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.

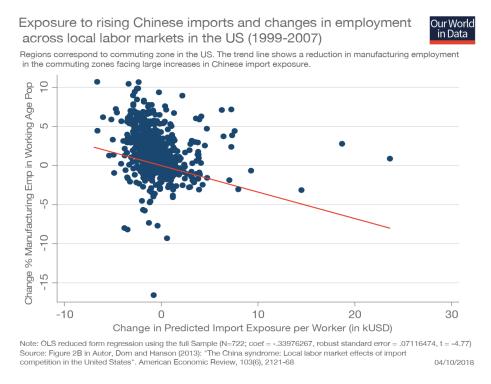


Figure 1: Autor et. al (2013)

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.

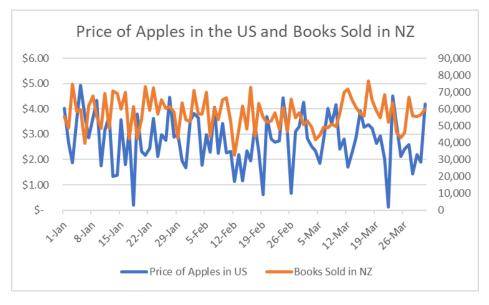


Figure 2: Line Chart with Two Axes

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.

File	Home	nsert Page L	ayout Formulas Data Review	View Automate Help													Comme	ents 🖻 Share
	Table Recomm			a SmartArt Screenshot ~ Recommended Charts @ ~	□· 八 ▲· 尚		art 3D Map ~		olumn Win/	Slicer Ti		Eink Comme	nt Text He Box ~ & F	ader WordArt Si	gnature Object	$\prod_{\text{Equation Symbol}} \Omega$		
	Tables		Illustrations	C1010	Scatter		fis Tours	Spa	arkines	Filter	irs .	inks Conner		Text	Line -	Symbols		
C3	v 1	$\times \checkmark f_{\mathbf{x}}$	Price of Apples in US		10	la 9 l. 1												
	А	В	C	D	•.		G	н		1		К	L	м	N	0	Р	Q
1	~	0		U	13 8		0					K	-	IVI		0		ų
2		Price of	Apples in the US and Nun	nher of Books Sold in NZ	×.	\times												
3		Date	Price of Apples in US	Books Sold in NZ	Bubble													
4		1-Jar		55,462						E	Books	Sold in	ΝZ					
5		2-Jar		48,835	•	33												
6		3-Jar	S 1.87	74,633			90,000											
7		4-Jar	\$ 3.60	58,332	E Mon	e Scatter Charts	80,000					•						
8		5-Jar	\$ 4.94	59,464			70,000			· · · ·			•					
9		6-Jar	\$ 3.73	39,726			60,000		• •	•				<u> </u>				
10		7-Jar	\$ 2.86	61,583			50,000		•	• •••				- T				
11		8-Jar	\$ 3.66	67,738						•								
12		9-Jar		55,537			40,000						•					
13		10-Jar					30,000		•									
14		11-Jar		69,161			20,000											
15		12-Jar		48,466														
16		13-Jar					10,000											
17		14-Jar					0											
18		15-Jar		59,750				\$-	\$1.00	\$2	2.00	\$3.00	\$4.00	\$5.00	\$6.00			
19		16-Jar		69,763														
20		17-Jar		42,328														
21		18-Jar		61,207														
22		19-Jar																
23 24		20-Jar 21-Jar		53,848														
24		21-Jar 22-Jar		59,103														
26		23-Jar	\$ 3.62	72,313														

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.

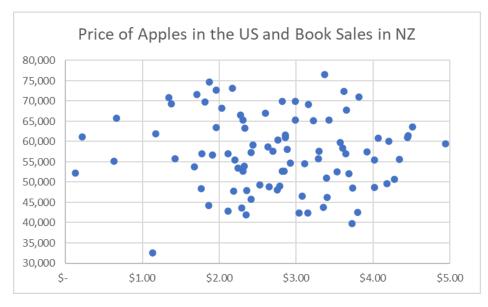


Figure 4: Edited Scatter Chart

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.

