Project Output

```
> knitr::opts_chunk$set(echo=FALSE)
>
> library(plotly)
> library(orca)
>
> #import avengers data from github
> path =
'https://raw.githubusercontent.com/fivethirtyeight/data/master/avengers/ave
ngers.csv' #raw
> avengers.data = read.csv(path)
>
> #Data table used in this project
> #Use Module3 as a guide
>
> head(avengers.data)
> tail(avengers.data)
>
> #results='hide', fig.keep='all'
> #include=FALSE #hide everything from chunk
>
> #### Part 1A
> #Do analyses, at least one categorical variable, show plots
>
> #categorical - Gender, Death, mosaic
> tab.sex.death = table(avengers.data$Gender, avengers.data$Death1)
> mosaicplot(tab.sex.death, color=c("#4d384c", "#9d98ae"), main="Avenger
Deaths")
>
> #### Part 1B
> #Do analyses, at least one numerical variable; show plots
>
> # font list info
> font.info = list(family = "Arial Black", size=18, color = "#000000")
> # set axes
> x.ax = list(title = "Year", titlefont=font.info)
>
> #numerical - Year, boxplot
> num.boxplot = plot_ly(data=avengers.data, x = ~Year, type = "box",
marker=list(color= "#903A19"),
+                        color=" ", colors="#903A19") %>%
layout(xaxis=x.ax)#change to y for vertical
> num.boxplot
>
> #### Part 2
> #Do analyses, at least one set of two or more variables; Show plots
> #scatterplot appearances, years since joining
>
> # font list info
> font.info = list(family = "Arial Black", size=18, color = "#000000")
> # set axes
> x.ax = list(title = "Years since joining", titlefont=font.info)
> y.ax = list(title = "Appearances", titlefont=font.info)
>
> #plot scatterplot
> pt2.scat.sex = plot_ly(data=avengers.data, x = ~Years.since.joining, y =
~Appearances,
+   color= ~Gender, colors= c("#4f7298", "#96b596")) %>% layout(xaxis=x.ax,
yaxis=y.ax) # #34595a, marker=list(color= "#49796b"), #6b8eb6
> pt2.scat.sex
>
> # save as static image
> #orca(pt2.scat.sex, width=3.4*300, height=2.5*300)
#"Pt2_scatt_yearsappear.jpg"
>
> #### Part 3
> #Examine distribution for one numerical variable
> #numerical - Years.since.joining, histogram
>
> #Show histogram of time each avenger joined the league
> # font list info
> font.info = list(family = "Arial Black", size=18, color = "#000000")
> # set axes
> x.ax = list(title = "Years since joining", titlefont=font.info, dtick=10,
ticklen=7,
+             tickwidth=2, tickfont=list(size = 15))# dtick=20000
> y.ax = list(title = "Frequency", titlefont=font.info, ticklen=7,
tickwidth=2,
+              tickfont=list(size = 15))
>
> # plot hist
> hist.fig = plot_ly(x= ~avengers.data$Years.since.joining, type=
"histogram",
+            alpha=0.8, marker=list(color= "#8e3c52")) %>%
layout(xaxis=x.ax, yaxis=y.ax)
>
> # save as static image
> #orca(hist.fig, width=4.5*300, height=3.5*300) #"Pt3_hist_yearsjoin.jpg"
>
> hist.fig
>
> #### Part 4
> #### Question 2B
> #Draw 1,000 samples of sizes 10, 20, 30, 40; plot hist of sample means,
2x2, Years.since.joining
>
> samples = 1000
> samp.sz10 = 10
> samp.sz20 = 20
> samp.sz30 = 30
> samp.sz40 = 40
>
> # list of zeros
> xbar10 = numeric(samp.sz10)
> xbar20 = numeric(samp.sz20)
> xbar30 = numeric(samp.sz30)
> xbar40 = numeric(samp.sz40)
>
> # draw samples of data, add to list
> #samp size 10
> for (i in 1:samples){
+   xbar10[i] = mean(sample(avengers.data$Years.since.joining, samp.sz10,
replace=FALSE))
+ }
>
> #samp size 20
> for (i in 1:samples){
+   xbar20[i] = mean(sample(avengers.data$Years.since.joining, samp.sz20,
replace=FALSE))
+ }
>
> #samp size 30
> for (i in 1:samples){
+   xbar30[i] = mean(sample(avengers.data$Years.since.joining, samp.sz30,
replace=FALSE))
+ }
>
> #samp size 40
> for (i in 1:samples){
+   xbar40[i] = mean(sample(avengers.data$Years.since.joining, samp.sz40,
replace=FALSE))
+ }
>
> #Show the histogram of the sample means.
