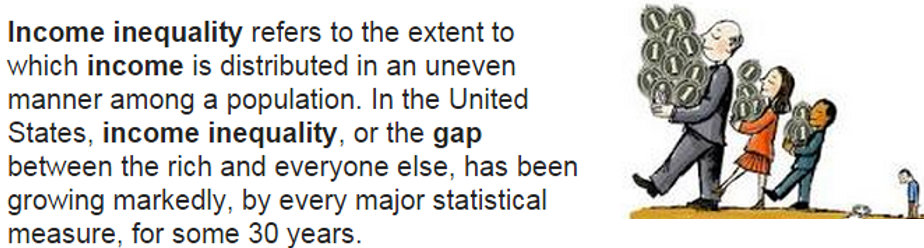
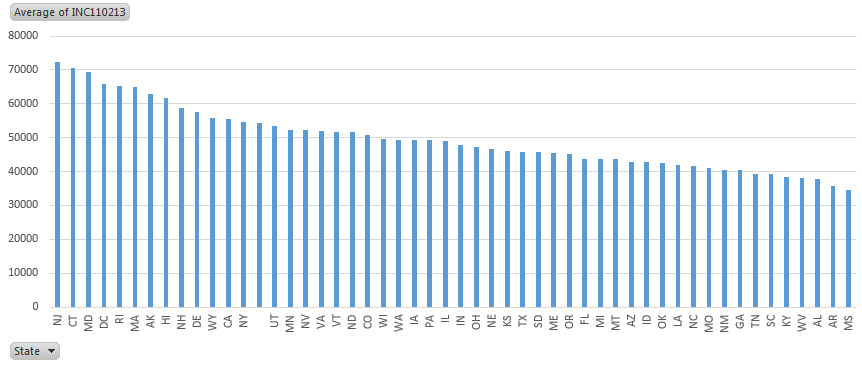
SUMMARIES AND VISUALIZATIONS

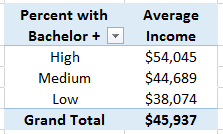
The notion of income inequality has received considerable attention in recent years. The gap between those that have a lot of money and those that don’t continues to increase in the United States.



The United States Census Bureau releases aggregated data on a regular basis for states and counties in the United States. Consider the following graph that shows Median Household Income (for years 2009-2013) by State. We see that NJ has the highest household income and Mississippi has the lowest at about $35,000.



Forbes has recently published several articles centered on whether or not a college education is worth it. For many, a college education is necessary to find their first job. However, other graduates are over qualified for their job and are hampered by a substantial amount of student loan debt. The following table suggest that when a large percentage of county residents have a bachelor’s degree or more, the typical household income is over $50,000. When this percentage is low, the typical income level drops to $38,000 per household.



The data from this handout is provided by the United States Census Bureau. This data will need to be imported into Excel. The information needed for our analysis is not contained in a single file, but two different files. The auxiliary information contained in the FIPS Codes dataset will need to be merged with the dataset before summaries and visualization can be constructed.

Procedural Steps

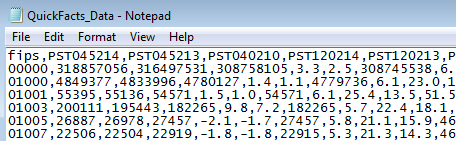
1. Download data and FIP Codes files from the US QuickFacts website
2. Merge the FIPS Code information with the data
3. Create new variables for County and State
4. Construct various summaries and visualizations in Excel

Data Technologies

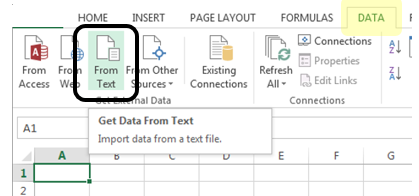
1. Import file into Excel
2. Summaries and Visualizations through PivotTables

|  |  |
| --- | --- |
| Data Source | |
| Address | <http://quickfacts.census.gov/qfd/download_data.html> |
| Description | US QuickFacts   |  |  | | --- | --- | | The US QuickFacts dataset contains aggregate information for several variables at the County and State level. The variable Median Household Income (2009-2013) will be the focus of this investigation.  [Download Data Dictionary](http://quickfacts.census.gov/qfd/download/DataDict.txt) | http://www.seo.com/wp-content/uploads/2009/10/census.jpg | |

First, let us consider the DataSet.txt file from their website. This data may not appear to have much structure. This data has been provided to us in a format known as a comma delimited or comma separated value format, i.e. csv format.



