

Lab 7: Tableau and Categorical Data Analysis

Javier E. Flores

April 10, 2019

Introduction

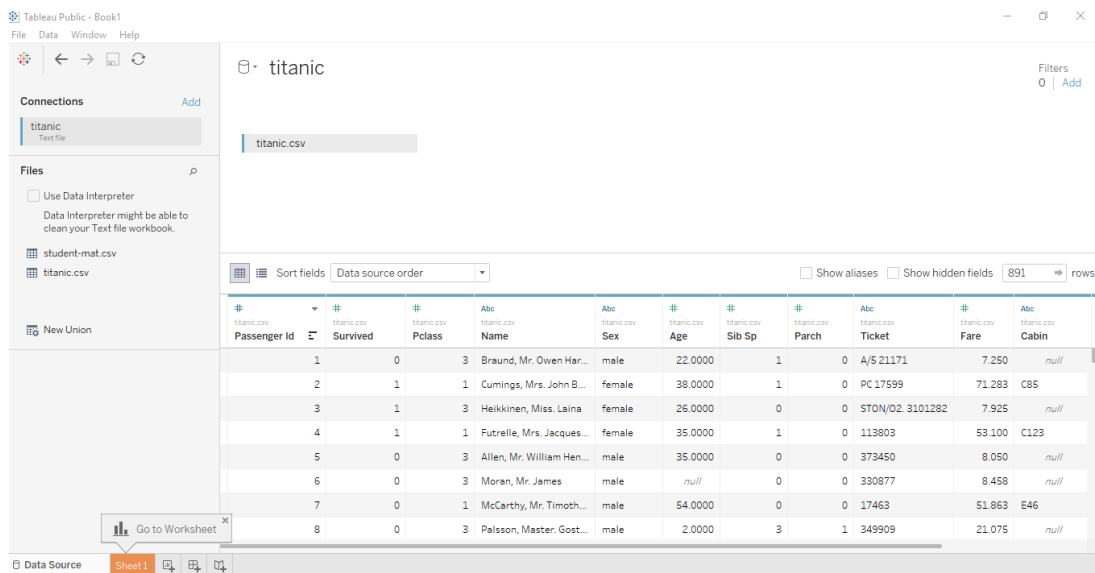
In today's lab, we will learn and apply some of Tableau's features in the visualization of categorical data. Throughout this lab, we will be working with two datasets, both of which may be somewhat familiar. The first is [data](#) on the RMS Titanic, which we briefly explored when learning about Simpson's paradox. The [second dataset](#) is the complete version of the Student Alcohol Consumption data that was introduced the very first day of class. Codebooks for each of the datasets may be found [here](#) (Titanic) and [here](#) (Student).

As you work through the lab with your group, you will be asked to answer several questions. Please submit your responses (as a group) in a single, separate document. Include the original questions as well as your group's response in the final submission. All generated figures must also be included.

