Final Examination FS19

*Name:*____

12/3/2019

Important

Due Date: Noon on Thursday 12th of December

Submission: Compile your answers in this RMarkdown file. Upload it, and supporting files, and a pdf rendering into the appropriate spot in Canvas.

Rules: As this is a take home test, there can't be any restrictions I can enfornce. I do ask that you do your own work. DO NOT give your R to another by email, usb drive, or the like. Do not let people take pictures of your R code.

Time of Last Revision: 8:00 am 4 December, 2019

Example - Bar Chart

I want to begin the final with an example for you. This example will demonstrate several nice features available to us in the {ggpubr} package.

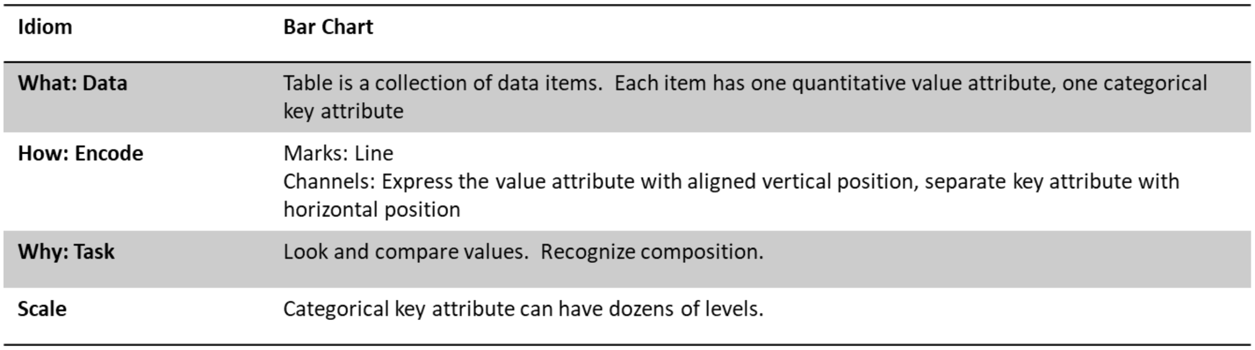
The format of these visualization exercises may include the following:

1. A data file or (R dataset)
2. A professional quality visualization addressing the task including a text caption offering some explanation or observation
3. An explanation of the idiom in the form of "What? Why? How?" table. (see box to follow)

We will call these "What-Why-Hows"

Note: I believe that a good test should teach you new things and not just stress your internal organs into rupture. Hence, I will often tell you to look up and use particular commands from various packages.

```
> theme_set(theme_pubr() + theme(legend.position = "right"))
>
> diam_df <- diamonds %>% group_by(cut) %>% dplyr::summarise(counts = n())
```



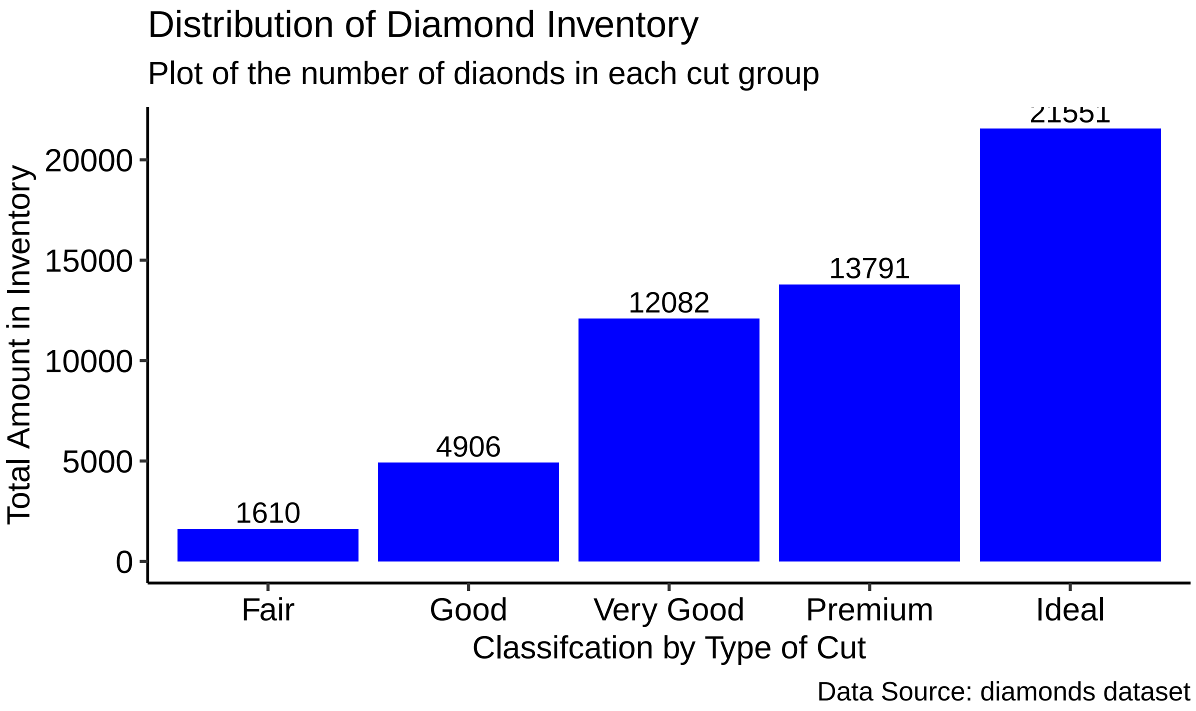
Idiom	Bar Chart
What: Data	Table is a collection of data items. Each item has one quantitative value attribute, one categorical key attribute
How: Encode	Marks: Line Channels: Express the value attribute with aligned vertical position, separate key attribute with horizontal position
Why: Task	Look and compare values. Recognize composition.
Scale	Categorical key attribute can have dozens of levels.

