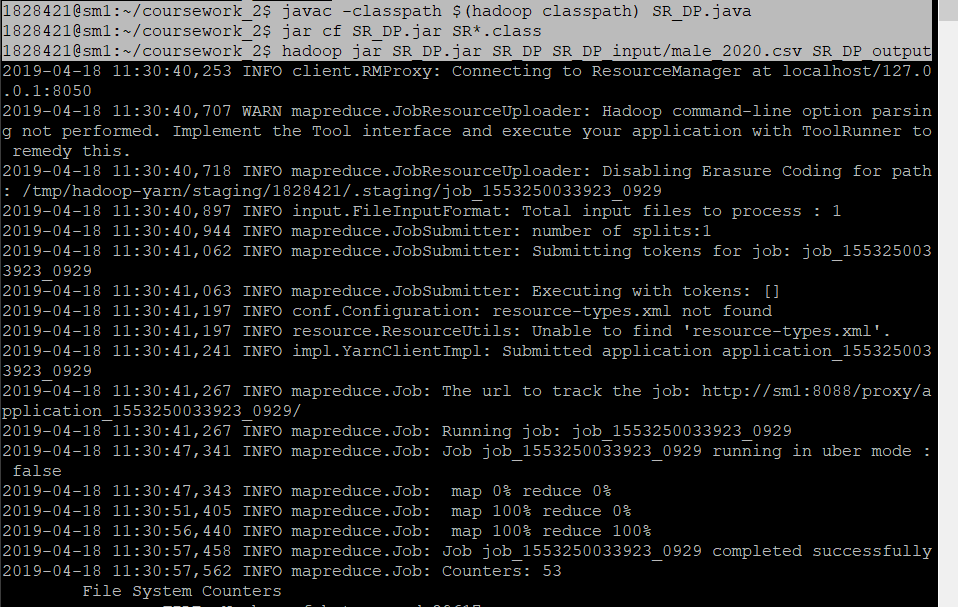


Making input -output files and putting file in it.

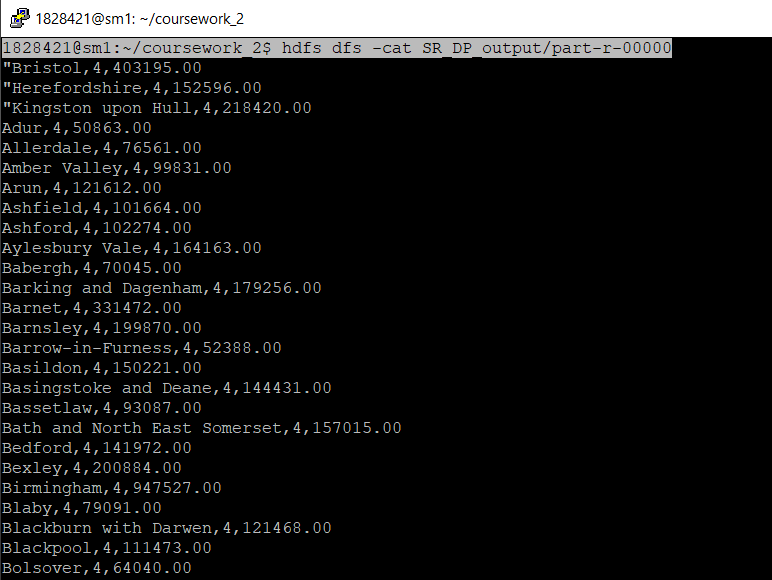




Compiling Java file, making jar file and running.

