## 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.

File Data Window Help												-	o ×
$\ast  \leftarrow \rightarrow \square \ \bigcirc$	⊖• titar	nic											Filters 0 Add
Connections Add													
titanic Text file	titanic.csv												
Files p													
Use Data Interpreter													
Data Interpreter might be able to clean your Text file workbook.													
student-mat.csv													
	I Sort fields Data source order 🔹												
III titanic.csv	💷 📃 Sort fi	ields	Data sour	ce ord	ler	Ŧ				Show	aliases Show hide	ien fields 8	391 ⇒ ro
⊞ titanic.csv	III III Sort fi	ields -	Data sour	ce ord	ler #	- Abc	Abc	#	#	#	Abc	ten fields 8	391 ⇒ ro Abc
III titanic.csv	# ttanic.csv Passenger Id	ields T	Data soure # titanic.csv Survived	ce ord	# titanic.csv Pclass	Abc titanic.cov Name	Abc titanic.csv Sex	# titanic.csv Age	# titanic.csv Sib Sp	# titanic.csv Parch	Aliases Show hidd Abc titanic.csv Ticket	# titanic.csv Fare	Abc titanic.csv Cabin
⊞ titanic.csv ∰ New Union	Sort fi     Sort fi     ttanic.csv     Passenger Id	E 1	Data source # titanic.csv Survived	ce orde	ler # titanic.cav Pclass 3	Abc titanic.csv Name Braund, Mr. Owen Har	Abc titanic.csv Sex male	# titanic.csv Age 22.0000	# titariic.csv Sib Sp	# titanic.csv Parch	Abc ttanic.csv Ticket	# titanic.csv Fare 7.250	Abc titanic.csv Cabin
∰ titanic.csv ∰ New Union	Sort fi # ttanic.csv Passenger Id	E 1 2	Data source # titanic.csv Survived	0 1	titanic.cav Pclass 3	Abc titaric.cov Name Braund, Mr. Owen Har Cumings, Mrs. John B	Abe titanic.csv Sex male female	# titanic.csv Age 22.0000 38.0000	# titanic.csv Sib Sp 1 1	# titanic.csv Parch	Abc thanic.csv Ticket 0 A/5 21171 0 PC 17599	# titanic.csv Fare 7.250 71.283	Abc titaric.csv Cabin null C85
⊞ btanic.csv ∰ New Union	i Sort fi	E 1 2 3	Data source # titanic.csv Survived	ce orde	# titanic.csv Pclass 3 1 3	Abc ttaric.cov Name Braund, Mr. Owen Har Cumings, Mrs. John B Heikkinen, Miss. Laina	Abc titanic.csv Sex male female female	#           titanic.csv           Age           22.0000           38.0000           26.0000	# titanic.csv sib Sp 1 1 1 0	# titanic.csv Parch	Abc         Show hide           Abc         Ticket           D         A/5 21171           D         PC 17599           D         STON/02. 3101282	# titanic.csv Fare 7.250 71.283 7.925	Abc titanic.csv Cabin null C85 null
⊞ titanic.csv ∰ New Union	Sort fi	E 1 2 3 4	# titanic.csv Survived	0 1 1	# titanic.cav Pclass 3 1 3 1 3	Abc ttanic.cav Name Braund, Mr. Owen Har Cumings, Mrs. John B Heikkinen, Miss. Laina Futrelle, Mrs. Jacques	Abc titanic.cov Sex male female female female	#         thanic.csv           Age         22.0000           38.0000         26.0000           35.0000         35.0000	# titanic.csv Sib Sp 1 1 0 0	# tkanic.csv Parch	Abc         Show hidd           Abc         Blank.ctv           Ticket         A/5 21171           0         PC 17599           0         STON/02.3101282           113803	# ttanic.csv Fare 7.250 71.283 7.925 53.100	Abc titanic.csv Cabin null C85 null C123
⊞ titanic.csv ∰ New Union	Sort fi	1 2 3 4 5	# titanic.csv Survived	0 1 1 0	ler # titanic.csv Pclass 3 1 3 1 3 3 1 3 3	Abs ttaric.cov Name Braund, Mr. Owen Har Cumings, Mrs. John B Heikkinen, Miss. Laina Futrelle, Mrs. Jacques Allen, Mr. William Hen	Abc titanic.csv Sex male female female female male	#         thanic.sv           Age         22.0000           38.0000         38.0000           26.0000         35.0000           35.0000         35.0000	H titanic.csv Sib Sp 1 1 0 0 1 1 0	# titanic.csv Parch	Abc         Show hidd           Hamic cav         Ticket           0         A/5 21171           0         PC 17599           0         STON/02. 3101282           0         113803           0         373450	# ttanic.csv Fare 7.250 71.283 7.925 53.100 8.050	Abc titanic.csv Cabin C85 null C123 null
⊞ titanic.csv ₽2 New Union	tanic.csv Passenger Id	E 1 2 3 4 5 6	Data sour	0 1 1 1 0 0 0	Her ttanic.cav Pclass 3 1 3 1 3 3 3 3 3	Abs tlancies Name Braund, Mr. Owen Har Cumings, Mrs. John B Heikkinen, Miss. Laina Futzelle, Mrs. Jacques Allen, Mr. William Hen Moran, Mr. James	Abc titaric.cov Sex male female female female male	#           thanic.csv           Age           22.0000           38.0000           26.0000           35.0000           35.0000           35.0000           null	# titanic.csv sib Sp 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	# Show	Abc         Show hide           Band Cav         Ticket           A/5 21171         PC 17599           STON/02.3101282         113803           373450         330877	## ## Fare 7.250 71.283 7.925 53.100 8.050 8.458	Abc titanic.csv Cabin null C85 null C123 null null
⊞ titanic.csv ₽2 New Union	H Sort fi	E 1 2 3 4 5 6 7	Data sourd	ce orde 1 1 0 0 0 0 0	# thank.cov Pclass 3 1 3 1 3 3 1 3 3 1 1 3 3 1 1 1 3 3 1	An Anno. 2007 Name Braund, Mr. Owen Har Curnings, Mrs. John B Heikkinen, Miss. Laina Futrelle, Mrs. Jacques Allen, Mr. Williamens Moran, Mr. Janens McCarthy, Mr. Timoth	Abc titaric.cov Sex male female female male male male	#           thank.csv           Age           22.0000           38.0000           26.0000           35.0000           35.0000           35.0000           mull           54.0000	# titanic.csv Sib Sp 1 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0	H thank.cav	Atc         Atc           ttanic.csv         Titket           0         A/521171           0         PC17599           5TON/02.3101282           113803           373450           330877           17463	# ttanic.cv Fare 7.250 71.283 7.925 53.100 8.050 8.458 51.863	Abo titanic.cav Cabin null C85 null C123 null C123 E46

