Lecture Note #13: Charts Part #3

BUSI 201: Business Data Analysis

Topic 1. Scatter Charts

Scatter charts are frequently used to visualize relationships between two numerical variables. Each of the two axes on the chart will represent different numerical information about a single data point. For instance, see Figure 1 below.

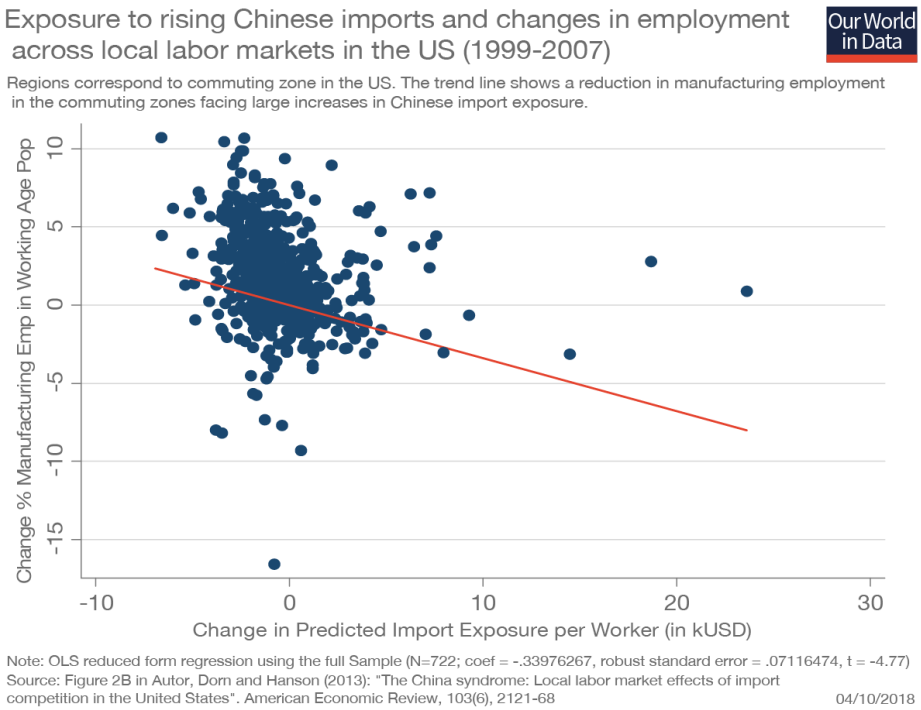


Figure 1: Autor et. al (2013)

Each data point about a commuting zone consists of (at least) two numbers; the estimated import exposure to Chinese manufacturing (on the horizontal axis), and the change in manufacturing employment in said commuting zone (on the vertical axis). Each dot on the chart represents the two numbers, and the red line that passes through the data is meant to represent the overall relationship between the two numerical variables each dot represents. In this specific example, we can conclude that the exposure to Chinese manufacuring is negatively correlated to manufacturing employment in the region.

Basic Scatter Charts

Navigate to the worksheet `SCATTER-01` of workbook `BUSI201-LEC14-Workbook.xlsx`. This worksheet contains synthetic data on the average price of apples (per lbs) in the United States, and the number of books sold in New Zealand over some period of time. One way to visualize this data may be to represent it as a line chart with two variables.

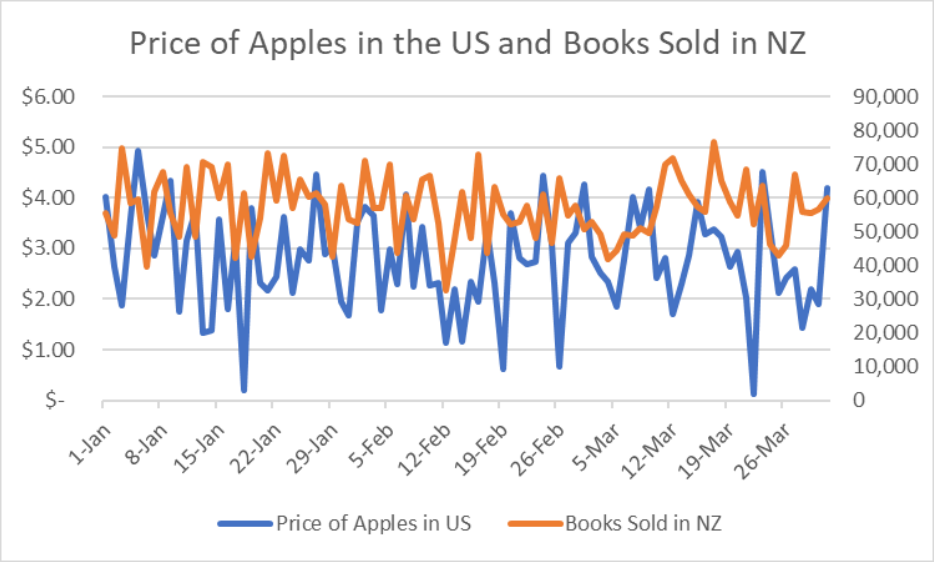


Figure 2: Line Chart with Two Axes

If we were interested in the historic trend of these two sequences, Figure 2 may be suitable. However, if we were interested in any potential correlation between the two variables, we would want to use a scatter chart. Select the data in the red box, navigate to insert, and select the basic scatter chart in the orange box in Figure 3.

