# 6CS030 Worksheet One – 10%

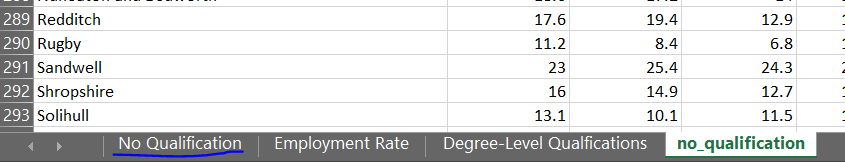
# Hand-out: Week 3. Due: Week 4

|  |  |
| --- | --- |
| **Student Number:** | 1828421 |
| **Student Name:** | Prakash |

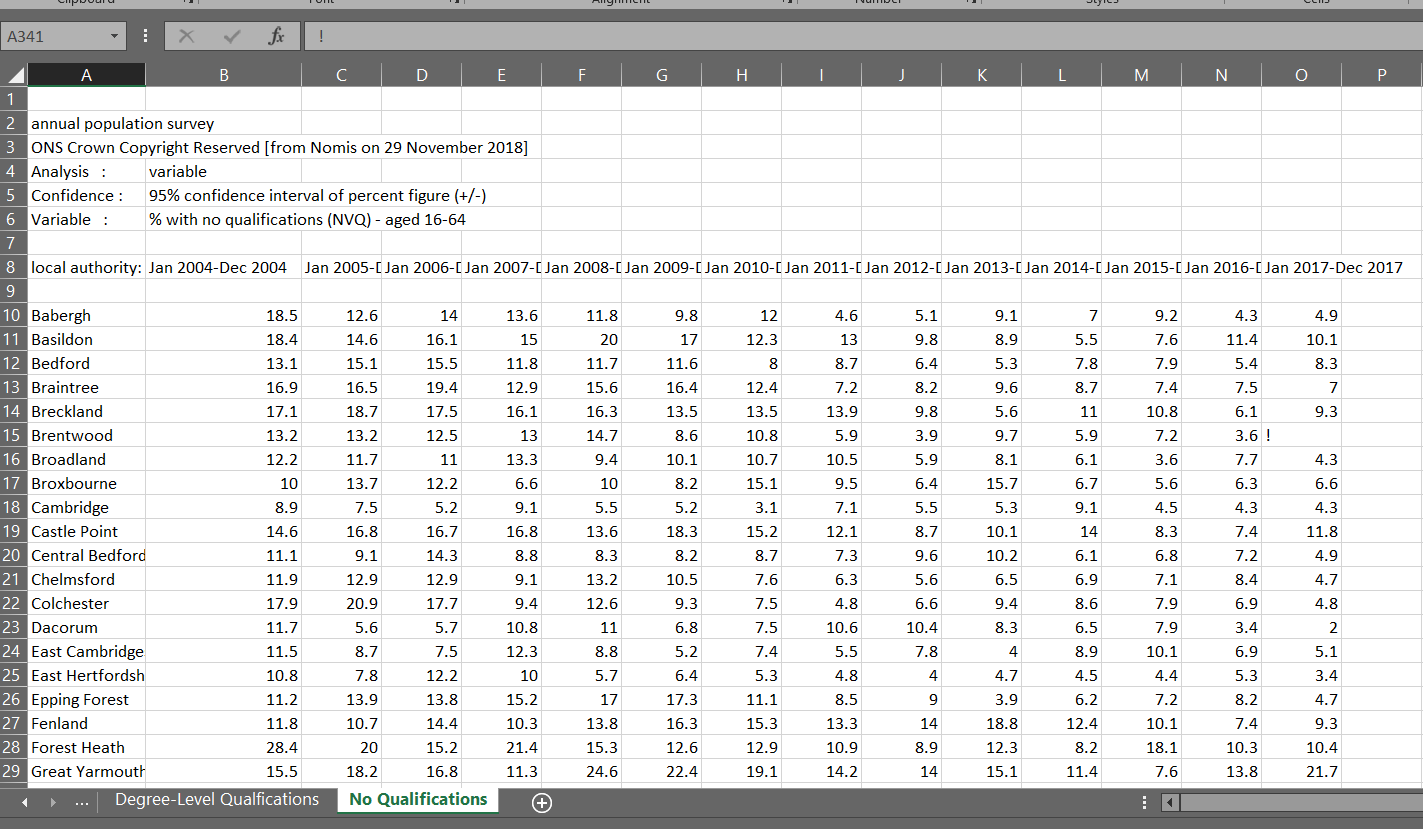
|  |  |
| --- | --- |
| **My Remainder Value:** | 2 |

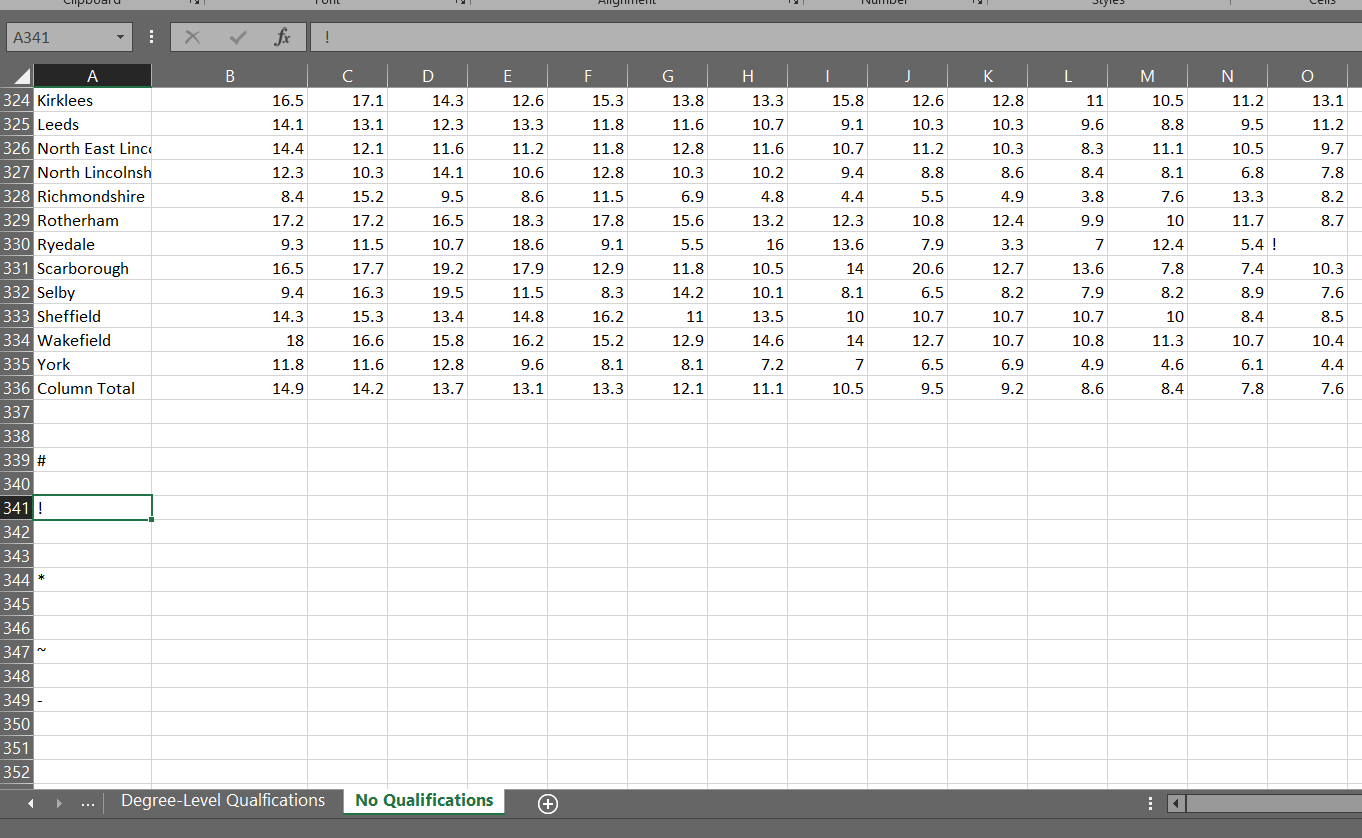
1. **Cleaning data:**

As per my remainder value, 2, my dataset is No Qualification.

