PSET2_ParthDesai

Parth Desai

2023-02-08

R Markdown

Part 1

Part 1.1

```
vec1 <- seq(1:1000)
set.seed(12345)
vec2 <- sample(vec1, 1000, replace = FALSE, prob = NULL)
dat <- data.frame(vec1, vec2)
```

Part 1.2

```
dat_pos2 <- which(dat[ ,'vec2'] == 2, arr.ind = TRUE)
dat_pos47 <- which(dat[,'vec2'] == 47, arr.ind = TRUE)
dat_pos290 <- which(dat[,'vec2'] == 290, arr.ind = TRUE)
dat_pos812 <- which(dat[,'vec2'] == 812, arr.ind = TRUE)
```

Part 1.3

```
dat$vec2[dat_pos2] <- NA
dat$vec2[dat_pos47] <- NA
dat$vec2[dat_pos290] <- NA
dat$vec2[dat_pos812] <- NA
```

Part 1.4

```
colnames(dat) = c("caseid", "wage")
```

Part 1.5

```
mean(as.numeric(dat$wage), na.rm = TRUE)
```

```
## [1] 501.3544
```

```
median(as.numeric(dat$wage), na.rm = TRUE)
```

```
## [1] 501.5
```

```
sd(as.numeric(dat$wage), na.rm=TRUE)
```

```
## [1] 288.3622
```

Part 1.6

```
summary(dat)
```

```
##      caseid            wage       
##  Min.   :   1.0   Min.   :   1.0  
##  1st Qu.: 250.8   1st Qu.: 251.8  
##  Median : 500.5   Median : 501.5  
##  Mean   : 500.5   Mean   : 501.4  
##  3rd Qu.: 750.2   3rd Qu.: 750.2  
##  Max.   :1000.0   Max.   :1000.0  
##                   NA's   :4       
```

```
dat2 = subset(dat, wage != 'NA')
summary(dat2)
```

```
##      caseid            wage       
##  Min.   :   1.0   Min.   :   1.0  
##  1st Qu.: 251.8   1st Qu.: 251.8  
##  Median : 501.5   Median : 501.5  
##  Mean   : 501.0   Mean   : 501.4  
##  3rd Qu.: 751.2   3rd Qu.: 750.2  
##  Max.   :1000.0   Max.   :1000.0  
```

Part 2

```
CAcity <- read.csv("CAcities.csv")
library(ggplot2)
CAcity_ordered <- CAcity[order(CAcity$pop2020, decreasing = FALSE), ]
```

Part 2.1

```
for ( x in 1:length(CAcity$city)) {
  print(CAcity[x, 1])
}
```

```
## [1] "Anaheim"
## [1] "Bakersfield"
## [1] "Fresno"
## [1] "Long Beach"
## [1] "Los Angeles"
## [1] "Oakland"
## [1] "Sacramento"
## [1] "San Diego"
## [1] "San Francisco"
## [1] "San Jose"
```