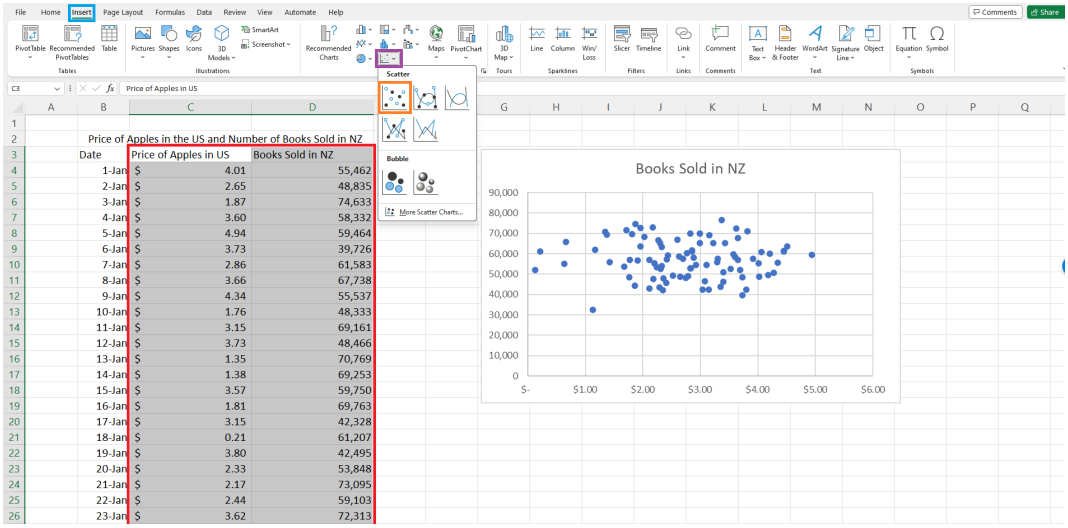


Figure 3: Generating Basic Scatter Charts

Before we actually start analyzing the chart we just generated, we should think: "would the price of apples in the United States tell us anything about the number of books sold in New Zealand?" The answer should probably be "No." Figure 4 is a slightly edited version of the scatter chart we created in Figure 3, which depict the relationship between the two variables.

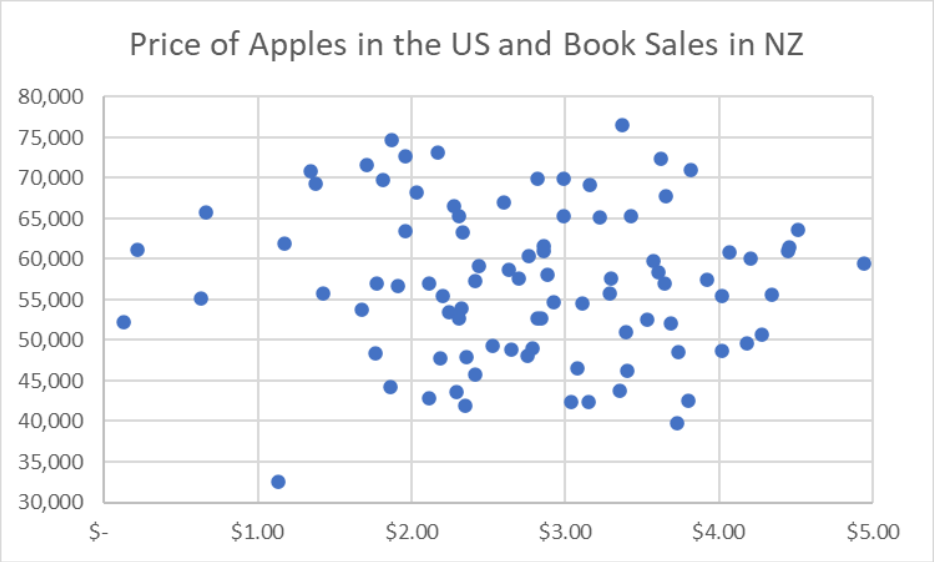


Figure 4: Edited Scatter Chart

As expected, there is not much of a trend that we notice from the scatter chart. To see a counter example, navigate to `SCATTER-02` worksheet, which contains synthetic data on the price of houses and their distance to a local power plant. Due to the risk of pollution, we should expect that the closer to a power plant, the price of houses should be lower. The default chart we get when we create a scatter chart using the data gives us Figure 5.

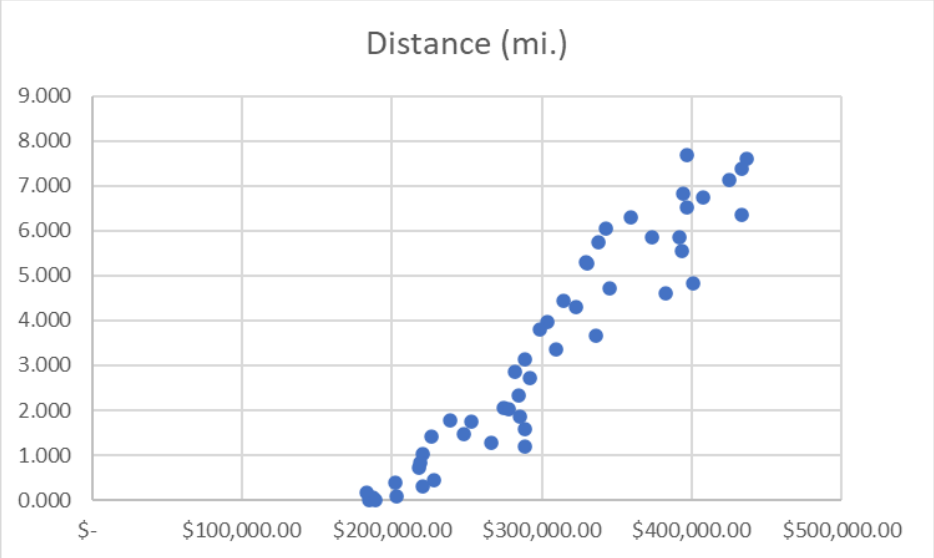


Figure 5: Default Scatter Chart: `SCATTER-02`

Editing Scatter Charts

From Figure 5, we can tell that the further away from the power plant, the housing prices tend to be higher. However, there are some edits that may make the scatter chart slightly more visually pleasing. Follow the steps illustrated in Figure 6. First, right click on the chart, and select `Select Data`. Then, `Edit` the Legend Entries, and match the color-coded entries.