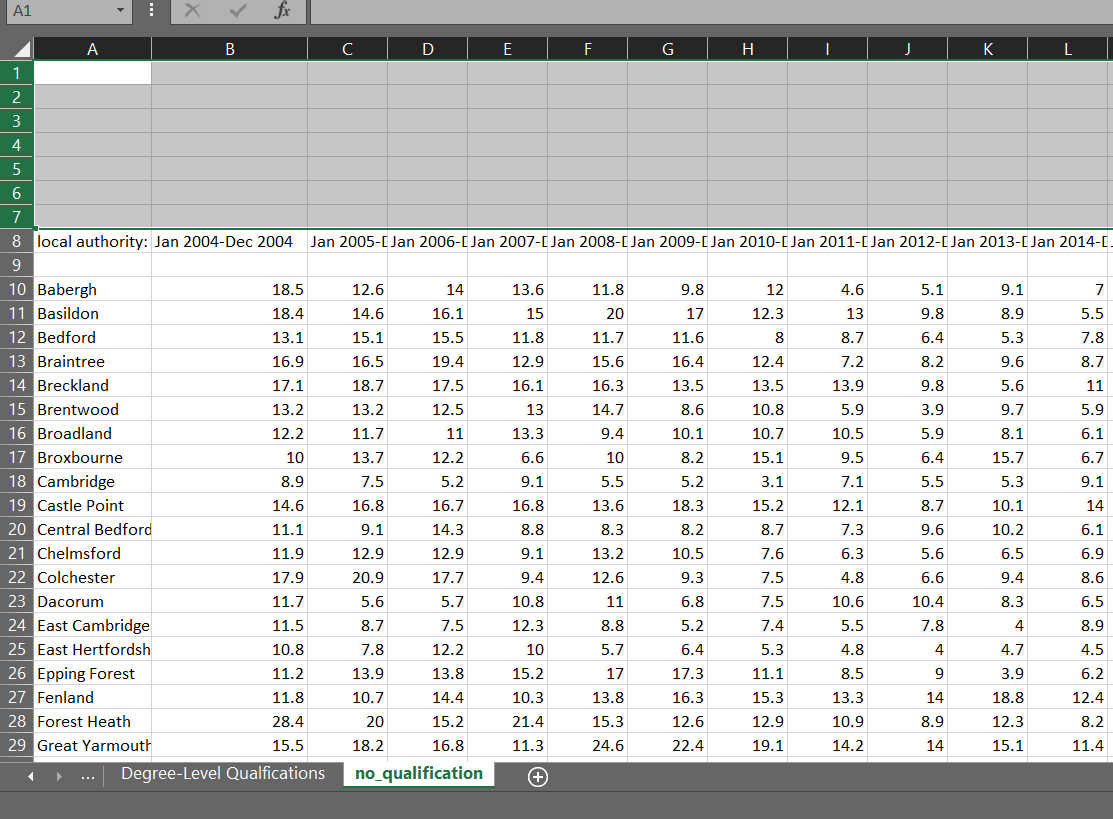
Before Cleaning data, it is good to make copy of it.

Default data representation before cleaning are:

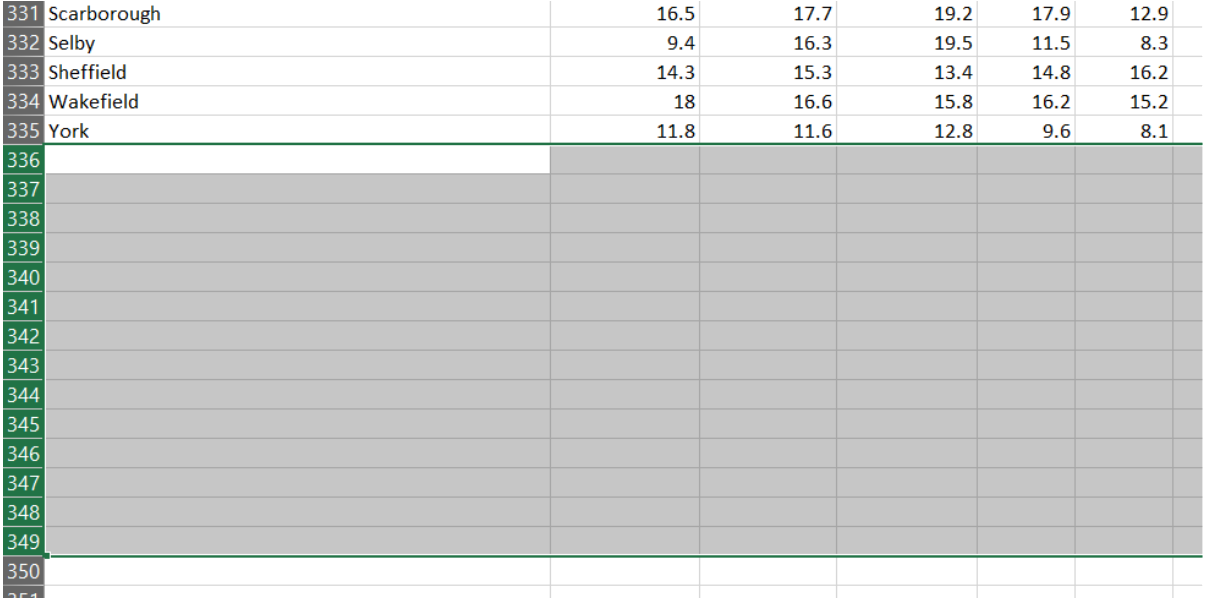
*Upper Header Section:*

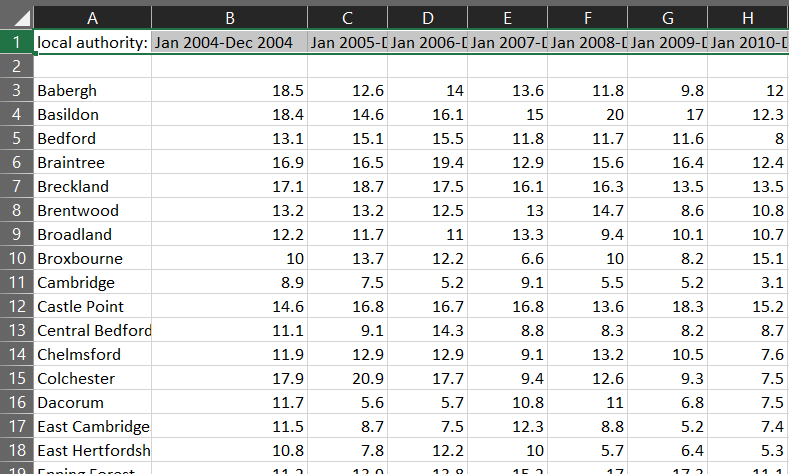
*Lower Footer Section:*

It is necessary to clean the data. Several cleaning activities will be carried out.