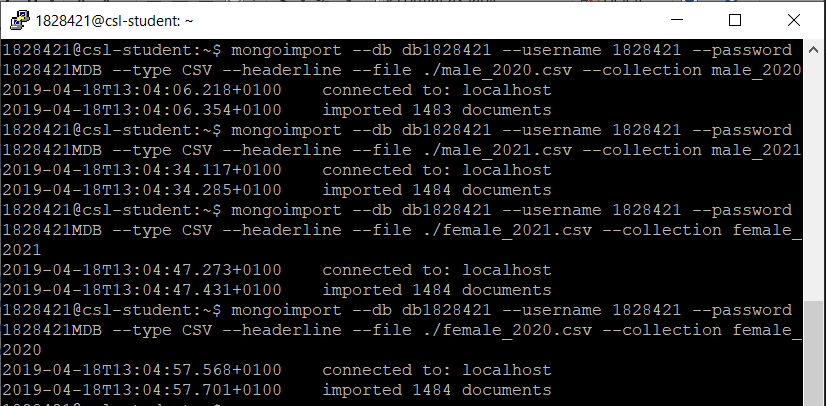


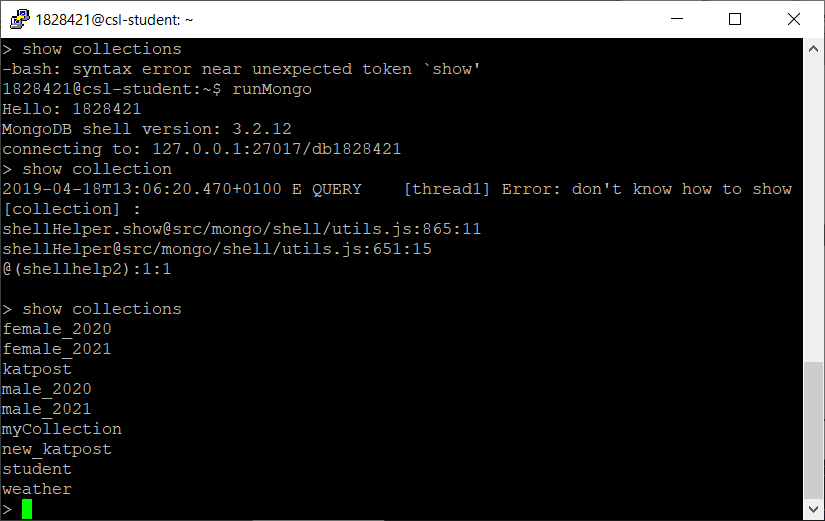
Result of male 2020

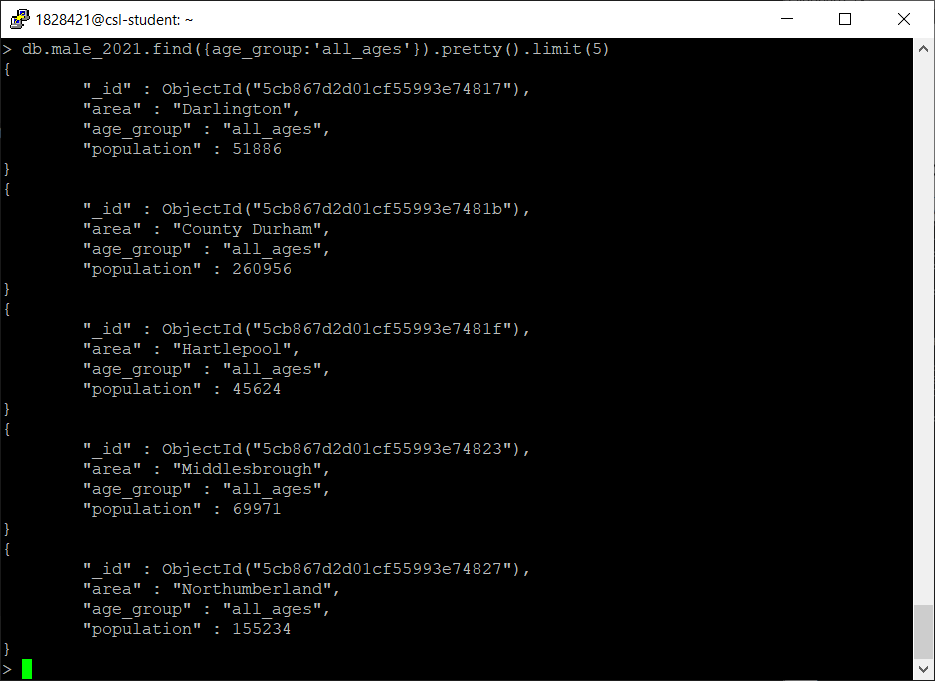


Mongo DB:

