Stats 140SL Midterm

Kitu Komya

November 11, 2016

Reading and Cleaning the Data

```
# loading packages
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

```
# reading in the data
library(haven)
demo <- read_sas("lecture2.sas7bdat")
dim(demo)
```

```
## [1] 102299     25
```

We have read it in correctly, since there are 102,299 observations and 25 variables.

```
# looking into data
summary(demo)
```

```
##      PUMA             AREANAME           SERIALNO            SPORDER
##  Length:102299      Length:102299      Length:102299      Min.   : 1.000
##  Class :character   Class :character   Class :character   1st Qu.: 1.000
##  Mode  :character   Mode  :character   Mode  :character   Median : 2.000
##                                                           Mean   : 2.348
##                                                           3rd Qu.: 3.000
##                                                           Max.   :20.000
##
##       AGEP           CIT                COW                JWMNP
##  Min.   : 0.00   Length:102299      Length:102299      Min.   :  1.00
##  1st Qu.:21.00   Class :character   Class :character   1st Qu.: 15.00
##  Median :39.00   Mode  :character   Mode  :character   Median : 30.00
##  Mean   :39.39                                         Mean   : 31.92
##  3rd Qu.:57.00                                         3rd Qu.: 45.00
##  Max.   :94.00                                         Max.   :141.00
##                                                        NA's   :57627
##      JWRIP           JWTR               MAR                MIG
##  Min.   : 1.00   Length:102299      Length:102299      Length:102299
##  1st Qu.: 1.00   Class :character   Class :character   Class :character
##  Median : 1.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 1.19
##  3rd Qu.: 1.00
##  Max.   :10.00
##  NA's   :62717
##      MIL                RELP               SCH
##  Length:102299      Length:102299      Length:102299
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
```

```
##
##
##
##      SCHL               SEX                WKHP           WKL
##  Length:102299      Length:102299      Min.   : 1.0    Length:102299
##  Class :character   Class :character   1st Qu.:32.0    Class :character
##  Mode  :character   Mode  :character   Median :40.0    Mode  :character
##                                        Mean   :37.6
##                                        3rd Qu.:40.0
##                                        Max.   :99.0
##                                        NA's   :49387
##      WKW                ESR               HICOV
##  Length:102299      Length:102299      Length:102299
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      PINCP            SCIENGP              WAOB
##  Min.   :  -5800   Length:102299      Length:102299
##  1st Qu.:   5000   Class :character   Class :character
##  Median :  20000   Mode  :character   Mode  :character
##  Mean   :  38811
##  3rd Qu.:  48000
##  Max.   :1161000
##  NA's   :16840
```

```
head(demo)
```

```
## # A tibble: 6 x 25
##    PUMA                                              AREANAME  SERIALNO
##   <chr>                                                 <chr>     <chr>
## 1 03701 Los Angeles County (North/Unincorporated)--Castaic 000000385
## 2 03701 Los Angeles County (North/Unincorporated)--Castaic 000000385
## 3 03701 Los Angeles County (North/Unincorporated)--Castaic 000002968
## 4 03701 Los Angeles County (North/Unincorporated)--Castaic 000002968
## 5 03701 Los Angeles County (North/Unincorporated)--Castaic 000002968
## 6 03701 Los Angeles County (North/Unincorporated)--Castaic 000002968
## # ... with 22 more variables: SPORDER <dbl>, AGEP <dbl>, CIT <chr>,
## #   COW <chr>, JWMNP <dbl>, JWRIP <dbl>, JWTR <chr>, MAR <chr>, MIG <chr>,
## #   MIL <chr>, RELP <chr>, SCH <chr>, SCHL <chr>, SEX <chr>, WKHP <dbl>,
## #   WKL <chr>, WKW <chr>, ESR <chr>, HICOV <chr>, PINCP <dbl>,
## #   SCIENGP <chr>, WAOB <chr>
```

Based on the information given, some of the variables need to be of other class types. Let's change that.