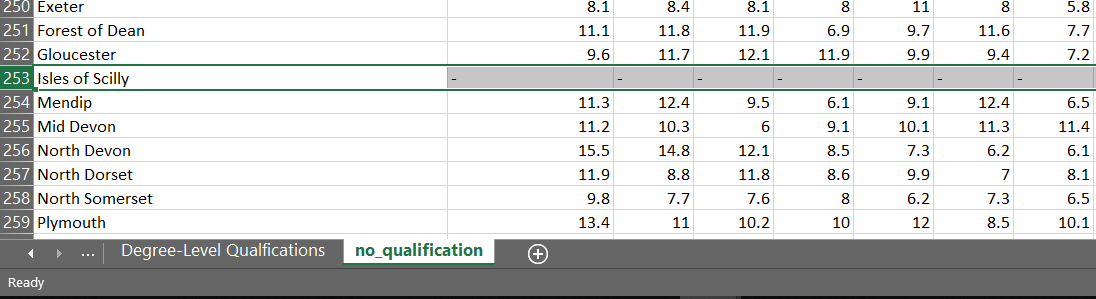
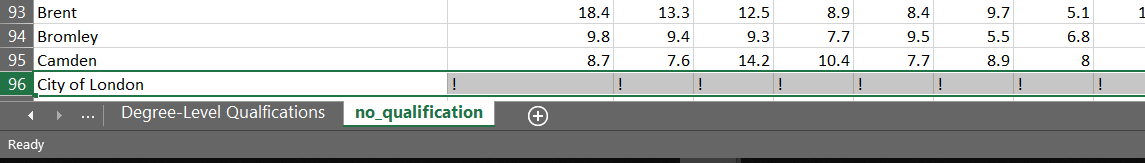
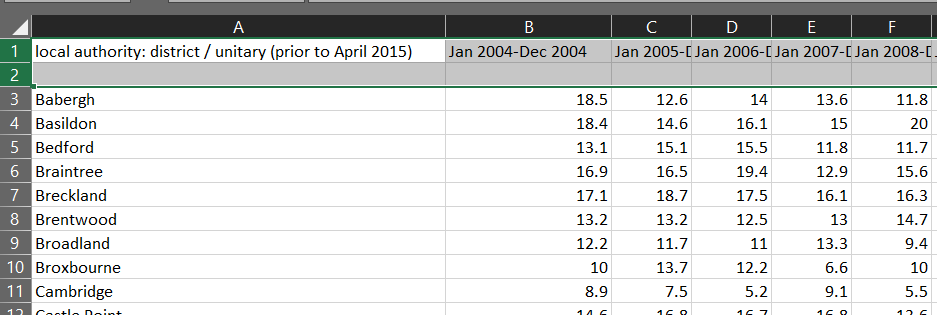
While importing data in oracle database, the data should be in tabular format. Renaming the table name into no\_qualification. The given data has some header section which can create problem. So, removing header section of the data up to 7th row.

Similarly, erasing data from lower section from 336th row to 349th row.



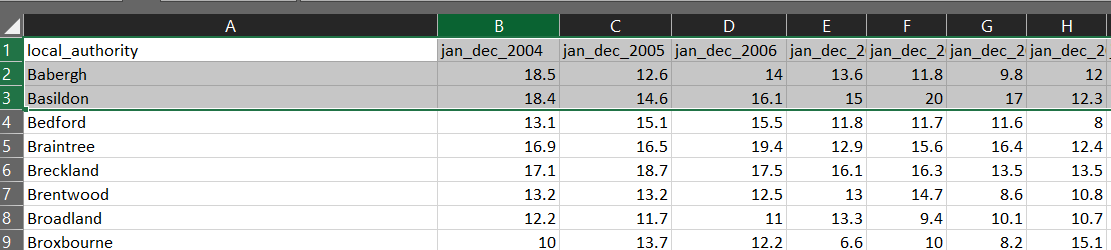
Upper 7 rows are empty so to maintain that, moving the table to A1.

Now, the table is in structured format. But still there are some missing values which can create problem in database.

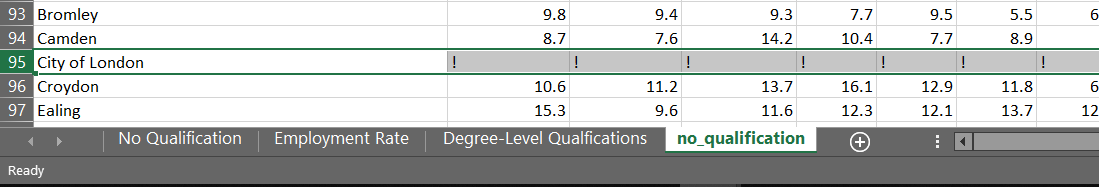
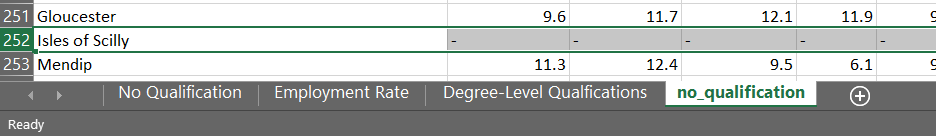


The column names, empty rows or #, ! -, values can be problematic.

So, removing empty row, renaming the columns name and representing in graph format for analysis.



Removing row of data of Isles of Scilly, South Bucks and city of London which has no values at all.



Some values are missing in some cities.

Last two years values of Melton are missing

Year 2011, 2012, 2016 and 2017 values are missing in city West Somerset.

Here are some values of South Bucks are missing which is represented as 0.

There are other cities whose values are missing. So, these things create problem, therefore it must be maintained. There are two different technique to add missing value.

1. Average: Taking average of the whole column and assigning the value in missing cell.

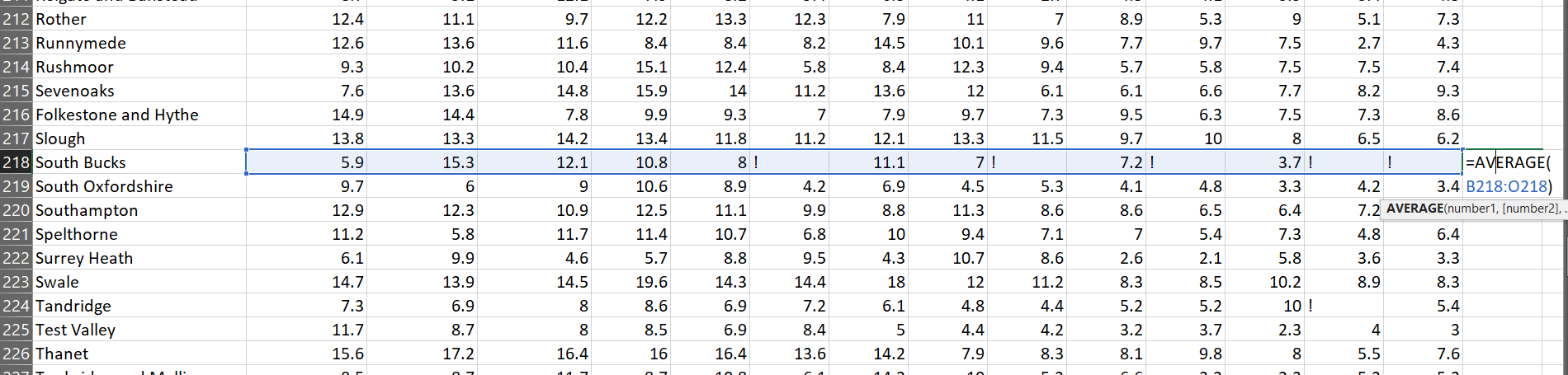
Average is useful if more than one values are missing in the column.

1. Median: Taking median of preceding and succeeding shell of the missing value and assigning the median value in missing cell.

