Big Data Analytics

Isara Anantavrasilp

Lecture 12: Data Analysis with Pivot Table

Pivot Table

- **Pivot Table**: Data table that summarizes of more extensive tables
- The summaries may include
  - Sum
  - Average
  - Other stats

Pivot Table Example

Singapore Residents by Sex – June – Annual / data.gov.sg

Year	Type	Residents
1957	Total Residents	1445929
1957	Total Male Residents	762760
1957	Total Female Residents	683169
1958	Total Residents	1518800
1958	Total Male Residents	797600
1958	Total Female Residents	721200
1959	Total Residents	1587200
1959	Total Male Residents	830800
1959	Total Female Residents	756400
1960	Total Residents	1646400
1960	Total Male Residents	859600
1960	Total Female Residents	786800

Questions

- How many residents does Singapore have in 1957?
- How many male and female does Singapore have in 1958?
- How many residents does Singapore have between 1957 – 1960?
- Compare male and female residents from 1957 – 1960.

Pivoting Data

- Such questions are very difficult and time-consuming for human
- Fortunately, virtually all spreadsheet software have pivoting features
- Software that is specialized in multi-dimensional data (**Data Cube**) analysis called **Business Intelligence**

Pivot Table in Excel

- Download the following file: https://data.gov.sg/dataset/resident-population-by-ethnicity-gender-and-age-group
- There are two files with and without age groups
- Open the one **with** age group in Excel

	A	B	C	D
1	year	level_1	level_2	value
2	1957	Total Residents	0 - 4 Years	264727
3	1957	Total Residents	5 - 9 Years	218097
4	1957	Total Residents	10 - 14 Years	136280
5	1957	Total Residents	15 - 19 Years	135679
6	1957	Total Residents	20 - 24 Years	119266
7	1957	Total Residents	25 - 29 Years	111726
8	1957	Total Residents	30 - 34 Years	89925

- You will see that the data are **stacked**
- See following page for more detail on **transpose** and **stack data**
- https://www.extendoffice.com/documents/excel/4235-excel-transpose-and-stack.html

Pivot Table in Excel

- Change the column names
  - level_1 -> Ethnic Group
  - level_2 -> Age Group
  - values -> Residents

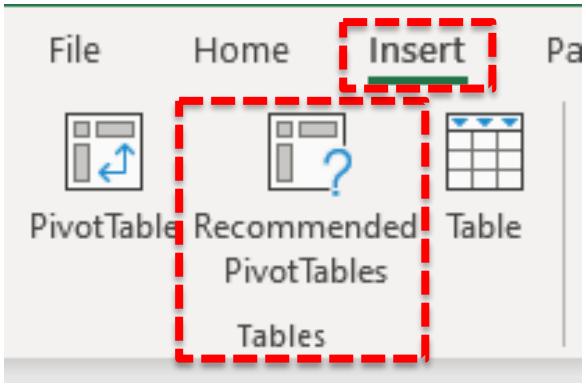
	A	B	C	D
1	Year	Ethnic Group	Age Group	Residents
2	1957	Total Residents	0 - 4 Years	264727
3	1957	Total Residents	5 - 9 Years	218097
4	1957	Total Residents	10 - 14 Years	136280
5	1957	Total Residents	15 - 19 Years	135679
6	1957	Total Residents	20 - 24 Years	119266
7	1957	Total Residents	25 - 29 Years	111726