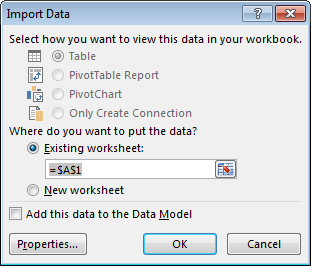
Excel has the ability to directly import this type of file. This process is started by selecting Data > From Text.



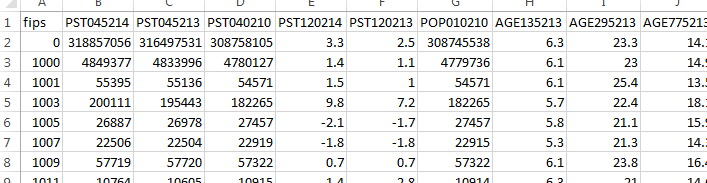
A Text Import Wizard window will show up. Proceed through this wizard by specifying the following in each step.

|  |  |
| --- | --- |
| In Step 1 of 3, select Delimited | Specify that the Delimiter is Comma and click Finish |

The last step is to tell Excel where you’d like the dataset to be placed. You can specify the cell location in an existing workshop or select New Workshet.



Click OK and the contents of the DataSet.txt should be successfully imported into Excel. The following snippet is given for reference.

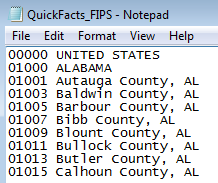


Unfortunately the only reference to county is through the Federal Information Processing Standard (FIPS) code provided in Column A. A FIPS codes is a five-digit code which uniquely identifies counties and county equivalents in the United States. States are given FIPS codes as well.

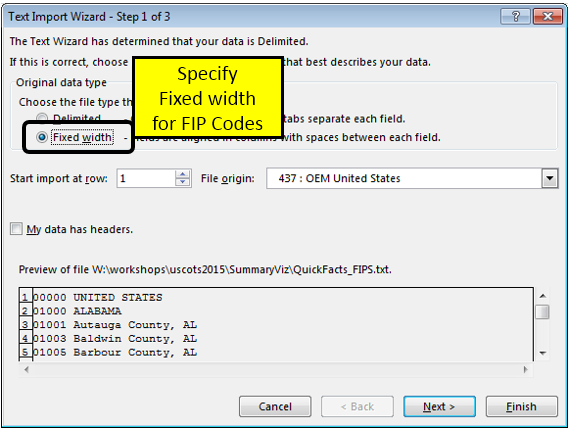
The FIPS\_CountyName.txt file contains the information necessary to relate a FIPS code to a county or state name. Before the file is imported, we must insert a column for the contents of this file.

|  |  |
| --- | --- |
| Right click on Column A and select Insert | An empty column should be provided |

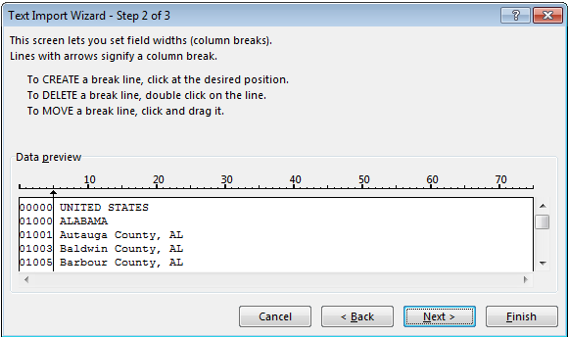
The FIPS\_CountyName.txt file format is a different format than file containing the data. In particular, the first 5 digits contain the FIPS code. A comma is used to separate the County Name from the State Name.



