Harvard Kennedy School - API 201z

Problem Set #1

Nolin Greene

April 18, 2020

Preparations:

Operating System & Software Used:

- MacBook Pro (Retina, 13-inch, Late 2013)
- macOS Catalina 10.15.4
- RStudio Version 1.2.5003

Options & Packages Loaded:

```
rm(list=ls())

options(scipen = 999)

library(readxl);library(dplyr);library(ggplot2);library(plotly)
library(tidyr);library(knitr);library(kableExtra);library(stringr)
```

Question #1: Case Study - Pine Street Inn

Load Data:

```
d<-read_excel("Pine Street Inn Length of Stay Data - Solutions.xls",
              sheet = 1, cell_cols(1:2))

colnames(d)<-c("n","los")
```

Analysis:

1.1: The mean length of stay at Pine Street Inn is 26 **days**.

1.2: The median length of stay at Pine Street Inn is 3 **days**.

1.3: The maximum length of stay at Pine Street Inn is 727 **days** and the minimum length of stay is 1 **day**.

1.4: The 75th percentile length of stay at Pine Street Inn is 17 **days.** The 95th and 99th percentiles are 65 **days** and 138 **days** respectively

1.5: There are 171,905 **bednights** represented in the dataset.

1.6: There are 6556 **guests** represented in the dataset.

Table 1: Guest and Bed Night Statistics by Length of Stay Category

Number of Guests	Number of Bed Nights	Fraction of Guests	Fraction of Bed Nights	Avg B
721	4,973	0.11	0.03	
1,177	7,328	0.18	0.04	
1,048	21,007	0.16	0.12	
721	53,832	0.11	0.31	
288	84,765	0.04	0.49	
3,955	171,905	NA	NA	

1.7

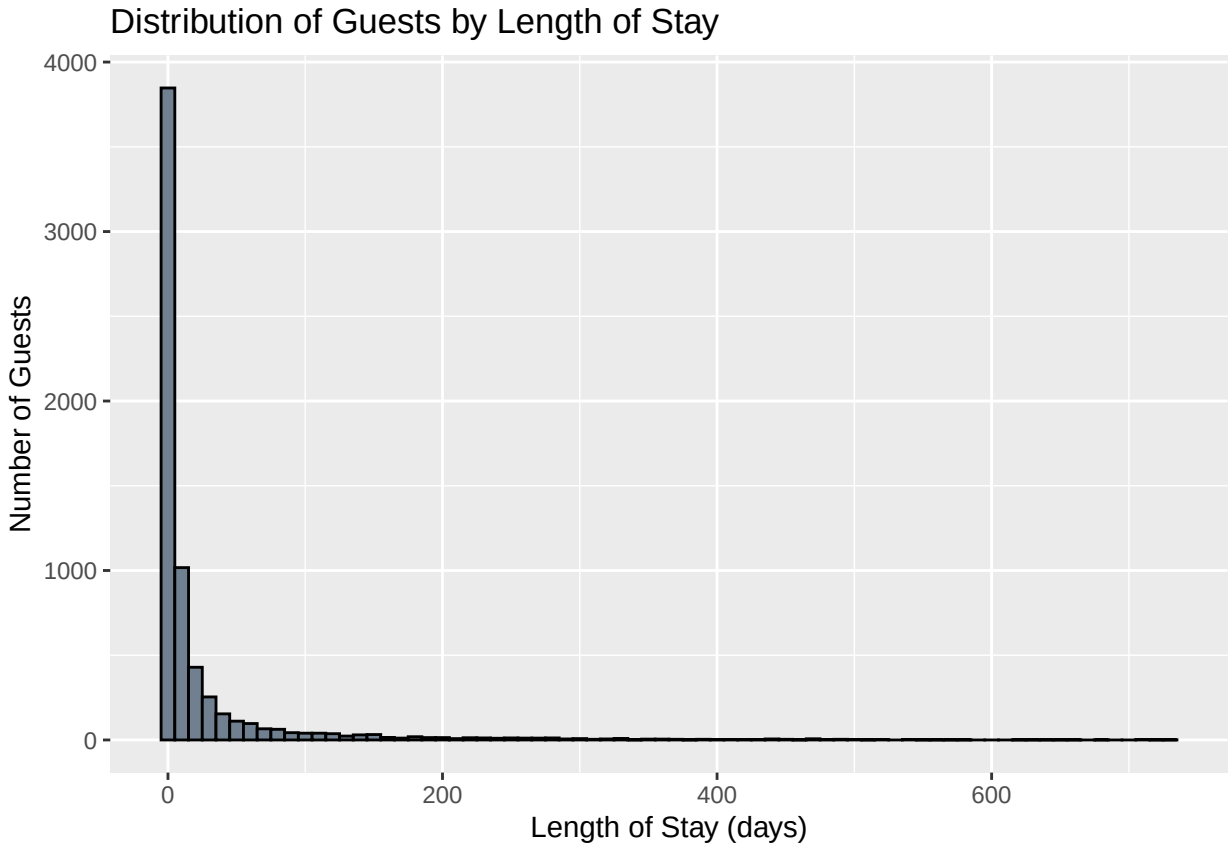
Summary Statistics for PSI Length of Stay