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:

Tableau Public - Book1 File Data Worksheet Dashboard Story Analysis Map	Format Window Help	- 0 ×
	L ↓ F ℓ · ℓ · T A Standard • Standard •	The Show Me
Data         Analytics         +           Image: The second	III Columns III Rows	
Dimensions Ⅲ P ▼ Abc Cabin Abc Embarked Abc Name	Sheet 1 Drop field here	
Passenger Id     Abe Sex     Abe Sex     Abe Ticket     Abe Measure Names     Color     Size     Text     Detail     Tooltp		
Measures + Age + Fare + Parch + Polass + Sib Sp + Survived + Number of Records + Measure Values	Drop field here Sea	At a second

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.

Tableau Public - Book1 File Data Worksheet Dashboard Story Analysis Ma	p Format Window Help	- 0 ×
	1월 1종 🖉 · // · T 🛠 Standard 🔹 월요 · 🖵 여운	Show Me
Data Analytics + Pages	iii Columns Survived Sex	
🔄 student-mat 🎼 🙀	III Rows V SUM(Number of R. V	
Dimensions Ⅲ P ▼ Acc Cabin Acc Embarked Acc Name # Passenger Id # Polass Acc Tacket Acc Tacket Acc Measure Names # Guid Automatic # Coolin # Survived Acc Tacket Acc Measure Names	Sheet 1	Survived
Measures     # Age       # Fare     # Parch       # Parch     # Sils Sp       # Number of Records     # Measure Values	200 100 0 female male female male	

- 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.