Loading CSV file:



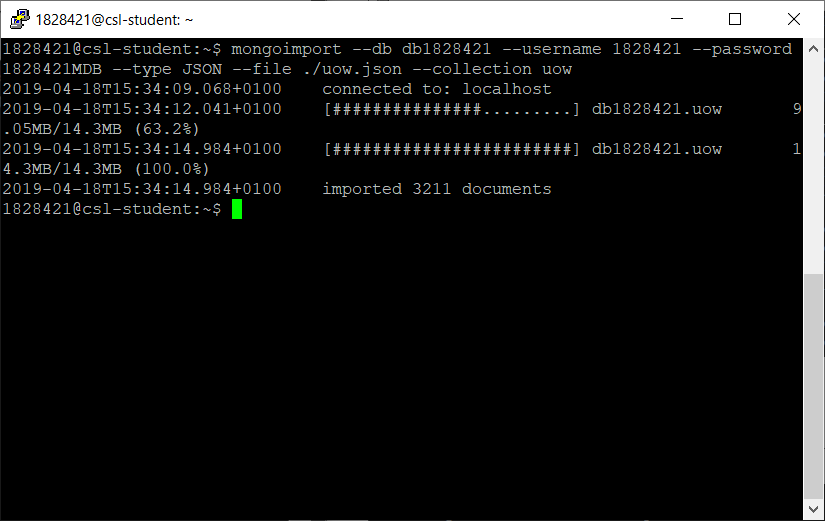


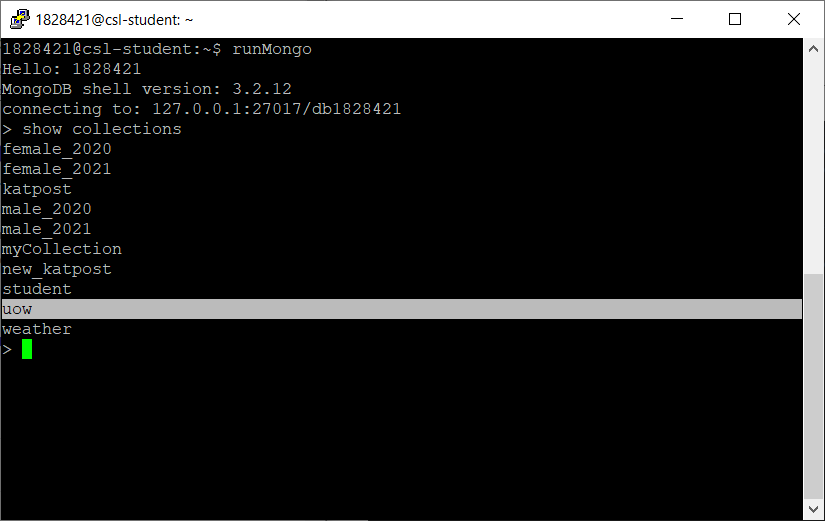


Loading JSON:

