Lab 11 - Data, Aesthetics, & Geometries

Your Name Here

November 9, 2017

Complete the following exercises below. Knit together the PDF document and commit both the Lab 11 RMD file and the PDF document to Git. Push the changes to GitHub so both documents are visible in your public GitHub repository.

1. Which variables in your dataset are you interested in visualizing? Describe the level of measurement of these variables and what type of geography you think is appropriate to represent these variables. Give your reasoning for choosing the `geom_()` you selected.

```
#creating barchart
options(repos = c(CRAN = "http://cran.rstudio.com"))
install.packages("ggplot2")
```

```
##
## The downloaded binary packages are in
##  /var/folders/dm/pz9f8jgs199g21t077m_9nrh0000gn/T//RtmpfSJzYT/downloaded_packages
```

```
library("ggplot2")

setwd("/Users/marakage/Desktop/Honors/STAT/Labs")
#install.packages("dplyr")
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
# Read in your data with the appropriate function
#Download main tables from Census
AsianAlone <- read.csv("DEC_10_SF2_PCT1_with_ann.csv")
Education <- read.csv("../Datasets/Education/ACS_15_5YR_S1501_with_ann.csv")
Totalpopulation <- read_csv("../Datasets/TotalPopulation/ACS_10_5YR_B01003_with_ann.csv",skip=0)
```

```
## Parsed with column specification:
## cols(
##   GEO.id = col_character(),
##   GEO.id2 = col_character(),
##   `GEO.display-label` = col_character(),
##   HD01_VD01 = col_character(),
##   HD02_VD01 = col_character()
## )
```

```
AsianAlone <- AsianAlone[-1,]
prop_data <- AsianAlone %>%
  mutate(Count = as.numeric(as.character(D001)),
         POPGROUP.id = as.numeric(as.character(POPGROUP.id))) %>% # Make new count variable which is nu

  group_by(POPGROUP.id) %>% # Group by ethnicity/subgroup
  summarise(Total_pop = sum(Count)) %>% #Sum within subgroup
  filter(POPGROUP.id != 031) %>% #Remove asian and other
  mutate(Total_Asian = sum(Total_pop), #Create total asian pop variable
         proportion = Total_pop/Total_Asian) # Creating proportion of subgroup of total Asians


#transform from factor to numeric
Data_Asian_Alone <- AsianAlone %>%
  filter(GEO.id != "Id") %>%
  filter(POPGROUP.id != "012" & POPGROUP.id != "031") %>% #removing the Asian Alone and Asian in combin
  mutate(Count = as.numeric(as.character(D001))) #creating a new column for the population number as nu

data_subset_total <- Totalpopulation %>%
  filter(GEO.id != "Id") %>%
  mutate(TotalCount = as.numeric(as.character(HD01_VD01)))


#barchart
mergedata <- data_subset_total %>%
  left_join(Data_Asian_Alone, by="GEO.id2") %>%
  select(-GEO.id.y, -ends_with("label.y")) %>%
  mutate(totalprop = Count/ TotalCount)
```

```
## Warning: Column `GEO.id2` joining character vector and factor, coercing
## into character vector
```

```
#total number of tracts that doesn't have Asian Alone or in combination
sum(is.na(mergedata$Count))
```

