SI649 Fall 2016 – Lab 1-handout

Sep 12/13, 2016 -- DUE SUNDAY, SEPTEMBER 18th at MIDNIGHT

Tableau is used to generate interactive data visualizations and makes very common visualizations very easy. Our focus today is on creating simple visualizations using Tableau. Your GSI (Licia) will give you a brief introduction about Tableau. After the demonstration, you and your lab partner will complete a set of group tasks followed by an individual problem.

Lab Logistics:

1. We have shared 2 documents with you. The **lab1-handout (direct link: http://goo.gl/DUu994)** and the **lab1-answer-sheet (direct link: http://goo.gl/hc6YRQ)**. When you work on the lab, please copy the lab1-answer-sheet (File > Make a Copy), and edit that document for answering questions (you may find it easiest to share the document between your partner and yourself and then make another copy when you're ready to work on the individual part). After completing the lab, please save the answer sheet as uniqName_lab1.pdf (e.g., eadar_lab1.pdf) and upload it to Canvas in the assignments tab.
2. **Though some of you may get through this, we do not expect you to finish this entire lab in class.** Work through as much as you can today and then finish the rest at home to turn in for next week (upload through Canvas by midnight on Sunday, Sep. 18th).
3. Please work with your lab partner in class (though note that the last question is individual). You can continue to work with them on the assignment outside of class. You may not share Tableau workbooks with other groups/individuals (though you can talk through solutions). In class, you can work on one computer (just don't forget to share the Tableau workbook before you leave class today) or you can work on your own laptops. Whatever is more effective for you…

You will upload the following items to the canvas:

☐ the pdf version of the answer sheet
☐ the tableau workbook of the group tasks (each person should upload one, even if you worked on it together)
☐ the tableau workbook of the individual tasks
☐ data used for the individual project (if it's not embedded in the workbook)

■ marks tasks that will be demonstrated during the presentation. When you work on the lab, try to replicate these steps, and then complete tasks that are marked with ☐. Make sure you record your answers on your answer sheet.

Group Task: Making an Investment in the Movie Industry

Scenario: you are a millionaire and you want to invest in the movie industry. However, you are not sure which director, which movie, or what movie genre that you want to invest your money in. In order to make a wise investment, you obtained some data about movies. You want to create visualizations to help you to make wise decisions.

Tableau basics

Please make sure you have Tableau installed and running. You can get a copy at: http://www.tableau.com/tft/activation

Activation Key **TDZA-15FF-86B0-0AAF-6283** (do not share! And please only install one copy)

Make sure you can run Tableau. If you need more information on Tableau take a look at:

- Videos: http://www.tableau.com/learn/training
- Books:
  - http://proquest.safaribooksonline.com/search?q=tableau
  - http://link.springer.com/book/10.1007/978-1-4842-1934-8

----------Short Glossary----------

Sheet: A sheet is a singular chart or map in Tableau.

Dashboard: A dashboard is a canvas for displaying multiple sheets at a time and allowing them to interact with each other.

Workbook: A workbook is the entire Tableau file containing your sheets and dashboards.

Measure: A variable from the dataset that is meant to be aggregated. (This means it should be a number that it makes sense to do math with: sum, average, and so on.) Measures are often continuous data. Examples include GPA, sales, quantity, quota, height, and salary.

Dimension: A categorical variable from the dataset that is used to slice and dice the data into different categories. Dimensions are often discrete data. Examples include country, gender, student ID, and name.

Filter: A filter is used to limit what data is being displayed on the sheet. Visible controls for a filter on a sheet or dashboard are called Quick Filters.

Tooltip: Tooltips are text boxes that appear when hovering over a mark on a sheet in order to give more information. The text and text formatting in them are easily edited through the Marks card.

Marks card: The Marks card is the tool used to create a sheet that controls most of the visual elements in a sheet. Using the Marks card, you can switch between different chart types (bar, line, symbol, filled map, and so), change colors and sizes, add labels, change the level of detail, and edit the tool tips.

Rows and Columns Shelves: The Rows shelf and the Columns shelf is where you determine which variables will go on what axis. Put data you want displayed along the X-axis on the Columns shelf and data you want displayed on the Y-axis on the Rows shelf.

Source version at: http://goo.gl/DUu994 Answer Sheet: **http://goo.gl/hc6YRQ**

(http://www.dummies.com/programming/big-data/big-data-visualization/tableau-for-dummies-cheat-sheet/)

Group Task: The movie dataset

Step 1) The Basics

■Download the datasets from canvas (under week 1 / Lab) open the Tableau app and load the data.

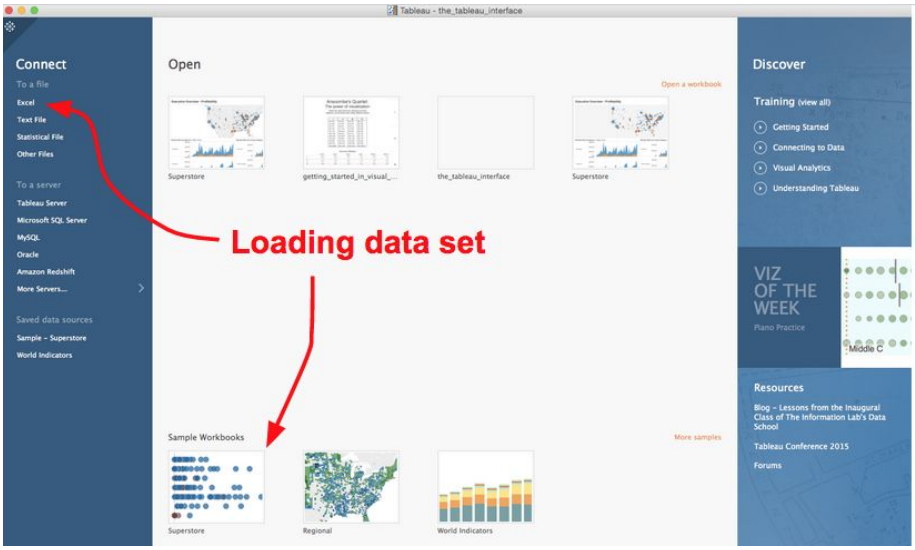


Figure 1

Once you load it, you will see something like this:

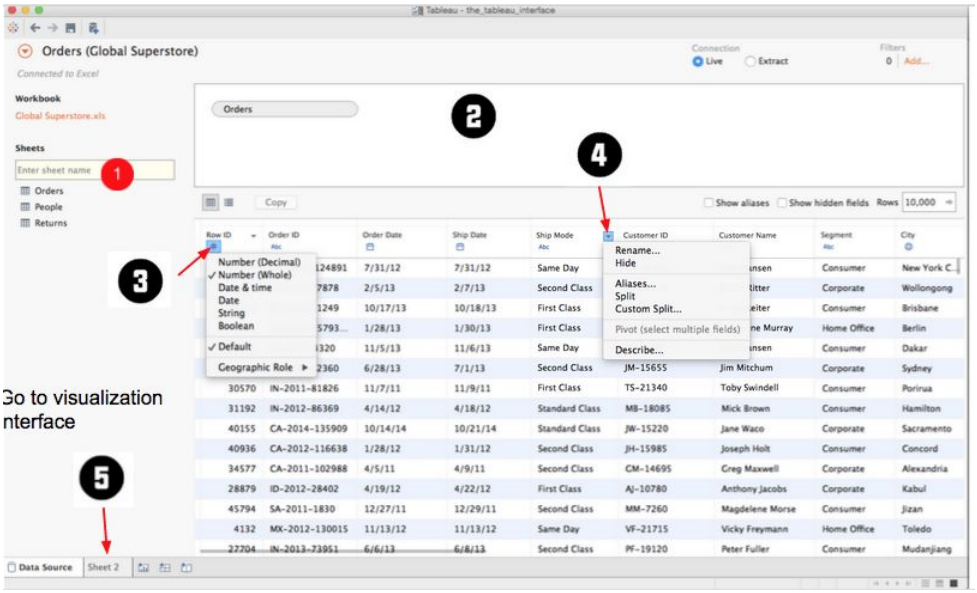


Figure 2

■You can view and edit the data under the "Data Source" Tab.

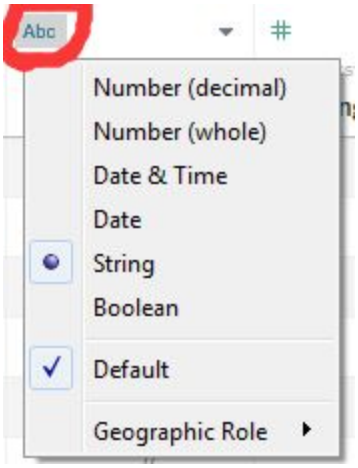
***Question 1*: Which movie has the highest production budget? Which movie has the lowest production budget?**

■1.1 Sort the dataset by production budget (Figure 3)

***Question 2*: How do I sort the dataset by releasing date?**

■2.1 Convert the "Release Date" from a *String* to a *Date*. (Figure 4)

(Figure 3)



(Figure 4)

■2.2 Sort the table by "Release Date".

***Question 3:*Which is the most profitable movie?**

■ 3.1 Sort the dataset by the "worldwide gross"

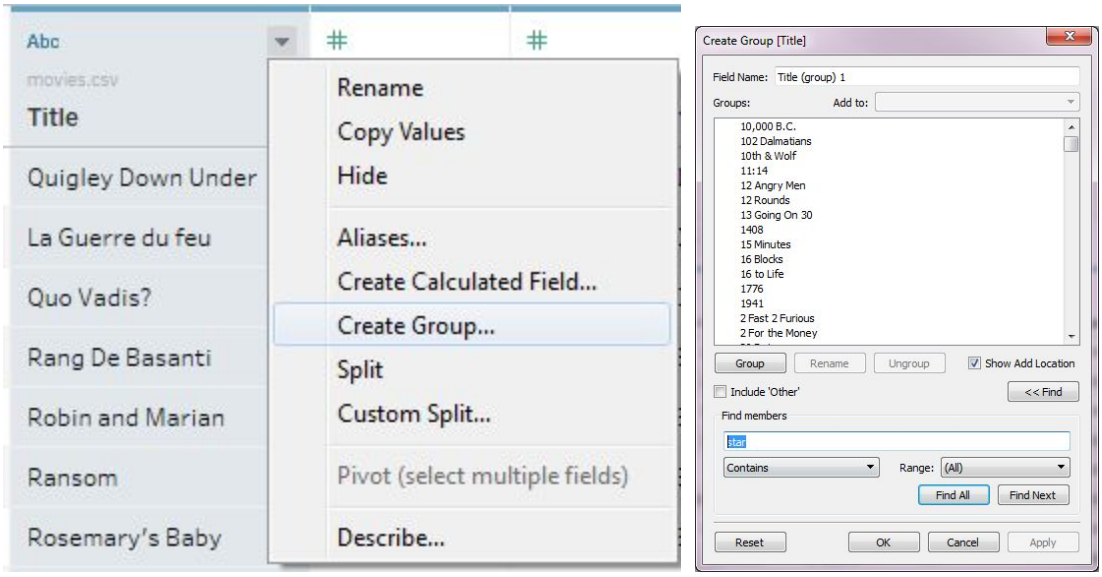
■ 3.2 Create a calculated field called "Revenue" that shows the actual profit that these movies make (hint: Revenue= Worldwide Gross-production budget, assuming that we only care about movies that target the international market) (Figure 5)

(Figure 5)

☐3.3 Which has the largest difference between worldwide gross and U.S. gross?

