My general data analysis steps – **ACM PVF**

1. Ask questions about data
2. Clean data
3. Model data
4. Create pivot tables (Excel) using questions in step 1
5. Visualise data using questions in step 1
6. Display forecast from your visualisations to answer the questions in step 1

* Undo a cell range formatted as table: Right click any cell -> Table -> Convert to Range
* Add filter options to a cell range not formatted as table: Data tab -> Sort & Filter group -> Filter
* Sum all numbers in a particular col or row without entering a formula, select the last empty cell at the end of that row or col and use the shortcut **Alt + =**
* Add custom lists for autofill in excel: File menu -> Options -> Advanced menu -> General menu -> Edit Custom Lists
* Display formulas instead of results in worksheet**: Ctrl + `** (accent key) **or** File menu -> Options -> Advanced menu -> Display options for this worksheet menu -> Show formulas in cells instead of their calculated results
* Freeze pane: Click row underneath column headers -> View tab -> Window group -> Freeze Panes -> Freeze Panes
* Enter current date: **Ctrl + ;**
* Flash fill: **Ctrl + e**
* View the endpoints of your data range or table selected: **Ctrl + .**
* Make same change across worksheets containing same data: Hold Ctrl and select worksheet tabs below -> make change to one cell -> check to ensure it’s applied across other worksheets.
* Multiple data sorting: Select table -> **Data** tab -> **Sort** **&** **Filter** group -> **Sort**
* The subtotals feature of excel only applies to ranges and not tables
* **Generating random values**
* Generate random whole numbers: =**ROUND(RAND()\*35,0)** 35 can be replaced with any number
* Generate random whole numbers within a range: **RANDBEETWEEN()**
* Generate random decimal numbers: **RAND()**
* Generate random decimals within a range**: RAND()\*[diff btn lower and upper bound value]+[lowerboundvalue]**
* **NB:** To prevent generated numbers from being changed upon any update in the worksheet, copy generated numbers -> right click -> **Paste** **Special** -> **Paste Values (V)** to paste right back into the copied range.

* **Statistical functions**
* **COUNTIF**: count the number of rows/cells in col A where col A=abc (1 col under consideration)
* **COUNTIFS**: count the number of rows/cells in col A where col B=abc, col C=def, D=ghi, etc (2+ cols under consideration).
* **SUMIF**:
  + Find the total x in col A where col A=abc. (1 col under consideration: **SUMIF with 2 args**)
  + Find the total x in col A where col B=abc. (2 cols under consideration ie. A,B hence **SUMIF with 3 args**)
* **SUMIFS**: Find the total x in col A where col B=abc, col C=def, col D =ghi, etc. (3+ cols under consideration)
* **AVERAGEIF:** 
  + Find the average value in col A where A=abc (1 col under consideration: **AVERAGEIF with 2 args**)
  + Find the average value in col A where col B=abc (2 cols under consideration: **AVERAGEIF with 3 args**)
* **AVERAGEIFS:** Find the average value in col A where col B= abc, col C=def, col D=ghi, etc. (3+ cols under consideration)

* **Lookup functions**
* **Vlookup** is used to search and return a particular category’s corresponding data value which is in the table. Example a student table containing their names and index numbers. We can use vlookup to search and return the index number of a student called John
* Formula: =vlookup([whose value are we searching for ie. lookup value], [in which table are we searching], [the item we want to pull out, what is its column number starting from 1?], [do we want the exact match for that lookup value ie. 0 or we want an approximate match ie. 1])
* **NB:** One constraint of VLOOKUP is that it can only look for values on the left-most column in the table array. If your lookup value is not in the first column of the array, you will see the #N/A error.
* **NB:** Exact match and approximate match perform completely different functions. Both are used to display an item’s corresponding value from a table. Exact match is best suited to lookup values which already exist in the table and have their corresponding value attached to them. Approximate match is best suited to looking for the corresponding of lookup values which are not the in table, thus they’d rather give you an approximate result value for that lookup value. Now since exact match looks into the table and gives you the specific result recorded for the lookup value then that’s easy. However, to get an approximate match for a lookup value, some conditions have to be met.
  + the approximate match generates its approximate value from data on the already existing lookup values and their corresponding values. The way it works is simple. Lets say we have a table like below

|  |  |
| --- | --- |
| Surveys completed | Bonus earned (%) |
| 0 | 0 |
| 20 | 30 |
| 60 | 80 |
| 100 | 100 |