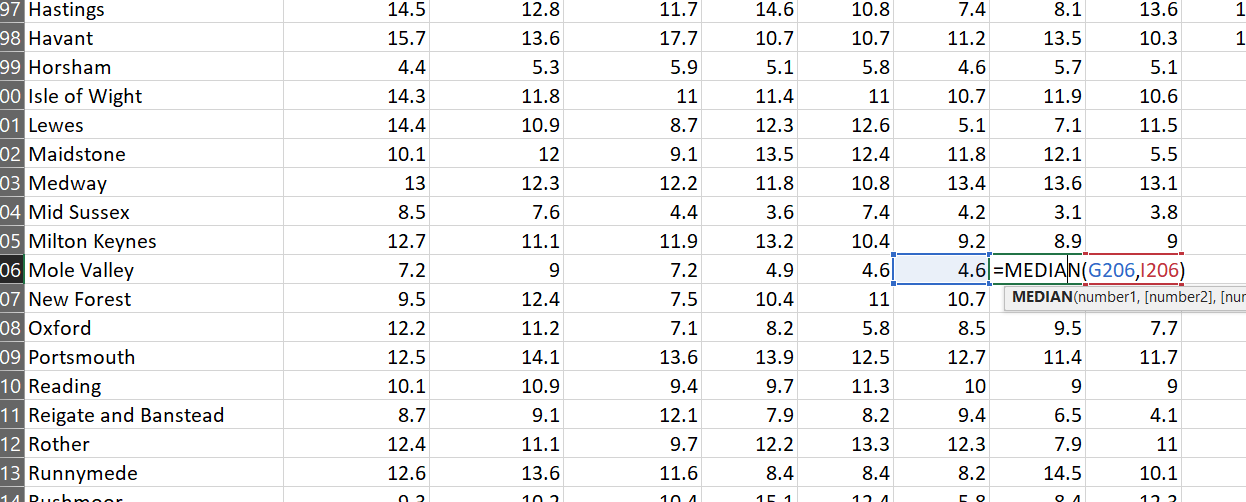
Median is useful if only few values are missing in the column.

In some cities, values are null. So, if many values are null average is suitable but if only one value is null then median is suitable.

In South Bucks, many values are missing so taking average and assigning value in it.

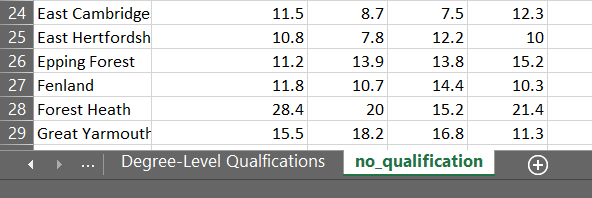


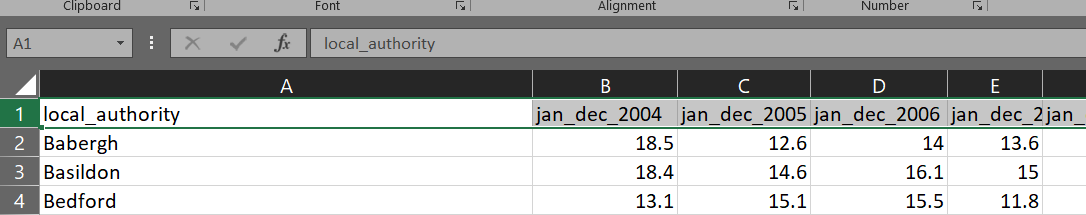
Obtained average value is assigned in ! values of South Bucks.

If few data are null then Median is taken and asingned.

Similarly adding vales in another empty cell.

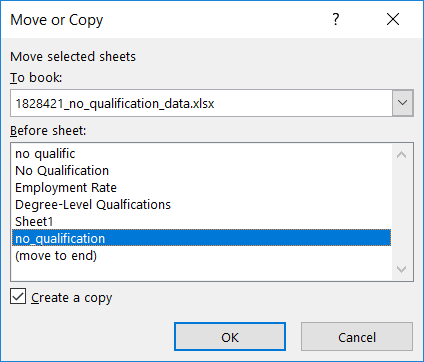
1. **Formatting and Importing**

At first renaming Data Name “No Qualification“ to “no\_qualification”.

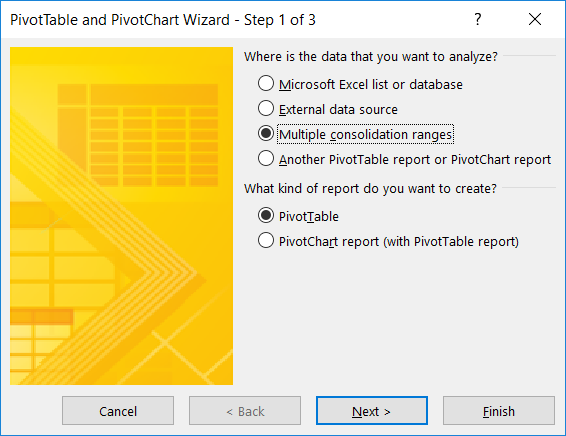
Columns name should be formatted well before importing it in the data base.

Now the data format is appropriate to import in oracle. But still the size of column is large and confusing which can be maintained by converting the table into pivot table.

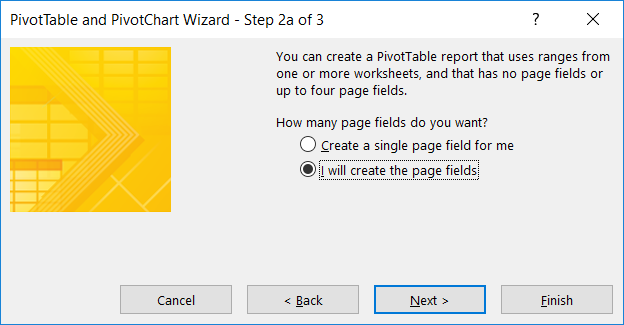
For this at first creating a copy of the page



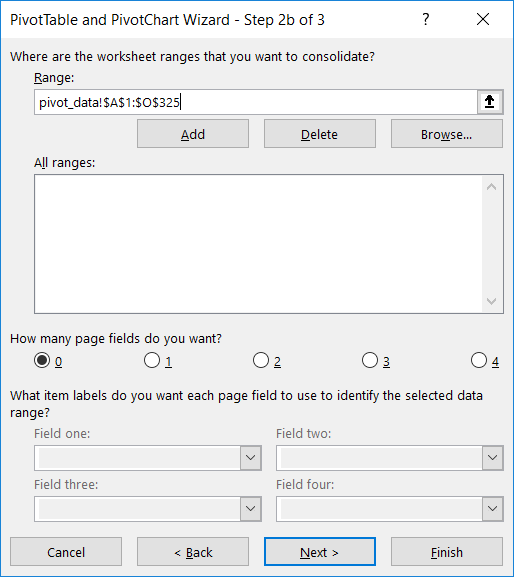
Now renaming the file name as pivot\_data. Pressing Crtl + D and then P which gives the dialog box.



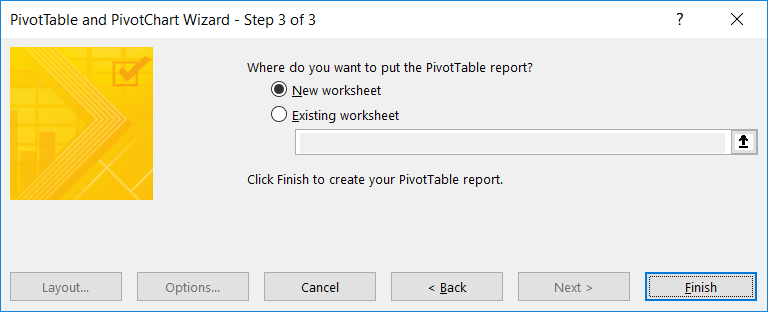
Selecting multiple consolidation ranges and Pivot table then hitting next.



Select I will create the page field.