The green box should have the title of the chart, the orange box contains the numeric values that go on the horizontal axis, and the purple box contains the numeric values that populate the vertical axis. Series X refers to the x axis, which is conventionally the horizontal axis, and Series Y refers to the y axis, which is typically the vertical axis.

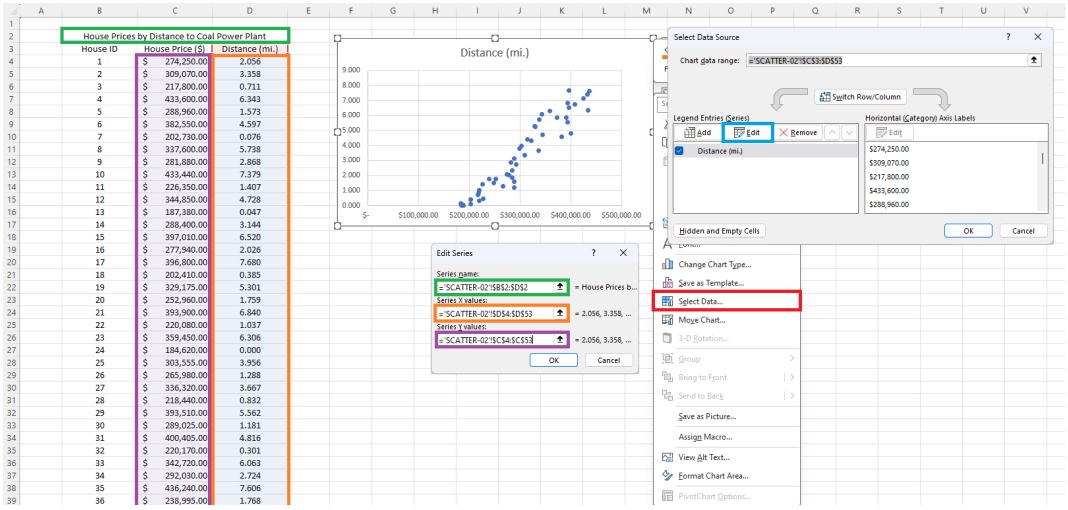


Figure 6: Editing Scatter Chart

To further customize the chart, we will edit the scale of the vertical axis. Right click the vertical axis labels in the red box, and select `Format Axis`. Then select the Axis Options in the orange box, click on Axis Options again in the blue box, and then change the bounds of the axis in the purple box.

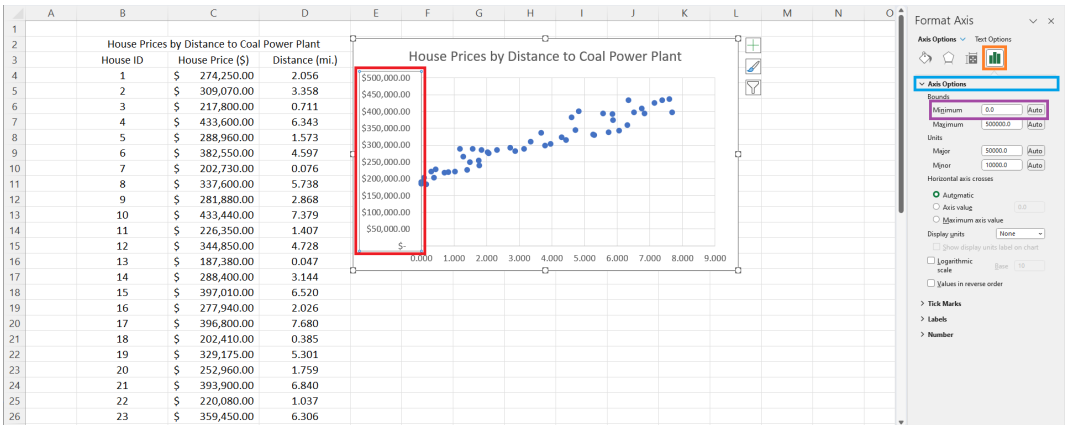


Figure 7: Editing Scatter Chart

Applying all changes in the previous page, we get the chart in Figure 8. By editing the chart slightly, the message is more clear. Of course we can add axis subtitles, and adjust the horizontal axis a bit more, and maybe add trendlines to make the point a bit more salient.

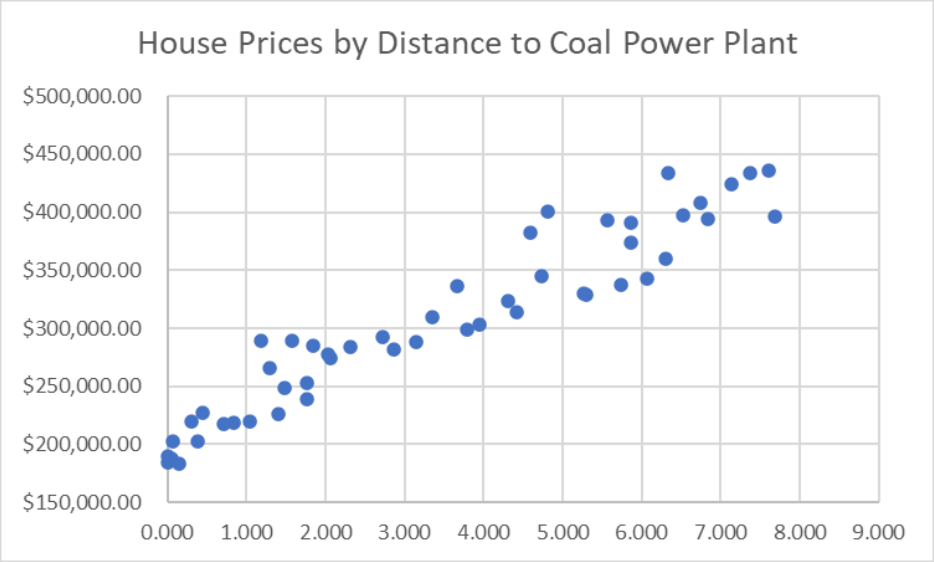


Figure 8: Edited Scatter Chart: SCATTER-02

The following chart in Figure 9 adds a few more edits to the chart in Figure 8; A) adding a trendline, B) editing the property of trendlines, C) editing the marker colors. We will cover these topics in the following section.