Figure 1: I made this table in PowerPoint with 12pt font. I copied it and pasted as a picture then did a Save As Picture to create an image I could import into this document.

```
>
> diam_df
```

```
## # A tibble: 5 x 2
##   cut       counts
##   <ord>      <int>
## 1 Fair        1610
## 2 Good        4906
## 3 Very Good  12082
## 4 Premium    13791
## 5 Ideal      21551
```

```
> diam.bar.p <- ggplot(diam_df, aes(x = cut, y = counts)) + geom_bar(fill = "blue", stat = "identity")
+     vjust = -0.3)
>
> diam.bar.p <- diam.bar.p + labs(title = "Distribution of Diamond Inventory", subtitle = "Plot of the
+     caption = "Data Source: diamonds dataset", x = "Classifcation by Type of Cut", y = "Total Amount
>
>
>
> text.bar <- paste("This chart shows stock increases dramatically for the higher quality cuts.  This i
>
> text.bar.p <- ggparagraph(text = text.bar, face = "italic", size = 11, color = "black")
>
> ggarrange(diam.bar.p, text.bar.p, ncol = 1, nrow = 2, heights = c(3.5, 0.5)) + theme_pubclean()
```



This chart shows stock increases dramatically for the higher quality cuts. This is fortunate as we can expect a higher profit margin for the Premium and Ideal cuts.

Texas Housing

Explore the dataset “txhousing” in {ggplot2}.

“Information about the housing market in Texas provided by the TAMU real estate center, http://recenter.tamu.edu/.”

What-Why-How 1: Histogram

Include a What-Why-How for the histogram idiom.

Plot 1: Histogram

Create a quality histogram of the median price of the houses being sold. Make the x-axis show currency and not in scientific notation. Give it a reasonable title.

Plot 2: Density

Building upon the previous chart, overlay an opaque density on the histogram.

Graduate Students Only Plot 1: Visual investigation of inventory problem

Create to subgroups. The A subgroup will be the subset of housing residing in cities beginning with the letter “A”. Create an “S” subgroup as well. On the same plot, over the the densities. In a text paragraph beneath the histograms, discuss if there seems to be any real difference in the probabilities of selling homes in those cities.

Colorblind Flowers?

In this problem set, you will create scatterplots without and with grouping, and finally with regression analysis. We will use the ever popular iris dataset. We place the restriction on the problem that the visualizations should be colorblindness friendly.

For your colors, you are to use the {viridis} R package. It uses beautiful colors that are printer-friendly and uniform, and easy to use.