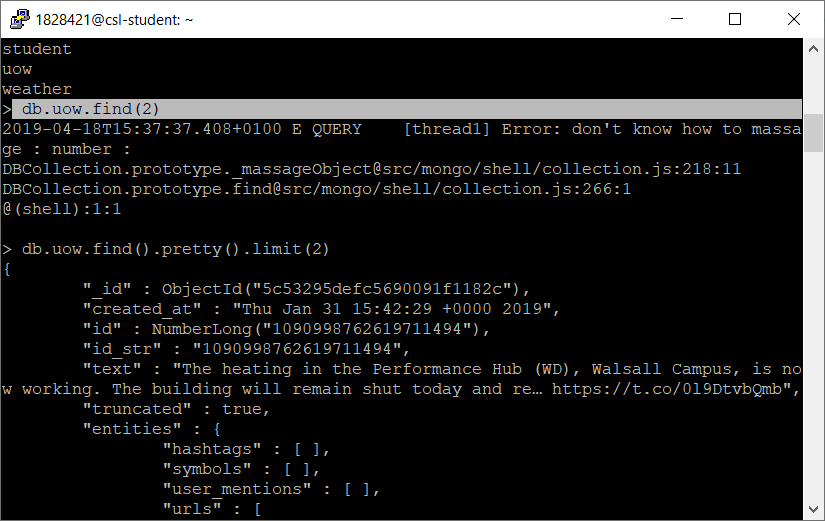


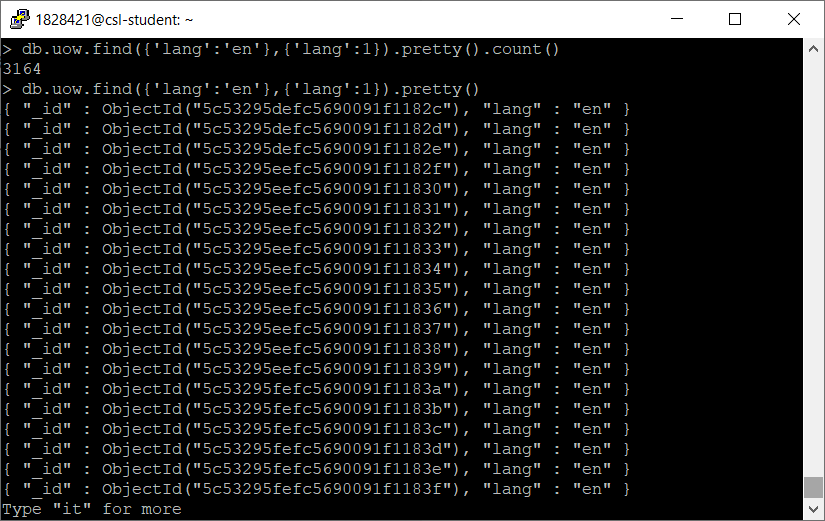
Making collection of json uow





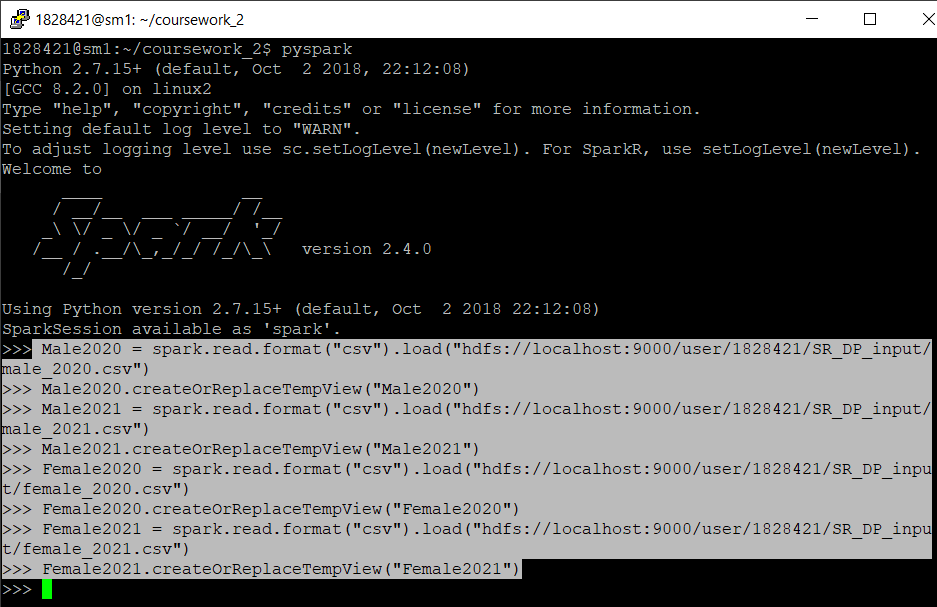