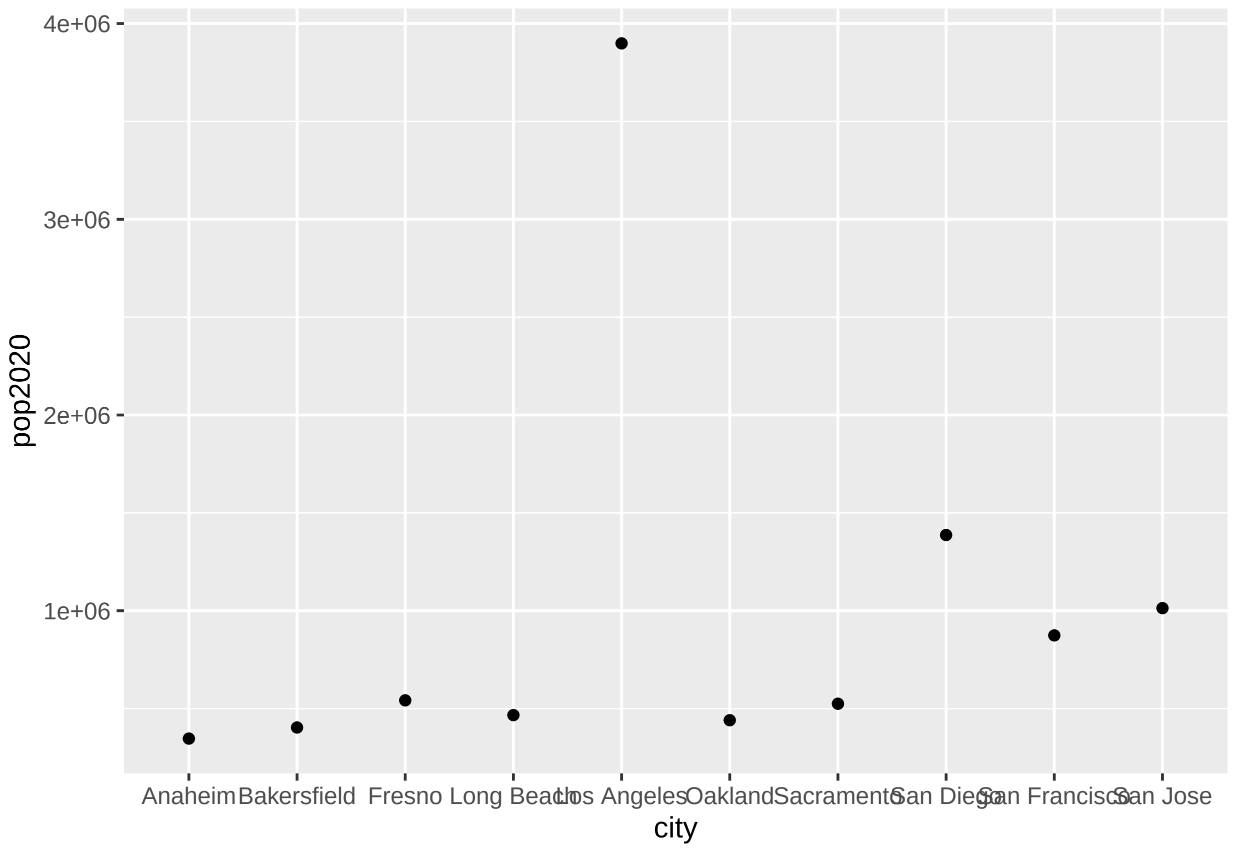
Part 2.2

```
for (x in 1:length(CAcity_ordered$city)) {
  print(CAcity_ordered[x, 1])
}
```

```
## [1] "Anaheim"
## [1] "Bakersfield"
## [1] "Oakland"
## [1] "Long Beach"
## [1] "Sacramento"
## [1] "Fresno"
## [1] "San Francisco"
## [1] "San Jose"
## [1] "San Diego"
## [1] "Los Angeles"
```

