Harvard Kennedy School

API 201Z: Fall 2020

Problem Set #1 - March 26, 2020

Nolin Greene

Question #1: Case Study - Pine Street Inn

Data and Package Loading (output masked for readability purposes)

```
rm(list=ls())

options(scipen = 999)

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

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

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

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 **171905 bednights** represented in the dataset.

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

1.7

```
d%>%
  filter(los<=3) %>%
  summarize(n=n(), bednights=sum(los))
```

```
## # A tibble: 1 x 2
##       n bednights
##   <int>     <dbl>
## 1  3322      4973
```

```
d%>%
  filter(los<=10 & los>3) %>%
  summarize(n=n(), bednights=sum(los))
```

```
## # A tibble: 1 x 2
##       n bednights
##   <int>     <dbl>
## 1  1177      7328
```

```
d%>%
  filter(los<=35 & los>10) %>%
  summarize(n=n(), bednights=sum(los))
```

```
## # A tibble: 1 x 2
##       n bednights
##   <int>     <dbl>
## 1  1048     21007
```

```
d%>%
  filter(los<=150 & los>35) %>%
  summarize(n=n(), bednights=sum(los))
```

```
## # A tibble: 1 x 2
##       n bednights
##   <int>     <dbl>
## 1   721     53832
```

```
d%>%
  filter(los>150) %>%
  summarize(n=n(), bednights=sum(los))
```

```
## # A tibble: 1 x 2
##       n bednights
##   <int>     <dbl>
## 1   288     84765
```