Displaying first two uow data



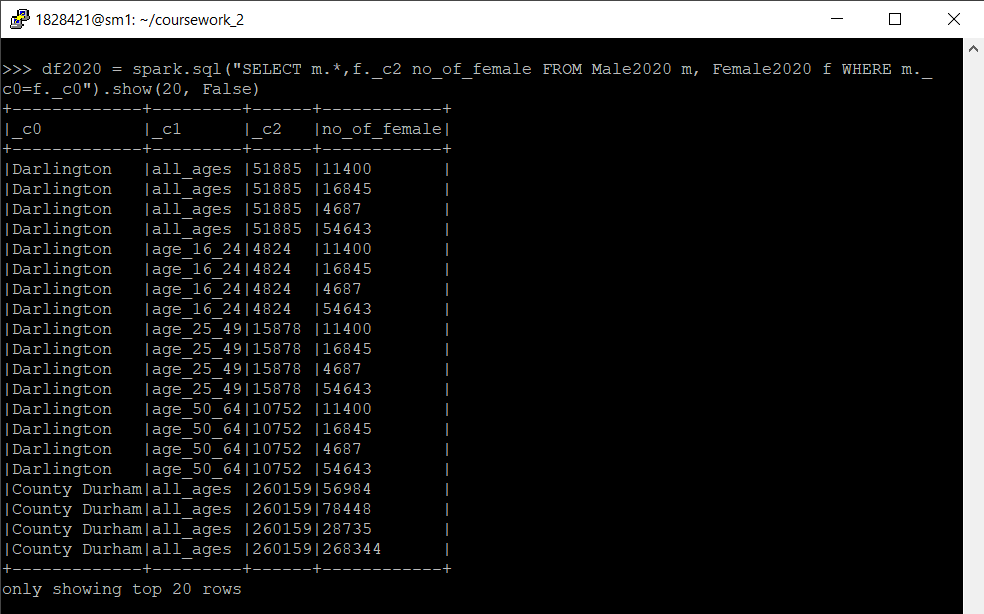


Py-Spark:

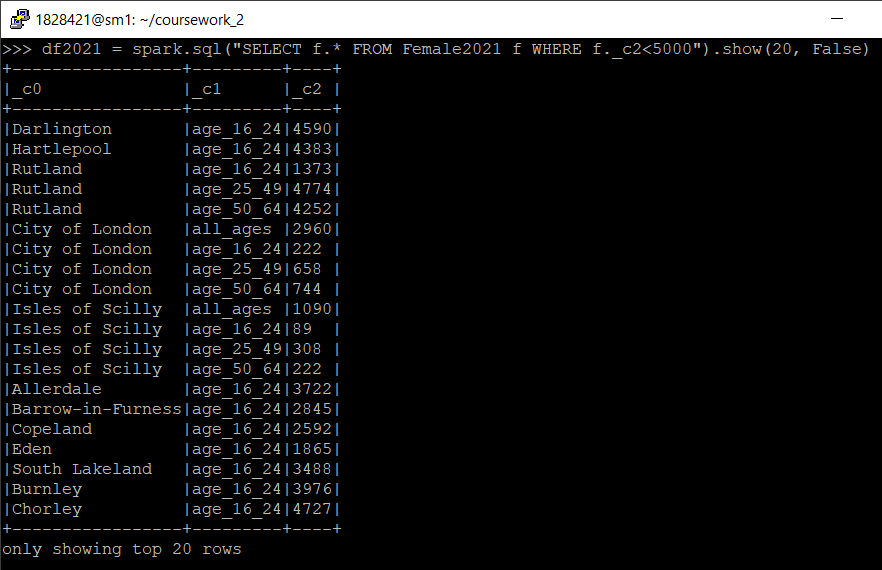
Loading csv and converting into view:



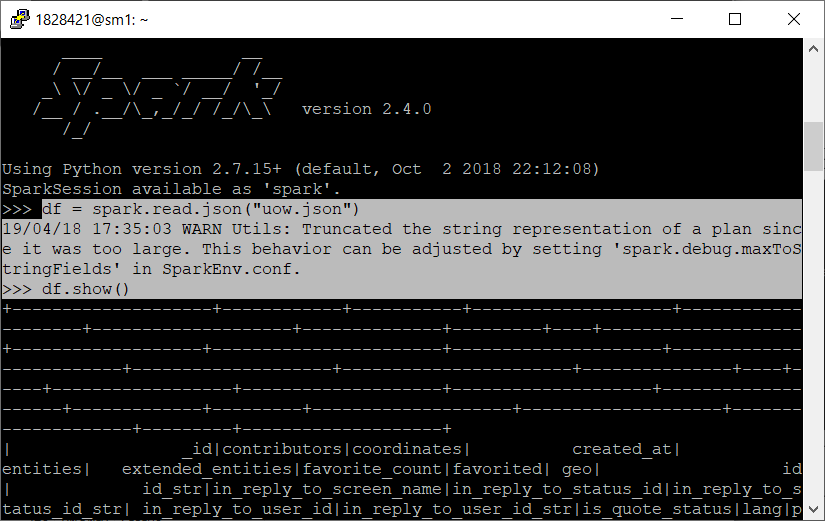
Joining two tables of male and female of 2020