In Step 1 of the Text Import Wizard, Fixed width should be selected.

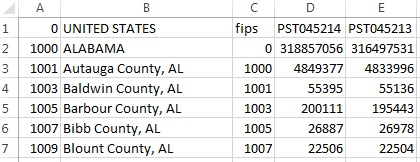


In Step 2, specify that the first five column of each row should be separated from the remaining information. Click Next.

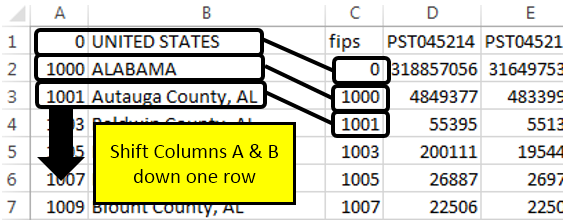


In the Import Window, specify you want the information placed in cell A1.

The information from the FIPS\_CountyName.txt file should now be placed into the first two columns.

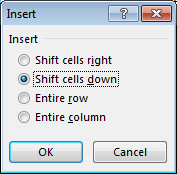


Unfortunately, the FIPS\_CountyName.txt file did not contain a header row like the DataSet.txt file did. Thus, all rows in columns A and B will need to be shifted down one row.

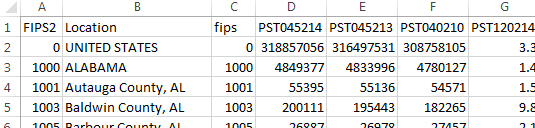


Note: You could import the FIPS\_CountyName.txt file a second time and specify the locations in the Import Window to be A2 instead of A1.

To add a row for only Columns A and B, highlight cells A1 and B1, right click, and select Insert. Specify Shift cells down to insert a row at the top of columns A and B.

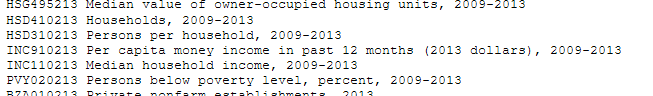


Specify variable names for these new columns. FIPS2 and Location were used in my dataset.



Getting Summaries Using PivotTables

This dataset contains several variable or fields. The names of these fields are abbreviated substantially as is often the case. A **data dictionary** is often provided with complex datasets. This dictionary contains detailed information about each variable. The dictionary for this data is provided in the DataDict.txt file. A review of this file informs us that INC110213 contains the Median Household Income, which is the variable of interest here.



Select Insert > PivotTable (Data > PivotTable on a MAC). Construct a pivot table using the following structure.

|  |  |
| --- | --- |
| Structure for PivotTable | Outcome |

|  |  |
| --- | --- |
| It appears the average median household income value is a little over $46,000. The count identifies that there were 3,195 observations used when computing this average.  However, the value for the United States, the other States, and Washington DC have been incorrectly included in this average. |  |

If summaries are to be computed only on county level data, then a new variable should be created to identify whether or not the information provide in that row from a country.

|  |  |  |
| --- | --- | --- |
| FIPS Code | Location | Description |
| 0 | UNITED STATES | FIPS code for US is 0 |
| 1000 | ALABAMA | FIPS code for AL |
| 1001 | Autauga County, AL | 1st county in AL |
| 1003 | Baldwin County, AL | 2nd count in AL |
| 1099 | Monroe County, AL |  |
| *skipped* |  | *1100 is skipped as hundreds denotes a state for most instances (see note below)* |
| 1101 | Montgomery County, AL |  |
| 1133 | Winston County, AL |  |
| 2000 | ALASKA | FIPS code for AK |
| 2013 | Aleutians East Borough, AK | 1st county equivalent in AK |
| Note: Four exceptions include: 2100 Haines Borough, AK; 51600 Fairfax City, VA; 51600 Newport News City, VA; 51800 Suffolk City, VA | | |

Creating a New Variable for County

The following will be used to create a new variable to identify whether or not the row contains information for a county.

