**Lab Week #7 – Data Analysis and Visualization**

**Response rate**

Survey agreement – right now all yeses. This may be a threat to the validity of the survey. Phone call surveys usually have a 10 – 30% response rate. Did the people you came to all agree to take the survey? Record the “no” answers as well.

Two ways of calculating response rate:

* Pivot table

participateAgreement -> copy to a new sheet -> create a column called count -> populate the count column with 1s -> insert pivot table

* Sort / filter

Usually response rate is a just a number in the report (you don’t have to have a pie chart of it). So one easy way is to sort and record the number, especially for the questions that ask the participant to elaborate.

unionActivityDetail -> sort A to Z (freeze the first row and set “my data has headers”)-> record the number of “SKIPPED”. 40% people have described the detail of union activity.

You can also use “filter”.

**Clean up data**

Sort

Home tab -> Sort and Filter (on the right) -> custom sort -> check on “my data has headers” (important)

Freeze the top row

We want to keep the top row (survey question IDs) visible all the time. We can freeze the top row and keep it always visible while scrolling the rest rows.

Select the first row by clicking the row number “1” (make sure it’s highlighted) -> View tab -> Freeze panes (in the middle of the panel) -> Freeze Top Row

**Format commuteChoice to minutes**

commuteChoice (Column C)

Save a copy, use xlsx format. Otherwise when you save the csv file during the process, there will be a warning message every time you save it.

Copy and paste Column B and Column C to a new sheet. We don’t want to operate in the original sheet. Also, always keep a copy of the original file. Column B here is an identifier column, so you have a reference of the order. What is better is a column of IDs.

Goal: convert commute time to minutes.

Add spaces to the cells that don’t have space between number and text.

Sort Column B from A to Z (expand the selection).

Data tab -> Text to Columns (in the middle) ->Delimited -> Next -> General / Text/ Text ; Destination; Make sure the three columns are all highlighted -> finish

Sort Column C from smallest to largest (expand the selection).

F1 = IF(C1<3,C1\*60, C1)

Note C1<3 here is a judgment by common sense. Have to be careful and double check.

The syntax for the **IF function** is:

IF( condition, [value\_if\_true], [value\_if\_false] )

Populate the rest of Column F (double click “+” or drag).

Data type conversion:

Quick tip: left aligned – text; right aligned – number.

Text to number: value function

VALUE( text )

Number to text: text function

TEXT( value, format )

|  |  |
| --- | --- |
| **=TEXT(A1, "$#,##0.00")** | would return "$7,678.87" |
| **=TEXT(A1, "0")** | would return "7679" |
| **=TEXT(A2, "0.0")** | would return "123.7" |
| **=TEXT(1277, "#,##0")** | would return "1,277" |
| **=TEXT(C1, "yyyy/mm/dd")** | would return "2003/12/12" |
| **=TEXT(C1, "mmm dd, yyyy")** | would return "Dec 12, 2003" |

**Summary statistics**

sum, median, average

In Microsoft Excel, the **SUM function** adds all numbers in a range of cells and returns the result.

The syntax for the **SUM function** is:

SUM( number1, [number2, ... number\_n] )

In Microsoft Excel, the **MEDIAN function** returns the median of the numbers provided.

The syntax for the **MEDIAN function** is:

MEDIAN( number1, [number2, ... number\_n] )

In Microsoft Excel, the **AVERAGE function** returns the average (arithmetic mean) of the numbers provided.

Calculate the median of commuteMinutes.

The syntax for the **AVERAGE function** is:

AVERAGE( number1, [number2], ... [number\_n] )

Calculate the median of commuteMinutes.

The mean of numeric-categorical numbers: streesLevel.

Data tab -> text to column -> fixed width.

The mean of stress level is 2.45. Be careful how to explain this number.

**Create charts**

**How to choose chart type:**

* Bar chart

Bar graphs are used to compare things between different groups or to track changes over time. However, when trying to measure change over time, bar graphs are best when the changes are larger. Also good for multiple (sub)datasets.

* Pie chart

Pie charts are best to use when you are trying to compare parts of a whole. They do not show changes over time.

* X-Y plot

X-Y plots are used to determine relationships between the two different things. The x-axis is used to measure one event (or variable) and the y-axis is used to measure the other. If both variables increase at the same time, they have a positive relationship. If one variable decreases while the other increases, they have a negative relationship. Sometimes the variables don't follow any pattern and have no relationship. More used for quantitative analysis.

* Line graph

Line graphs are used to track changes over short and long periods of time. When smaller changes exist, line graphs are better to use than bar graphs. Line graphs can also be used to compare changes over the same period of time for more than one group.

**Add more than one data series to the chart**

Let’s say you want to explore the difference and relationship of gender and perceived injury risk. (IDs: genderChoice and InjuryRisk)

We’ve learned how to show the distribution of injuryRisk, but how to break down by gender?

Copy the two columns to a new sheet. Filter -> separate the data -> add “count” column (make sure there are all categories of choices) -> make two pivot tables

Insert column graph

1. Perceived injury risk by gender (count)

Right click on the blank of the plot area -> select data ->Add ->Series name (Male)-> Series values (select count) -> Horizontal (Category) Axis Labels -> Edit -> Select categories ->OK

Same steps, add Female.

1. Perceived injury risk by gender (percentage)

D2=B2/$B$7

In Excel, there are two types of reference:

e.g. B2 (without dollar sign): relative reference

e.g. $B$7(with dollar sign): absolute reference

**Label and format**

Chart Tools tab -> design tab -> chart layouts (on the left) -> Layout 2, 4, 5

Add a title: Insert tab -> text box (on the right)

**Bubble plot – show the relationship between two variables**

workSatisfaction and supervisorRelationship

Copy the two columns to a new sheet. Add count column. Add count 0 for no response as well. Insert pivot table ->drag supervisorRelationship and workSatisfaction to Row labels, make sure supervisorRelationship is on top. Drag Count to values.

Create three columns: supervisorRelationship (x), workSatisfaction(Y), and count.

Number code them. (text to column). Select the three columns (without headers), insert bubble plot.



**Text prompt: Key words**

unionDifferenceDetail



unionActivityDetail



State of the Union Address, 2002 vs. 2011



It seems that Ohmage doesn’t exclude the function words (articles, conjunctions, etc.) first when calculating the frequency of a certain word. Be careful when using this feature.

unionDifferenceDetail – bargain, represent, collective

Use find function in Excel to see how many times a word/phrase appears.

Ctrl + F, or Home tab -> find and select (on the right) -> find

We can see that 6 people mentioned represent, and 2 mentioned collective bargain.

**Mapping**

**Fix the missing lat longs**

**Map zip code**

Add the shapefile of all the zip codes of LA County.

Add the table to Table of Content (the side panel).

Join table to LA County. Right click on zip\_poly -> Joins and Relates -> Join

Use the common column (zip code in this case) as the basis of join.

Color code by job types, work status (full time/ part time/ contractor…), etc. Is there a relationship between job types (income) and where people live? How will it affect commute time and stress level? How about work status?

**Google Fusion Table review**

**Data visualization**

Powerful tools:

Microsoft PowerPoint, Adobe Illustrator/PhotoShop

Too much, too little, or just right

**\*\*\*\*\*\*\*\*Exercise/Assignment\*\*\*\*\*\*\*\*\***

Submit the two column charts, one bubble chart, and the zip code map. Also, prepare two more graphs, or one graph and one map, of data analysis. At least one of them should be showing multivariate variable relationship. Submit the excel file (.xlsx) to CCLE.