***Question 4:* How do I find all movies that contain specific strings in their names?**

■ 4.1 Create a group of all movies that contain "star" in their names. (Figure 6)

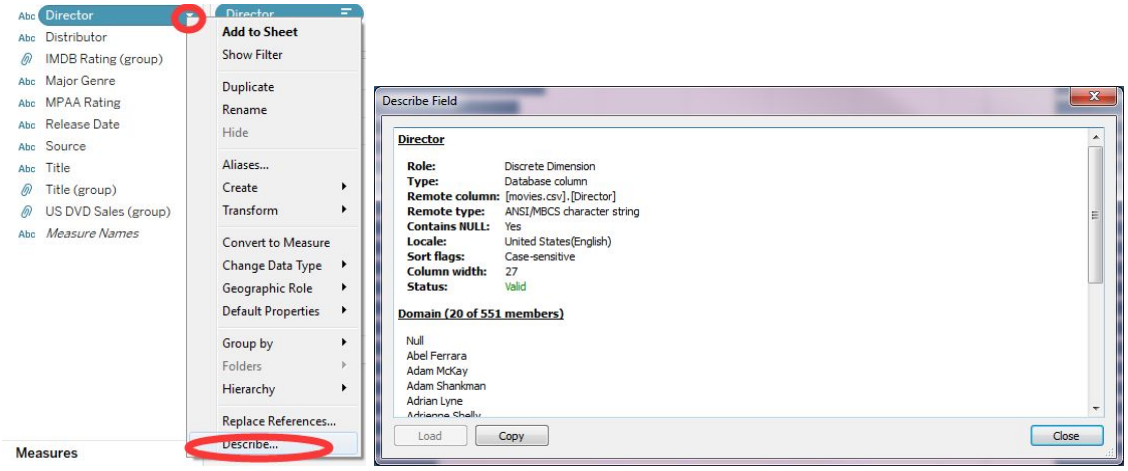


(Figure 6)

☐4.2 Create a group of all movies that that contain the string "dragon" in their names.

***Question 5*: How many directors are included in the dataset?**

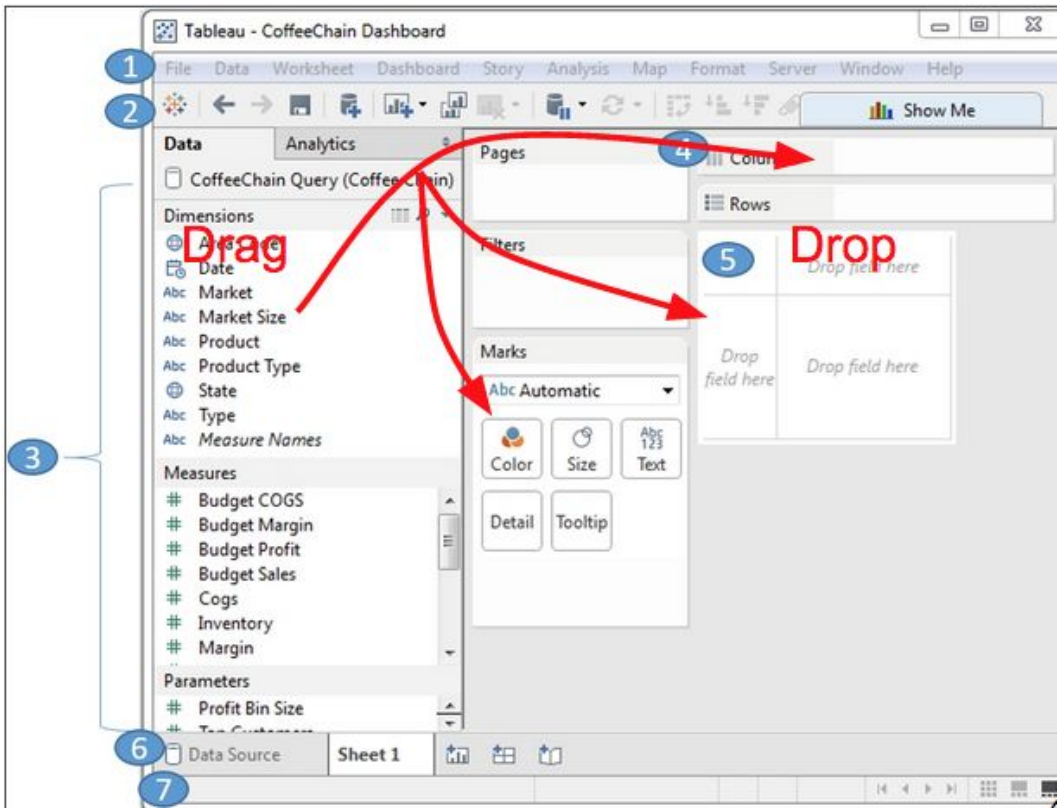
■5.1 Click on the triangle next to " Director"-> select "Describe" . The "Describe Field" window shows the answer. (Figure 8)



(Figure 8)

Step 2) Create Simple Visualizations

After doing basic data manipulations, we are going to create some simple visualizations. Click on the "Sheet 1" tab and you will see a layout similar to Figure 7.