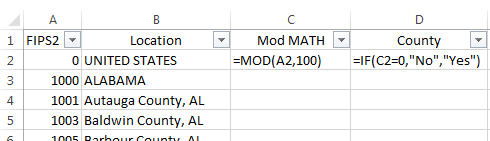
Insert a new column after Location. Type the following formula into cell C2.

Cell C2: =MOD(A2,100)

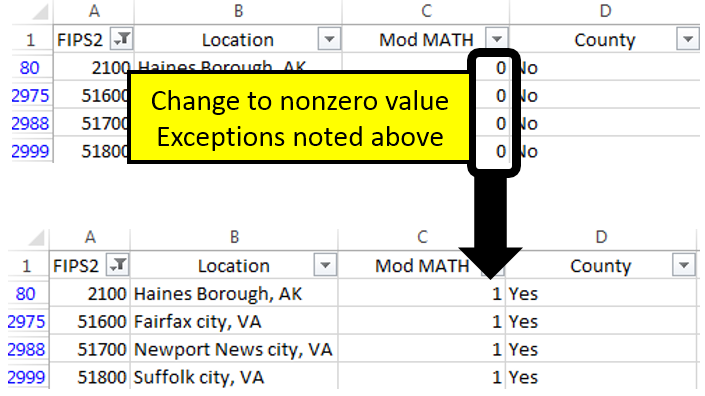
Modular arithmetic is used because it is know that for all but four counties this function will produce a nonzero value. After this formula is entered, place the cursor back into cell C2 and double click on the lower-right corner. This will autofill the formula for all remaining rows.

In cell D2 enter the following formula which simply checks whether or not the value in Cell C2 is 0. If this value is 0, then we know that for all but four FIPS codes this value will not be a county. “No” is returned when the condition being checked is TRUE and “Yes” when the condition is False.

Cell D2: =IF(C2=0,”No”,”Yes”)



Before proceeding, the four exceptions to the rule that “hundreds denotes state FIPS code” should be fixed. This can be done by applying a filter to the FIPS code column and select values 2100, 51600, 51700, and 51800. Simply change the value in Column C to a non-zero value, e.g. change them to 1.



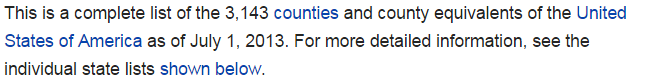
Questions

1. Use PivotTables to verify that the average median household income across all counties in the United States is about $46,000.

|  |  |
| --- | --- |
| PivotTable structure | Outcomes |

1. There is a Wikipedia page that lists information regarding all counties and county equivalents in the United State. The following text is from this Wikepedia page.

Note: The table provided on the Wikipedia page is missing one entry. FIPS Code 51515 Bedford City, VA is missing. Please don’t ask how I discovered this!



*Source*: <http://en.wikipedia.org/wiki/List_of_United_States_counties_and_county_equivalents>

Does it appear that the average computed above is using the correct number of rows? Discuss.

1. The County = No has 52 entries; however, there are only 50 states. Determine why there are 52 rows labeled as “No” by our procedure.

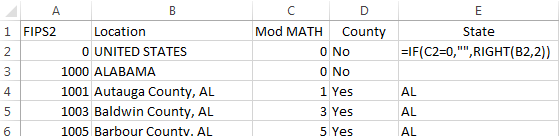
Creating a New Variable for State

A process similar to labeling Counties can be used to create a new variable for State abbreviations, i.e. AL, AK, etc. This abbreviation is present in Column B when the row consists of a county. In fact, this abbreviation is always the last two digits in the string. The =RIGHT() function will be used to pull off the last two digits, when appropriate, from column B.