Pivot Table in Excel

- Select all columns

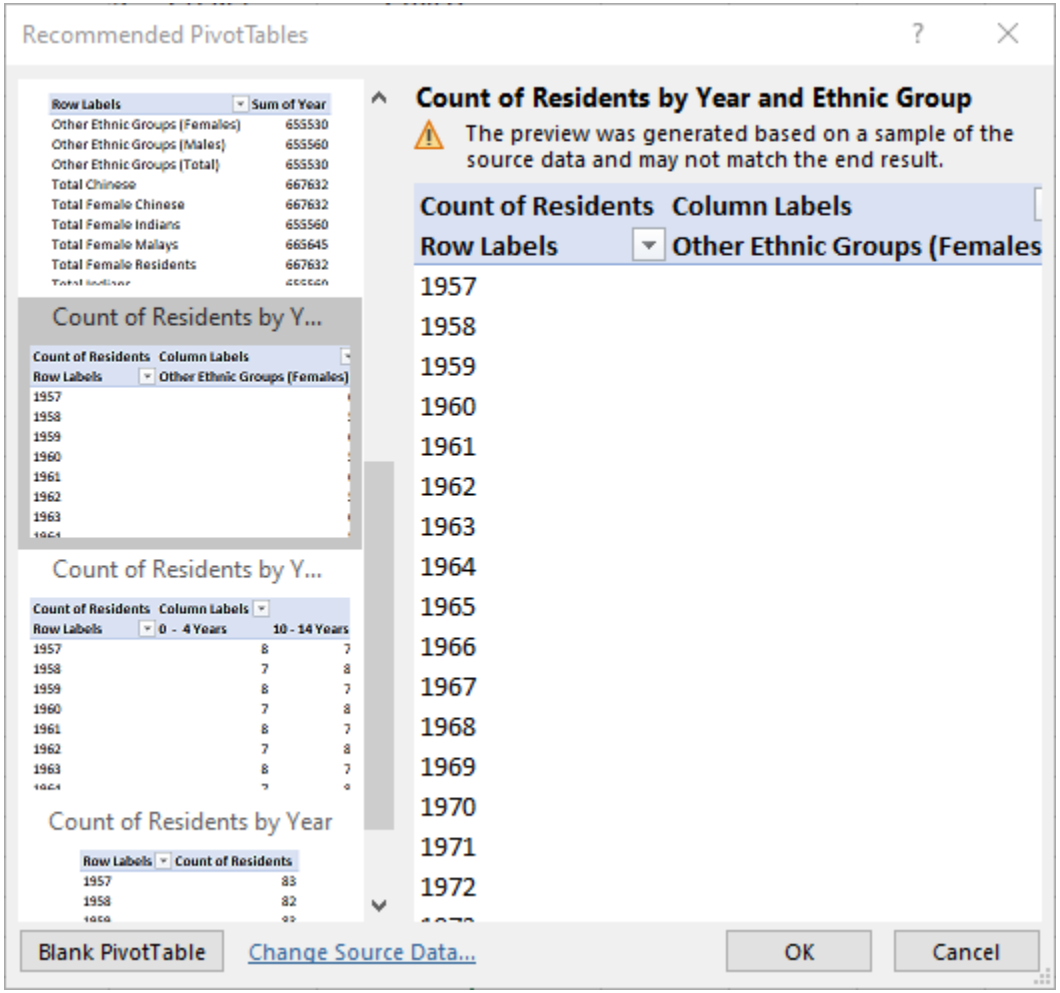
	A	B	C	D
1	Year	Ethnic Group	Age Group	Residents
2	1957	Total Residents	0 - 4 Years	264727
3	1957	Total Residents	5 - 9 Years	218097
4	1957	Total Residents	10 - 14 Years	136280
5	1957	Total Residents	15 - 19 Years	135679
6	1957	Total Residents	20 - 24 Years	119266

- And go to menu **Insert** and select **Recommend PivotTables**

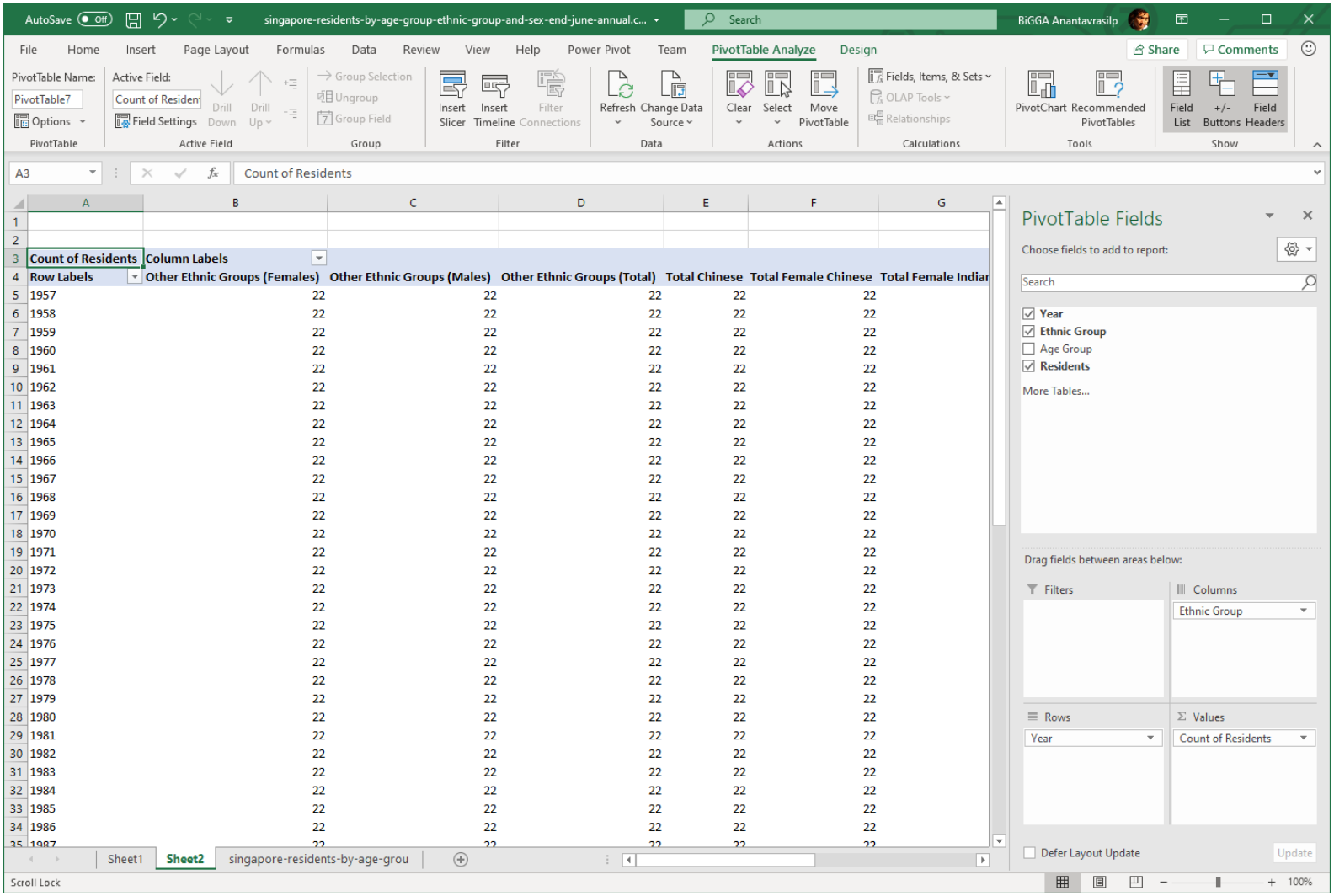


Pivot Table in Excel

- Excel will ask for preferred pivot table
- We want to start from **Year**, so we can select **Count of Residents by Year and Ethnic Group**