(Figure 7)

1. Menu
2. Toolbar
3. Sidebar for data and analytics
4. Shelves (e.g. pages, marks, rows)
5. Canvas to visualize the data
6. Tabs of different data sheet
7. Status bar to navigate sheets

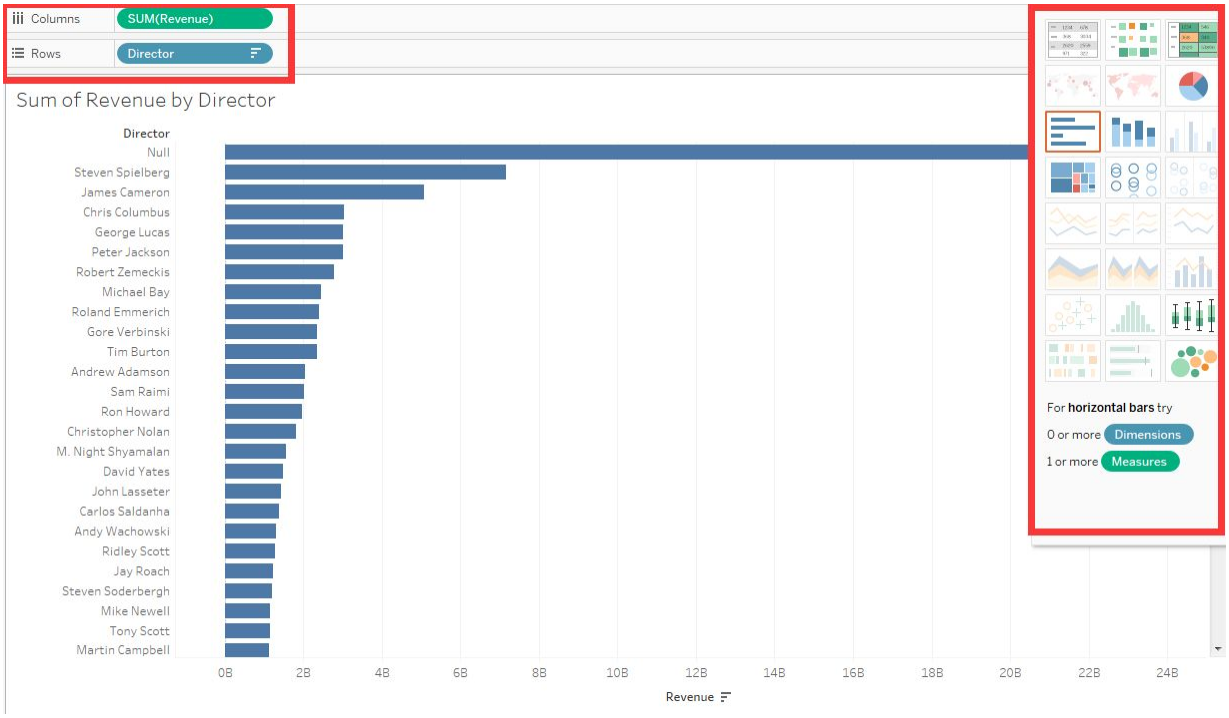
The side bar has two important blocks, **measures** and **dimensions**.

Dimensions are values that determine the **level of detail at which measures are aggregated.** Basically you will have a "mark" for each dimension (e.g., states in the US). You can think of them as slicing the measures or creating groups into which the measures fit. The combination of dimensions used in the view defines the view's basic level of detail. Measures are **values that are aggregated.** They can still be summed, averaged, counted, or they can have a minimum or maximum.

We will see some examples in Question 6 and 7.

Question 6: Who is the most successful director?

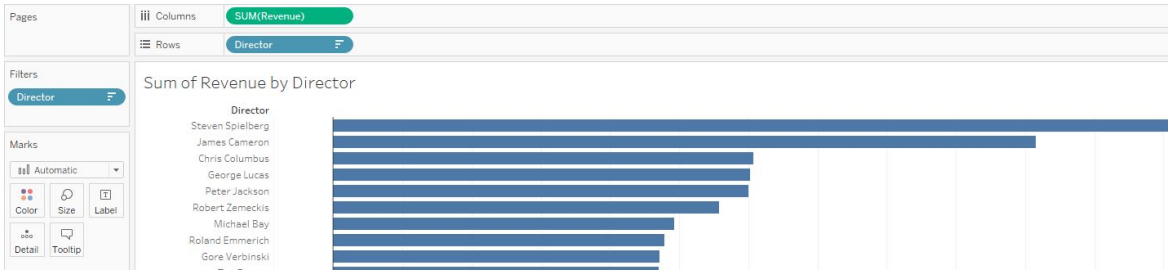
■ 6.1 Drag “Director” to the row , drag “Revenue” to the column. The “Show Me” menu let you create different types of charts (Figure 9).



(Figure 9)

■6.2 Which director has the highest sum of revenue? (Try to create a different visualizations and sort the revenue)(Figure 10)