```
demo$PUMA <- as.numeric(demo$PUMA) # as recommended by professor
demo$SERIALNO <- as.factor(demo$SERIALNO) # households should be discrete
demo$CIT <- as.factor(demo$CIT) # citizenship status should be discrete
demo$COW <- as.factor(demo$COW) # class of worker should be discrete
demo$JWRIP <- as.factor(demo$JWRIP) # vehicle occupancy should be discrete
demo$JWTR <- as.factor(demo$JWTR) # means of transportation to work should be discrete
demo$MAR <- as.factor(demo$MAR) # marital status should be discrete
demo$MIG <- as.factor(demo$MIG) # migrant status should be discrete
demo$MIL <- as.factor(demo$MIG) # military status should be discrete
```

```
demo$RELP <- as.factor(demo$RELP) # relationship status should be discrete
demo$SCH <- as.factor(demo$SCH) # school enrollment should be discrete
demo$SCHL <- as.integer(demo$SCHL) # educational attainment should be continuous because there is order
demo$SEX <- as.factor(demo$SEX) # gender should be discrete
demo$WKL <- as.factor(demo$WKL) # when last worked should be discrete
demo$ESR <- as.factor(demo$ESR) # employment status recode should be discrete
demo$HICOV <- as.factor(demo$HICOV) # health insurance coverage should be discrete
demo$SCIENGP <- as.factor(demo$SCIENGP) # field of degree in science and engineering should be discrete
demo$WAOB <- as.factor(demo$WAOB) # world area of birth should be discrete

# we are removing variable WKW, at the discretion of the professor:
demo$WKW <- NULL
```

There. We have cleaned all of the variable's classes. Now analyses will make more sense. Now, however, as seen by the summary function, there were quite a few NA cells within each variable. Let's just remove all entries that contain any NA values, for the sake of time and simplicity (in the future, we could use regression methods and probability to replace NA values with values the cells are most likely to be).

```
demo <- demo[complete.cases(demo), ]
dim(demo)
```

```
## [1] 39582    24
```

Yikes! We have trimmed our data from 102,299 observations to 39,582 observations. That's nearly a 40% reduction! Again, however, this choice is justified because of our limited time and desire for quality analyses from a full dataset.

Analyzing the Data

As a passionate Bruin, Trojans are my life-long nemesis. Therefore, I am interested in comparing Westwood (UCLA) with Boyle Heights (USC). By searching using the filter method while viewing the dataframe, we see that the PUMA of Westwood is 3729 and the PUMA of Boyle Heights is 3744.