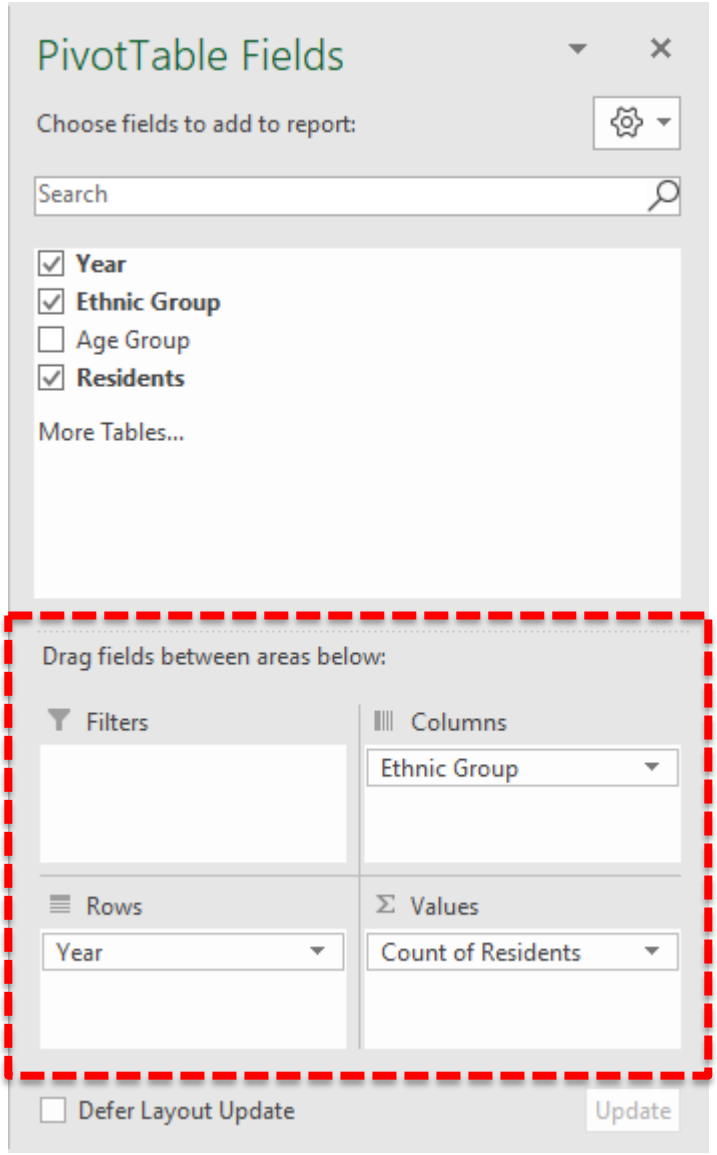


(Confusing) Pivot Table!



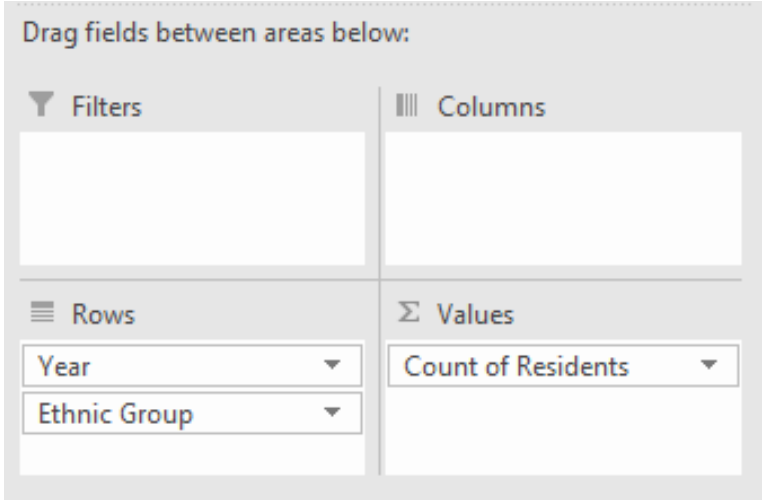
Select what to see

- In our data, the info that we want to know is the **Number of Residents**
- Other fields are only attributes
  - They only specify the constraints or **groups** of data
- Thus, we have to tell Excel which field is the **value** and which are **data groups**