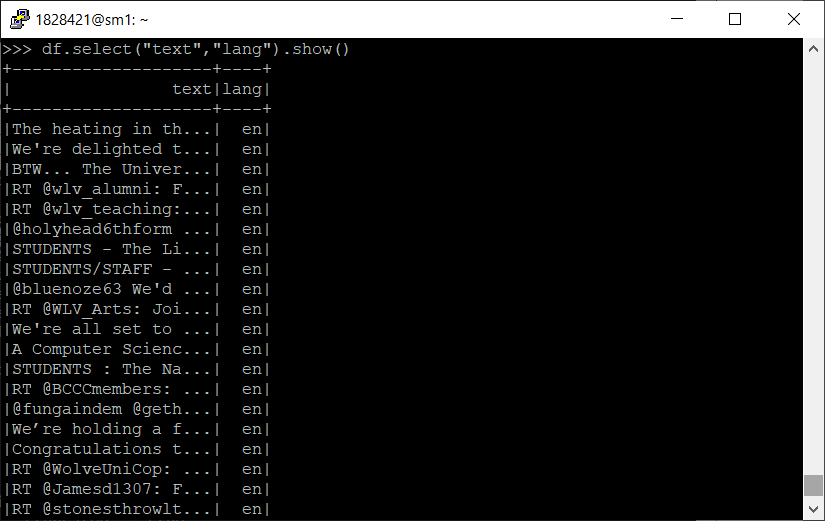


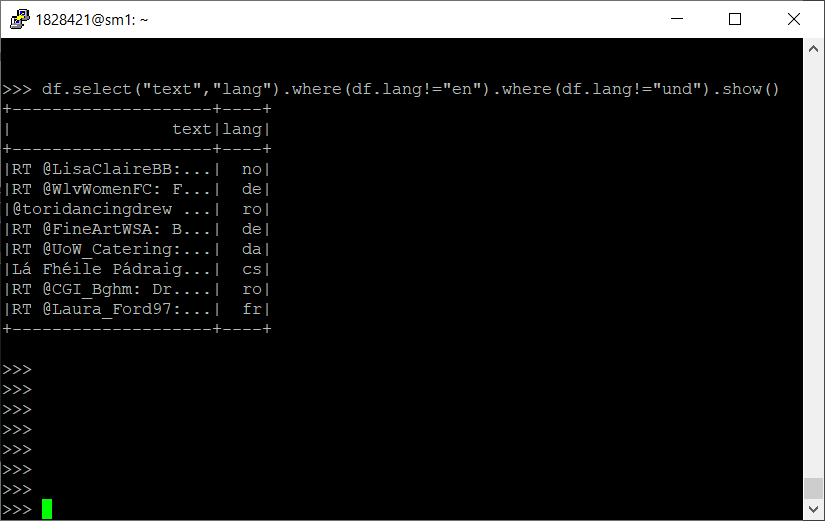
Female of 2021



JSON with py-spark



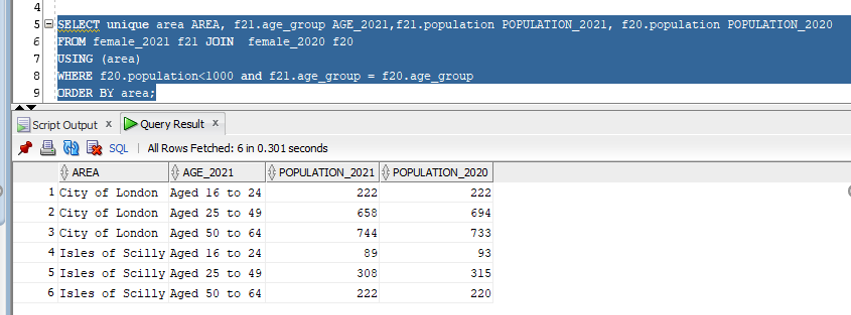




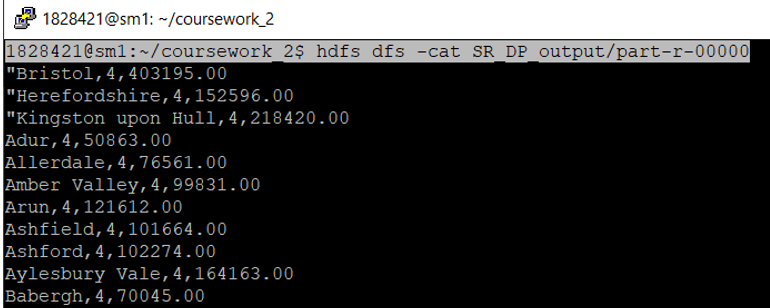
## Analysis of the data:

Data analysis is most important part to understand about the data and extract information out of it. Two different years is selected 2020 and 2021. From these two years, male and female are selected so that the comparison of male of both years can be analyzed. So, with the female. Total population can be extracted adding total population of male and female.

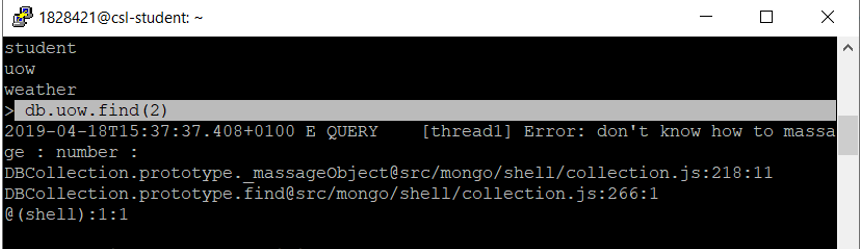
Oracle Query displays the result of all aged group population and selecting female of 2020 and 2021 and joining by area and only displaying which population is less than 1000.



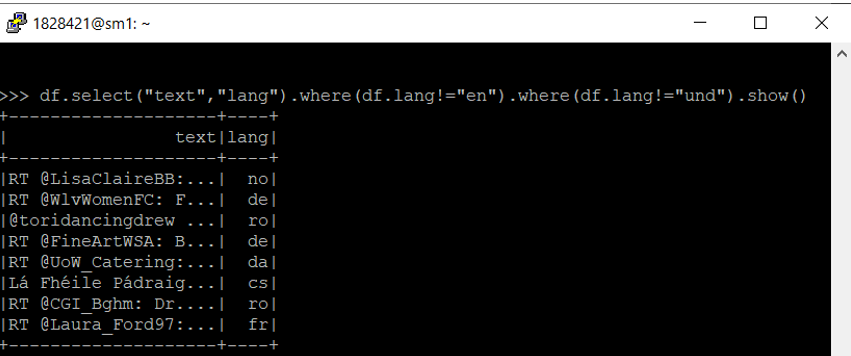
In Map-Reduce, population is added as per the area and gives number of area and total population of the respective area.