1.8:

```
d<-d %>%
  mutate(bin = case_when(
      los<4 ~ "3 Days or Less",
      los>3 & los<11 ~ "4 to 10 Days",
      los>10 & los <36 ~ "11 to 35 Days",
      los>35 & los<151 ~"36 to 150 Days",
      los>150 ~ "151 Days or More"),
    bin = factor(bin, levels = c("3 Days or Less","4 to 10 Days","11 to 35 Days",
                 "36 to 150 Days", "151 Days or More")))

count<-group_by(d, bin)%>%
  summarise(n = n(), los=sum(los))

count<-mutate(count,
       Clients = count$n/sum(count$n),
       Bed_Nights = count$los/sum(count$los))

count<-count[c(1,4,5)]

g<-gather(count, stat, percent, -bin)

p<-ggplot(g,aes(x=bin, y=percent, group = stat, color=stat))+
  geom_line()+
  theme(legend.title = element_blank()) +
  labs(x="Length of Stay", y="Percent of Total", title = "Total Clients and Bednights by Length of Stay
  scale_x_discrete(labels = function(x) str_wrap(x, width = 8))
```

1.9:

```
ggplot(d, aes(x=los))+
  geom_histogram(colour="black", fill = "slategray", binwidth = 10)+
  labs(x="Length of Stay (days)", y="Number of Guests", title="Distribution of Guests by Length of Stay
```



1.9: Simply by looking at the mean, one might infer that it is common for a PSI guest to spend 3-4 weeks in shelter. However, upon calculating additional statistics (median, IQR, historgram), we see that the distribution of length of stay is heavily right skewed, with a small number of guests having very long stays. This leads me to believe that Pine Street faces a very severe Pareto Principle, with a small number of guests occupying an extreme proportion of the shelter's total bed stays.

Question #2: State Spending Data

A. The total direct expenditure was $3.147tr. The total spent on Elementary and Secondary Education was $565bn, the total spent on Health was $84bn and the total spent on Corrections was $72.6bn.

B. The data are organized into columns separated by spaces. There appears to be five columns, with each row separated on a different line. The state code is contained in the row's first two characters (Delaware = "08") and the item code is contained in characters with position 5-7 (Construction for Primary & Secondary Education = "F12").

C. This chunk downloads and extracts our zip file, and creates a data frame from it.

```
temp <- tempfile()
download.file("http://www2.census.gov/govs/local/11statetypepu.zip",temp)
state_exp <- read.table(unz(temp, "11statetypepu.txt"))
unlink(temp)
```

This chunk cleans and tidys our data.

```
colnames(state_exp)<-c("govtype","itemcode","amount", "cv", "yr")
state_exp$govtype<-formatC(state_exp$govtype, width = 3, format = "d", flag = "0")
state_exp<-transform(state_exp, state = substr(govtype,1,2), gov_level=substr(govtype, 3,3))
```

Table 2: Summary Statistics for State Per Capita Spending

	Mean	StdDev	Median	IQR
Health	265.1	113.6	235.9	127.9
Education	1,870.6	491.4	1,803.0	361.6
Corrections	221.1	68.9	200.4	83.5

```
state_exp<-filter(state_exp, gov_level == 1)
state_exp<-select(state_exp, -govtype, -yr, -gov_level)
state_exp$amount<-state_exp$amount/1000
```

This chunk creates three clear categories of spend and groups the rows by that spend and each state.

```
state_exp<-state_exp %>%
  mutate(cat = case_when(
      itemcode=="E32" | itemcode=="F32" | itemcode=="G32" ~ "Health",
      itemcode=="E12" | itemcode=="F12" | itemcode=="G12" ~ "Education",
      itemcode=="E04" | itemcode=="F04" | itemcode=="G04" |
      itemcode=="E05" | itemcode=="F05" | itemcode=="G05" ~ "Corrections"))
state_exp<-filter(state_exp, cat %in% c("Health", "Education", "Corrections"))
state_exp <- state_exp %>%
  group_by(state,cat)%>%
  summarize(sum = sum(amount))
```

The above code found rows whose item codes are related to Health, Education or Corrections and summed all expenditures in the dataset by those groupings. The result is a simple data frame displaying how much was expended across the country for those three categories (in millions).