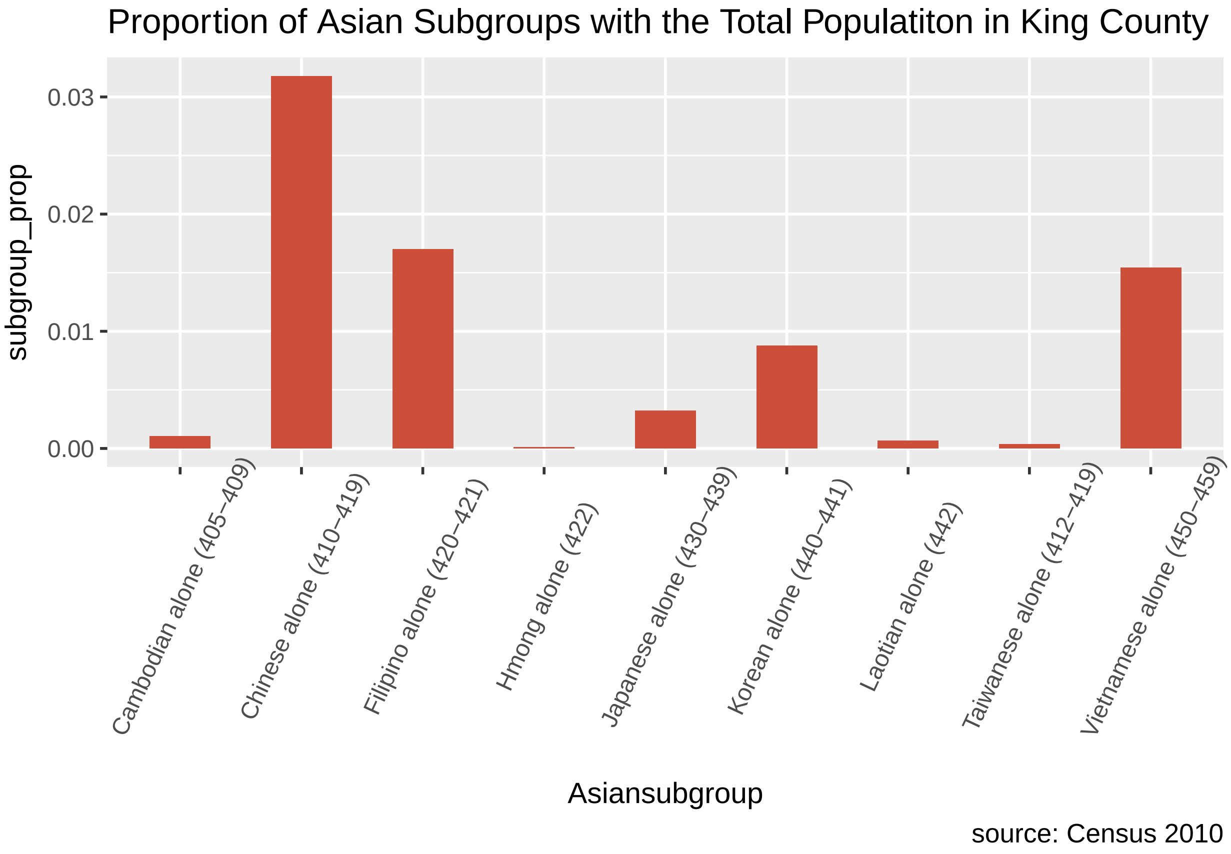
```
## [1] 117
```

```
submergedata <- mergedata %>%
  filter(POPGROUP.id != "012" & POPGROUP.id != "031")

#barchart table ready
barpop <- submergedata %>%
  mutate(Asiansubgroup = `POPGROUP.display.label`) %>%
  group_by(POPGROUP.id, Asiansubgroup) %>% #join all the tract population by same subgroup
  summarise(subpoptotal = sum(Count)) %>%
  mutate(subgroup_prop = round(subpoptotal/sum(data_subset_total$TotalCount), 5))#new column with the p
```

```r
#creating barchart
install.packages("ggplot2")
```

```
##
## The downloaded binary packages are in
##  /var/folders/dm/pz9f8jgs199g21t077m_9nrh0000gn/T//RtmpfSJzYT/downloaded_packages
```

```r
library("ggplot2")

ggplot(barpop, aes(x=Asiansubgroup, y=subgroup_prop)) +
  geom_bar(stat="identity", width=.5, fill="tomato3") +
  labs(title="Proportion of Asian Subgroups with the Total Populatiton in King County",
       caption="source: Census 2010") +
  theme(axis.text.x = element_text(angle=65, vjust=0.6))
```