Use the minimal theme from {ggpubr}

What-Why-How 2: 2-d scatterplot

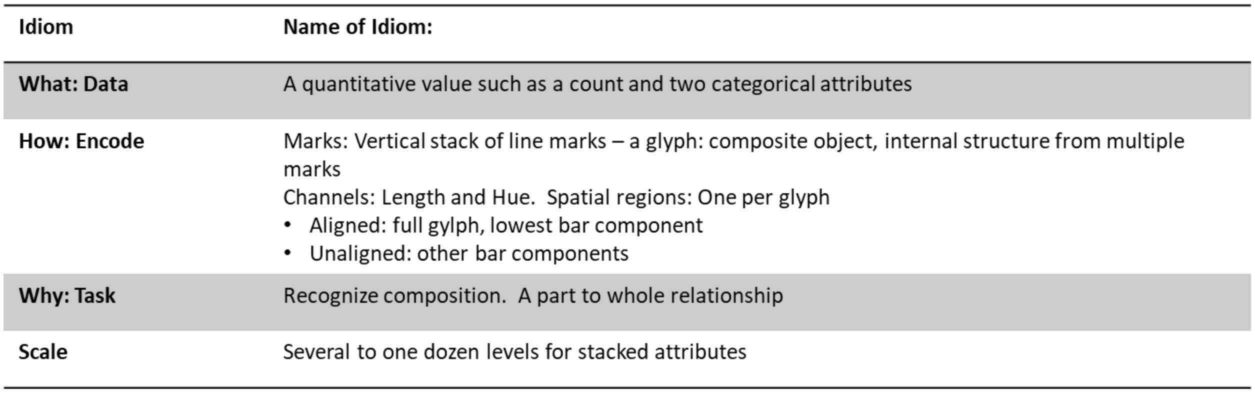
Insert the visual idiom description for a 2-d scatterplot.

Plot 3: Scatterplot

Using a minimal theme, create a scatterplot of Sepal.Length against Sepal.Width. Create visual grouping of the species suitable for colorblind people. Overlay a regression line with standard error bands. Make sure the axises have physical units listed.

Plot 4: Type of Dot Plot

It would seem to me another approach to aiding those whose are colorblind is to use changes in size of the dots. So for this plot, retain the coloration from the viridis palette, but also change the size of the circle for each species. That is there will be three size circles, each of a constant colorblind-friendly color. Make sure the legend makes clear the two-fold way of visualizing the grouping. A text sentence might be helpful.



Idiom	Name of Idiom:
What: Data	A quantitative value such as a count and two categorical attributes
How: Encode	Marks: Vertical stack of line marks – a glyph: composite object, internal structure from multiple marks Channels: Length and Hue. Spatial regions: One per glyph • Aligned: full gylph, lowest bar component • Unaligned: other bar components
Why: Task	Recognize composition. A part to whole relationship
Scale	Several to one dozen levels for stacked attributes

Figure 2: Write the name of this visualization in the space provided using red ink.

Plot 5: Facetting

Yet another way to help is to seperate the scatterplots. For this plot, facet three scatterplots in a row. One scatterplot per species. For this one, include the regression line and error bands around each subplot.

Plot 6: Grouping

We could also highlight the groups for the colorblind (or really any) viewer by putting an enclosing figure about clusters of points, such as an ellipse, around clusters of points. There are many ways and means of doing this, but try the stat_ellipse() function.

Graduate Students Only Plot 2: Edge Densities

Explore the ggscatterhist() plot in {ggpubr}. Create a scatterplot as above without the regression line and error bands. Colorize the groups by species using colorblind safe colors. Finally have the upper edge and right edge present density histrographs to reflect the marginal distributions.

A data item of Many Hats

A data item consisting of a quantitative value such as a count and two categorical attributes is simple enough to wear at least a dozen different visual idioms.

Plot 7: Mystery Idiom

In the R package {vcdExtra} there is a dataset called GSS.

> General Social Survey– Sex and Party affiliation Description Data from the General Social Survey, 1991, on the relation between sex and party affiliation.

To prepare us in reading the visualization, present the head and structure of the file. Once you have identified this visual idiom give a quality visualizations with title, and descriptive axes. Use colors from the correct type of Brewer color palette. Report the totals in the vis gylph. [Hint: {ggpubr} has a nice wrapper.]

Plot 8 - Playing Dodge ball

Repeat the above but dodge the bars rather than stack them.

Plot 9 geom_count

{ggplot2} has introduced a new geom called "geom_count". It is a mixture of dot plot and tile plot. See the examples in https://ggplot2.tidyverse.org/reference/geom_count.html.