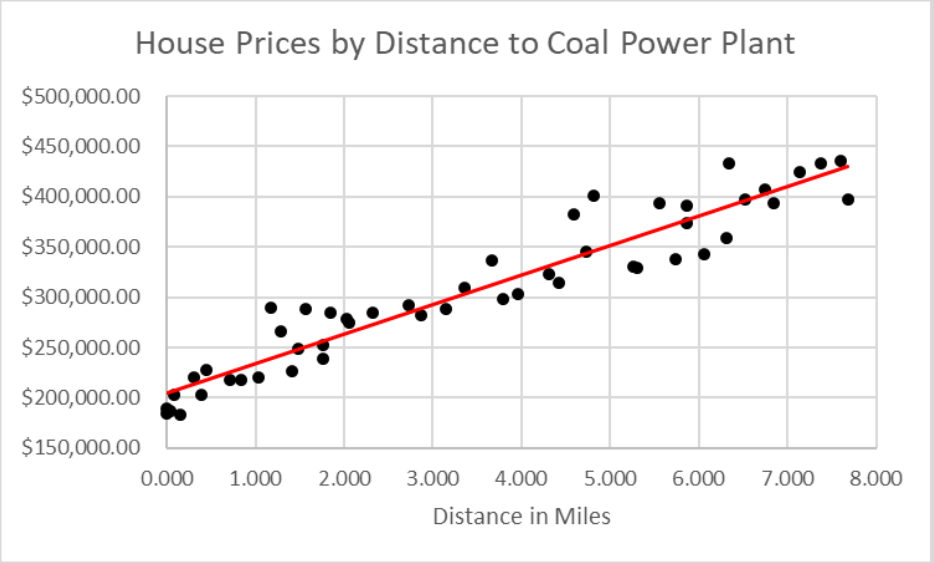


Figure 9: Edited Scatter Chart: SCATTER-02

Trendlines in Scatter Charts

The trendline represents the line which results in the smallest estimation error, assuming that the true data generating process follows a linear function. If you are interested in interpreting data to analyze the trend in a more precise method, or in learning how to distinguish between correlation and causation, take a sequence in Econometric analysis.

Back to our topic... Trendlines can be added by left clicking on the chart, left clicking on the + mark that pops up in the upper right hand side corner, and selecting `Trendline`. In our case, the default trendline added by this means will be a blue dotted line. Since we already have blue points as the markers of our scatter chart, it is not easy to identify the trendline.

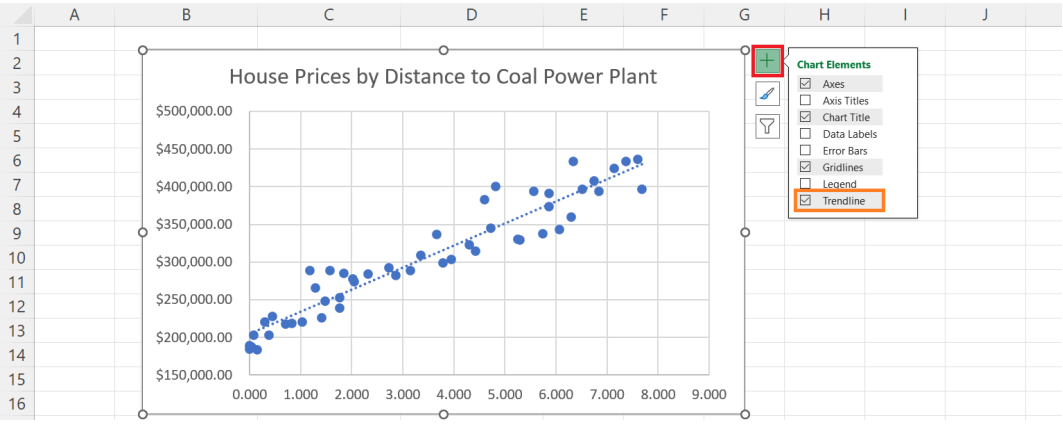


Figure 10: Adding Trendlines

To edit the properties of the trendline, left click the newly generated trendline, right click on the trendline, and then select `Format Trendline` as shown in Figure 11. Then, use the options that pop up to the right to change the property of the trendline. The blue box can be used to change the color of the trendline, the orange box can be used to change the line width, and the purple box can be used to change the dashed line to a solid line.