* + if we wanted the bonus for someone who completed 60 surveys, we can easily use the exact match because its corresponding value is directly in the table.
  + What if we wanted to know the bonus for someone who completed 40 surveys. Would it be 30 or 80?. This is where approximate match becomes very important.
  + First and foremost, to perform an approximate match, the lookup values (in this case, Surveys) should **always** be sorted in ascending order as shown in the table above.
  + Next, it is important to understand how approximate match uses a range to calculate the result of a lookup value. The table below explains the overall mechanism

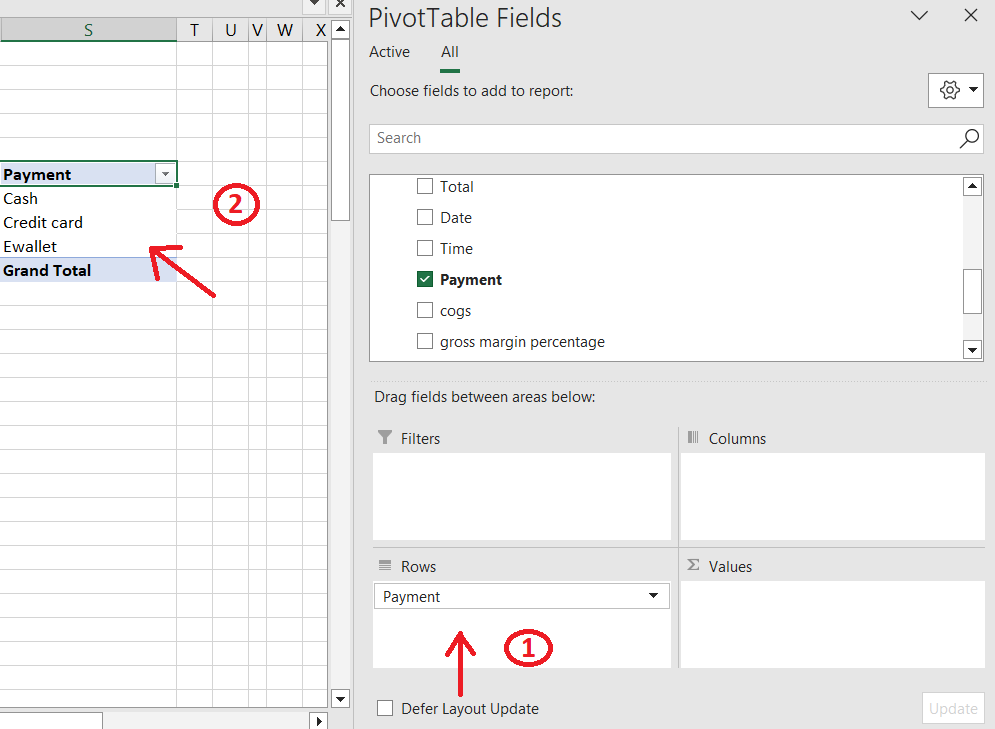
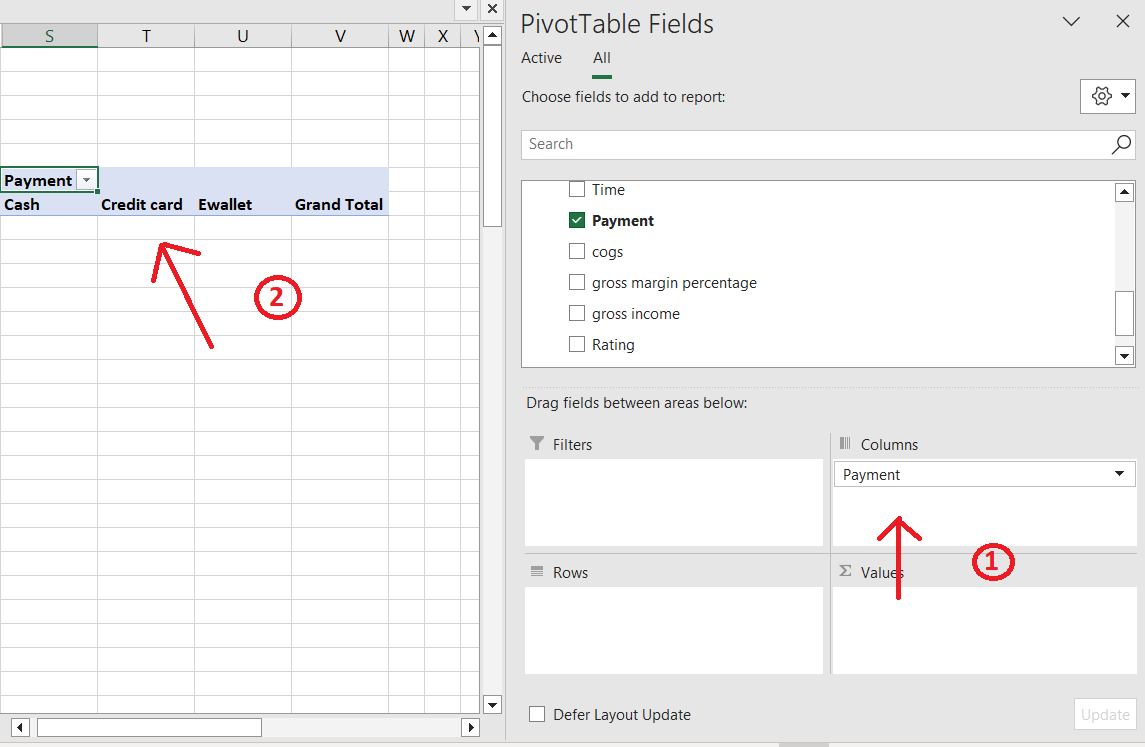
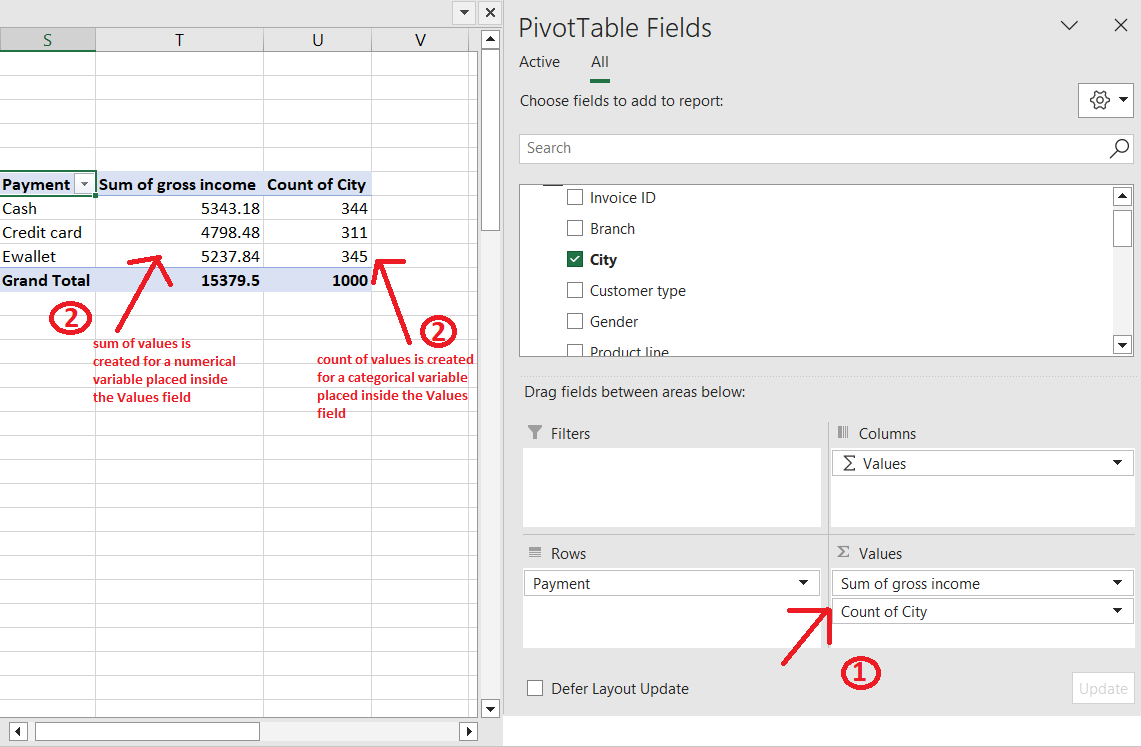
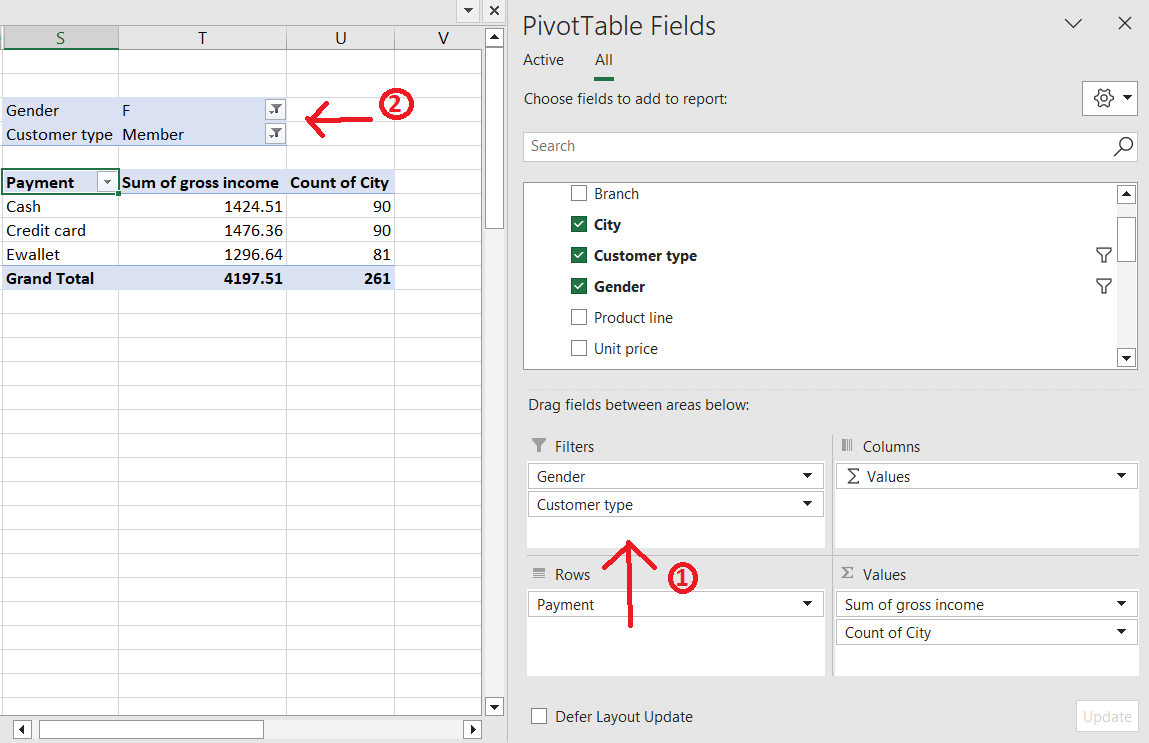
|  |  |  |
| --- | --- | --- |
| Surveys completed | Bonus earned (%) | How approximate match produces results |
| 10 | 5 | 10 <= survey <20 |
| 20 | 30 | 20 <= survey < 60 |
| 60 | 80 | 60 <= survey < 80 |
| 100 | 100 | 100 <= survey < 100 |