D. The categories of E32, F32 and G32 represent "Current Operations", Construction" and "Other Capital Outlays" for "other" health expenditures not otherwise captured in other item codes.

E. One reason we need to look at combined state and local expenditures is because different states have different frameworks for allocating governmental responsibilities within policy areas. For example, some states might take on the majority of the financial and administrative burden for maintaining corrections, while other states might cede that authority to the local level.

Adding on State Names

```
statenames<-read.csv("statenames.csv", colClasses = c(state = "character", names = "character", pop = "
state_exp<-left_join(state_exp, statenames, by = "state")
spendclean<-filter(state_exp,state != "00")[,2:5]
```

F. The data are now represented in millions. We converted them from thousands to millions by dividing the original value by 1000.

G

```
spendclean<-spendclean %>%
  mutate(per_capita = sum*1000000/pop)
```

H

I (i) The formula for the variance of health spending can be denoted by the following:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

I (ii) The variance is the sum of the squared differences of each observation and the mean, divided by the number of observations minus 1.

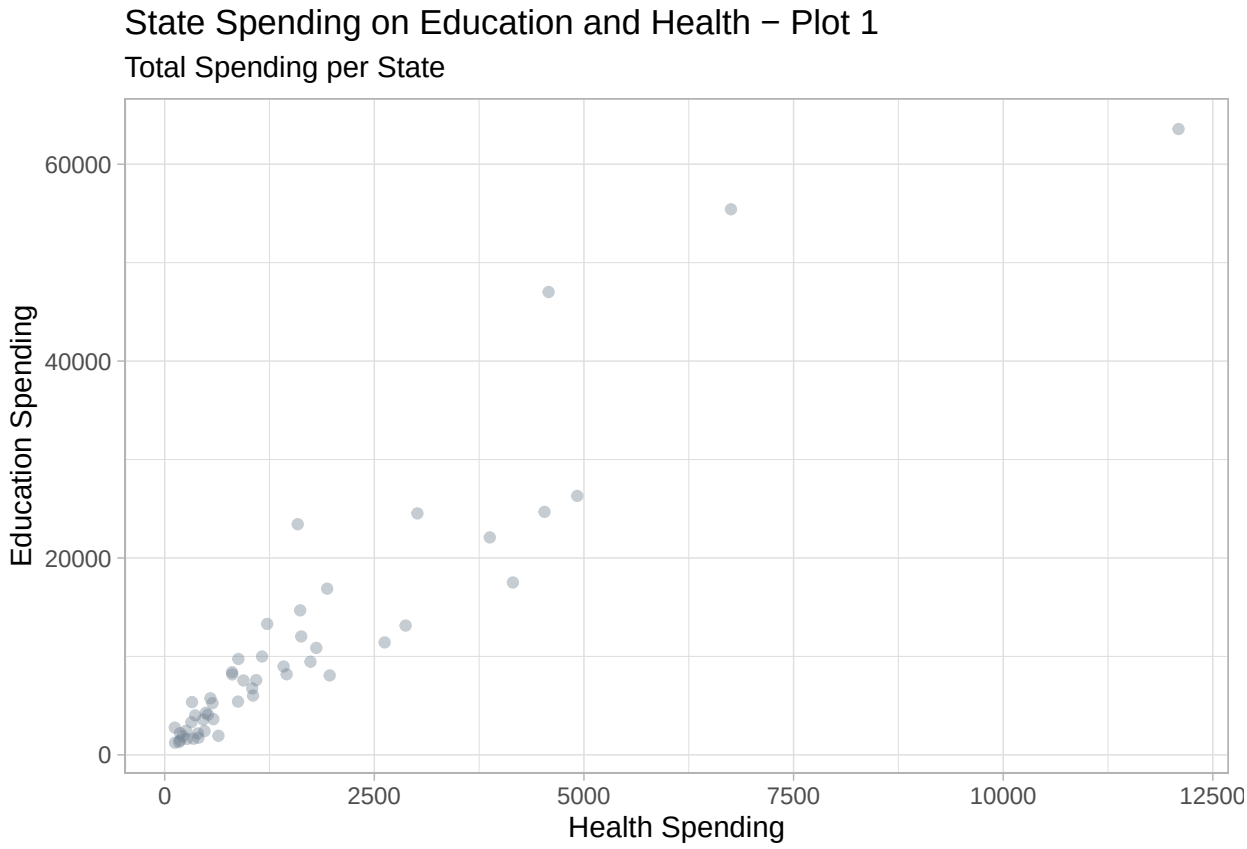
I (iii)

```
x1<-mean(pc_health)
x2<-pc_health-x1
x3<-x2^2

sum(x3)/length(x3)-1
```

```
## [1] 12653.7
```