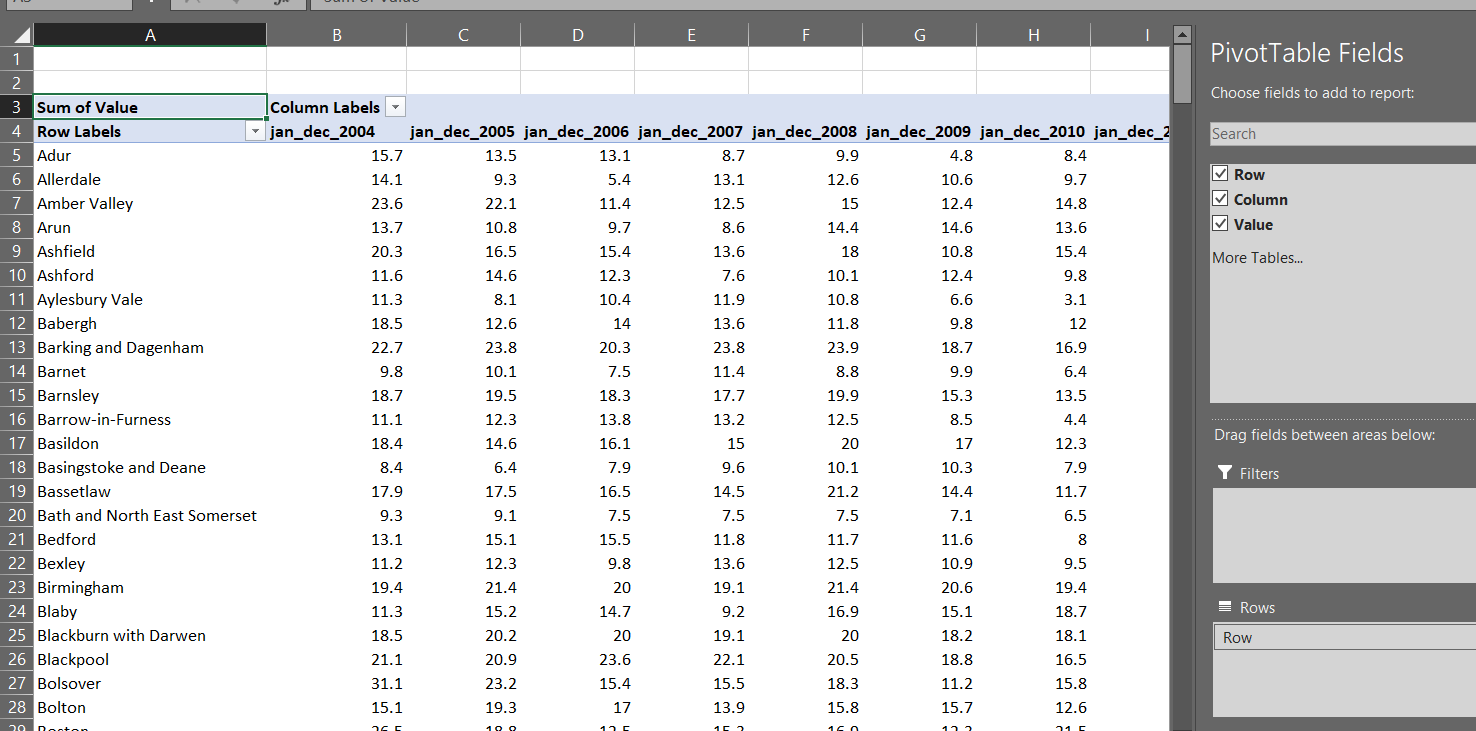


Selecting All ranges and hitting next.

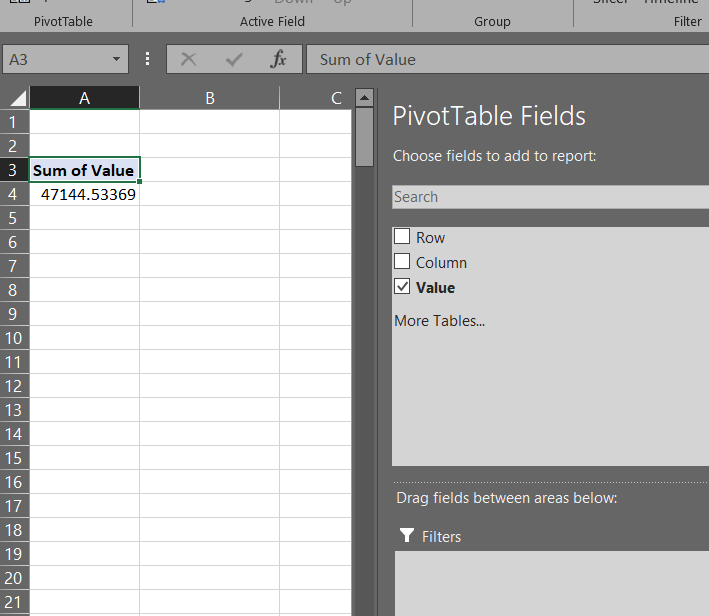


Leaving the default New worksheet and finish.

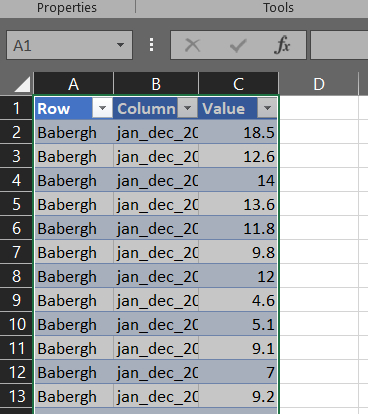
New sheet is displayed



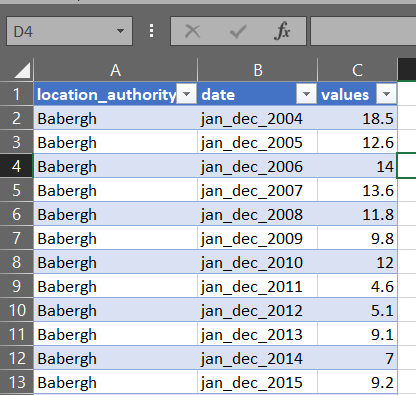
Unchecking the row and column which gives the total sum.



Double click on the value (cell A4).

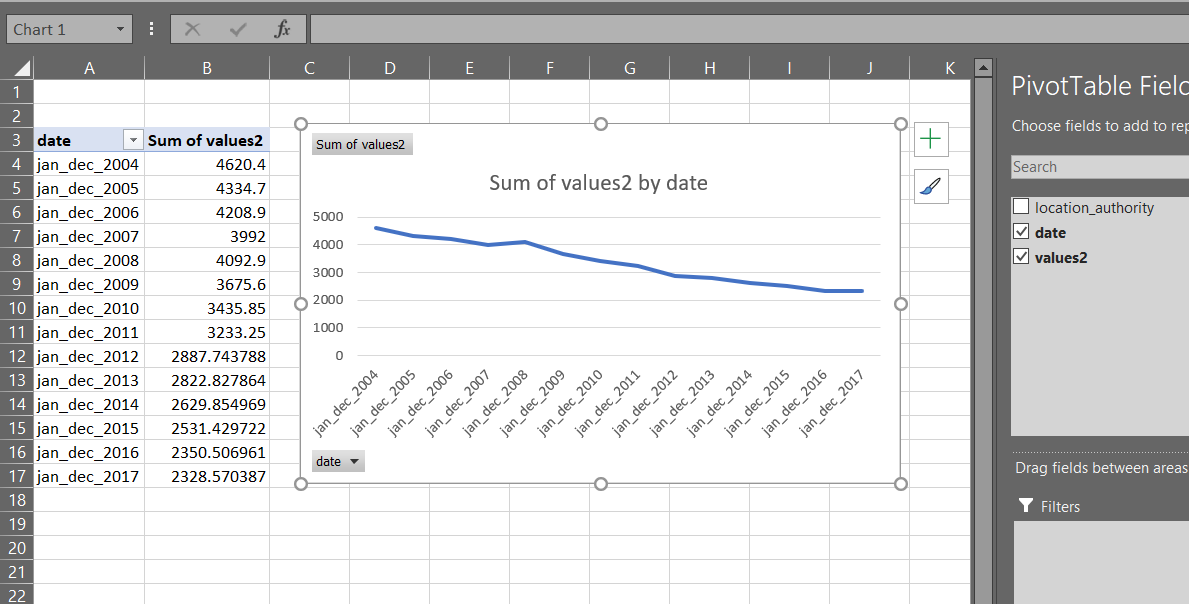


The new pivot table is ready. Renaming the row column and value as location\_authority, date and values.