```
sum(demo$PUMA == 3729) # ensuring enough Westwood observations
```

```
## [1] 763
```

```
sum(demo$PUMA == 3744) # ensuring enough Boyle Heights observations
```

```
## [1] 559
```

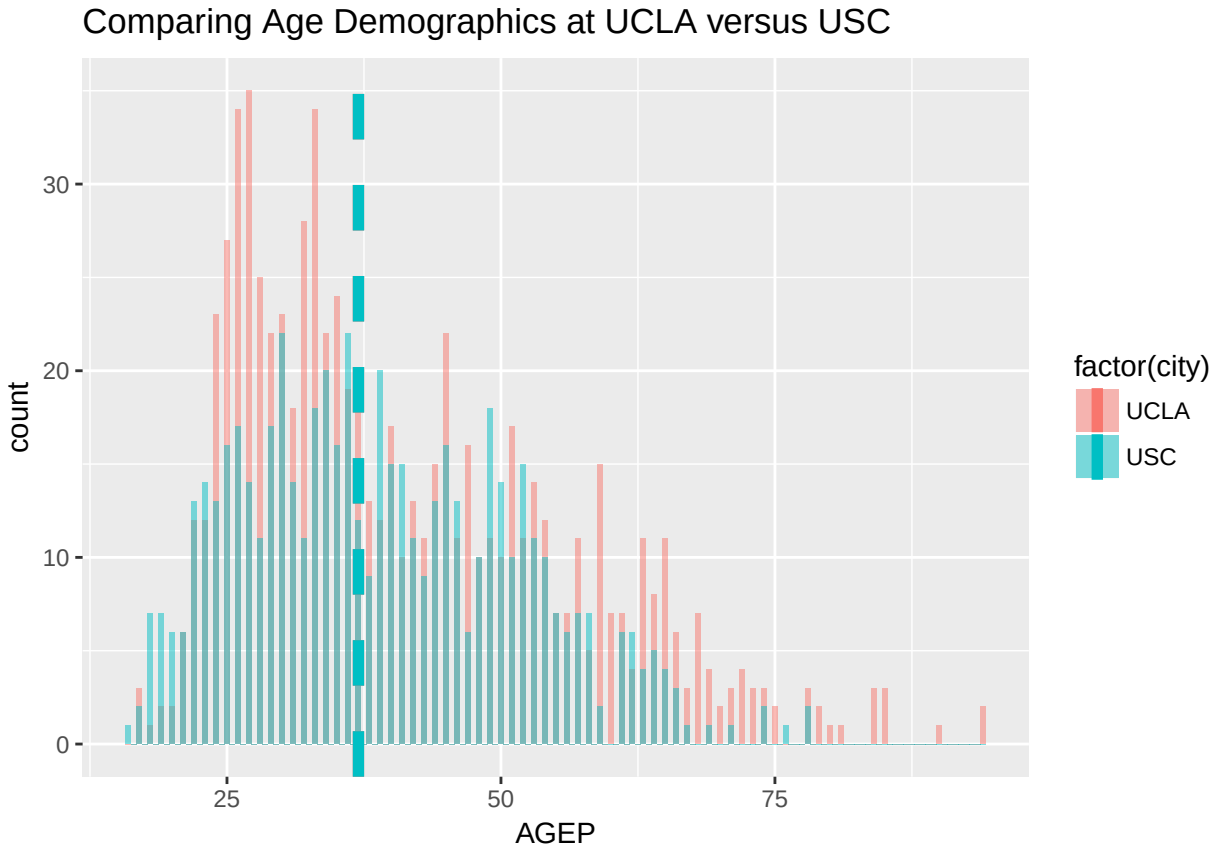
There are enough observations in both of the neighborhoods to continue analysis.

```
# let's subset dataframe to our 2 neighborhoods
demo_ww <- demo[demo$PUMA == 3729, ] # westwood
demo_usc <- demo[demo$PUMA == 3744, ] # usc
demo_sub <- rbind(demo_ww, demo_usc) # combined