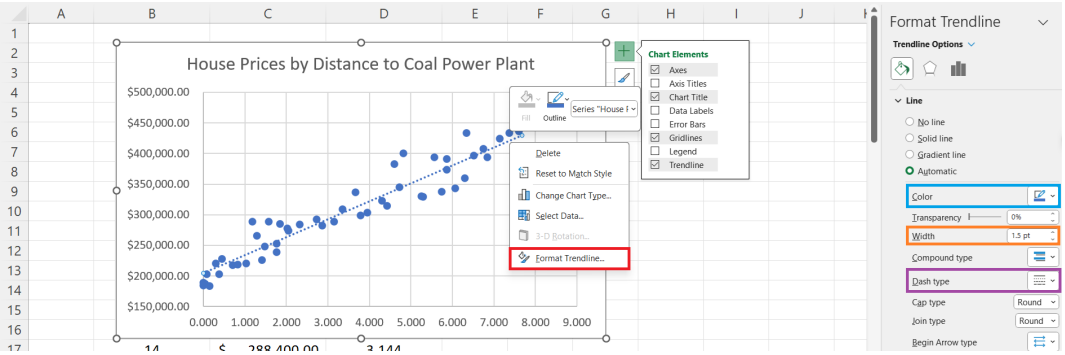


Figure 11: Editing Trendlines

One may change the other elements in the chart by following a similar process. We can assume that the true underlying data generating process is exponential, or logarithmic, or a polynomial of a higher degree. We may change the color and size of the markers, etc.

Negative Correlation

Navigate to worksheet `SCATTER-03`, which contains real-world data on the total fertility rate of a country, alongside with the average years of education that the female population of said country completes between the ages of 15 to 49. Conventional wisdom tells us that as gender equality is advanced, and women gain more opportunities to education, the fertilty rate of a country tends to fall.

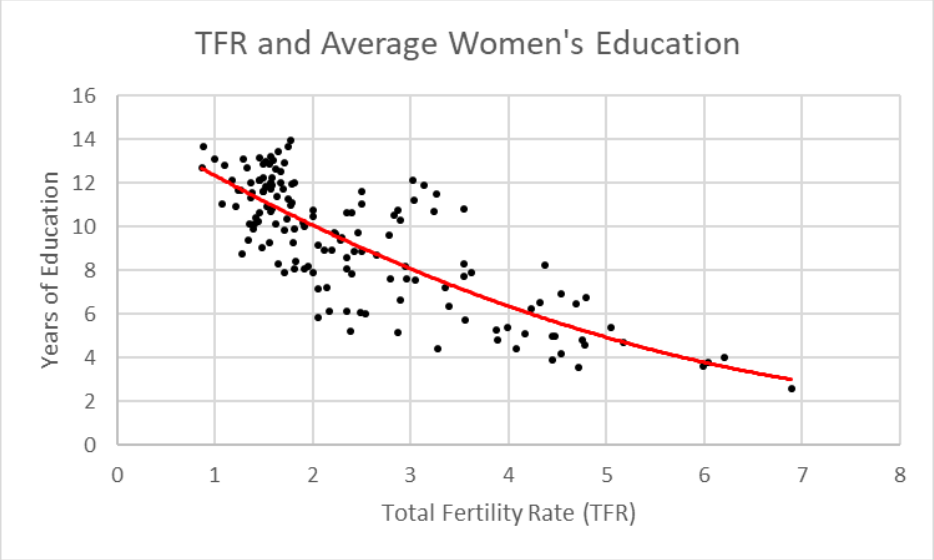


Figure 12: Adding Trendlines

Topic 2. Spinoff: Bubble Charts

Bubble charts are similar to scatter charts, but adds another dimension to our scatter charts. Specifically, the scatter chart plots a single dot that represents the numerical variables we plot on the horizontal and vertical axis. Meanwhile, the bubble chart adds a dimension by varying the size of each dot by the numerical value of a third variable.

To try out plotting a bubble chart, navigate to the worksheet `BUBBLE-01`, which has real world data on the share of carbohydrates out of the total caloric intake of various countries. Alongside this information, we have information on the GDP per capita, and the population estimates. Figure 13 shows a bubble plot with the bubble size representing the estimated total population of the country.

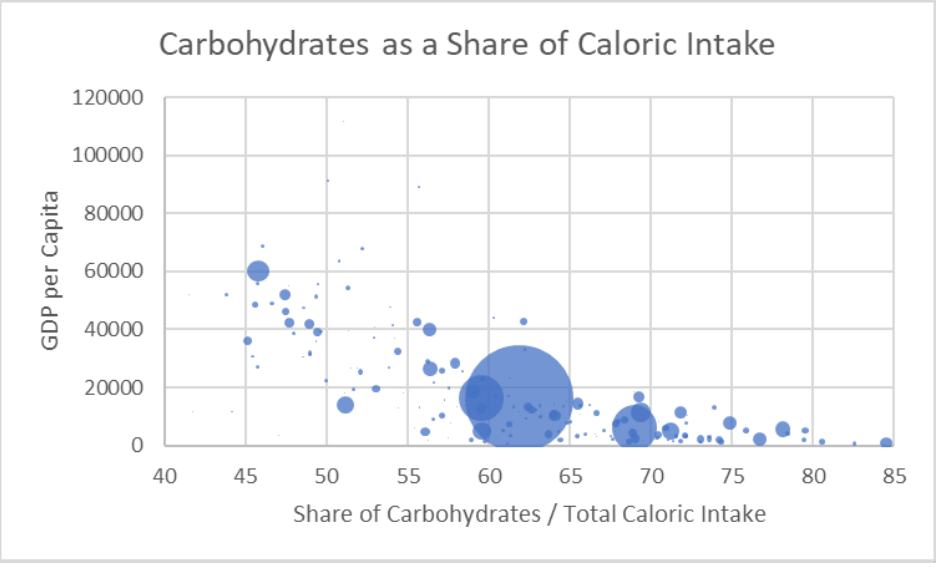


Figure 13: Bubble Chart

One of the reasons that we see some "bunching" on the bottom of the chart is due to some of the smaller countries with extremely high per capita GDP values such as Luxembourg, Qatar, or the UAE. One way to deal with this situation may be to use logarithmic scaling on the vertical axis. See Figure 14 for the bubble chart with a logarithmic axis, and how to apply the logarithmic scale to an axis.