Select what to see

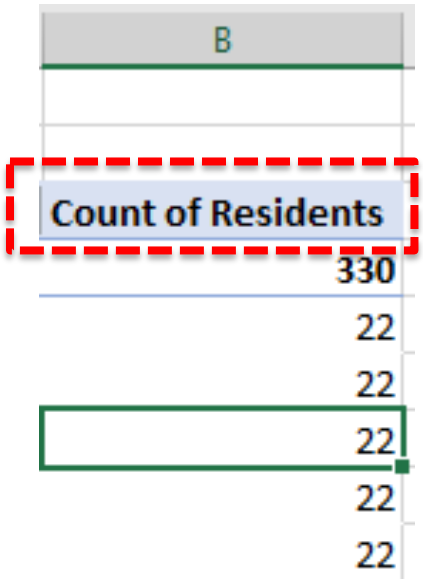
- Drag **Year** and **Ethnic Group** to **Rows**
- Drag **Count of Residents** to **Values**
- Note that the data are now group by **Year** and **Ethnic Group**

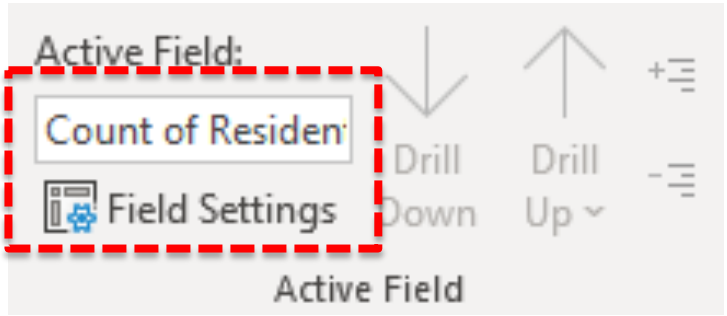


Row Labels	Count of Residents
1957	330
Other Ethnic Groups (Females)	22
Other Ethnic Groups (Males)	22
Other Ethnic Groups (Total)	22
Total Chinese	22
Total Female Chinese	22
Total Female Indians	22
Total Female Malays	22
Total Female Residents	22
Total Indians	22
Total Malays	22
Total Male Chinese	22
Total Male Indians	22
Total Male Malays	22
Total Male Residents	22
Total Residents	22
1958	330
Other Ethnic Groups (Females)	22
Other Ethnic Groups (Males)	22
Other Ethnic Groups (Total)	22
Total Chinese	22
Total Female Chinese	22

Change the Type of Data Fields

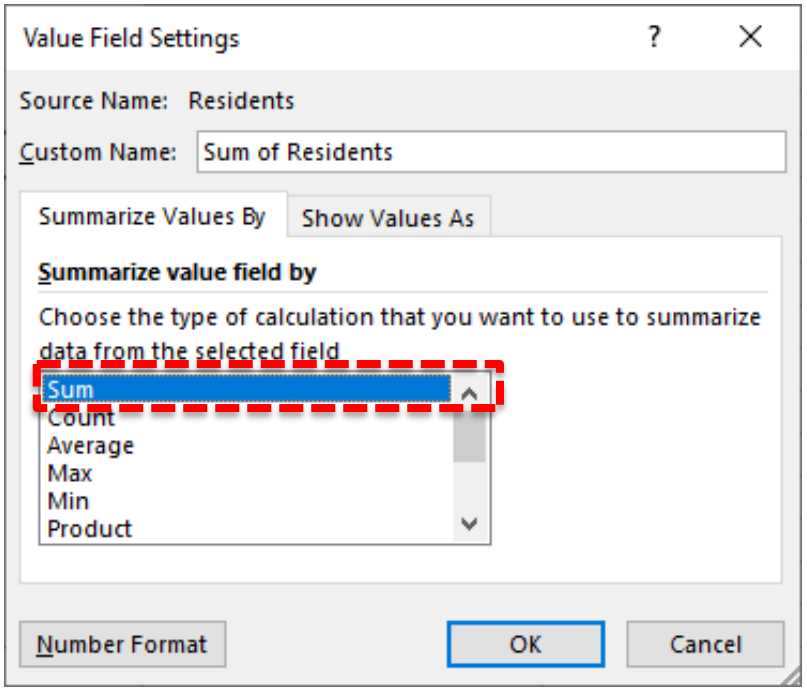
- The **Values** is not the sum of **Residents**, but the counts
- Let's change to sum by
  - Select any row in Count of Residents column (Col B)
  - Make sure Active Field is the same
  - Click Field Settings





Change the Type of Data Fields

- Select Summarize value field by Sum, instead of Count
- You should see sum of all residents in each data group