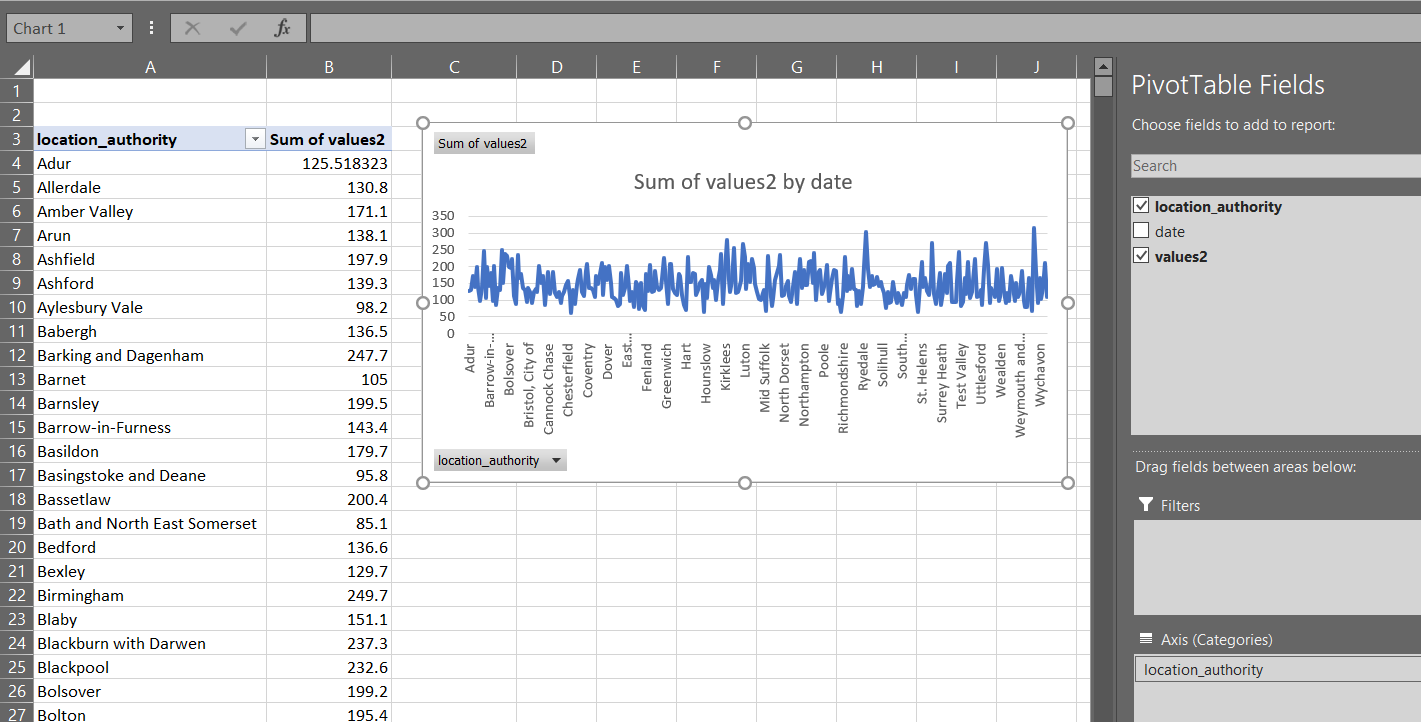


Renaming the sheet name as pivot\_table

The obtained data representation:

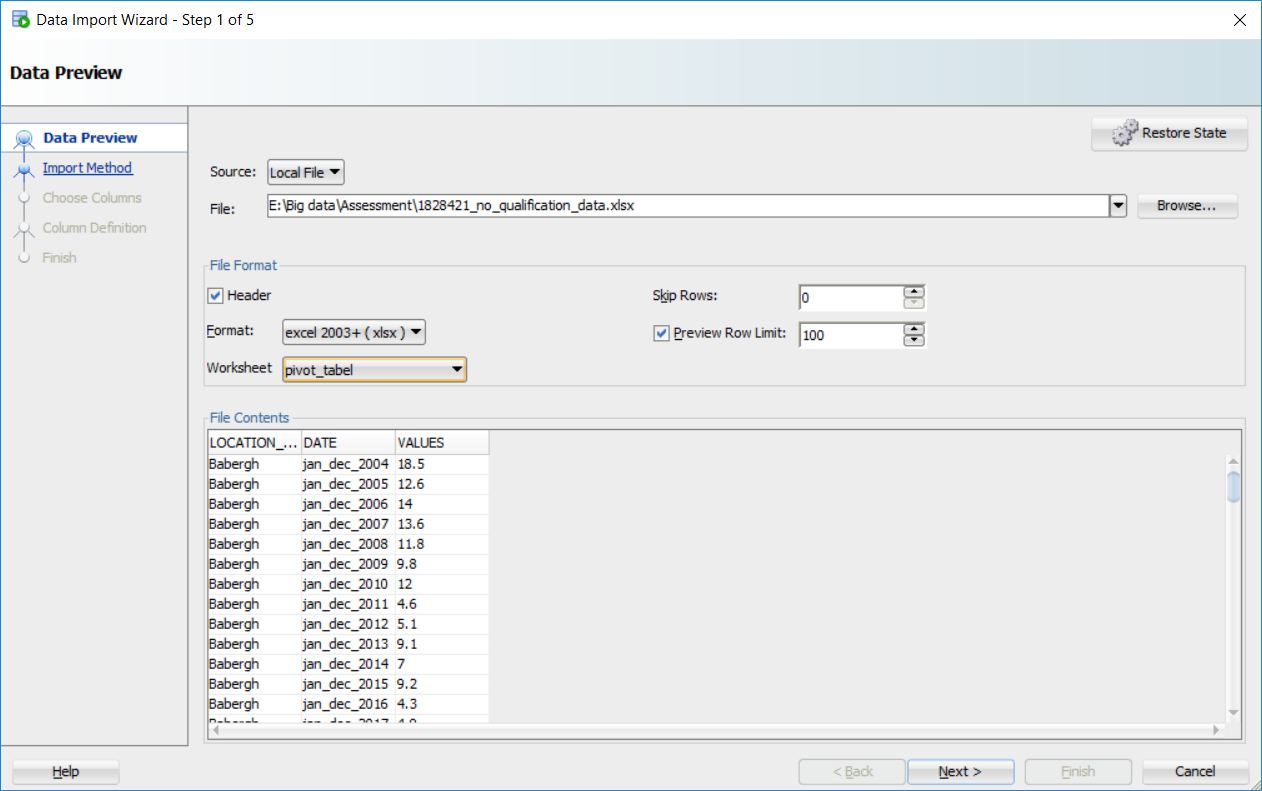


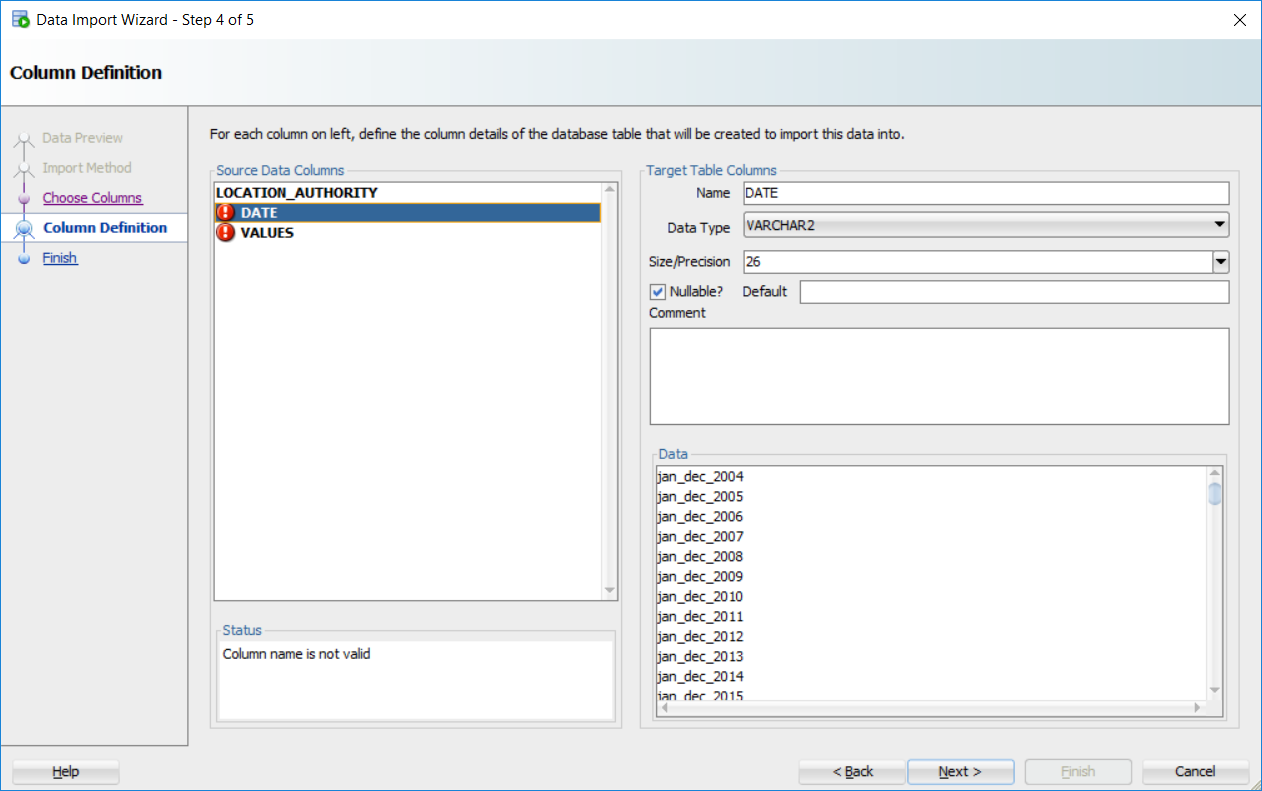
Without location\_authority



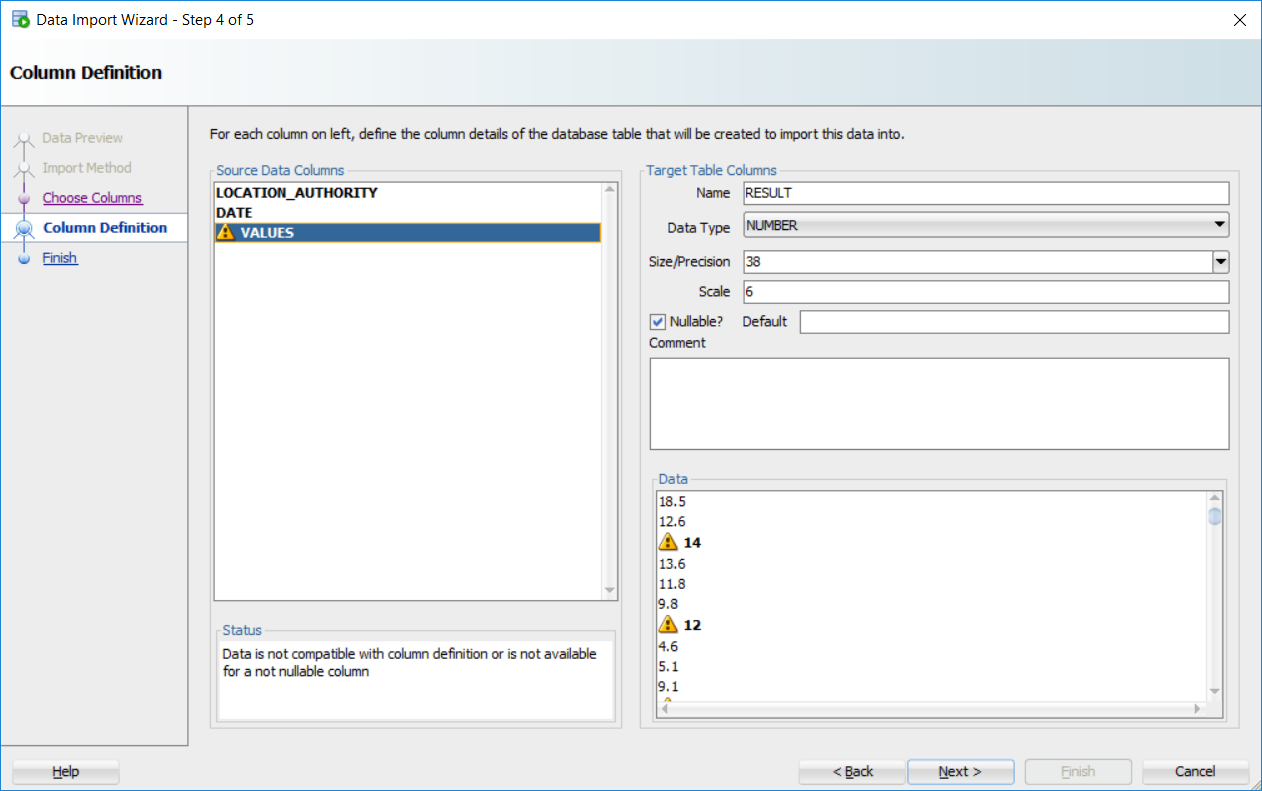
Without date field, since there are huge locations, so the line graph is huge.

Now the data is ready to import in oracle.



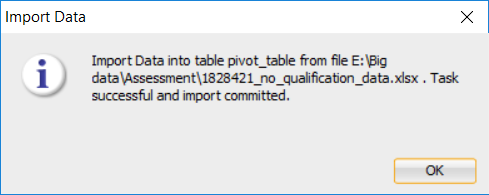


local\_authority is column name is checked not null and increasing the size/precision from 38 to 70 since location name are larger than 38. The date and values column is not valid because it is sql keywords of oracle which can create problem while extracting data. So, changing date as year and values as result.