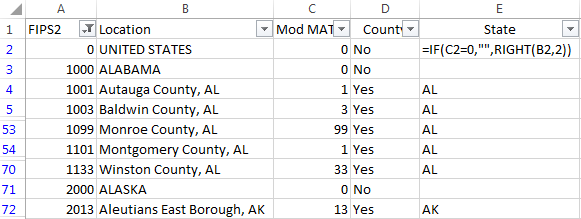
Cell E2: =IF(C2=0 , ”” , RIGHT(B2,2) )

* IF statement used to check whether or not column C contains a 0
* IF cell C2 is zero, then an empty string is returned, i.e. nothing is returned
* IF cell C2 is non-zero, then use the =RIGHT() function to pull off the last two digits

The following snippet is provided for reference.



Verify the content is correct for the new variable State for several rows.



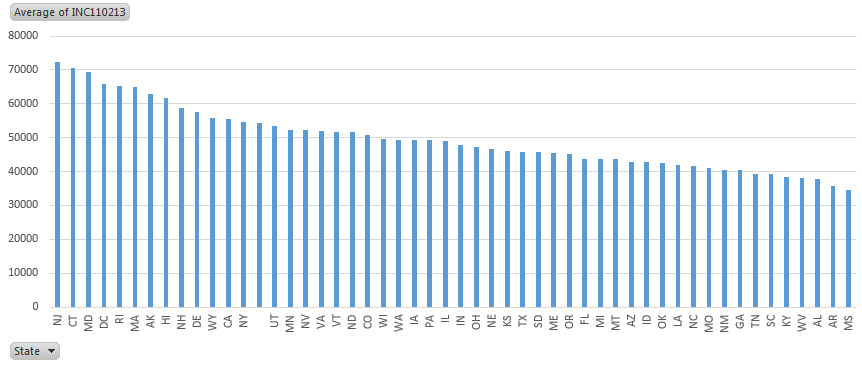
After this new variable is created, PivotTables can be used to create summaries by State.

|  |  |
| --- | --- |
| Summary Statistics by State | Structure used for PivotTable to the left |

The PivotTable can be sorted by any column – simply place your cursor in the column to be sorted and select Data > Sort.

|  |  |
| --- | --- |
| Sort inside a PivotTable | Average Household Income sorted from largest to smallest. |

Place your cursor in the PivotTable. A variety of charts can be created from the Charts menu. The following pareto-type graphic displays the average household income from largest to smallest.

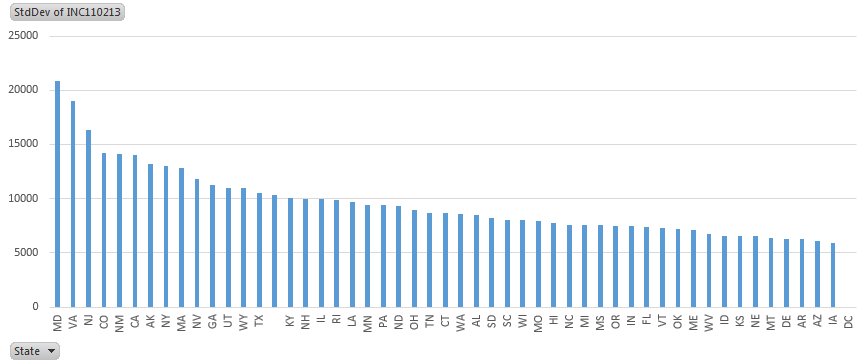


Questions

1. What is the average median household income for the state you live in? How does this value compare to others states?
2. Why is there a blank state label on the graphic above? Discuss.
3. Consider the following graphics. Why would standard deviation of median household income be a better measure for income disparity than an average? Discuss.

|  |  |
| --- | --- |
| Averages for MN and VA are similar | Standard deviation, i.e. average distance from data point to mean is considerably larger for VA. |

1. The pareto-type chart below show the standard deviation of incomes from highest to lowest. Identify states that appear to have low income disparity, i.e. incomes are similar across counties. What states appear to have high income disparity?