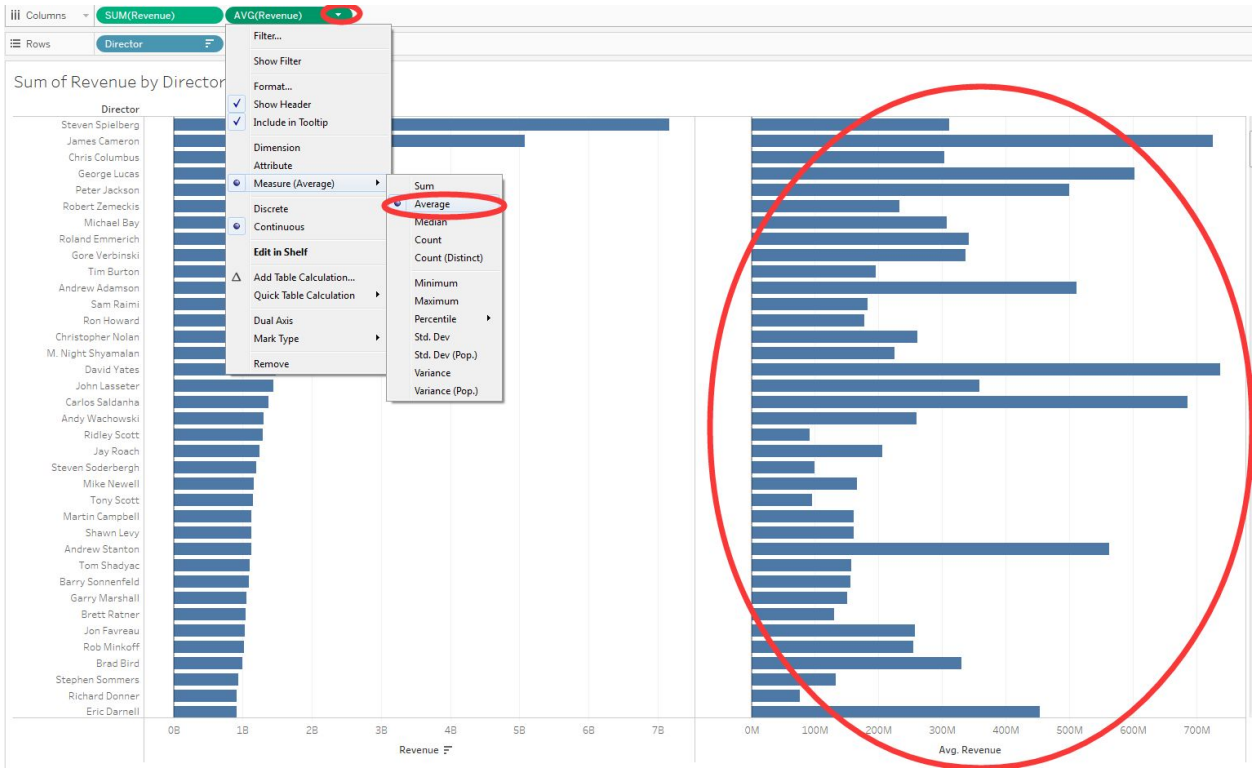
■6.3 Filter Null Values in the “Director” field by right-clicking on the “Null” value and excluding the values. (notice that a new filter is created)



(Figure 10)

■ 6.4 Drag “Revenue” to column and convert it to the measure of average (i.e. 

) (Figure 11).



(Figure 11)

■6.5 Add annotations.

■6.6 Edit the color of the bar by clicking on the “Color” option in the “Mark” shelf.

■6.7 Edit the Sheet1 title to be “Sum and Average of Revenue by Director”.

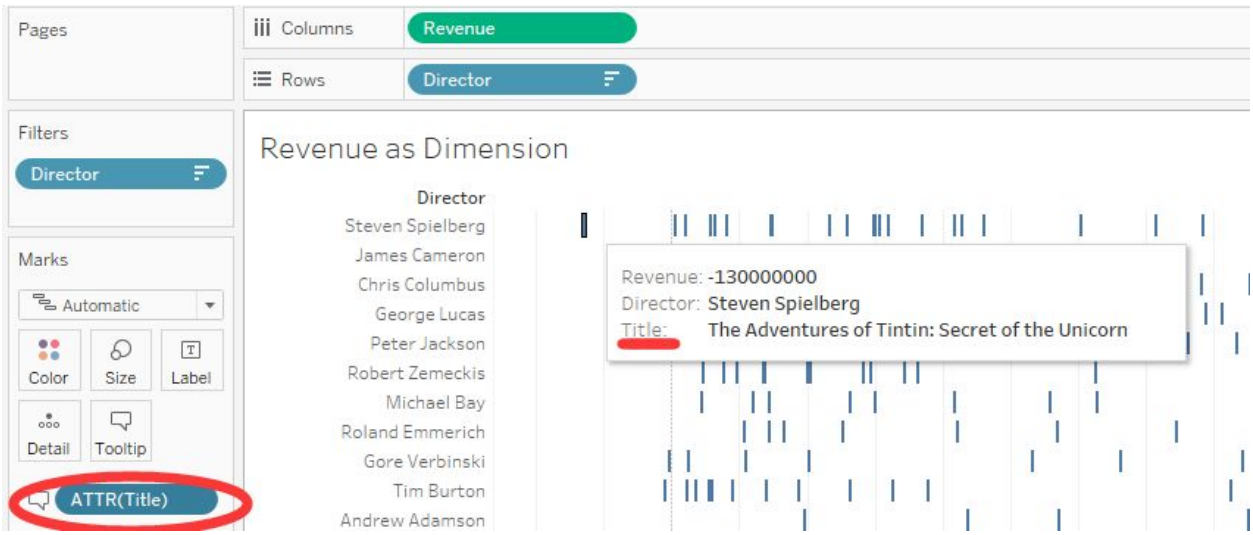
Question 7: **Is there any director who has never directed any profitable movie?**

■ 7.1 Create a new worksheet and name it as “Revenue as Dimension”.

■ 7.2 Drag “Director” to the Rows. Drag “Revenue” to Columns, and convert it to Dimension (Figure 12).

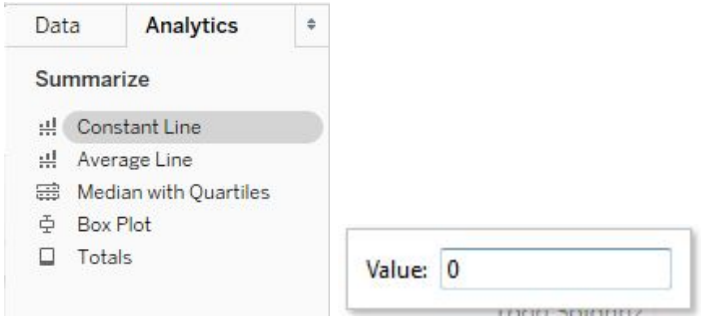
(Figure 12)

■7.3 Add “Title” to the “Mark” shelf as tooltip so that each revenue mark is associated with the movie title (Figure 13)