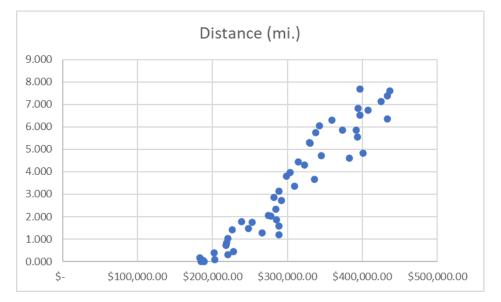


Figure 5: Default Scatter Chart: SCATTER-02

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.

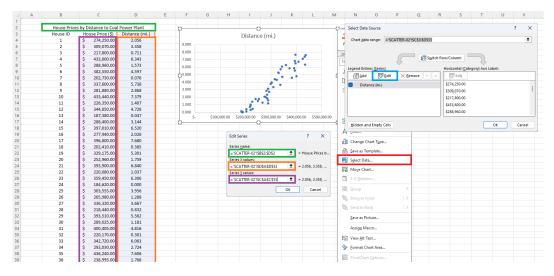


Figure 6: Editing Scatter Chart

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.

A	В	С	D	E	F	G	Н	1	J	K	L	М	N	0	Format Axis \sim \times
1	House Pric	es by Distance to Coal	Power Plant	ррррр							-1	Axis Options 🗸 Text Options			
3	House ID House Price (\$) Distance (mi.)				House Prices by Distance to Coal Power Plant										👌 😟 📓 🔳
4	1	\$ 274,250.00	2.056	\$500,000.	00						1				
5	2	\$ 309,070.00	3.358	\$450,000.	00						Y				Axis Options Bounds
6	3	\$ 217,800.00	0.711	\$400,000.					•	•••					Minimum 0.0 Auto
7	4	\$ 433,600.00	6.343	\$350,000.				•							Magimum 500000.0 Auto
8	5	\$ 288,960.00	1.573	\$300,000.				· ·	••						Units
9	6	\$ 382,550.00	4.597	r							0				Major 50000.0 Auto
10	7	\$ 202,730.00	0.076	\$250,000.											Minor 10000.0 Auto
11	8	\$ 337,600.00	5.738	\$200,000.											Horizontal axis crosses
12	9	\$ 281,880.00	2.868	\$150,000.	00						-				Automatic Axis value
13	10	\$ 433,440.00	7.379	\$100,000.	00										Axis value Maximum axis value
4	11	\$ 226,350.00	1.407	\$50,000.	00						-				Display units None *
15	12	\$ 344,850.00	4.728		-										
16	13	\$ 187,380.00	0.047		0.000 1	.000 2.000	3.000 4.0	00 5.000	6.000 7.0	00 8.000 9	.000				Logerithmic Base 10
17	14	\$ 288,400.00	3.144	0			0				0				scale Base 10
18	15	\$ 397,010.00	6.520												
19	16	\$ 277,940.00	2.026												> Tick Marks
20	17	\$ 396,800.00	7.680												> Labels
21	18	\$ 202,410.00	0.385												> Number
22	19	\$ 329,175.00	5.301												
23	20	\$ 252,960.00	1.759												
24	21	\$ 393,900.00	6.840												
25	22	\$ 220,080.00	1.037												
26	23	\$ 359,450.00	6.306												

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.

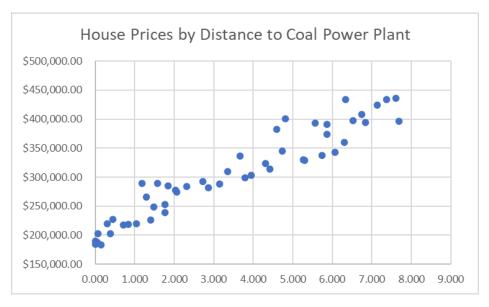


Figure 8: Edited Scatter Chart: SCATTER-02

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.

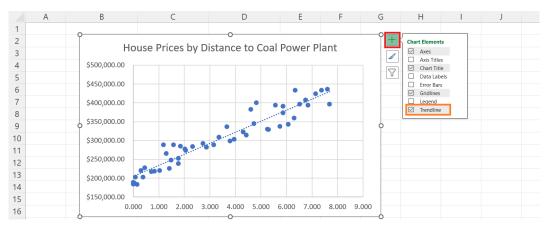


Figure 10: Adding Trendlines

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.

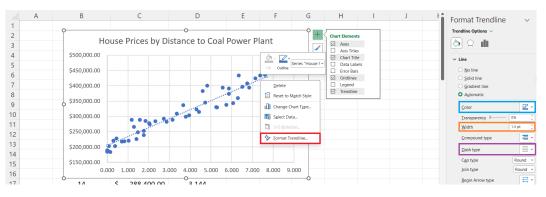


Figure 11: Editing Trendlines

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.