Use the diamond data set to create a visualization where there is a disk present for each combination of "color" (color as in the property of the gem), and "clarity". The size of the disk reflects the total proportion of the diamonds in this joint-category scaled by the number of diamonds in the entire set.

Visualize the data items described above. Intoduce a title and subtitle. Improve the x and y axis and see what you can do to the legend. Include a caption. Finally add a small paragraph of text explaining to the owners what patterns are seen to exist in their inventory visualization.

Pick an appropriate publication theme.

What-Why-How 3: geom_count

Insert the visual idiom description for geom_count.

Graduate Students Only Plot 3: {vcd}

This is a type of research question. I've given you the paper called "The Structplot Framework: Multi-way Contienegency Tables with vcd." I want you to research the presentation of the the simplest contingency table (2 categorical variables, and one frequence or count variable). I give you three datasets to work with: GSS, Mental, and Glass. I want you to give me back three completely different visualizations. Page 8 of the paper lists several options for you to explore. Make them look good so the differences in design can be readily seen.

Honest Maps

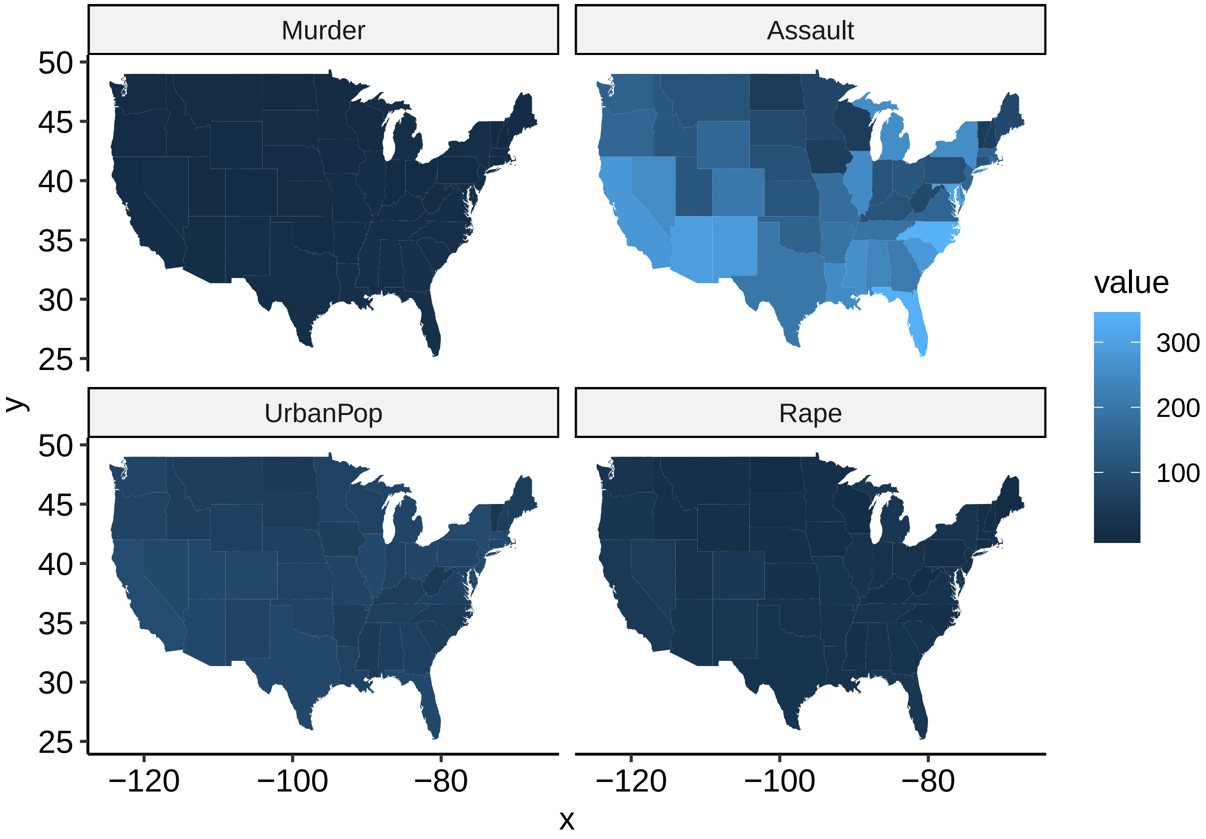
What-Why-How 4: Geom-Map

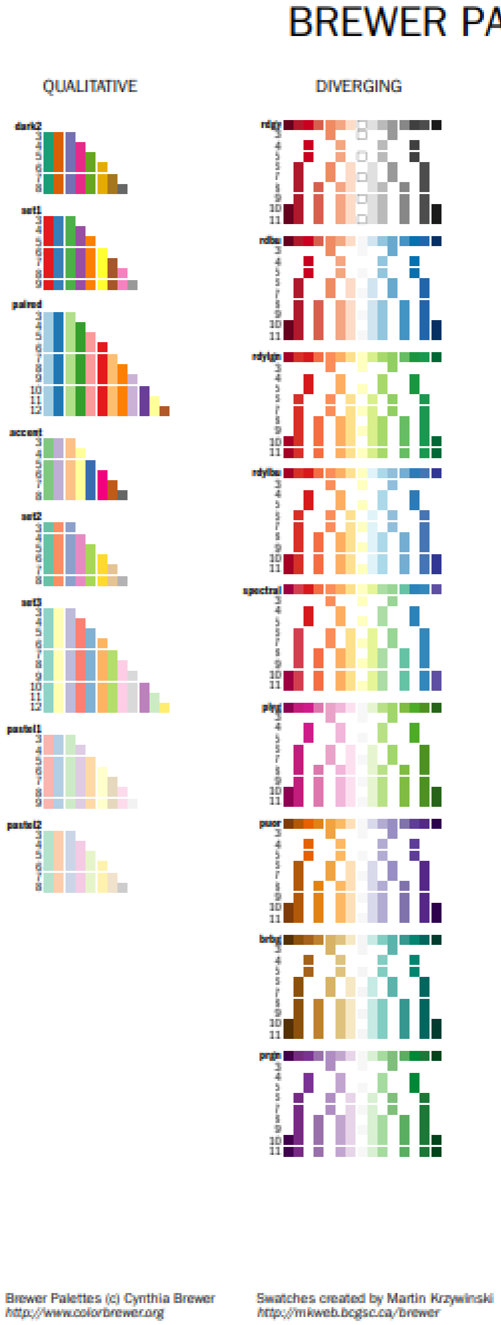
Insert the visual idiom description for {ggplot2} layer geom_map. In particular, focus on applications using geographic information such as the United States.

Plot 10 - Maps

This problem gives you the solution, sort of. Start with the example provided in the provided at the bottom of the listing. I repeat it for your convinence and reference.