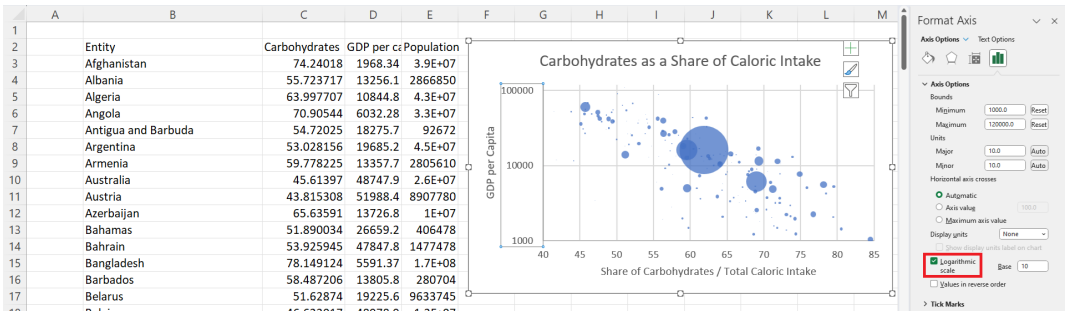


Figure 14: Bubble Chart with Logarithmic Axis

Topic 3. Pie Charts

Pie chars are typically used to visualize how some total amount is divided into its subcategories. For instance, pie charts are often used in demographic data to break down a total population by ethnicity, gender identity, nationality, employment status, etc. The larger the subgroup, the larger share of the "pie" it represents.

Navigate to the worksheet `PIE-01` for Illinois state data on the breakdown of the population by self-reported ethnicity. The first row (after the variable names are defined) reports data aggregated over the entire state, and the following rows each report the ethnicity breakdown by county. The pie chart in Figure 15 plots the breakdown of the self-reported ethnicity of residents of Illinois.

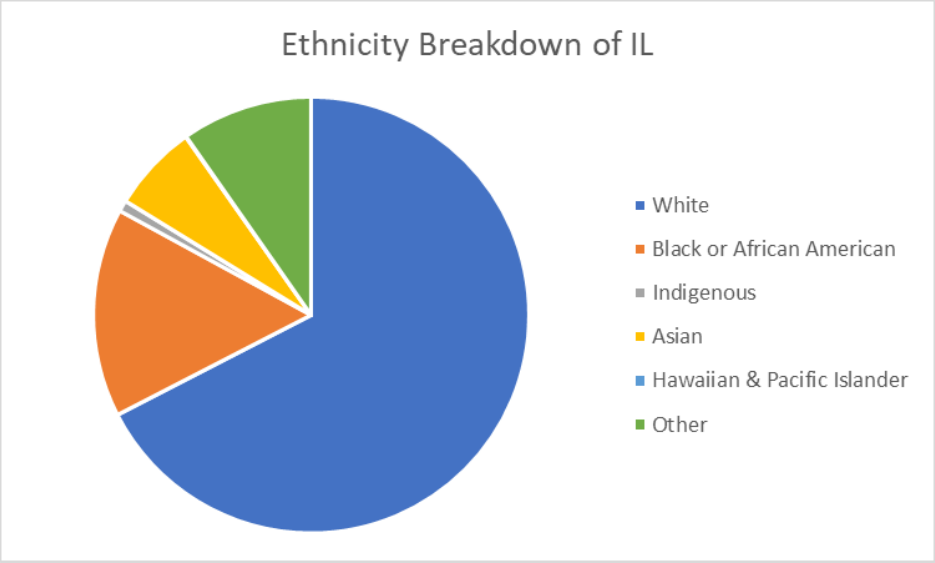


Figure 15: Population Breakdown of Illinois

There are some caveats when it comes to generating the pie chart above. To talk about them, it is best to first talk about when pie charts should be used. Pie charts are recommended when a total amount is partitioned into smaller distinct subgroups. Also, pie charts are similar to the "100% stacked" versions of the bar and line charts in the sense that it does not visualize the magnitude of each subgroup, just the shares. So if the objective is to visualize the size of each element, the pie chart is not your best choice.

	A	B	C	D	E	F	G	H	I	J	K
1					White	Black or African Ar	Indigenous	Asian	Hawaiian & Pacific	Other	
2		Name	Total Population	Total (One Race)	White (One Race)	Black or African Ar	American Indian a	Asian (One Race)	Native Hawaiian a	Some Other Race	Two or More Races
3		Illinois	12,812,508	11,667,524	7,868,227	1,808,271	96,498	754,878	4,501	1,135,149	1144984

Figure 16: First Row of Data

Notice that in the data, there are two "total" populations; one in column `C`, and another in column `D`, both in the red box. Under this scenario, we must choose the data to plot carefully whether the total we are adopting is the one in column `C` or column `D`, due to the existence of the `Two or More Races`. If we are using the data in the blue box, we should use column `D`, if we are combining it with `Two or More Races`, we should use column `C`.

Another point we wish to demonstrate is that the pie chart truly does not represent the magnitude of the items. Consider the following two pie charts, one representing the ethnicity breakdown of Cook county (Population: ~ 5 million) in Figure 17, and the same breakdown in Warren county (Population: ~ 15k) in Figure 18. Comparing between pie charts are rarely useful, as there may be scale differences that are not easily noticeable.