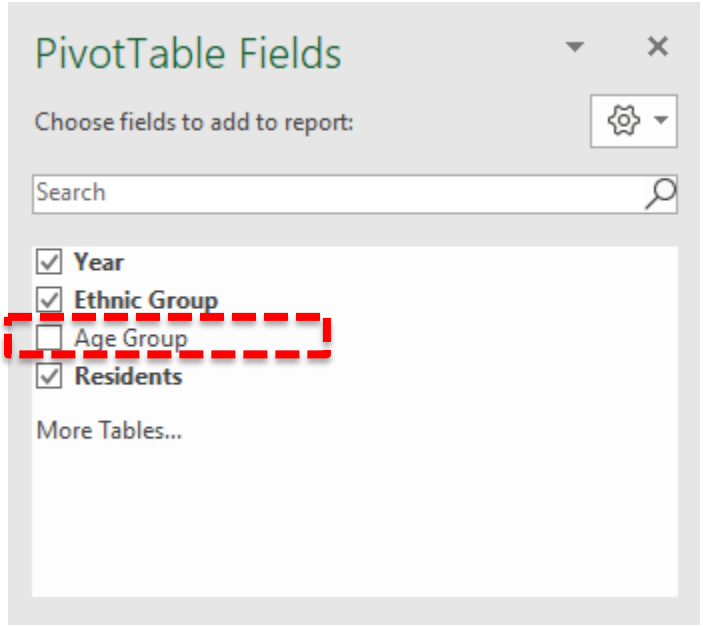
Row Labels	Sum of Residents
1957	6008812
1958	6210000
Other Ethnic Groups (Females)	15200
Other Ethnic Groups (Males)	14800
Other Ethnic Groups (Total)	30000
Total Chinese	1178100
Total Female Chinese	582700
Total Female Indians	42400
Total Female Malays	101400
Total Female Residents	741700
Total Indians	132400
Total Malays	212000

Mini Exercise (1)

- There are some redundant entries
  - The data include total residents of each Ethnic Group
  - They also include total residents of male and female
- Remove those multiple entries using filtering feature
  - clicking the down arrow symbol behind the column name

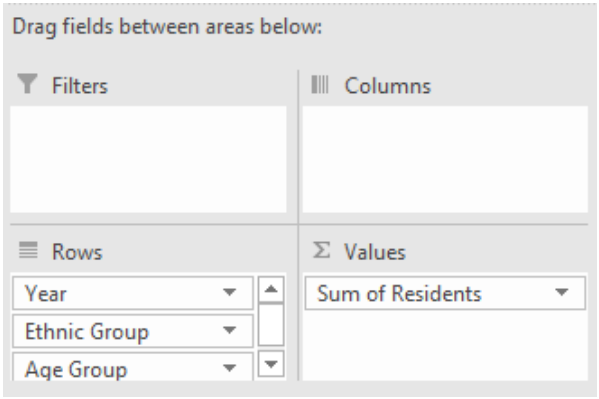
Mini Exercise (2)

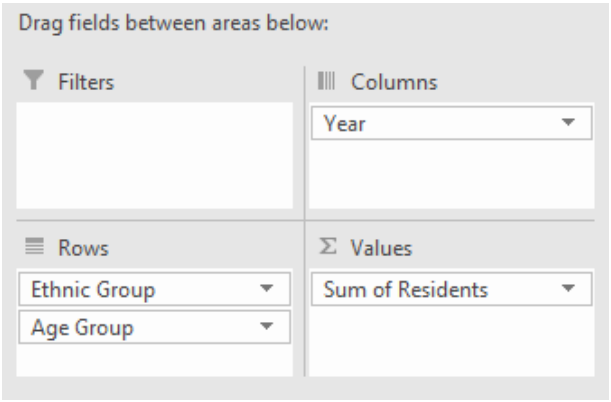
- Try to add / remove pivot table fields
- See if you can make sense of your data

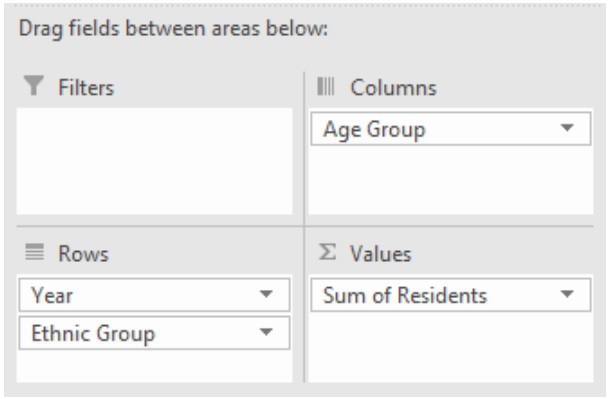


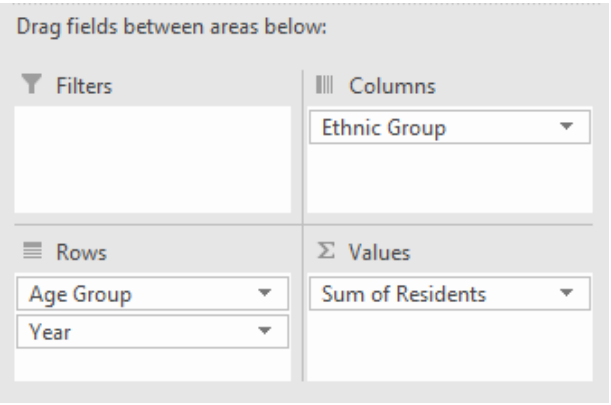
Mini Exercise (3)