Data Visualization with Tableau

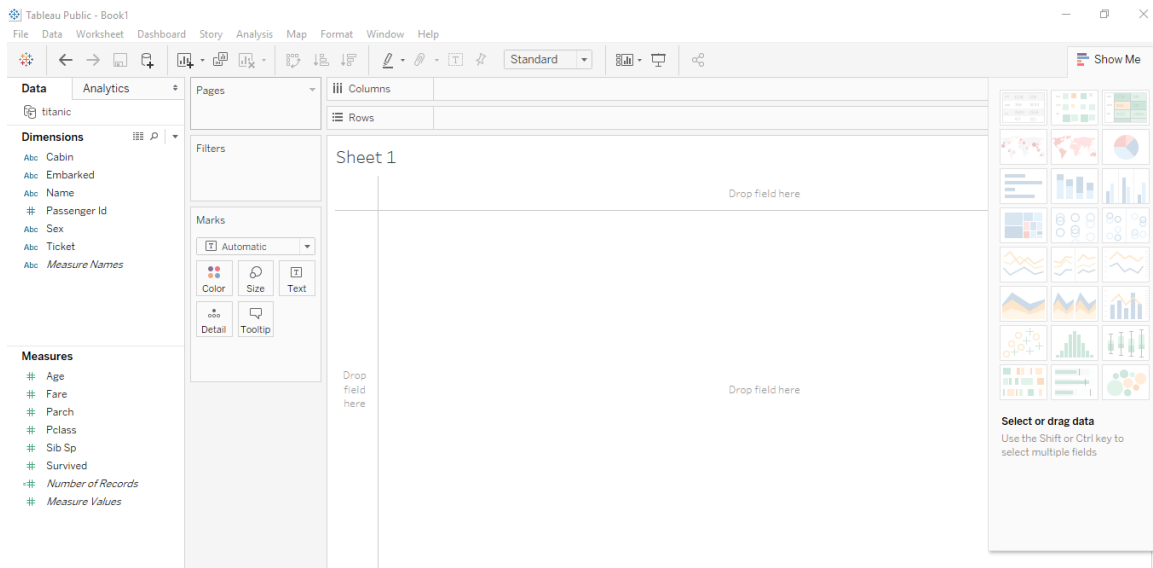
In our very first lab, we learned of several visualization methods and how they can be implemented in Minitab. For categorical data, these included bar charts, pie charts, clustered bar charts, and stacked bar charts. Using the Titanic data, we will see how these same figures (and more) can be created in Tableau.

Prior to generating any figures, we first need to load data into Tableau. To load data into Tableau, open Tableau and select "Text file" if the datafile is a .csv file (it is). A dialogue box should then appear from which you can search for and select the appropriate datafile. After successfully loading the data, you should see a "Data Source" screen which presents the raw data in a way similar to Minitab.



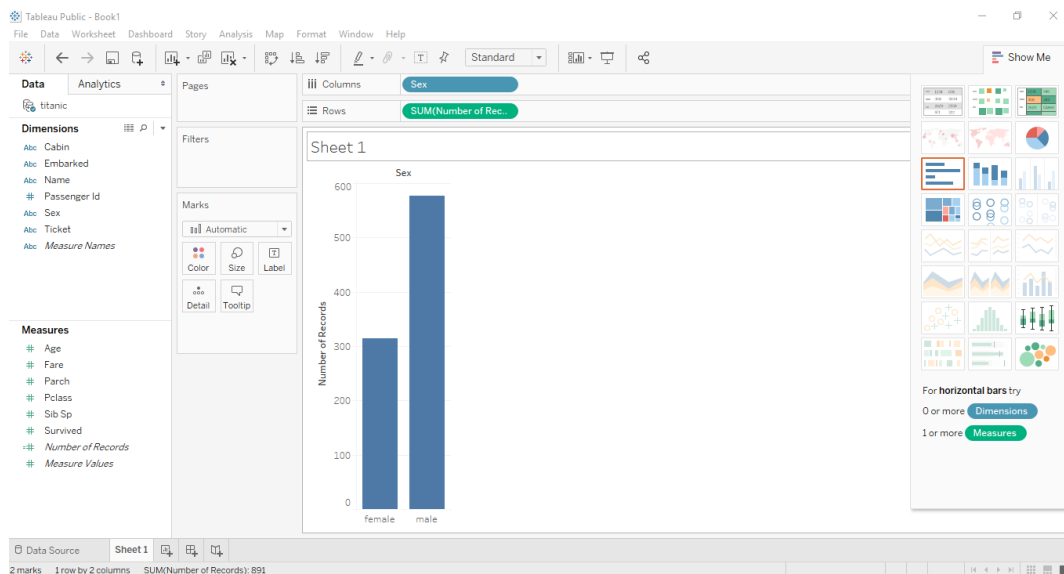
Passenger Id	Survived	Pclass	Name	Sex	Age	Sib Sp	Parch	Ticket	Fare	Cabin
1	0	3	Braund, Mr. Owen Har...	male	22.0000	1	0	A/5 21171	7.250	null
2	1	1	Cummings, Mrs. John B...	female	38.0000	1	0	PC 17599	71.283	C85
3	1	3	Heikkinen, Miss. Laina	female	26.0000	0	0	STON/O2 3101282	7.925	null
4	1	1	Futrelle, Mrs. Jacques...	female	35.0000	1	0	113803	53.100	C123
5	0	3	Allen, Mr. William Hen...	male	35.0000	0	0	373450	8.050	null
6	0	3	Moran, Mr. James	male	null	0	0	330877	8.458	null
7	0	1	McCarthy, Mr. Timoth...	male	54.0000	0	0	17463	51.863	E46
8	0	3	Palsson, Master. Gost...	male	2.0000	3	1	349909	21.075	null