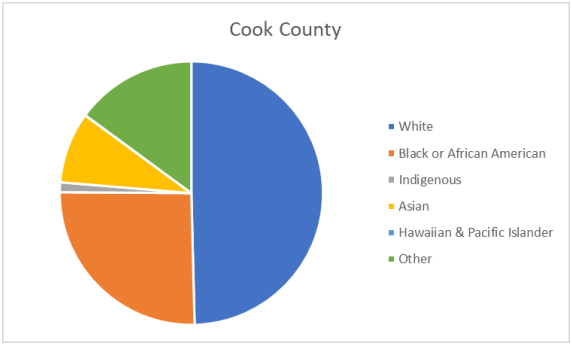


Figure 17: Ethnicity of Cook County

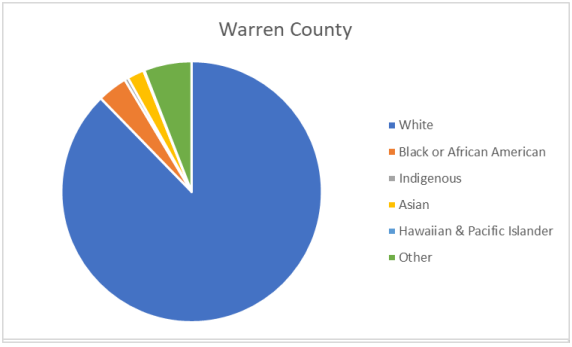


Figure 18: Ethnicity of Warren County

Minor Spinoffs: Treemap Chart and Doughnut Charts

Some minor spinoffs of the pie chart include the treemap chart and the doughnut chart. The treemap chart of the ethnicity breakdown of Illinois is displayed in Figure 19, and the doughnut chart equivalent is in Figure 20. There is no material difference between the pie, doughnut, or the treemap chart, so the choice is up to the individual responsible of visualizing the data.

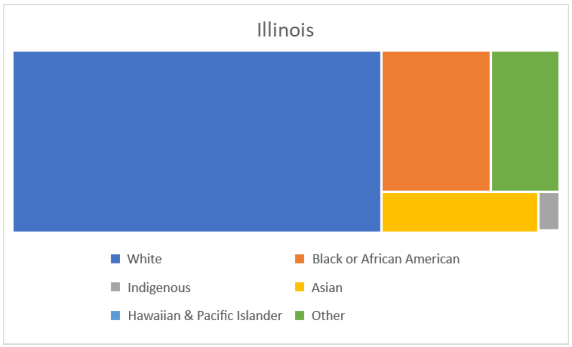


Figure 19: Treemap Chart of Illinois

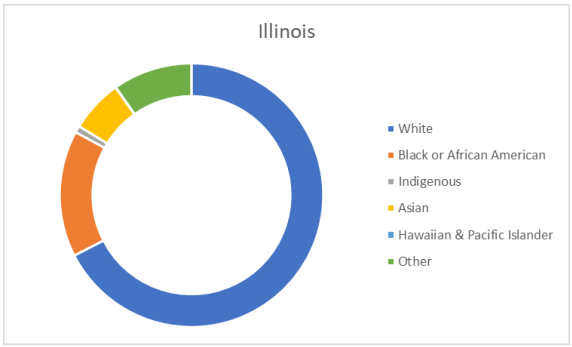


Figure 20: Doughnut Chart of Illinois

The "Too Many Subcategories" Problem

Sometimes, when there are too many categories in the data, a pie chart or a treemap chart will become too dense. For an example, navigate to the worksheet `PIE-02`. There are a total of 102 counties in the state of Illinois, and Figure 21 is a pie chart that attempts to show all 102 counties. Clearly, this is not an effective method of visualizing data.

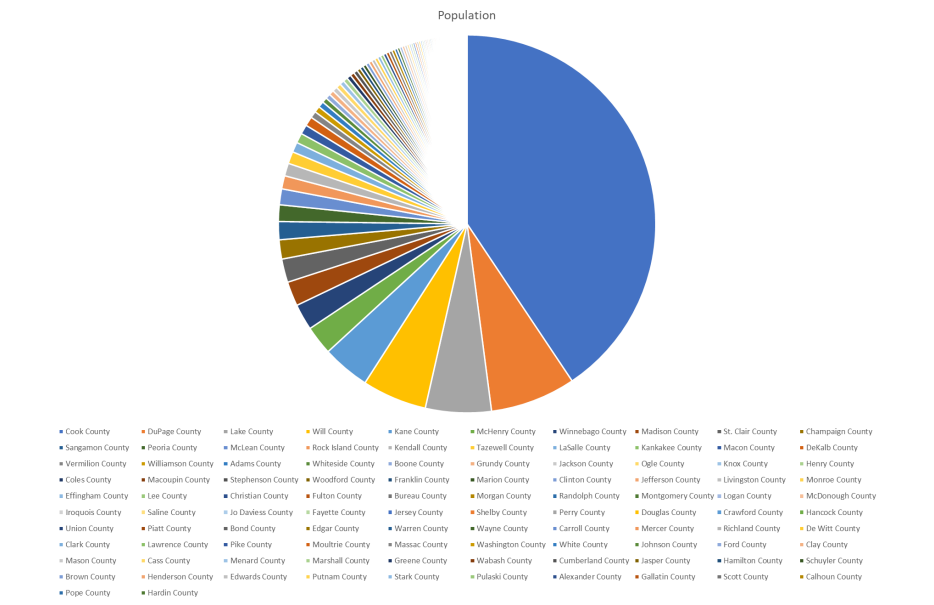


Figure 21: Congestion in a Pie Chart

One way to deal with this issue is to aggregate the smaller counties into one "Other Counties" category as shown in Figure 22.