* + Therefore for 40 completed surveys, the bonus will be 30%
  + **NB:** if our new lookup value whose result we want is less than the least value in the column (ie. the first lookup value), approximate match is #NA. Eg. If we wanted to find out the bonus points for 9 completed surveys, bonus would be #NA
  + Since it uses a range mechanism to calculate the result, approximate match only works for numerical lookup values
* Defining a vlookup without stating the match type(exact or approximate) defaults to approximate match.
* The lookup value for vlookup should always be unique. If there are multiple entries of the lookup value, vlookup gives the corresponding value for the lookup’s first occurrence.
* In vlookup, the two key aspects are the lookup value and its corresponding value. The position of the column for the corresponding value can never come before the column position of the lookup value. It wont work in vlookup
* **Hlookup:** we used vlookup because we were searching for a result which was inside a range laid out vertically. We use hlookup when the value we want to return is inside a range laid out horizontally.
* **Easiest way to decide whether to use vlookup or hlookup.**
* Assuming we have a lookup value A and we want to find its match B. If the title of the result we want to return is a column header then (vlookup). If the title of the result we want to return is a row header then (hlookup). eg. Table 1

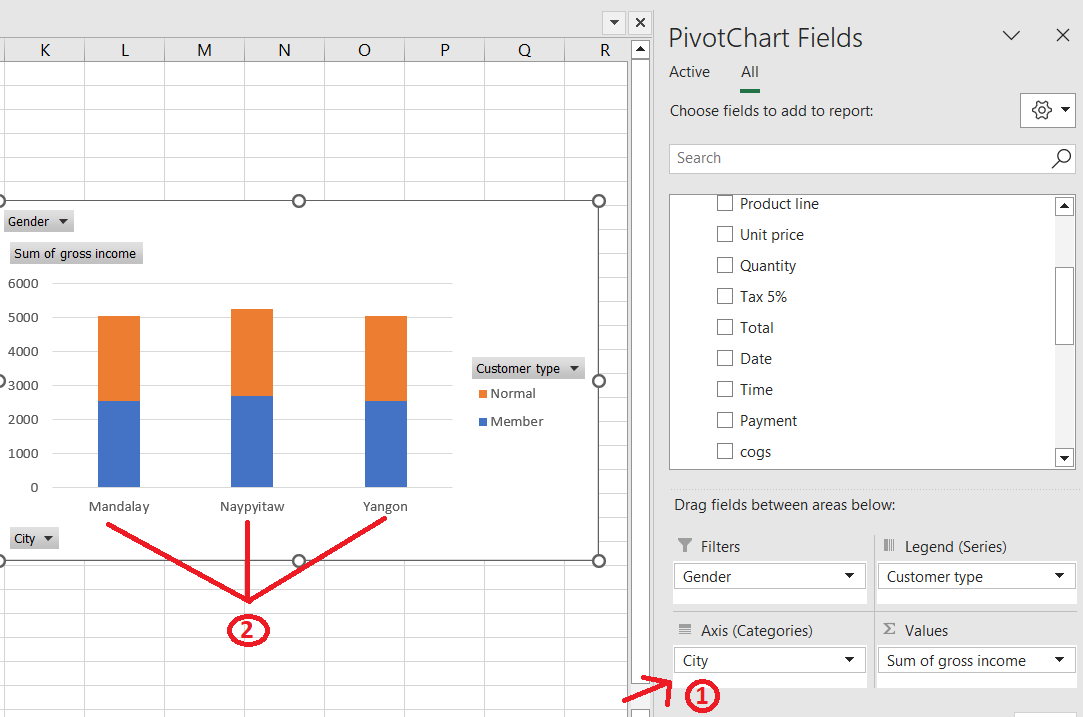
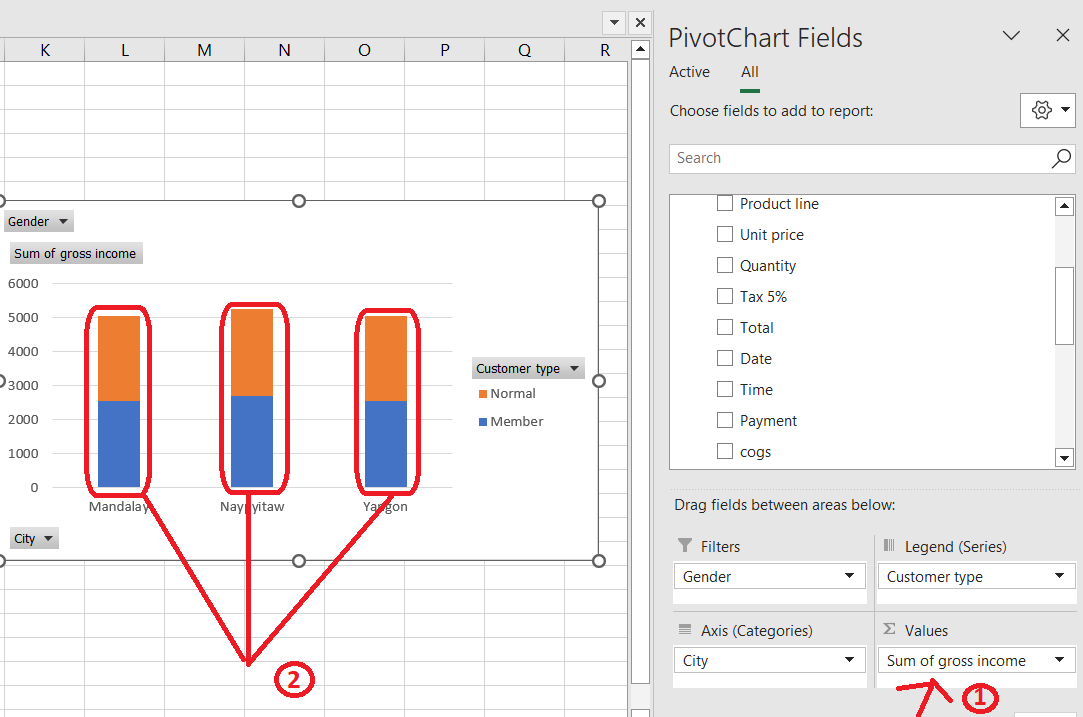
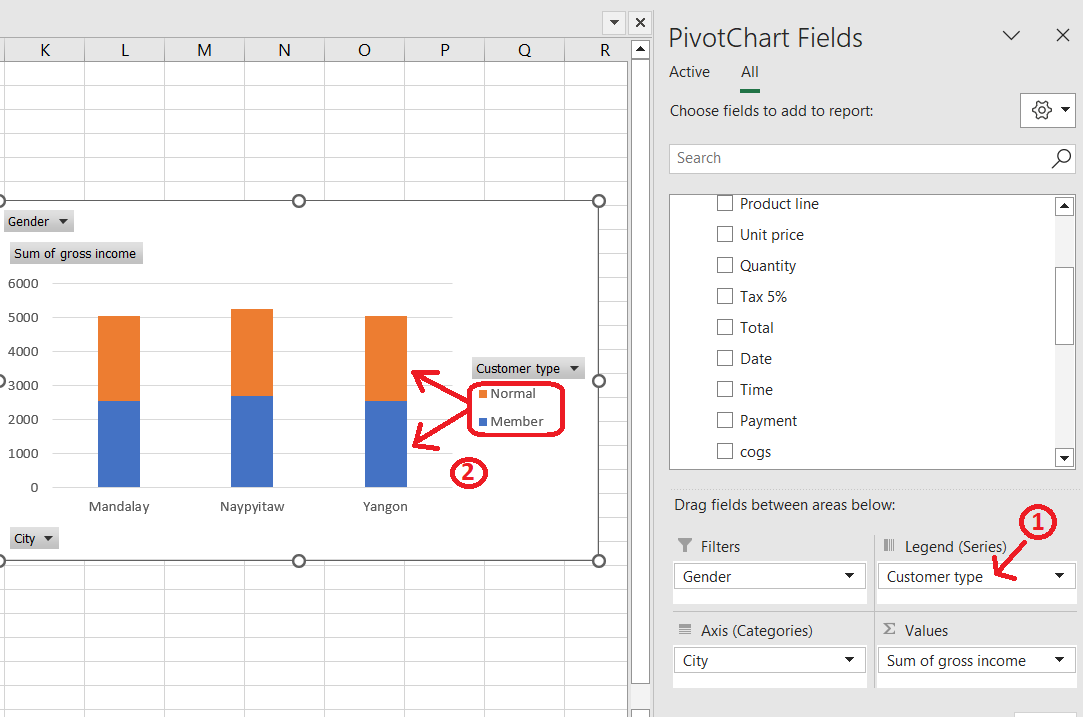
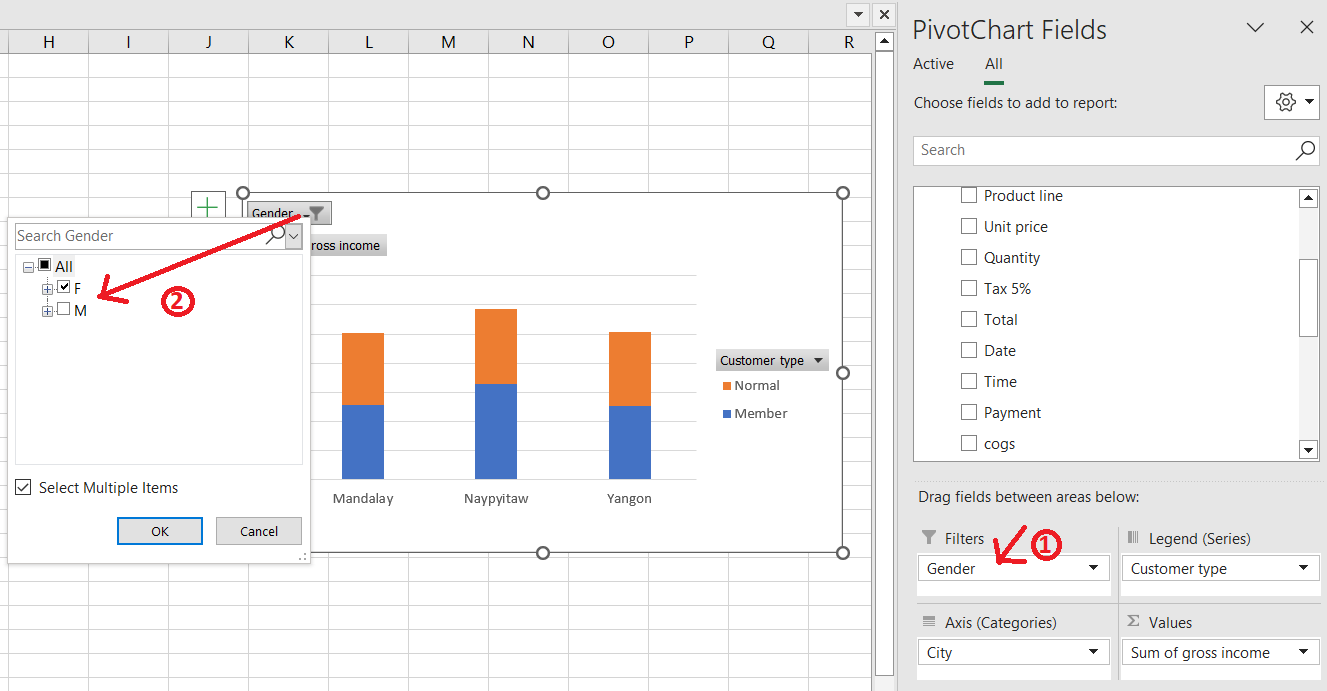
|  |  |
| --- | --- |
| Surveys completed | Bonus earned (%) |
| 10 | 5 |
| 20 | 30 |
| 60 | 80 |
| 100 | 100 |