- Change the fields in row and column and see how your table changes



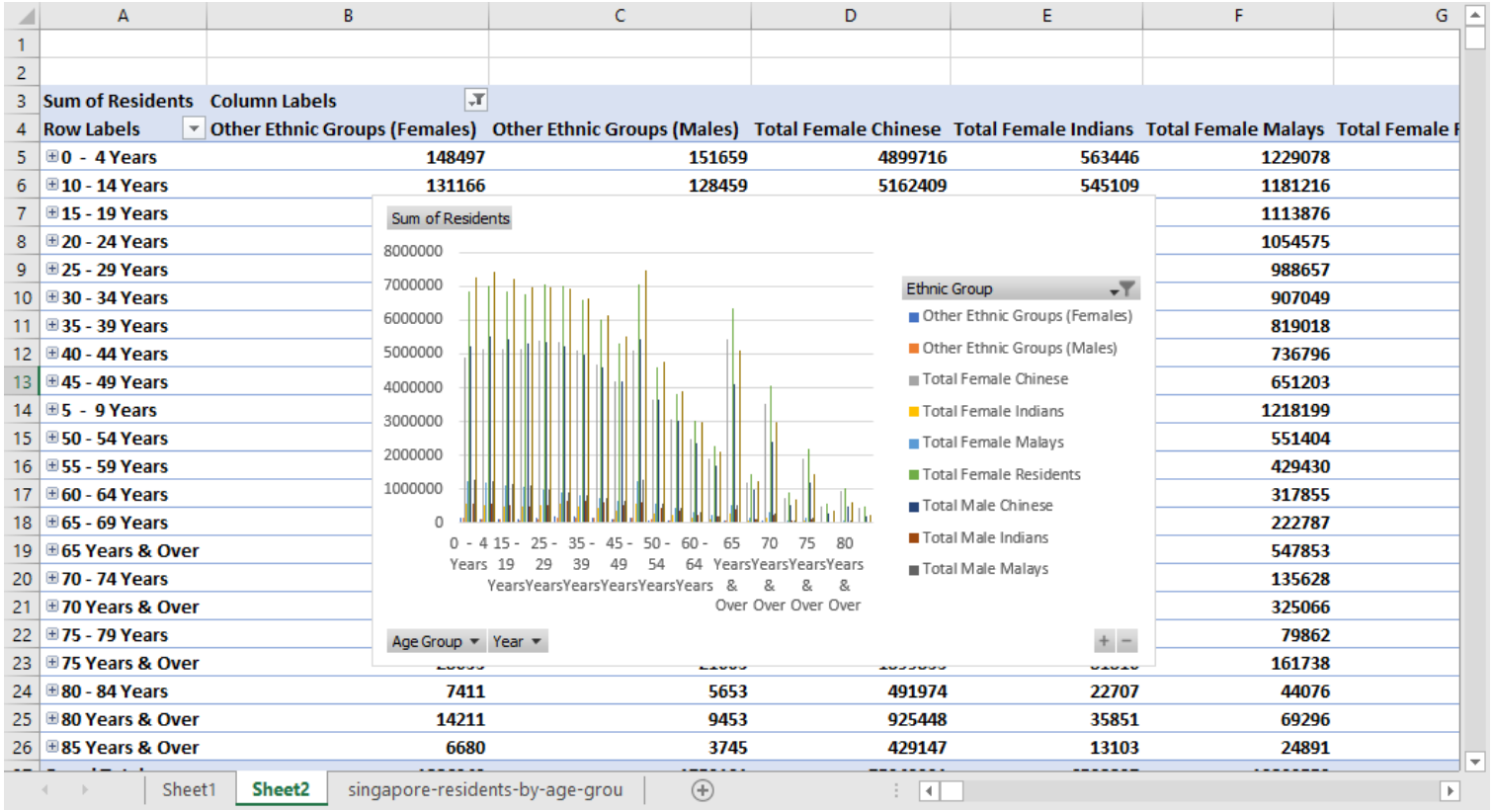






Mini Exercise (4)

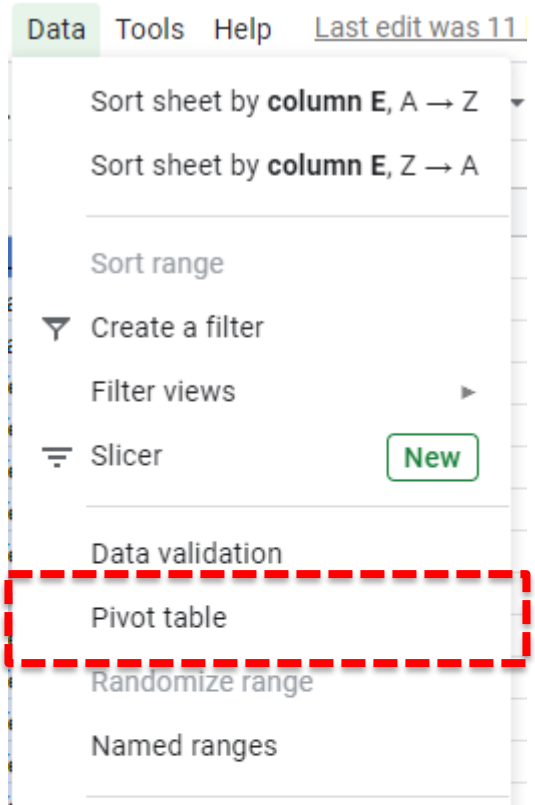
- Collapsing the groups to hide some information
- Then plot a graph