# create a new city variable
demo_sub$city <- ifelse(demo_sub$PUMA == 3729, "UCLA", "USC")

# re-name PINCP variable
names(demo_sub)[22] <- "Total Person's Income"
```

Now that we have subsetted the dataframe, let's make some plots.

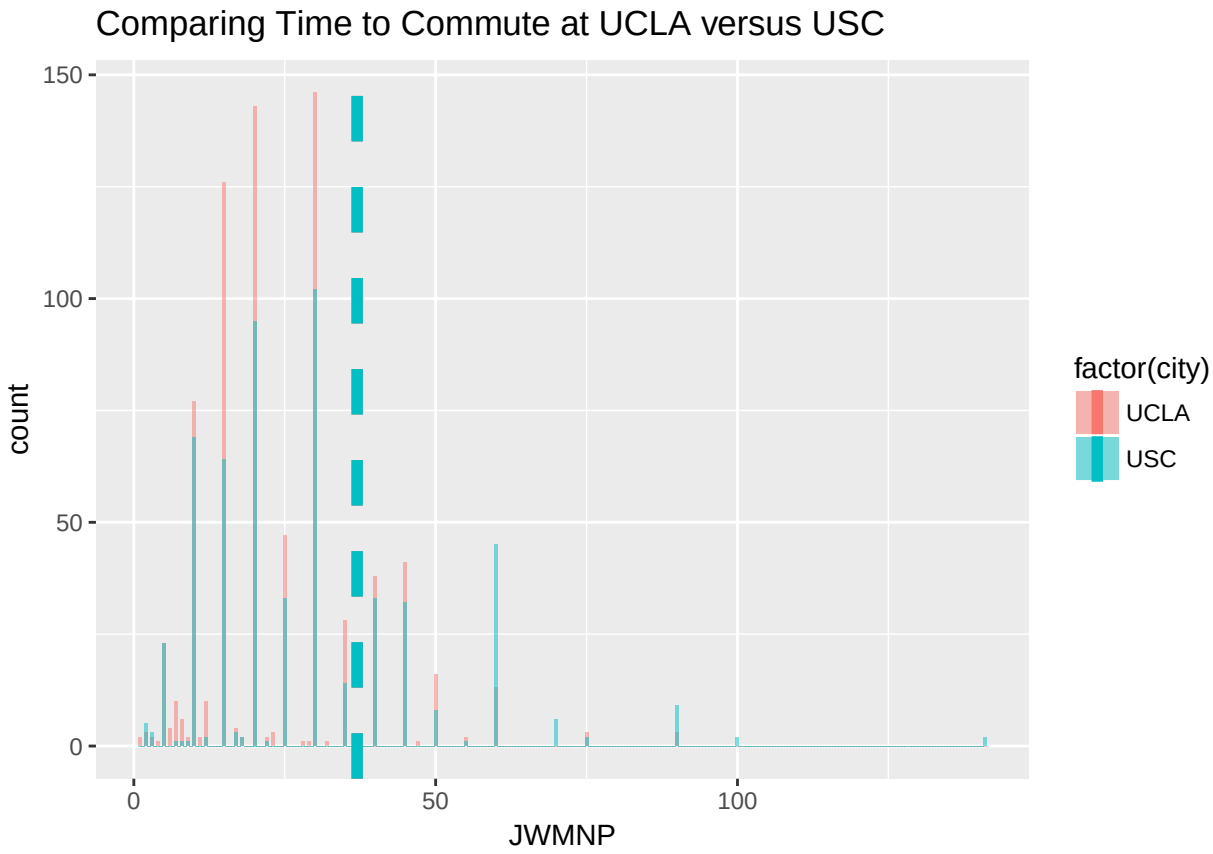
1

```
# let's compare ages in both cities
ggplot(demo_sub, aes(x = AGEP, fill = factor(city))) +
  geom_histogram(binwidth = 0.5, alpha = 0.5, position = "identity") + labs(title = "Comparing Age Demog
  geom_vline(data = demo_sub, aes(xintercept = median(demo_sub$AGEP),  colour = factor(city)),
             linetype = "dashed", size = 2)
```



The demographics are roughly the same, although UCLA has more inhabitants than USC. The medians are the exact same, hence the overlap.

```
# let's compare time to commute in both cities
ggplot(demo_sub, aes(x = JWMNP, fill = factor(city))) +
  geom_histogram(binwidth = 0.5, alpha = 0.5, position = "identity") + labs(title = "Comparing Time to C
  geom_vline(data = demo_sub, aes(xintercept = median(demo_sub$AGEP),  colour = factor(city)),
             linetype = "dashed", size = 2)
```



The commuting times are roughly the same, although UCLA has more inhabitants than USC. The medians are the exact same, hence the overlap.