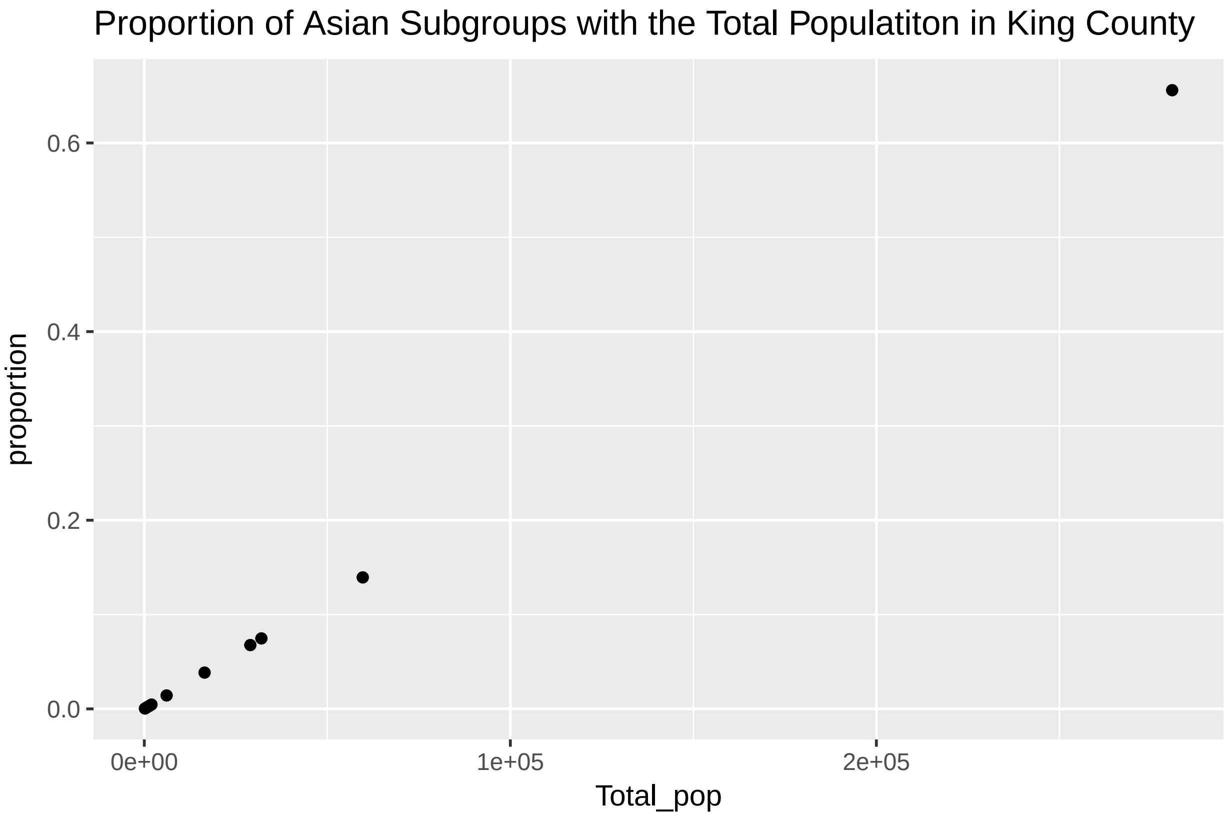
```r
# In this bar chart the x-axis is the Asian subgroups and the y-axis is the proportional Asian subgroup
```

2. Is your data in the proper format to visualize the data in the way you want? Why or why not? *If you need/want to change the structure of your data, do it below.*

```r
#For this visualization, I am taking the Asian Alone subgroup, to better convey the most marginal and l
```