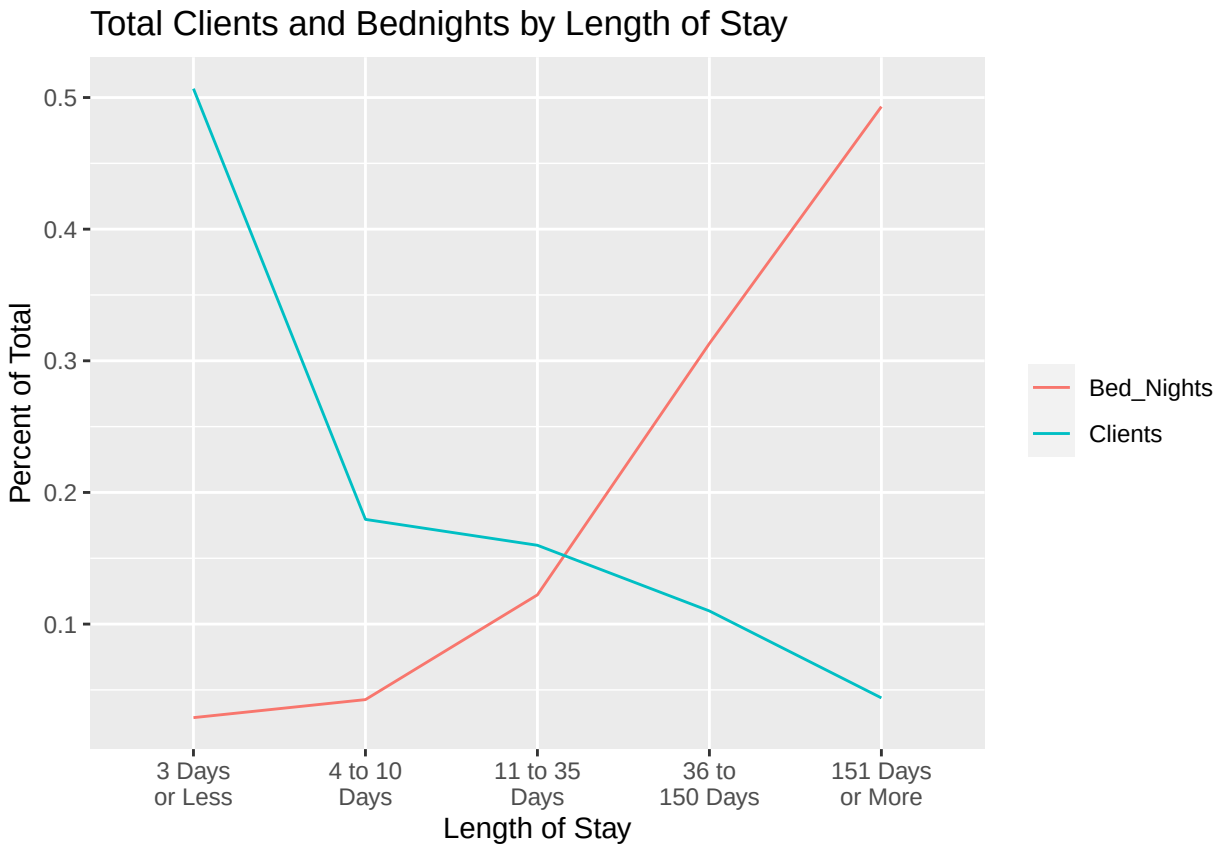
Summary Statistics for PSI Length of Stay

	Number of Guests	Number of Bed Nights	Fraction of Guests	Fraction of
3 Days or Less	721	4973	0.11	0.03
4 to 10 Days	1177	7328	0.18	0.04
11 to 35 Days	1048	21007	0.16	0.12
36 to 150 Days	721	53832	0.11	0.31
151 Days or More	288	84765	0.04	0.49
Total	3955	171905		

1.8:

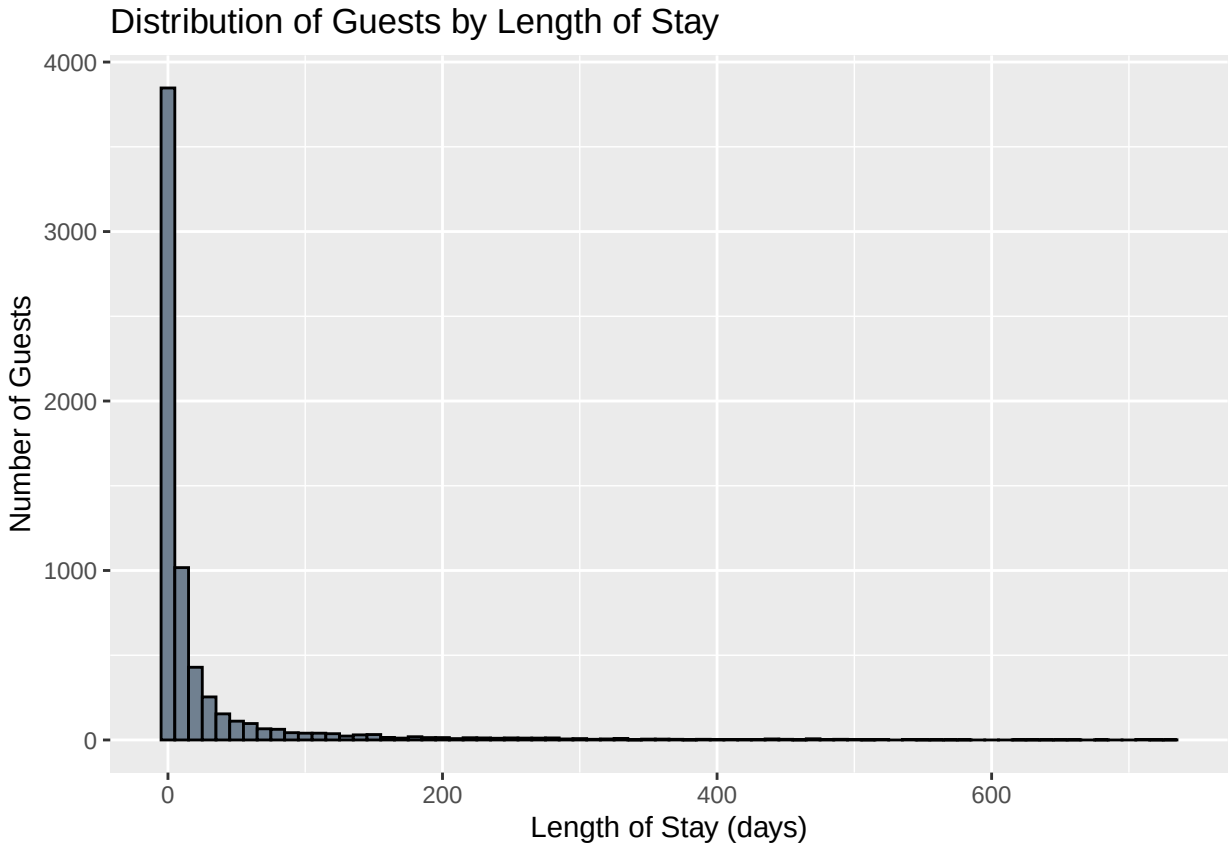
```
d<-d %>%
  mutate(bin = case_when(
      los<4 ~ "3 Days or Less",
      los>3 & los<11 ~ "4 to 10 Days",
      los>10 & los <36 ~ "11 to 35 Days",
      los>35 & los<151 ~"36 to 150 Days",
      los>150 ~ "151 Days or More"),
    bin = factor(bin, levels = c("3 Days or Less","4 to 10 Days","11 to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

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.

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

```
table(state_exp$yr)
```

```
##
##    11
## 30594
```

```
state_exp<-subset(state_exp, select = -yr)
state_exp<-filter(state_exp, govtype == 1)
state_exp<-subset(state_exp, select = -govtype)
```

```r
state_exp$amount<-state_exp$amount/1000
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 %>%
  group_by(cat)%>%
  summarize(sum = sum(amount))
```

```
## # A tibble: 3 x 2
##   cat             sum
##   <chr>         <dbl>
## 1 Corrections  73243.
## 2 Education   565284.
## 3 Health       82392.
```