Part 2.3

```
ggplot(data=CAcity, aes(x=city, y=pop2020)) + geom_point()
```



Part 3

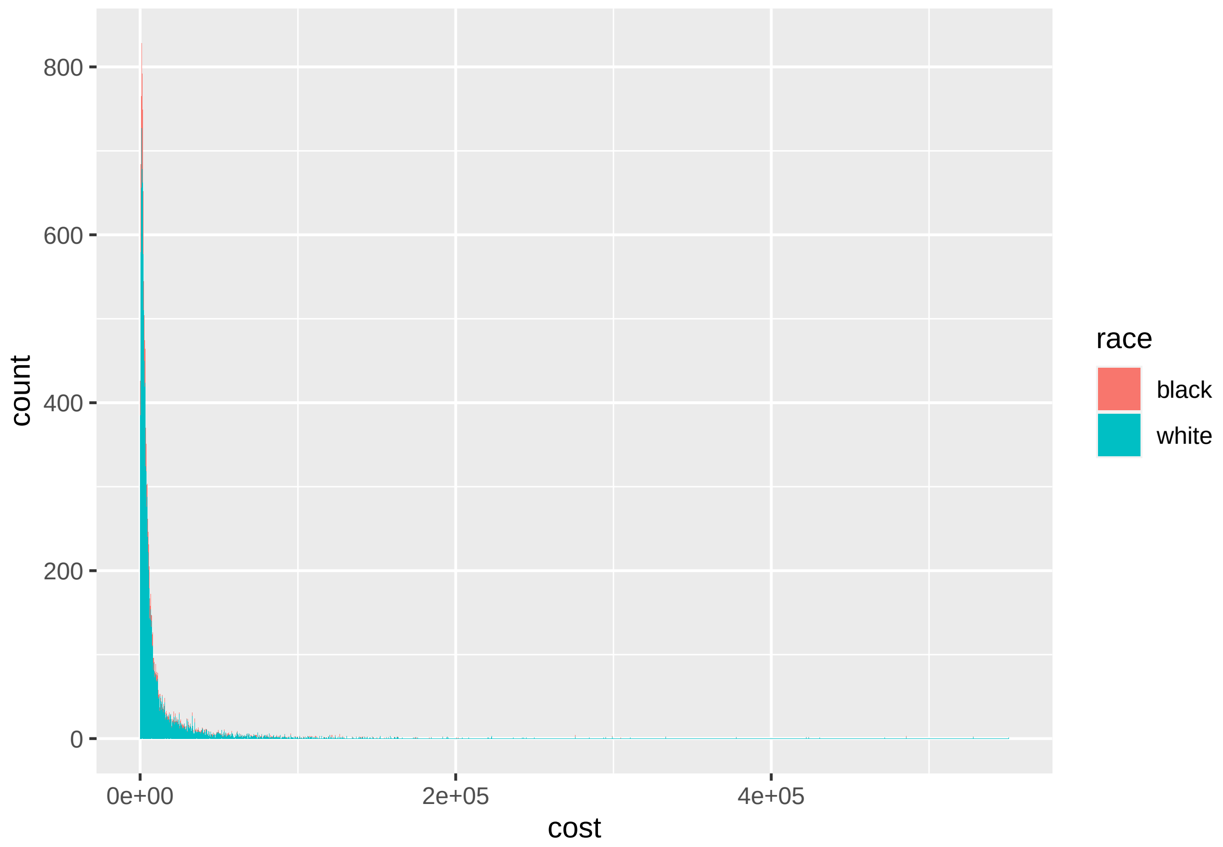
```
hdat <- read.csv('data_health_synth_small.csv')
```