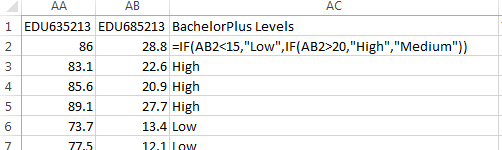
1. Notice, DC, i.e. Washington DC, on the chart above has no standard deviation. Why is this the case? Explain.

Task #1

For this task, the average income levels will be separated on a new variable called BachelorPlus Levels. This variable should be setup using the existing variable EDU685213 and with the following structure.

EDU685213 Bachelor's degree or higher, percent of persons age 25+, 2009-2013

* About 1/3 of counties have EDUC685213 less than 15, label these counties as Low
* About 1/3 of counties have EDUC685213 greater than 20, label these counties as High
* Label the remaining counties as Medium



Next, use this new variable to create the following PivotTable.

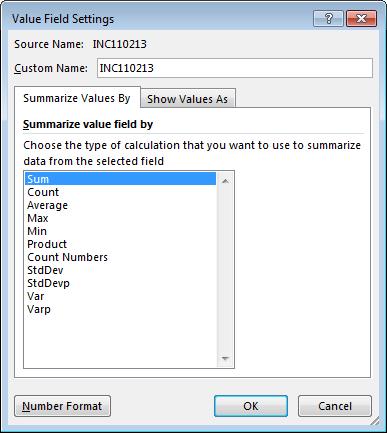
|  |  |
| --- | --- |
| PivotTable structure | Select County = Yes from the Filter to use only County data |
| Default order is incorrect. Right click  on Low and select Move > Down | Finished PivotTable |

Questions

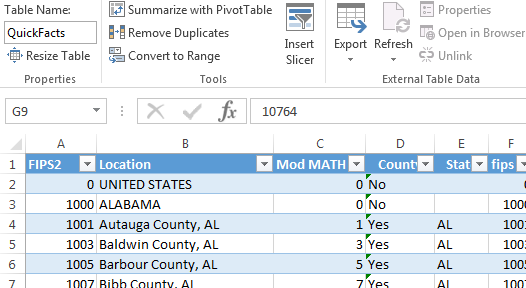
1. Does average income increase as the proportion of residents who have a Bachelors or higher increases? If so, discuss to what degree.

Using =AGGREGATE() and SLICER

There are limitation to what PivotTables is able to compute. The list of functions available can be found under the Summarize values By tab. Notable exceptions from this list include median or more generally percentiles. The =AGGREGATE() function and the SLICER feature in Excel can be used as an alternative to PivotTables.



To invoke the SLICER feature in Excel, convert the dataset to an Excel Table. Give the Table a name, e.g. QuickFacts.



To specify a SLICER, select Insert > Slicer.



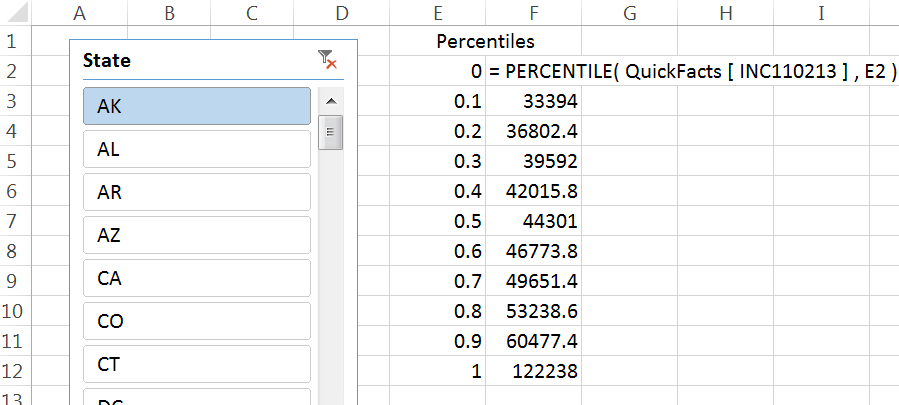
The following can be used to setup a SLICER for State.

|  |  |
| --- | --- |
| In the Insert Slicers window, specify the field (or fields) from which to construct the slicer. | If State is specified, then the following window is displayed |

In an empty column, enter a sequence of values from 0 to 1 by increments of 0.1. These will be used to compute percentiles for income.