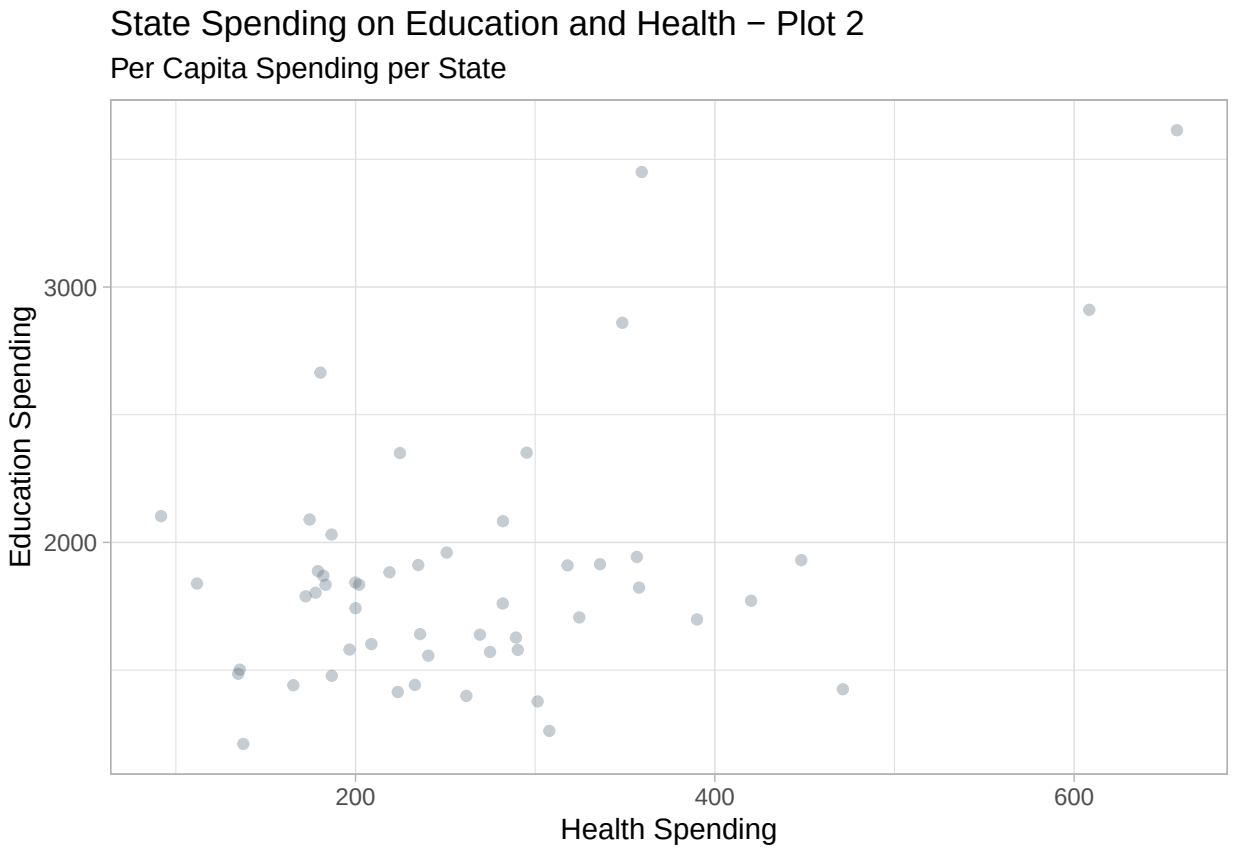
```
sqrt(sum(x3)/length(x3)-1)
```

```
## [1] 112.4887
```