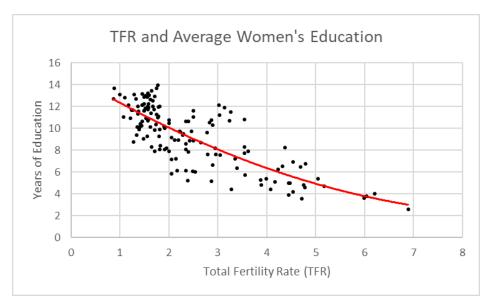


Figure 12: Adding Trendlines

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.

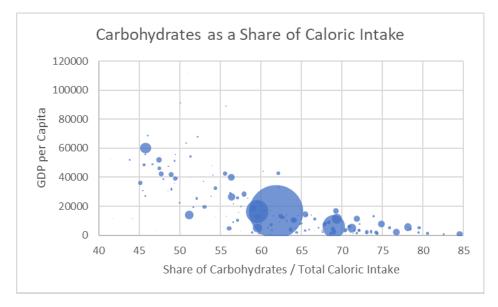


Figure 13: Bubble Chart

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.

	A	В	С	D	E	F		G	н	1	J	K	L	M	Format Axis	~ ×	
1						n									Axis Options V Text Option	16	
2		Entity	Carbohydrates G	DP per cal	Population	Ĭ							. +	Ĭ			
3		Afghanistan	74.24018	1968.34	3.9E+07			Carl	bohydrate	es as a Sh	are of (Caloric Int	ake 🛛		🔿 🛛 🔳 🛄		
4		Albania	55.723717	13256.1	2866850		10000	. 1						2	✓ Axis Options		
5		Algeria	63.997707	10844.8	4.3E+07		100000	,					Ľ		Bounds		
6		Angola	70.90544	6032.28	3.3E+07				1 7 3 2 1						Minimum 1000.0	Reset	
7		Antigua and Barbuda	54.72025	18275.7	92672	apita			1.1						Magimum 120000 Units	0 Reset	
8		Argentina	53.028156	19685.2	4.5E+07	Cap				1 - E - E - E - E - E - E - E - E - E -					Major 10.0	Auto	
9		Armenia	59.778225	13357.7	2805610	b - D	10000) <u> </u>	· ·						Minor 10.0	Auto	
10		Australia	45.61397	48747.9	2.6E+07	GDP p									Horizontal axis crosses		
11		Austria	43.815308	51988.4	8907780	10							1.1		O Automatic		
12		Azerbaijan	65.63591	13726.8	1E+07							•	•		Axis valug Maximum axis value		
13		Bahamas	51.890034	26659.2	406478							1 M.	•		-	one v	
14		Bahrain	53.925945	47847.8	1477478		1000	40	45 50	55	60 65	70	5 80	85	Show display units lab		
15		Bangladesh	78.149124	5591.37	1.7E+08			40				10		0.0	Logarithmic scale Base	: 10	
16		Barbados	58.487206	13805.8	280704		Share of Carbohydrates / Total Caloric Intake								Values in reverse order		
17		Belarus	51.62874	19225.6	9633745	ò				0				ó	> Tick Marks		
40		n tr	10 000017	10070.0	4.05.07										2 TICK IMARKS		

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.

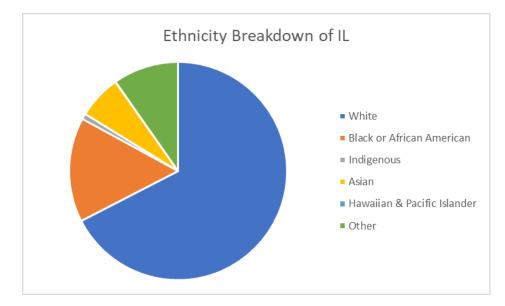


Figure 15: Population Breakdown of Illinois

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.



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: \sim 5 million) in Figure 17, and the same breakdown in Warren county (Population: \sim 15k) in Figure 18. Comparing between pie charts are rarely useful, as there may be scale differences that are not easily noticeable.