To begin generating figures, we need to edit a new worksheet by clicking "Sheet 1" shown in the bottom of the above figure. After doing so, you should see the following screen:



On the left of this screen is a field containing all of your variables, each grouped under either "Dimensions" or "Measures". The "Dimensions" group corresponds to all those variables which are categorical, while the "Measures" are all those that are quantitative.

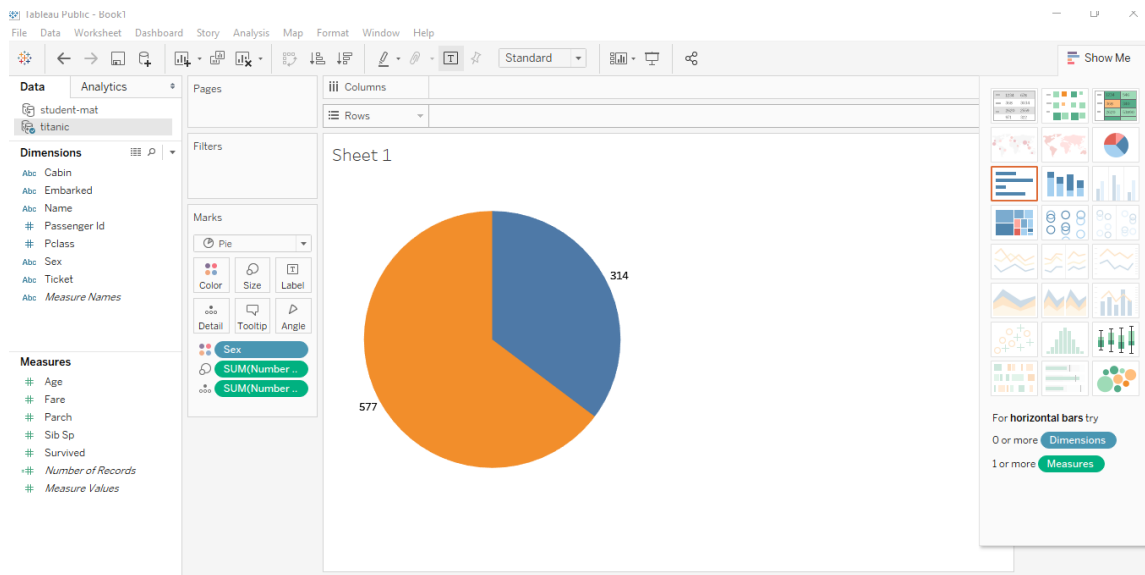
On the right side of this screen is where you will see several figure options visually listed under the "Show Me" tab. Hovering over each option will display the name of the figure and the data type necessary for its use. Hovering over the "horizontal bars" option (i.e. bar chart), we see that we need to specify one or more measures. To create a bar chart for a given categorical variable, we need to first drag and drop a categorical variable (from under "Dimensions") onto the "Columns" field. Next, from among the measures variables, drag and drop the "Number of Records" variable into the "Rows" field. Shown below is the bar chart for "Sex". Different aspects of this graph may be edited using any of the options listed under "Marks".