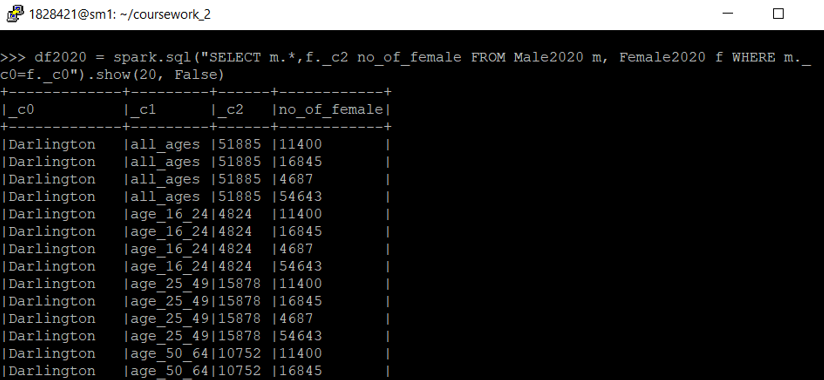
Loading and displaying result from json file in mongo and pyspark



In spark, also displaying text and language except English and undefined.



Similarly, CSV file is loaded and manipulated in both mongo db. and py spark. Instead of header option, default column is used. It displays, male2020 all col and female total population.



## Data Visualization:

Taking five different cities of male 2020 and displaying graph of age 16 to 24. Conty Durham has more population than usual.

Adding all male2020 population and analyzing age group, age from 25 to 49 has more population.

# Contribution

|  |  |  |
| --- | --- | --- |
| Name | Prakash Dahal | Rajan Sapkota (Sharma) |
| Report | Different references were discussed and finally, introduction, evaluation, matrix and conclusion were written with the agreement of both. | |
| Investigation: Oracle | Data cleaning, Pivot tables | Importing and SQL Query |
| Investigation: Mongo DB | Individual | Individual |
| Investigation: Hadoop | Individual | Individual |

For this report, both of us worked together discussing on different topic and ideas. Both of us selected different research papers, sites and discussed to complete this report. Practical is important in real world so practical was done individually except oracle. Though everything was discussed, some of the contribution is mentioned below:

# References

Chaokui Li, W. Y., 2014. *The distributed storage strategy research of remote sensing image based on Mongo DB.* Changsha, IEEE.

Chitresh Verma, R. P., 2016. *Cloud System and Big Data Engineering (Confluence).* Noida, India, IEEE.

Cyran, M., 2005. *Oracle Database Concepts,* s.l.: Oracle.

Gurjit singh Bhathal, A. S. D., 2018. *Big Data Solution: Improvised Distributions Framework of Hadoop.* Madurai, IEEE.

Huadong Dai, S. Z. L. W. Y. D., 2016. *Research and implementation of big data preprocessing system based on Hadoop.* Hangzhou, IEEE.

M. Sowmya, N. S., 2017. Big Data: An Overview of Features, Tools, Techniques and Applications. *International Journal of Engineering Science and Computing,* pp. 1364-13647.

Prabagaren, G., 2014. *Systematic approach for validating Java-MongoDB Schema.* Chennai, IEEE.

Richard L. Villars, D. V., 2014. *Building a Datacenter Infrastructure to Support Your,* s.l.: IDC.

Rick Greenwald, D. C. K., 2003. *Oracle in a Nutshell: A Desktop Quick Reference.* Sebastopol: O'Reilly Media, Inc.

Ruchi Bhardwaj, N. M. R. K., 2014. *Data analyzing using Map-Join-Reduce in cloud storage.* Solan, IEEE.

Sriramoju, S. B., 2017. *INTRODUCTION TO BIG DATA: INFRASTRUCTURE AND NETWORKING CONSIDERATIONS.* Warangal, India: Horizion Books.