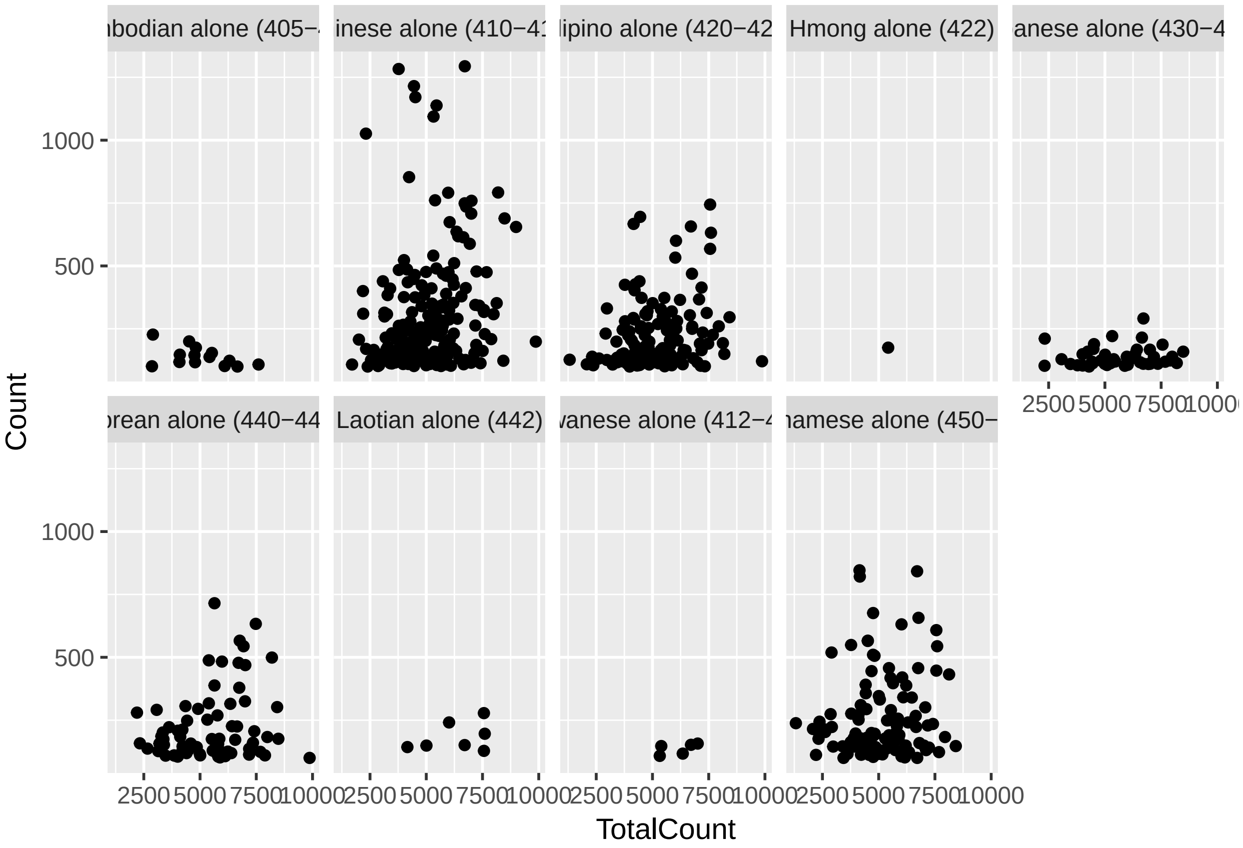
3. Create at least two different exploratory plots of the variables you chose using the skills we covered in class today. What types of mapping aesthetics did you choose and why? What do these plots tell you about your data?

```r
ggplot(prop_data) + geom_point(mapping = aes(x=Total_pop, y=proportion)) +
  labs(title ="Proportion of Asian Subgroups with the Total Populatiton in King County",
       caption="source: Census 2010")
```