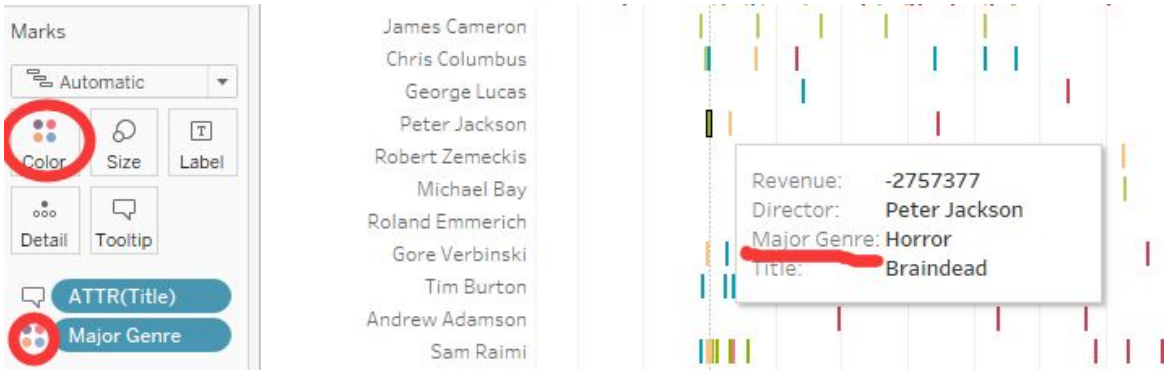
(Figure 13)

■7.4 Add a constant line to indicate movies that have negative revenue (Hint: Analytics-> Summarize->Constant Line-> Value=0) (Figure 14)



(Figure 14)

■ 7.5 Drag "Major Genre" to "Marks" shelf and encode it using color. (Figure 15)



(Figure 15)

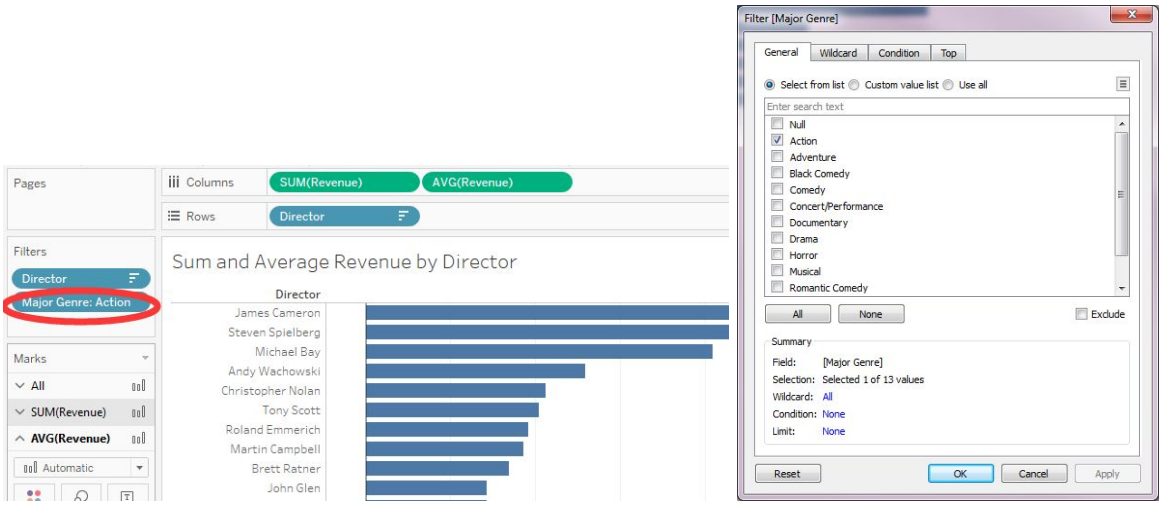
☐ 7.6 Are there directors who have never directed profitable movies? Who are they? (Hint, besides using revenue as a dimension, you can also use filters to complete this task)

Question 8: Can I get more information about a specific genre?

■ 8.1 Go back to the "Sum and Avg of Revenue" worksheet. Drag "Major Genre" to the "Pages" shelf, and display only movies from the genre "Action". (Figure 16)

(Figure 16)

■ 8.2 Drag the "Major Genre" to the "Filters" shelf, and display only movies from the Action genre (Figure 17).



(Figure 17)

☐ 8.3 If you want to find out which director has the highest average revenue for action movies, do you want to use the pages shelf or the filters shelf? Why?

☐8.4 Which director has directed the most profitable Comedy movie? (Hint: use

and filters/pages shelf)

☐ 8.5 Bonus Question: are movies contain the name "star" have higher average revenue than movies contain the name "dragon"?

***Question 9*: When is the best time to release a movie?**

■ Create a visualization of revenue by quarters.