Part 3.1

```
hdat <- na.omit(hdat)
```

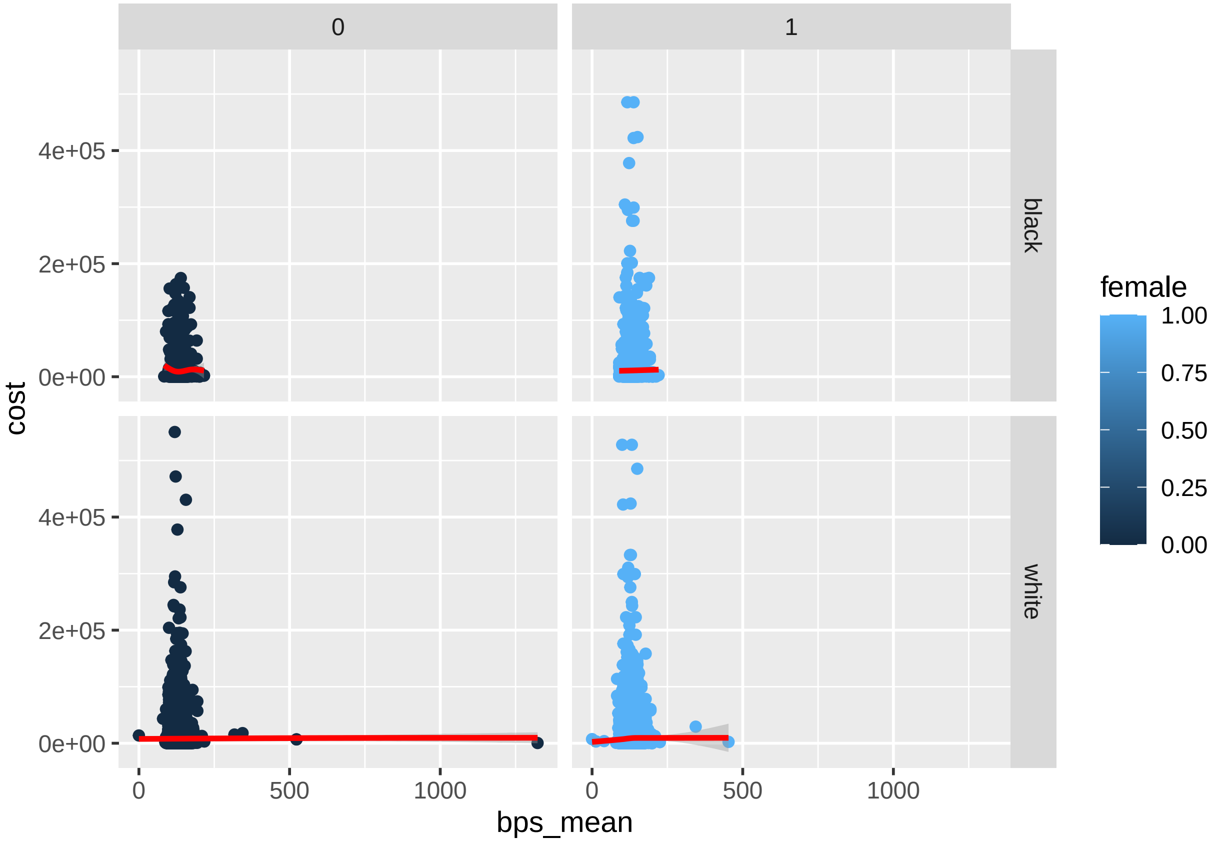
Part 3.2

```
h <- ggplot(data = hdat, aes(x=cost))
h + geom_histogram(binwidth = 100, aes(fill = race))
```



Part 3.3

```
s <- ggplot(data=hdat, aes(x=bps_mean, y=cost))
s + geom_point(aes(color=female)) + geom_smooth(color='red') + facet_grid(race~female)
```



Part 3.4

```
set.seed(12345)
cost_samp <- sample(hdat$cost, length(hdat$cost), replace = TRUE)
```

Part 3.5

```
mean(hdat$cost)
```

```
## [1] 8634.66
```

```
mean(cost_samp)
```

```
## [1] 8524.394
```

```
sd(hdat$cost)
```

```
## [1] 19123.94
```

```
((mean(hdat$cost) - mean(cost_samp))/sd(hdat$cost)) * 100
```

```
## [1] 0.5765863
```

```
## The value of cost_samp lies within one standard deviation of the cost variable of the
## original dataset. The variable cost_samp is actually only 0.576% off from the
## original value, thus making it a fairly accurate approximation.
## They are very similar.
```

Part 3.6

```
cost_samp_1000 <- c()
set.seed(12345)
for (x in 1:1000) {
  cost_samp_1000[x] <- sample((hdat$cost), length(hdat$cost), replace = TRUE)
}
```

Part 3.7

```
sd(cost_samp_1000)
```

```
## [1] 14166.6
```

Part 3.8

```
my_sampsd_function <- function(inputvec){
  cost_samp_1000sd <- c()
  for (x in 1:1000) {
    cost_samp_1000sd[x] <-sample((inputvec), length(inputvec), replace = TRUE)
  }
  return(sd(cost_samp_1000sd))
}
```

Part 3.9

```
set.seed(12345)
my_sampsd_function(hdat$cost)
```

```
## [1] 14166.6
```

Part 3.10

```
set.seed(12345)
my_sampsd_function(hdat$bps_mean)
```

```
## [1] 14.1612
```