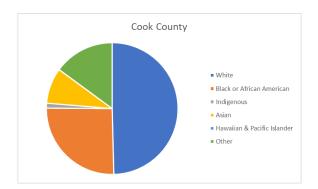


Figure 17: Ethnicity of Cook County

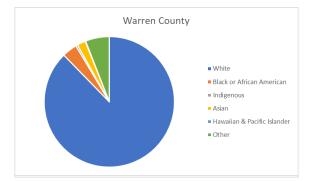


Figure 18: Ethnicity of Warren County

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.

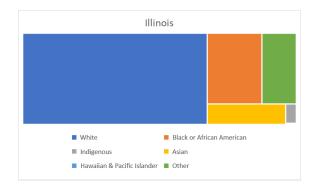


Figure 19: Treemap Chart of Illinois

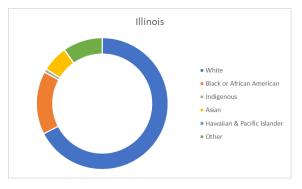


Figure 20: Doughnut Chart of Illinois

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.

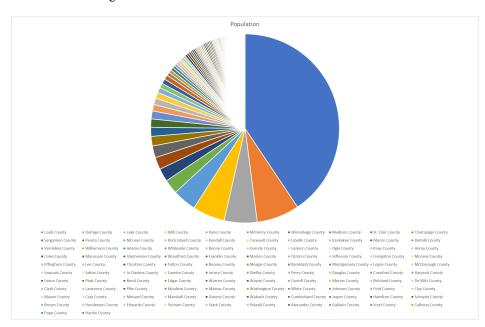


Figure 21: Congestion in a Pie Chart

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

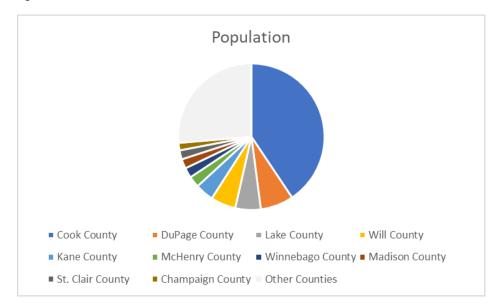


Figure 22: Generating "Other Counties"

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.

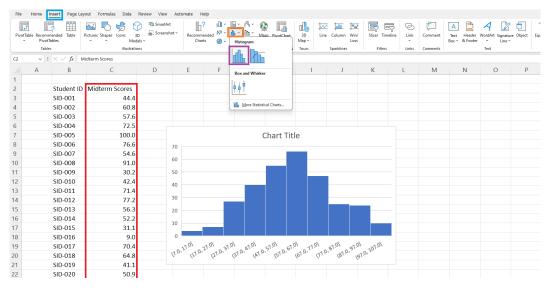


Figure 23: Histogram of Midterm Scores

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.

	A B	C	D E	F	G H	I	J K	L L	М	N O	Format Axis		~
1													Ť
2	Student ID	Midterm Scores	р———		0						Axis Options V Text Options		
3	SID-001	44.4			Chart Title			별) 🔿 🄉 🛄		
4	SID-002	60.8	70					1			✓ Axis Options		
5	SID-003	57.6	60										
6	SID-004	72.5	60								Bins		
7	SID-005	100.0	50								By Category Automatic		
8	SID-006	76.6	40					Delete			O Bin width	10.0	0
9	SID-007	54.6	9 30					Reset to N	latch Style		O Number of bins	10	
10	SID-008	91.0	30					A Font			Ogerflow bin		Reset
11	SID-009	30.2	20					Change C			Underflow bin	10.0	Reset
12	SID-010	42.4	10					🔢 Sglect Dat			> Tick Marks		
13	SID-011	71.4	10					3-D Rotati	on		> Number		
14	SID-012	77.2	ř.					Add <u>M</u> ajo	r Gridlines				
15	SID-013	56.3	\$ 10.0	11 120.0, 30.01 130.0, 40.01	40.0,50.01 (50.0,60.0) (60.0,60.0)	10.01 80.01 9	0.01 100.01	Add Migo	r Gridlines				
16	SID-014	52.2	(10,0,	120:0, 130:0, 1	10,0, 150,0, 160,0,	(70.0, 180.0)	(90.0)	🐓 <u>F</u> ormat Au	is				
17	SID-015	31.1	Ċ.	1	μ. U	1		- d					

Figure 24: Editing Bin Definitions

Topic 5. Maps and Surfaces

INCOMPLETE VERSION: UPDATED NOTES COMING SOON