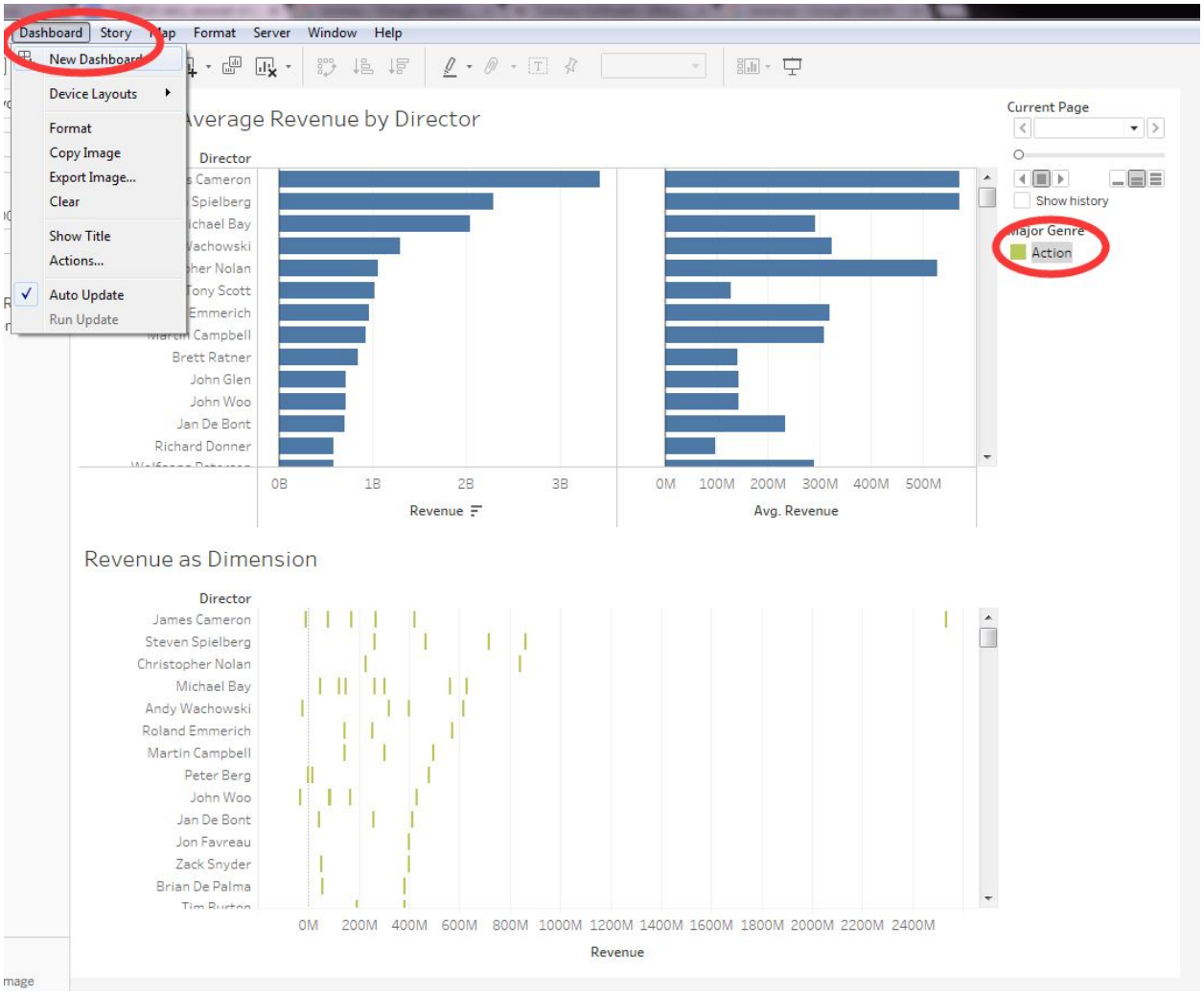
Step 3: Putting it together

***Question 10*: How do I compare two visualizations in the same view?**

■ 10.1 Create a dashboard to host both worksheets that we have created (e.g. "Sum and Avg of Rev" and "Rev as Dimension") (Figure 18).

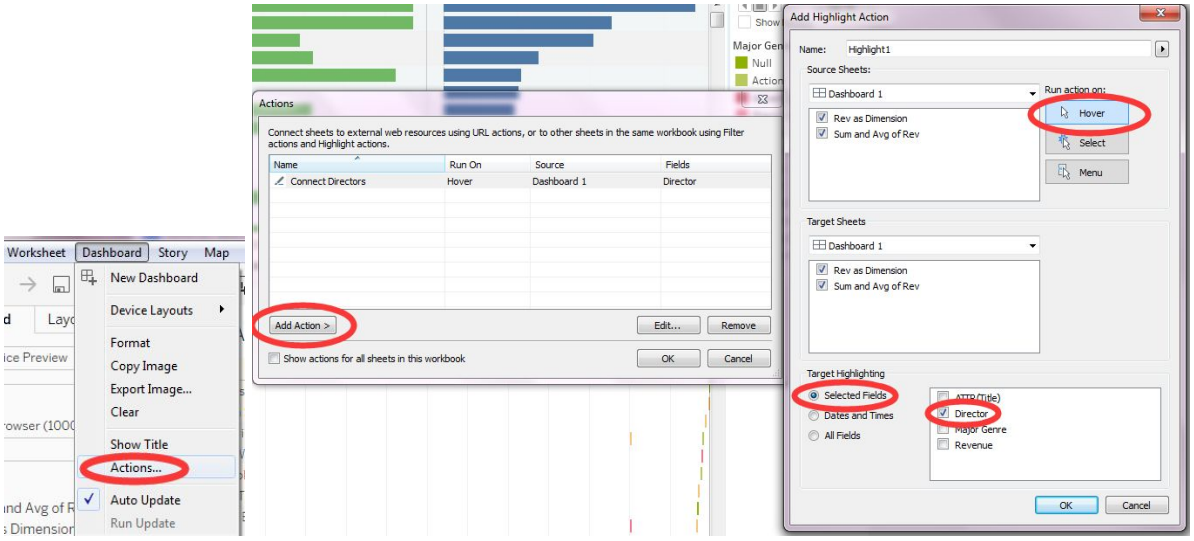
■10.2 Select "Action" from the legend so that only action movies are displayed in the (Revenue as Dimension visualization) (Figure 18).

Source version at: http://goo.gl/DUu994 Answer Sheet: **http://goo.gl/hc6YRQ**



(Figure 18)

■ 10.3 Add a highlight action: “Dashboard”-> “Action”-> “Add highlight”. In the pop up window, add a “Hover” action for the field of “Director”. (Figure 19).