Cell F2: =PERCENTILE( QuickFacts[INC110213] , E2 )

Enter this =PERCENTILE() function into cell F2 as shown. Copy this down for the remaining cells.



|  |  |
| --- | --- |
| Try the Slicer | |
| Click AK to get percentiles for State = AK | Click AL to get it’s percentiles |
| Note: The Slicer fails to produce the desired because =PERCENTILE ignores the fact that some rows should be excluded from it’s calcuations. | |

The =AGGREGATE() function in Excel should be used in Excel when certain rows should be excluded from the requested calculations. =AGGREGATE() is a more complete version of the =COUNTIF() functions used in the previous handout.

A explanation of the required arguments for the =AGGREGATE() function are briefly discussed here.

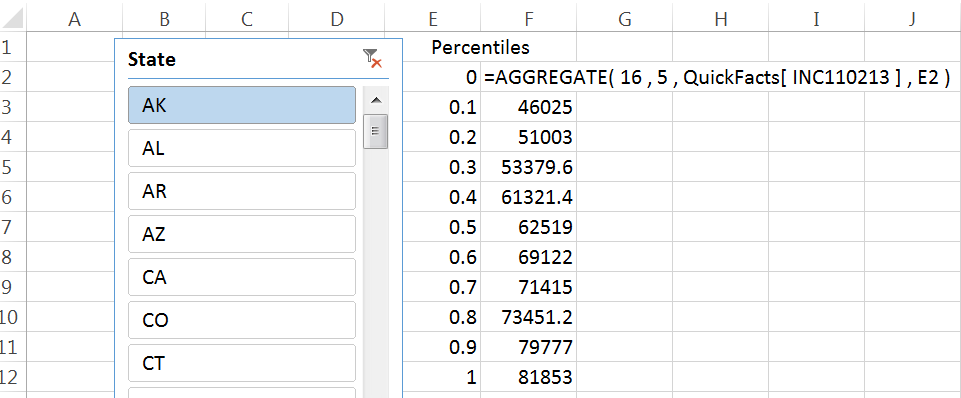
=AGGREGATE(function\_num, options, array, [k])

|  |  |
| --- | --- |
| The function\_num is a number from the following list. 16 should be used for percentiles | The options value should be selected from the following list. 5 will ignore hidden rows in it’s calcuations. |

Replace the =PERCENTILE() function used above with the following function.

Cell F2: =AGGREGATE(16, 5,QuickFacts[INC110213],E2)

Copy this down for the remaining percentiles.



The slicer can be used to specify any state or a collection of states.

|  |  |
| --- | --- |
| Selecting State = PA |  |
| Selecting State = MN |  |
| Use Ctrl to select multiple States, i.e. neighboring states to MN shown here  {IA, MN, ND, SD, WI} |  |

Task #2

For this task, apply a SLICER on the median household income, i.e. variable INC110213. Use the slicer to verify that the following 5 counties in United States have the highest and lowest median household income.

|  |  |
| --- | --- |
| Counties with highest median income | Counties with lowest median income |

Questions

1. What are the median household income values for each set of counties listed above?
2. There are 3143 counties, so the richest 31 counties would represent the top 1%. How many of the top 31 counties are from VA? How about MD?

Task #3

There are seven counties surrounding the Twin Cities that are known locally as the “metro counties.” A map of these counties is provided here.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Map of Metro Counties surrounding the Twin Cities, i.e. Minneapolis - St. Paul | List of Metro Counties   |  | | --- | | Anoka County, MN | | Carver County, MN | | Dakota County, MN | | Hennepin County, MN | | Ramsey County, MN | | Scott County, MN | | Washington County, MN | |

Create necessary variables and then use PivotTables to find the average median household income for Metro = No and Metro = Yes counties in MN. How does the average income compare across these two geographic regions? Discuss.