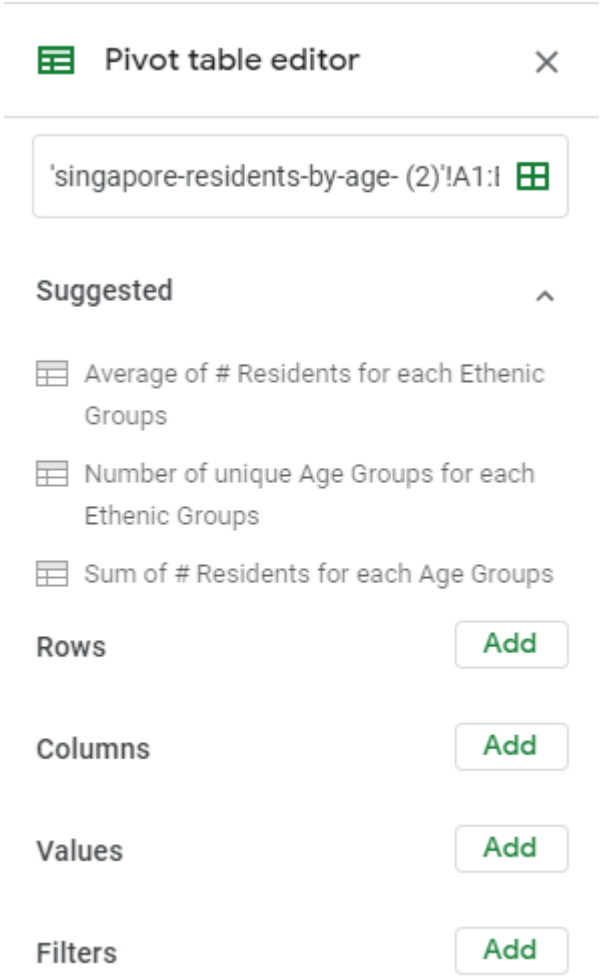


Sum of Residents	Column Labels					
Row Labels	Other Ethnic Groups (Females)	Other Ethnic Groups (Males)	Total Female Chinese	Total Female Indians	Total Female Malays	Total Female F
0 - 4 Years	148497	151659	4899716	563446	1229078	
10 - 14 Years	131166	128459	5162409	545109	1181216	
15 - 19 Years					1113876	
20 - 24 Years					1054575	
25 - 29 Years					988657	
30 - 34 Years					907049	
35 - 39 Years					819018	
40 - 44 Years					736796	
45 - 49 Years					651203	
5 - 9 Years					1218199	
50 - 54 Years					551404	
55 - 59 Years					429430	
60 - 64 Years					317855	
65 - 69 Years					222787	
65 Years & Over					547853	
70 - 74 Years					135628	
70 Years & Over					325066	
75 - 79 Years					79862	
75 Years & Over					161738	
80 - 84 Years	7411	5653	491974	22707	44076	
80 Years & Over	14211	9453	925448	35851	69296	
85 Years & Over	6680	3745	429147	13103	24891	