> # font list info
> font.info = list(family = "Arial Black", size=25, color = "#000000")
>
> # plot hist samp size 10
> hist.fig10 = plot_ly(x= ~xbar10, type= "histogram", alpha=0.8,
name="SampSize10")
>
> # plot hist samp size 20
> hist.fig20 = plot_ly(x= ~xbar20, type= "histogram", alpha=0.8,
name="SampSize20")
>
> # plot hist samp size 30
> hist.fig30 = plot_ly(x= ~xbar30, type= "histogram", alpha=0.8,
name="SampSize30")
>
> # plot hist samp size 40
> hist.fig40 = plot_ly(x= ~xbar40, type= "histogram", alpha=0.8,
name="SampSize40")
>
>
> # plot four plex
> samp.size.panel = subplot(hist.fig10, hist.fig20, hist.fig30, hist.fig40,
nrows=2)
> samp.size.panel
>
> #### Part 5
> #Show how various sampling methods can be used on data
> library(sampling)
>
> samp.sz = 50
> pop.sz = length(avengers.data$Appearances)
>
> #full data set - Appearances
> hist.fig.full = plot_ly(x= ~avengers.data$Appearances, type= "histogram",
alpha=0.8, name="Full dataset")
>
> #simple random sampling without replacement
> # set var to random sample list, samp.sz, tot num in pop
> simp.rand.wor = srswor(samp.sz, pop.sz)
> subset.srswor = subset(avengers.data$Appearances, simp.rand.wor != 0)
> # plot hist srswor
> hist.fig.srswor = plot_ly(x= ~subset.srswor, type= "histogram",
alpha=0.8, name="SRSWOR")
>
> #systematic sampling - equal
> #population num
> N = pop.sz
> #samp size
> n = samp.sz
> # sample at every k nums
> k = ceiling(N/n)
> # find the start num in the first k sample
> r = sample(k, 1)
> # find sequence for every k nums
> s = seq(r, by=k, length=n)
> # apply sytematic sample sequence to dataframe
> sys.samp.appear = avengers.data['Appearances'][s, ]
> # plot hist systematic sampling
> hist.fig.sys.samp = plot_ly(x= ~sys.samp.appear, type= "histogram",
alpha=0.8, name="Systematic Sampling")
```

```
>
> #systematic sampling - unequal probabilities
> # Calculate inclusion probabilities
> incl.prob = inclusionprobabilities(avengers.data$Appearances, samp.sz)
> # Sample drawn using systematic sampling, unequal probabilities
> ss = UPsystematic(incl.prob)
> sys.samp.appear.uneq.prob = subset(avengers.data$Appearances, ss != 0)
> # plot hist systematic sampling, unequal probabilities
> hist.fig.sys.samp.uneq = plot_ly(x= ~sys.samp.appear.uneq.prob, type=
"histogram", alpha=0.8, name="Unequal Systematic Sampling")
>
> # plot all figs in one fig
> rand.samp.panel = subplot(hist.fig.full, hist.fig.srswor,
hist.fig.sys.samp, hist.fig.sys.samp.uneq, nrows=4)
> rand.samp.panel
>
> #### Part 6
> #Implementation of any feature not mentioned above
> #2D histogram based on the scatterplot in Pt2; showing appearances, years
since joining
>
> # scatterplot with single variable
> pt2.scat = plot_ly(data=avengers.data, x = ~Years.since.joining, y =
~Appearances, marker=list(color= "#00baba"))
> #2plex scatterplot and 2D histogram
> pt2.scat.2d = subplot(pt2.scat %>% add_markers(), pt2.scat %>%
add_histogram2d()) #colorscale = "Blues"
> pt2.scat.2d
>
> #proportion of avengers that die return later in the series
> #scatterplot of death years since joining
> #boxplot as swarmplot
>
```