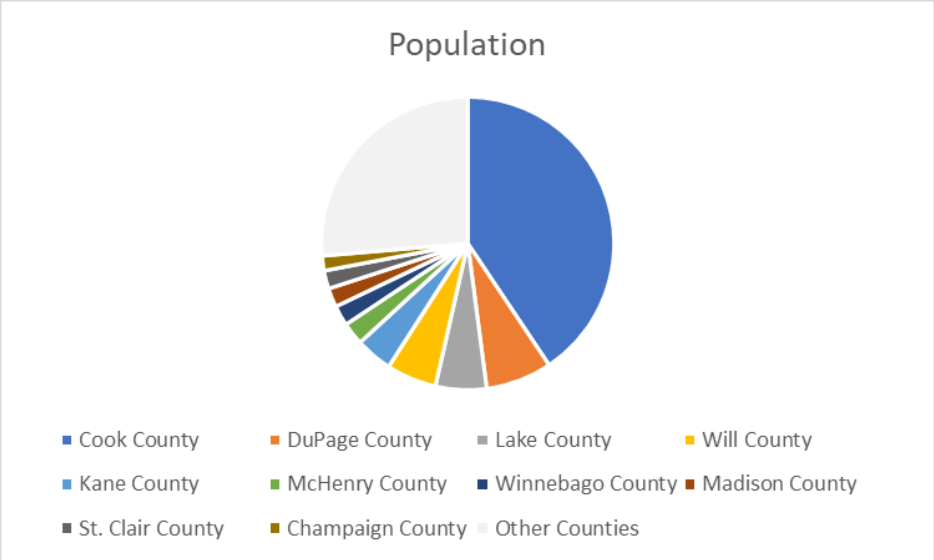
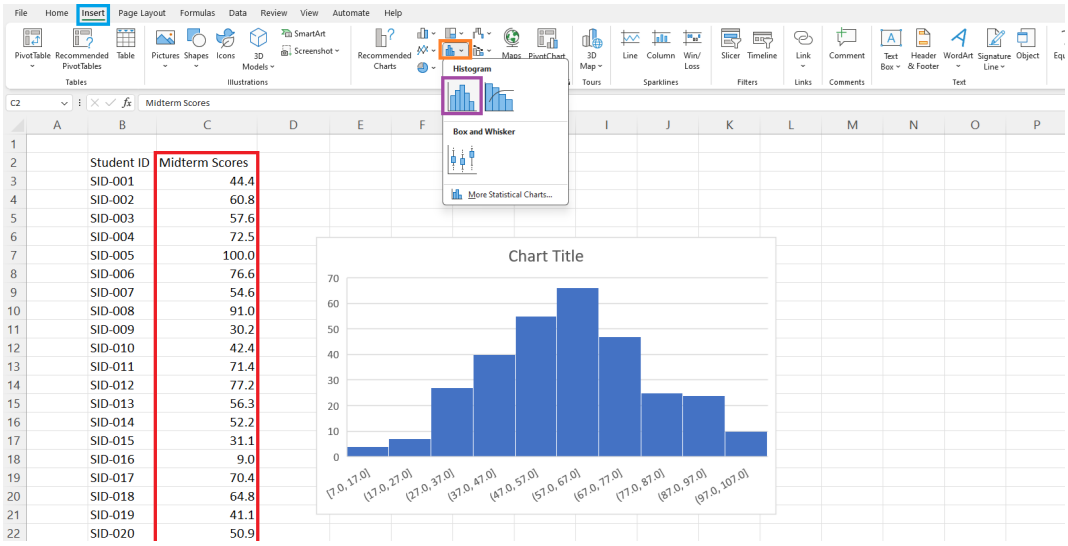


Figure 22: Generating "Other Counties"

Topic 4. Histograms

Histograms are used to visualize the distribution of a numeric variable using bars. Each bar's height depends on how many observations fall within each "bin." Each bin covers a certain range of numeric values that breaks down the horizontal axis into distinct portions. One of the most frequent use of histograms that you may come across is when the distribution of some test results are announced.

Navigate to the `HISTOGRAM` worksheet for a synthetic dataset on students' midterm exam scores. Select the midterm scores in the red box, and select the `Insert` tab, and plot a histogram following the orange box, and then the purple box in Figure 23.



Student ID	Midterm Scores
SID-001	44.4
SID-002	60.8
SID-003	57.6
SID-004	72.5
SID-005	100.0
SID-006	76.6
SID-007	54.6
SID-008	91.0
SID-009	30.2
SID-010	42.4
SID-011	71.4
SID-012	77.2
SID-013	56.3
SID-014	52.2
SID-015	31.1
SID-016	9.0
SID-017	70.4
SID-018	64.8
SID-019	41.1
SID-020	50.9

Figure 23: Histogram of Midterm Scores

The resulting histogram is not ideal in many dimensions, but the most important defect is the definition of the bins. In order to change the bins, right click the horizontal axis labels in the red box, and select `Format Axis`. Then, set up the bin widths to be 10 as shown in the orange box, and then set up the underflow bin to be 10 to match the blue box.

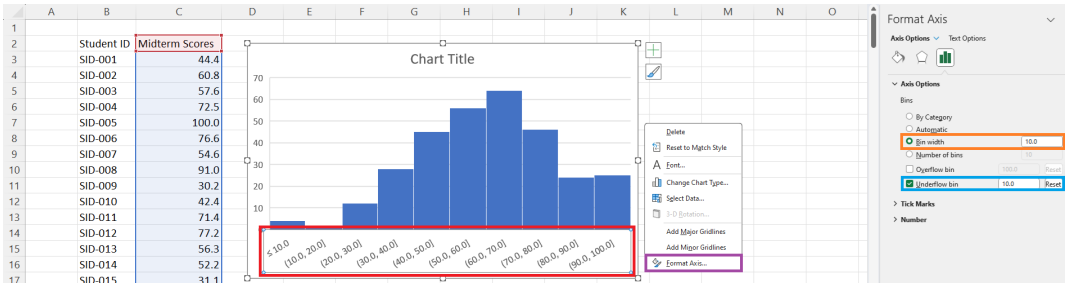


Figure 24: Editing Bin Definitions

Topic 5. Maps and Surfaces

INCOMPLETE VERSION: UPDATED NOTES COMING SOON