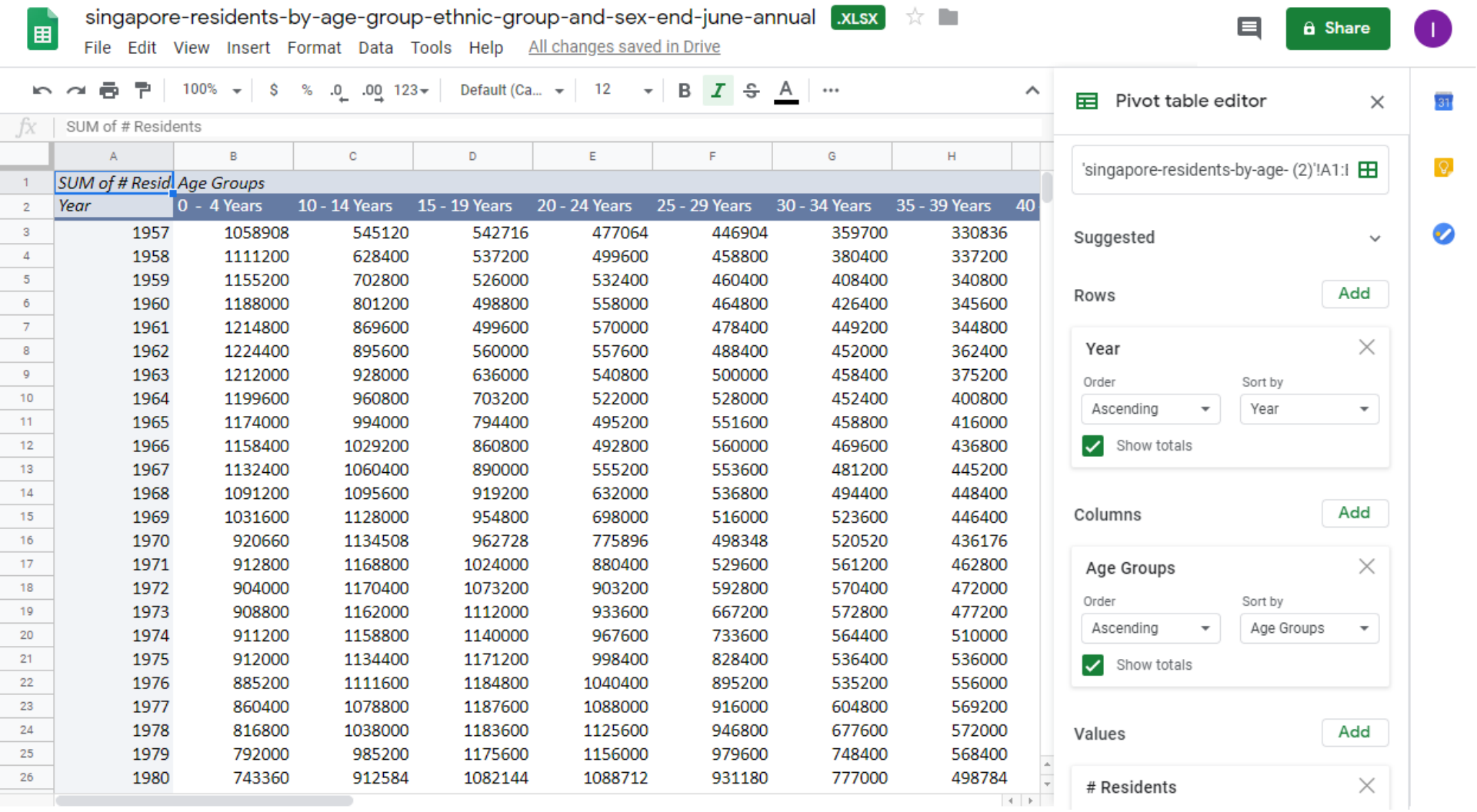
Pivot Table in Google Sheet

- Open the data file
- Select Data -> Pivot Table
- Pivot table editor will appear





Pivot Table in Google Sheet



SUM of # Resid	Age Groups							
Year	0 - 4 Years	10 - 14 Years	15 - 19 Years	20 - 24 Years	25 - 29 Years	30 - 34 Years	35 - 39 Years	40
1957	1058908	545120	542716	477064	446904	359700	330836	
1958	1111200	628400	537200	499600	458800	380400	337200	
1959	1155200	702800	526000	532400	460400	408400	340800	
1960	1188000	801200	498800	558000	464800	426400	345600	
1961	1214800	869600	499600	570000	478400	449200	344800	
1962	1224400	895600	560000	557600	488400	452000	362400	
1963	1212000	928000	636000	540800	500000	458400	375200	
1964	1199600	960800	703200	522000	528000	452400	400800	
1965	1174000	994000	794400	495200	551600	458800	416000	
1966	1158400	1029200	860800	492800	560000	469600	436800	
1967	1132400	1060400	890000	555200	553600	481200	445200	
1968	1091200	1095600	919200	632000	536800	494400	448400	
1969	1031600	1128000	954800	698000	516000	523600	446400	
1970	920660	1134508	962728	775896	498348	520520	436176	
1971	912800	1168800	1024000	880400	529600	561200	462800	
1972	904000	1170400	1073200	903200	592800	570400	472000	
1973	908800	1162000	1112000	933600	667200	572800	477200	
1974	911200	1158800	1140000	967600	733600	564400	510000	
1975	912000	1134400	1171200	998400	828400	536400	536000	
1976	885200	1111600	1184800	1040400	895200	535200	556000	
1977	860400	1078800	1187600	1088000	916000	604800	569200	
1978	816800	1038000	1183600	1125600	946800	677600	572000	
1979	792000	985200	1175600	1156000	979600	748400	568400	
1980	743360	912584	1082144	1088712	931180	777000	498784	

Being a web app, Google Sheet is not as powerful as Excel

Power of Data Analysis

- Pivot table is a very powerful tool
- Sometimes it is called **dynamic report**
- It allows the users to **drill-down** to see more details or **roll up** to see bigger pictures
- Many DBMS, ERP or data-related systems provide some pivoting features
- There are several specialized software design specifically to visualize and analyse data
  - Tableau, Power BI, Pentaho

Exercise

- Study and analyse selected dataset
- Try to understand any interesting information in the data
- Write up a report describing what you learn from the data
- Present the obtained knowledge or insight
- You are allowed to present your information any way you see fit
  - Charts, tables, text, infographic etc

Datasets – Data.og.th

- สถิติขนส่งทางอากาศปริมาณการจราจรทางอากาศของสายการบินต้นทุนต่ำปี 2557 https://data.go.th/DatasetDetail.aspx?id=b6b42ac5-a7e6-41df-af98-2fae9152ed35
- ตำแหน่งงานว่าง จำแนกตามอุตสาหกรรม และวุฒิการศึกษา ทั่วราชอาณาจักร ปี 58 – 59
  https://data.go.th/DatasetDetail.aspx?id=a6287b62-0c4a-427d-aad1-29c4809d07cb
  https://data.go.th/DatasetDetail.aspx?id=5179ef3e-b689-49a0-aed6-d268e857f012
- จำนวนผู้โดยสารรถไฟฟ้า
  https://data.go.th/DatasetDetail.aspx?id=2acfc0e1-7a59-4896-8347-6bdfae79cc81

Datasets – Gapminder

- Energy production / used per person in a country (not country total)
- GDP/Capita
- Unemployment rate

Data.gov

- Fruit and Vegetable Prices https://catalog.data.gov/dataset/fruit-and-vegetable-prices
- Accidental Drug Related Deaths 2012-2018 https://catalog.data.gov/dataset/accidental-drug-related-deaths-january-2012-sept-2015
- Average Daily Traffic Counts https://catalog.data.gov/dataset/average-daily-traffic-counts-3968f