J

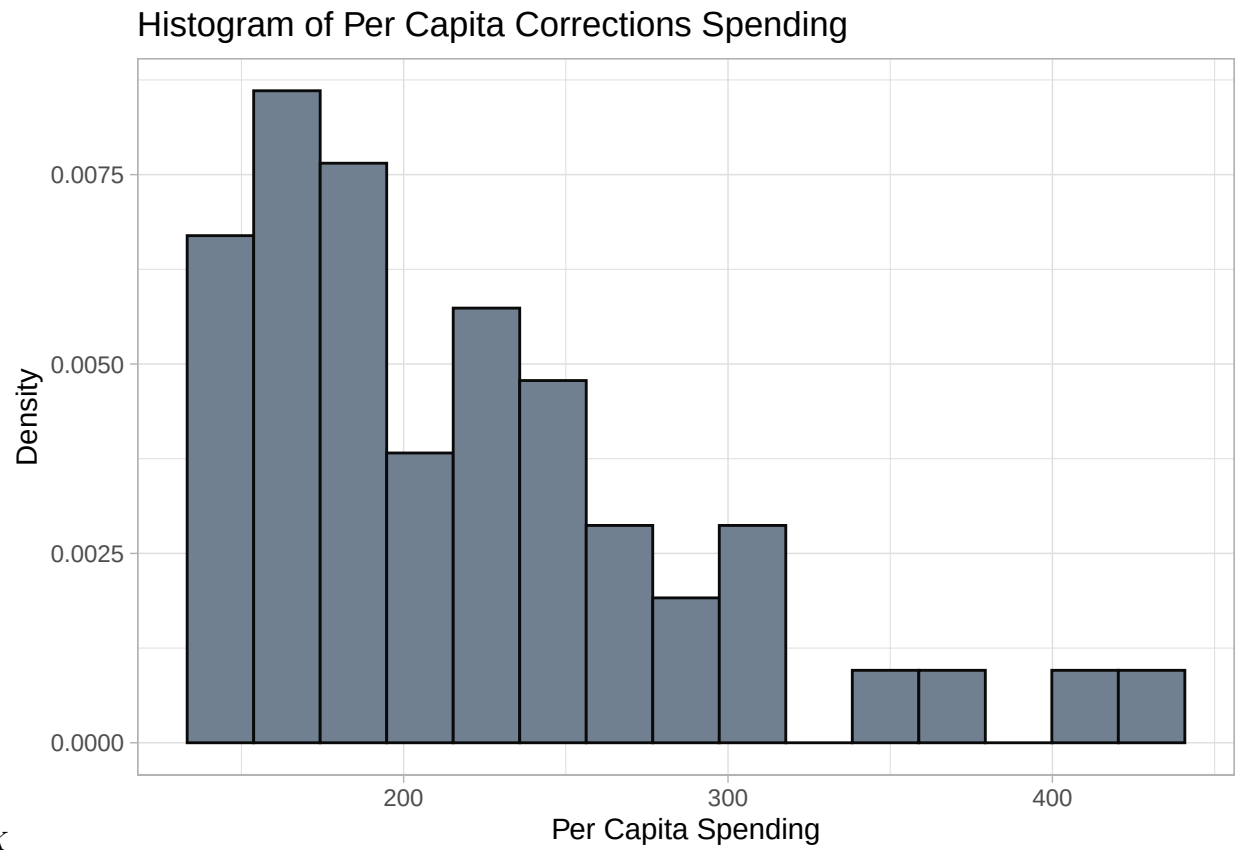
The plot shows a strong postiive correlation between total education and health spending. This is to be expeceted because the larger the population of the state, the more they are likely to spend on both education and health.



TThere is no immediate relationship between per capita Health spending and per capital Education spending (though a few outliers may skew the relationship slightly positive). The plot shows a strong postiive correlation between total education and health spending. This is to be expected though since different states may have different priorities with regards to health and education and may fund them at different levels relative to the size of their populations.

Table 3: Summary Statistics for Per Capita Corrections Spending - Inflated Alaska Example

	Mean	StdDev	Median	IQR
Corrections	409.3	1371.3	200.4	83.5



K

L

```
spendclean_alaska<-mutate(spendclean_wide, per_capita_Corrections = ifelse(names == "Alaska", 10000, pe

sum_stats_alaska<-data.frame(Mean = c(mean(spendclean_alaska$per_capita_Corrections)),
           StdDev = c(sd(spendclean_alaska$per_capita_Corrections)),
           Median = c(median(spendclean_alaska$per_capita_Corrections)),
           IQR = c(IQR(spendclean_alaska$per_capita_Corrections)),
           row.names ="Corrections")

kable(sum_stats_alaska,
      row.names = T,
      digits = 1,
      caption = "Summary Statistics for Per Capita Corrections Spending - Inflated Alaska Example",
      escape = F)%>%
kable_styling(bootstrap_options = c("striped", "hover", "condensed"))
```

The median and the IQR remain unchanged from the original per capita corrections summary, but the mean and the standard deviation have increased dramatically due to Alaska now being a considerable outlier in the dataset.