Proportion of Asian Subgroups with the Total Populatiton in King County

source: Census 2010

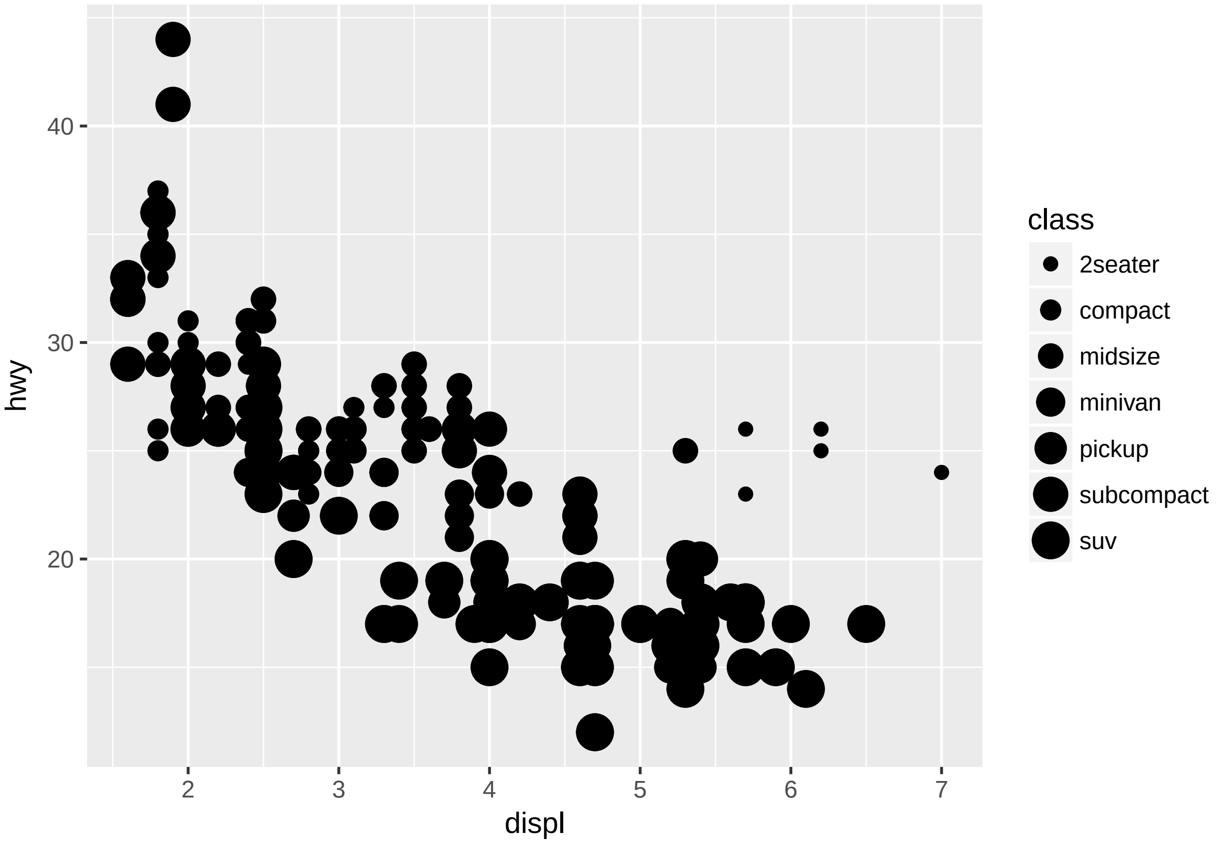
```
#Faceted scatterplot of each Asian subgroup. It shows the dispersion of each subgroup by tract and allo

ggplot(data = submergedata) +
  geom_point(mapping = aes(x = TotalCount, y = Count)) +
  facet_wrap(~ `POPGROUP.display.label`, nrow = 2)
```



4. Create at least three variations of the plots you've already made by modifying some of the arguments we covered in class (i.e. `position`, `scale`, `size`, `linetype` etc.). Do any of these modifications help you understand your data better? Why or why not? Do any of them create a misleading interpretation of the relationships between your variables? If yes, how so?

```r
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, size = class))
```

```
## Warning: Using size for a discrete variable is not advised.
```



```
#Scaling has been modified before per question 2.
```

5. From the plots you've created thus far, do any of them seem appropriate for a general audience? Why or why not? If so, what do you think you'd still need to do to make them more suitable as explanatory visualizations?

The second plot could be a good visualization of the population distribution by tract. To best visualize this information, I will seek help to create a map as the plot doesn't display where in King County the Asian population is located/or not (i.e Eastside, North Seattle, South Seattle) and will be interesting to see where the tracts are located versus just the density.