* + We’re looking for the bonus of a given survey. “Bonus earned (%)” is a column header (vertical) so we use vlookup (vertical lookup)
  + Table 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Surveys completed | 10 | 20 | 60 | 100 |
| Bonus earned (%) | 5 | 30 | 80 | 100 |

* + We’re looking for bonus of given survey “Bonus earned (%)” is a row header (horizontal) so we use hlookup (horizontal)
* Consider we have a lookup value and we want to find its result, if the result’s header is a row header, we use xlookup
* **Xlookup** is an advanced version of vlookup which is not affected by the constraint.
* Formula: =xlookup([whose value are we searching for ie. lookup value], [in which array is the lookup value located], [the item we want to pull out, what is its array])
* **Index and Match functions**
* The term “index” is often used in programming to find the position of an item in a list
* **NB:** Here in excel, this is rather than by the MATCH function ie. it finds the position of a given item.
* The INDEX function does the opposite. Ie. It is used to the find the item of a given position
* **Data Modelling**
* Power Pivot is used to create data models, establish relationships and create calculations in excel.
* **NB: Power Pivot group items are greyed out in excel if the file is not saved as an excel worksheet (xlsx). Formats like csv and others won’t work. Even if you opened a csv file and accessed power query from there... after loading the data back into the worksheet, its still greyed out as long as the file the data is being loaded into is saved as a csv. To be on the safer side, always open a blank excel workbook (which is by default opened as an xlsx file) and use power query’s ‘Get Data’ to load the csv/xlsx files into the blank workbook. That way the power pivot features don’t get greyed out**
* To add a table to the model: select cell in the table -> **Power** **Pivot** tab -> **Tables** group -> **Add** **to** **Data** **Model**
* Advantage of creating data models and establishing relationships is that when we set up our pivot tables and charts, we are able to fetch data from a different table into our table through the link established between the tables in the data model.
* **NB:** A rule of thumb is to always import pivot tables from the data model which adds all the tables of the model to the ‘fields’ section of the pivot rather than setting up pivot tables with the raw tables in the excel worksheet which gives you only that table’s columns for the ‘fields’ section of the pivot. To import pivot table from the data model: **Insert** tab -> **Tables** group -> **Pivot** **Table** -> **From** **Data** **Model**
* To create, update or delete relationships between tables. NB: second option is also used to access diagram view of relationships
  1. **Data** tab -> **Data** **Tools** group -> **Relationships**
  2. **Power Pivot** tab-> **Data Model** group-> **Manage**
* **Pivot tables**
* Use pivot tables to group your data categories before using the new grouped tables for visualisations
* **NB:** When creating a pivot table, note that in the Pivot table Fields, the variable you drag into the columns field is used as a column heading and the variable dragged into the rows field is used as the row heading. Below gives full explanation
* **Rows field (Place categorical variables or dates here):** If you drag a variable into the rows field, pivot table uses the distinct values of that variable to create the row headings. Eg.
* 
* **Columns field (Place categorical variables or dates here):** Similarly, if you drag a variable into the columns field, pivot table uses the distinct values of that variable to create the column headings. Eg.
* 
* **Values field (You can place categorical or numerical values here):** The values field gives us the totals (sum or counts) of the variable we placed there. The sum or counts can be changed into different formats like sum, counts, average, deviation, etc. By default, if we place a categorical variable X in the values field, it aggregates the count of the distinct values inside X according to the row heading or column heading variable. However, if we place a numerical variable X in the values field, it calculates sum of the values inside the variable X according to the row heading or column heading variable. Eg.
* 
* **Filters field (Place categorical variables or dates here):** After retrieving your summarized data in the pivot table, if you want to filter your results according to a particular condition, place that variable inside the filters field. Eg.
* 
* **Pivot Charts**
* While PivotTable is used to summarize, analyse, explore, and present summary data, Pivot Charts complement PivotTables by adding visualizations to the summary data in a PivotTable, and allow you to easily see comparisons, patterns, and trends
* A Pivot table’s fields are Rows, Columns, Values, Filters. A Pivot Chart’s fields are Axis (Categories), Legend (Series), Values, Filters.
* On windows, you can create a pivot chart directly without needing a set up pivot table. Adding your variables to the pivot chart fields automatically designs the chart while also building the pivot table.
* On mac however, you need to create a PivotTable first, and then insert a chart. Once that is done, the chart will behave like a PivotChart if you change the fields in the PivotTable Fields list ***(*resource 9*)***.
* **NB:** Available charts in Pivot Charts