Q1) In Tableau find a way to convert the variable "Pclass", which describes the class of each passenger, into a categorical variable. Create a bar chart displaying the number of passengers in each class. Be sure to add a label of counts above each bar indicating the exact amount of passengers represented in a given bar.

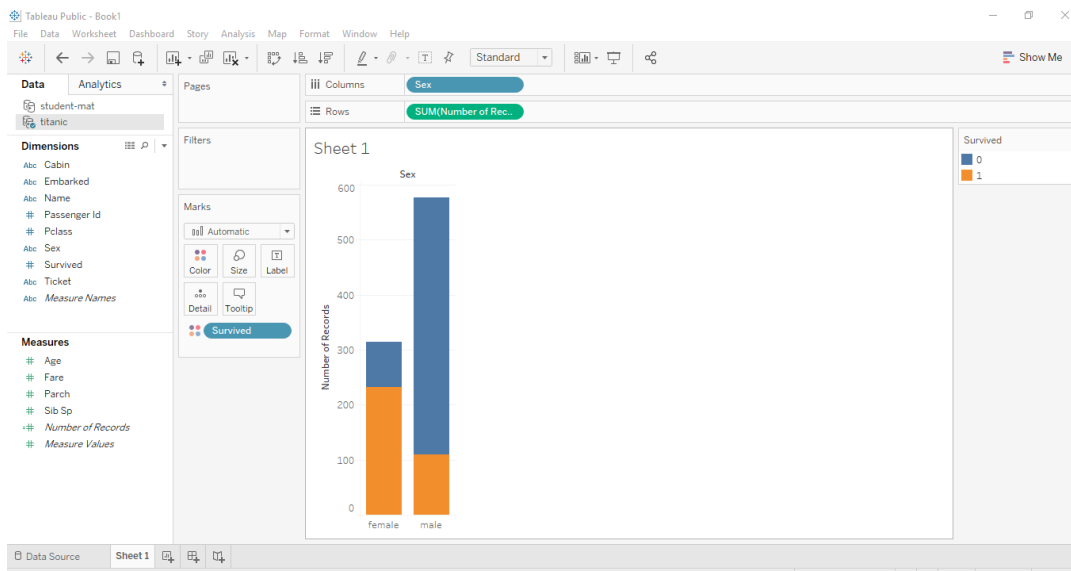
Q2) Drag and drop the variable "Survived" over the "Size" and "Color" options. Describe the change that this induced in your graph. From this graph, what can you say about survival across each class?

Creating a pie chart in Tableau is as simple as selecting the "pie chart" option from under the "Show Me" tab:

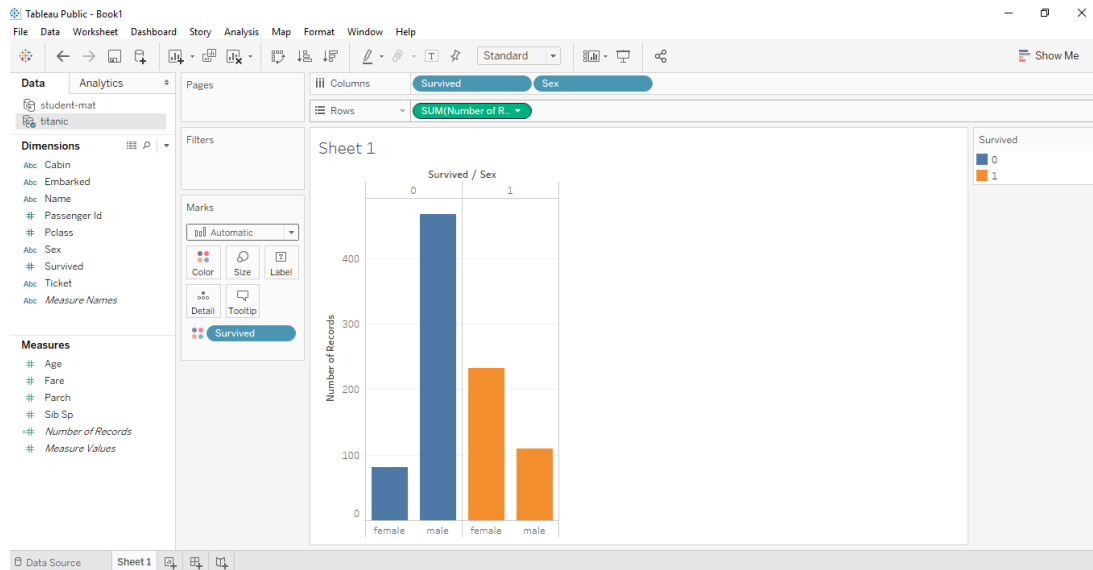


Q3) Create a pie chart for "Pclass". Label each section of the chart with the total counts and the corresponding percentage for each category.

To generate a stacked bar chart, first create a simple bar chart. Then, drag and drop the categorical variable of interest over the "Color" option. This variable will create the stacks within each existing bar according to the categories of the variable dropped over "Color". Shown below is the stacked bar chart for sex, with the size and color of each stack corresponding to the amount of individuals who died (0) or survived (1).



This stacked bar chart can easily be converted to a clustered bar chart by selecting "side-by-side" bars under the "Show Me" tab. Note that, for clustered bar charts, the ordering of the categorical variables listed in the "Columns" field affects which variable is treated as the clustering (outer) variable and which as the inner variable. In the following example, "Survived" is listed first and so is treated as the outer variable. "Sex" is the inner variable.



- Q4) Create a clustered bar graph with "Pclass" as the outer variable and "Survived" as the inner variable. What do you observe about survival across each class?
- Q5) Create a clustered bar graph with "Pclass" as the outer variable and "Survived" as the inner variable. Within each displayed bar, display stacks corresponding to the amount of male and females. Add labels of the number of records corresponding to each stack.
- Q6) Proportionally, which class had the most women? How does the proportion of women relate to the survival across each class? (I am not asking for you to perform a test, just what you might conclude).
- Q7) Using either the titanic or student alcohol data, explore three of the other data visualization options provided under the "Show Me" tab in Tableau. You may choose any variable(s) to use. Describe and interpret each of the three figures generated.
- Q8) According to [Pordata](#), a base of certified statistics on Portugal, the 2005 national proportions of educational attainment among those older than 15 were 13.6% for no schooling, 31.2% for primary education, 32.4% for 5th to 9th grade, 13.4% for secondary education, and 9.4% for higher education. Using Tableau,
- Generate figures appropriate for visualizing the educational attainment levels of father's and mother's separately (from the student alcohol consumption data).
 - Perform a statistical test to determine whether the distribution of educational attainment among fathers within our sample is consistent with national proportions. Be sure to state your null and alternative hypotheses, provide a test statistic, find a p-value, and interpret your results.
 - Repeat the same analyses for the educational attainment of mothers.
- Q9) Perform a statistical test to determine whether there is an association between a student's address type and their desire to pursue higher education. Be sure to state your null and alternative hypotheses, provide a test statistic, find a p-value, and interpret your results. Use Tableau to generate a figure which best communicates your findings. Describe and interpret the chosen figure.

Challenge (Optional)

The challenge for this lab is to replicate all of the Tableau-generated graphics from this lab using R. You may work on this challenge as a group, but I expect individuals to turn in separate copies of their own code and generated statistics/figures.

Turning in this challenge may prompt a brief assessment in which you will be asked to state the code used to generate any one of the results produced.

This challenge may be turned in and redeemed for credit at any point prior to the week of project presentations, which will be during the week before finals.