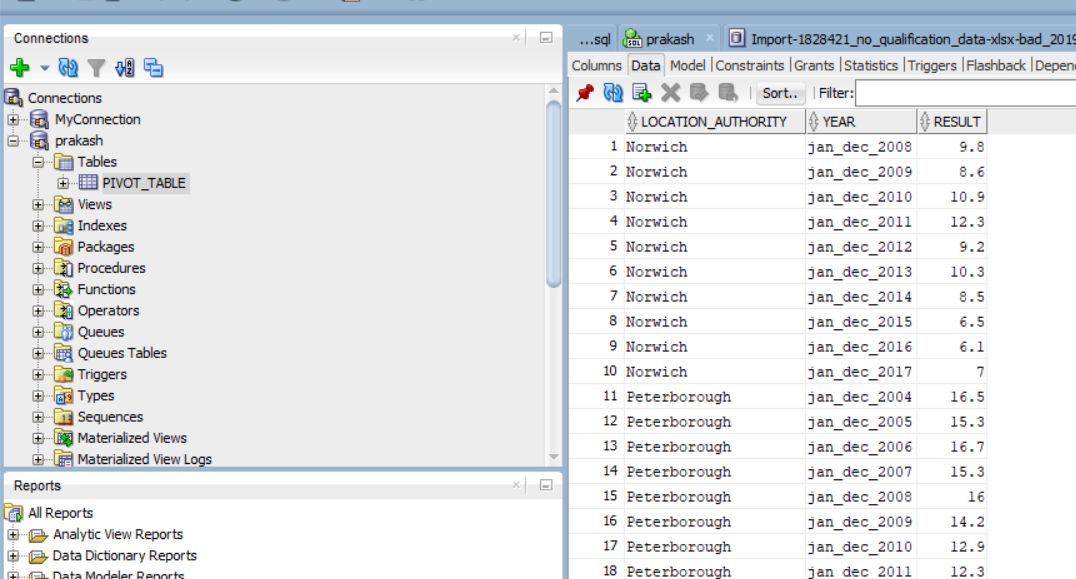


Values gives warning because of its datatype. Some values in result are integer and some are in decimal. Therefore, warning is given. If datatype is changed to varchar2 it won’t give any warning, but we must need its values in number format, neglecting the warning and leaving it as NUMBER datatype.

The default Data Type is in number format which gives warning because all values do not have decimal numbers. Since arithmetic operation is required leaving the data type as it is.



The structured table is imported successfully into database.



1. **SQL**

ROLLUP query presents data in hierarchical format where CUBE presents query in combination form. Therefore, more null values will be obtained from CUBE query.

Two sql queries which demonstrates an Online Analytical Processing (OLAP) query is presented below:

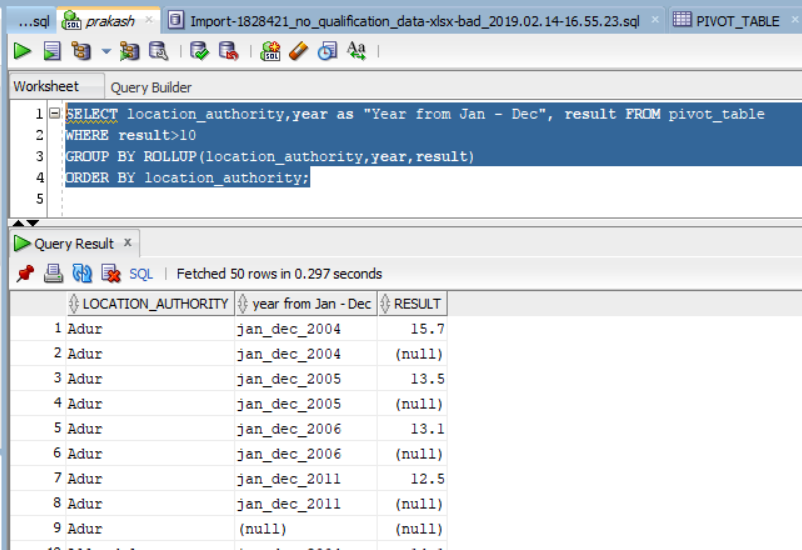
1. **ROLLUP Query:**

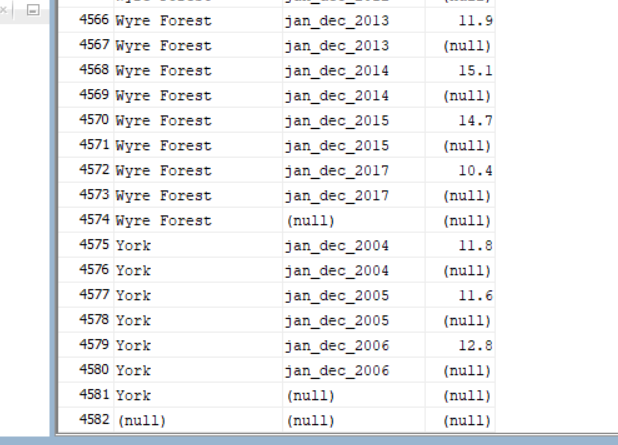
SELECT location\_authority,year as "Year from Jan - Dec", result FROM pivot\_table

WHERE result>10

GROUP BY ROLLUP(location\_authority,year,result)

ORDER BY location\_authority;



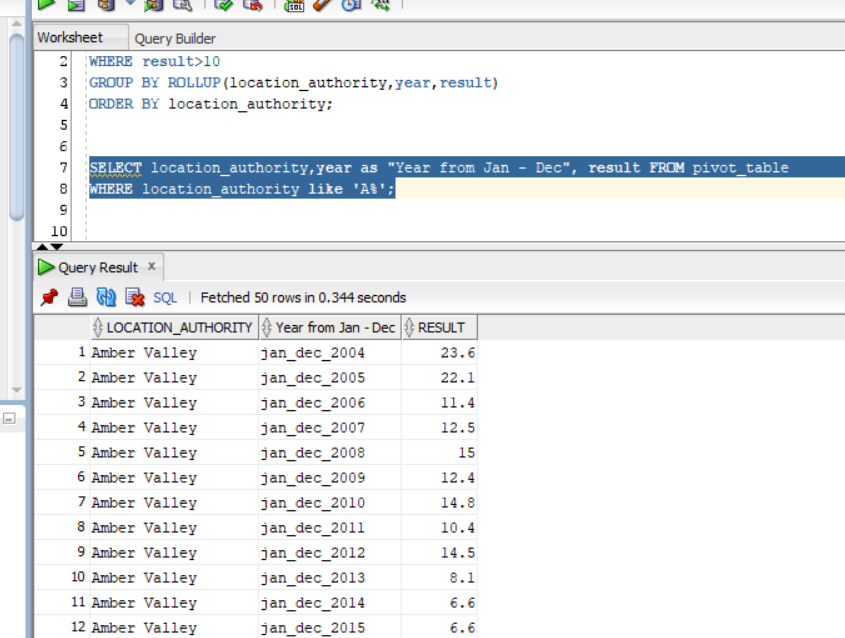


This query selects all three columns, year as given allies, from the table pivot\_table. It only displays the output if the result values are greater than 10. It is groped using ROLLUP function which gives result in hierarchy, which means result column can be null but if its upper column is null then it must be null similarly if location\_authority is null then year is also null, so the result column is also null. Finally, when it gives output, it is sorted in an ascending order (i.e. A-Z) by location\_authority column.

1. **Group By Query:**

SELECT location\_authority,year as "Year from Jan - Dec", result FROM pivot\_table

WHERE location\_authority like 'A%';



This query selects all three columns from pivot\_table of only the column location\_authority which starts with A.

1. **Advantages**

The advantage of using this approach in big data are explained below:

1. Data cleaning:

The data which we may not be in proper format but in this approach since the data should be in an appropriate format, it must be cleaned so Null, missing values can be assigned taking its average or median. Because of this data analysis can be effective.

1. **Disadvantages**

Disadvantage of this approach for using big data are given below:

1. Deletion of inappropriate data:

If the data is not in a structured format, then it must be cleaned. If there are null data, then it should be filled. But if many values are missing then it is difficult to assign values in it. So, it the whole row or column must be removed which may result the deletion of important values as well.