|  |  |
| --- | --- |
| **Charts which can be generated with pivot charts** | **Charts which cannot be generated with pivot charts** |
| * Pie chart * Bar chart * Combo chart * Line chart * Area chart * Surface chart * Radar chart | * Sunburst * Treemap * Waterfall * Funnel * Histogram * Boxplot * Scatter plot * Bubble chart * Map * Stock |

* **Axis field (Place categorical variables or dates here):** If you drag a variable into the axis field, pivot chart uses the distinct values of that variable to create the x-axis labels of the chart
* ****
* **Values field (You can place categorical or numerical values here):** The variable X placed here is what forms the composition of the graph. Eg the bars in a bar chart, the space in the pie chart, the lines of a line chart. If it’s a numerical variable, it calculates the sum of the values inside X for each category value on the x-axis. If it’s a categorical variable, it calculates the sum of the counts
* ****
* **Legends field (Place categorical variables or dates here):** For a variable X inside the values field, placing a variable Y inside the legends field groups the values in X according to the categorical values of Y. Eg. Your values field is Income sales, dragging Customer type into the legends field would segregate the sales values into those made by Normal Member customers like a stacked bar chart
* ****
* **Filters field (Place categorical variables or dates here):** After getting the graph you desire in the pivot chart, if you want to filter your results according to a particular condition, place that variable inside the filters field. Eg
* ****
* **Power Pivot**
* An Excel add-in you can use to perform powerful data analysis and create sophisticated data models

* **Macros**
* **NB:** a macro becomes non-functional if a column or row header changes. You’d have to update the header names in the VBA code or re-record the macro

Resources

1. <https://www.simplilearn.com/tutorials/excel-tutorial/data-analysis-excel>
2. <https://excelchamps.com/formulas/random-date-generator/>
3. <https://support.microsoft.com/en-us/office/relationships-between-tables-in-a-data-model-533dc2b6-9288-4363-9538-8ea6e469112b>
4. <https://support.microsoft.com/en-us/office/text-function-20d5ac4d-7b94-49fd-bb38-93d29371225c>
5. <https://support.microsoft.com/en-us/office/delete-a-pivottable-1de9b894-9178-43b3-b436-92e3ddb9175b>
6. <https://support.microsoft.com/en-us/office/sort-data-in-a-pivottable-or-pivotchart-e41f7107-b92d-44ef-861f-24430830450a#:~:text=In%20a%20PivotTable%2C%20click%20the,column%20you%20want%20to%20sort.&text=on%20Row%20Labels%20or%20Column,or%20Sort%20Z%20to%20A>
7. <https://support.microsoft.com/en-us/office/overview-of-pivottables-and-pivotcharts-527c8fa3-02c0-445a-a2db-7794676bce96#:~:text=You%20can%20use%20a%20PivotTable,comparisons%2C%20patterns%2C%20and%20trends>.
8. <https://www.geeksforgeeks.org/how-to-show-percentages-in-stacked-column-chart-in-excel/>
9. <https://support.microsoft.com/en-us/office/create-a-pivotchart-c1b1e057-6990-4c38-b52b-8255538e7b1c#OfficeVersion=macOS>
10. <https://support.microsoft.com/en-us/office/overview-of-pivottables-and-pivotcharts-527c8fa3-02c0-445a-a2db-7794676bce96#:~:text=You%20can%20use%20a%20PivotTable,comparisons%2C%20patterns%2C%20and%20trends>.