```
> # Better example
> library(maps)
> crimes <- data.frame(state = tolower(rownames(USArrests)), USArrests)
> crimesm <- reshape2::melt(crimes, id = 1)
>
> if (require(maps)) {
+     states_map <- map_data("state")
+     ggplot(crimes, aes(map_id = state)) + geom_map(aes(fill = Murder), map = states_map) + expand_limi
+         y = states_map$lat)
+
+     last_plot() + coord_map()
+     ggplot(crimesm, aes(map_id = state)) + geom_map(aes(fill = value), map = states_map) + expand_limi
+         y = states_map$lat) + facet_wrap(~variable)
+ }
```


Murder
Assault
UrbanPop
Rape
y
50
45
40
35
30
25
x
−120
−100
−80
value
300
200
100



Brewer Palettes (c) Cynthia Brewer
http://www.colorbrewer.org

Swatches created by Martin Krzywinski
http://mkweb.bcgsc.ca/brewer

Your job. Make this professional. Also you must use the Brewer color palette.
I think it is very difficult to distinguish between the blues. Why does the x axis have negative numbers? What do the y values mean. What is “Urban Pop”? Give it a title, subtitles and captions. Maybe a paragraph explaining. What is “value” on the legend.

This was an impressive visualization until we tried to use it for something.

Graduate Students Only Plot 4: Scaling Data

As mentioned in class, many maps showing a single attribute held by people across the US, are usually just a popultation density map. Using what numbers you can find, scale the data so you have “number of murders per capita” and not just number of murders. (I might be wrong and this data already shows this, but I have my doubts. That it is difficult to distinguish the blues doesn't help and the legend is meaningless doesn't help.)