```
# compare economic demographics
summary(demo_ww$PINCP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -5800   34800   62000   98653  100215 1161000
```

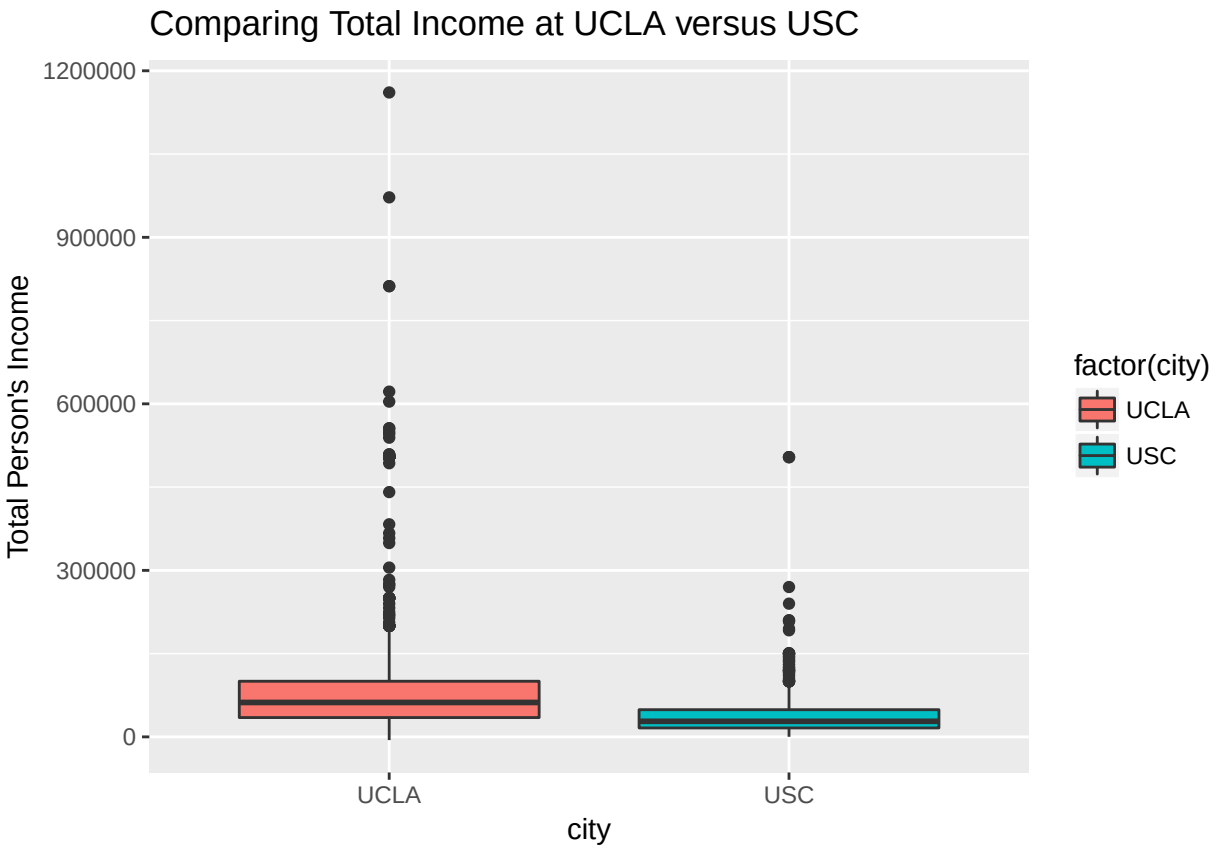
```
summary(demo_usc$PINCP)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     220   16100   28000   41595   48950  504000
```

Wow! As we can see, UCLA has a much higher median and mean in terms of income than at USC. Let's further explore this in the next question.

2

```
# let's plot neighborhood versus income
ggplot(demo_sub, aes(x = city, y = `Total Person's Income`, fill = factor(city))) + geom_boxplot() + lab
```



We can see that UCLA folks have a higher income than USC. . . whoo!

3

In order to statistically compare commute times or economic levels, just showing graphs is not enough. I would create a linear model within each of the two subsets for the cities and compare which factors affect each variable per city.

Another idea I would do is do ANOVA testing to compare the medians or means of the income or commuting time to see if they are statistically different from one another.