(Figure 19)

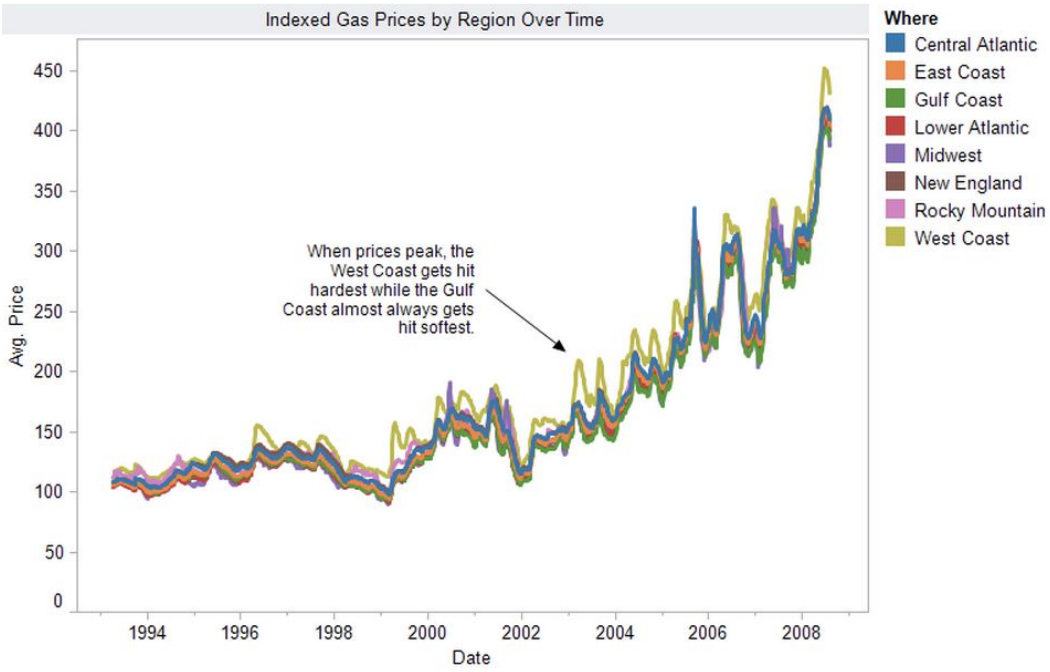
***Question 12*: Is it better to invest in action movies or in comedy movies? Which director would you invest in?**

(This is an open-ended question. Please write down the steps that you used to create the visualizations, and describe how these visualizations help you to make your decision)

***Question 13*: What's the relationship between IMDB ratings and the revenue?Are highly-rated movies more profitable?**

(Please write down the steps that you used to create the visualizations, and describe how these visualizations help you to make your decision)

Individual Task: Open ended exploration



(Borrowed from Maneesh Agrawala)

A wide variety of digital tools have been designed to help users visually explore data sets and confirm or disconfirm hypotheses about the data. The task in this assignment is to use existing software tools to formulate and answer a series of specific questions about a data set of your choice. After answering the questions you should create a final visualization that is designed to present the answer to your question to others.

Try these steps:

- ☐ **Step 1. Pick a domain that you are interested in.**

  Some good possibilities might be the physical properties of chemical elements, the types of stars, or the human genome. Feel free to use an example from your own research, but do **not** pick an example that you already have created visualizations for. Make sure you can find data to support your question. You should not be doing any data collection. Find an existing, clean, dataset in some easy-to-use format (see below).
- ☐ **Step 2. Pose an initial question that you would like to answer.**

For example: Is there a relationship between melting point and atomic number? Are the brightness and color of stars correlated? Are there different patterns of nucleotides in different regions in human DNA?

- ☐ **Step 3. Assess the fitness of the data for answering your question.**

Inspect the data--it is invariably helpful to first look at the raw values. Does the data seem appropriate for answering your question? If not, you may need to start the process over. If so, does the data need to be reformatted or cleaned prior to analysis? Perform any steps necessary to get the data into shape prior to visual analysis.

You will need to iterate through these steps a few times. It may be challenging to find interesting questions and a dataset that has the information that you need to answer those questions.

Data Sets

You should look for data sets online in convenient formats such as Excel or a CSV file. The web contains a lot of raw data. In some cases you will need to convert the data to a format you can use. Format conversion is a big part of visualization research so it is worth learning techniques for doing such conversions. Although it is best to find a data set you are especially interested in, here are pointers to a few datasets: goo.gl/BMbKLA and https://github.com/caesar0301/awesome-public-datasets.

Exploratory Analysis Process

After you have an initial question and a dataset, use Tableau to construct visualizations that provides an answer to your question. As you construct the visualization you will find that your question evolves - often it will become more specific. Keep track of this evolution and the other questions that occur to you along the way. Once you have answered all the questions to your satisfaction, think of a way to present the data and the answers as clearly as possible. In this assignment, you should use existing visualization software tools. You may find it beneficial to use more than one tool.

Before starting, write down the initial question clearly. And, as you go, maintain a notebook of what you had to do to construct the visualizations and how the questions evolved. Include in the notebook where you got the data, and documentation about the format of the dataset. Describe any transformations or rearrangements of the dataset that you needed to perform; in particular, describe how you got the data into the format needed by the visualization system. Keep copies of any intermediate visualizations that helped you refine your question. After you have constructed the final visualization for presenting your answer, write a caption and a paragraph describing the visualization, and how it answers the question you posed. Think of the figure, the caption and the text as material you might include in a research paper.

What to turn in!

At the end, we would like for you to turn in a Tableau file. Name sheets that contain "failed" experiments (e.g., "experiment 1: stocks performance by region"). Name sheets in which you have found interesting insights as "insights" (e.g., "insight 1: stock prices fall during war periods"). Each insight visualization should be well constructed and labeled (put annotations on the images if you want to call attention to specific findings). You should have at least 2-3 insight visualizations and 9-10 experiments. Aesthetics count here, please make sure you make good choices in colors, fonts, layouts, etc. Turn in:

1) **Documentation of what you did/tried (text or PDF). Put in screenshots as needed.**
2) **Your Tableau workbook file**
3) **Any data we need to open your file.**