Lab 2 - Intro to US Census and Mapping

Prof. Anthony Howell

In this lab you will practice working with the US census data searching for variables, loading them into R, performing basic calculations and data wrangling tasks, visualizing the data with figures and maps, and offer substantive feedback on the data analysis.

The topic for our analysis will be to compare the house-price-to-income ratio across US counties and over time (pre- versus post- financial crisius). The ratio tells the number of years it would take for the median income household to buy the median household price. Under healthy economic conditions, the rule of thumb is that a buyer can afford a house if its price is equivalent to a house-price-to-income ratio of 2.6. Read the Citylab report by Richard Florida for more background on the ratio and its importance and US county rankings.

Step 1:

- Load `ggplot2` Library which contains the `mpg` dataframe

```
#Edit me
```

```
library(ggplot2)
```

- Explore `mpg` Data using `head`, `str`, `summary`, and `names`

```
#edit me
```

```
head(mpg)
```

```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans  drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>  <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(~ f        18    29 p     comp~
## 2 audi         a4      1.8  1999     4 manua~ f        21    29 p     comp~
## 3 audi         a4      2    2008     4 manua~ f        20    31 p     comp~
## 4 audi         a4      2    2008     4 auto(~ f        21    30 p     comp~
## 5 audi         a4      2.8  1999     6 auto(~ f        16    26 p     comp~
## 6 audi         a4      2.8  1999     6 manua~ f        18    26 p     comp~
```

```
str(mpg)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    234 obs. of  11 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

```
summary(mpg)
```

```
##  manufacturer          model               displ            year
##  Length:234         Length:234         Min.   :1.600   Min.   :1999
##  Class :character   Class :character   1st Qu.:2.400   1st Qu.:1999
##  Mode  :character   Mode  :character   Median :3.300   Median :2004
##                                        Mean   :3.472   Mean   :2004
##                                        3rd Qu.:4.600   3rd Qu.:2008
##                                        Max.   :7.000   Max.   :2008
##       cyl           trans                drv                 cty
##  Min.   :4.000   Length:234         Length:234         Min.   : 9.00
##  1st Qu.:4.000   Class :character   Class :character   1st Qu.:14.00
##  Median :6.000   Mode  :character   Mode  :character   Median :17.00
##  Mean   :5.889                                         Mean   :16.86
##  3rd Qu.:8.000                                         3rd Qu.:19.00
##  Max.   :8.000                                         Max.   :35.00
##       hwy              fl               class
##  Min.   :12.00   Length:234         Length:234
##  1st Qu.:18.00   Class :character   Class :character
##  Median :24.00   Mode  :character   Mode  :character
##  Mean   :23.44
##  3rd Qu.:27.00
##  Max.   :44.00
```

```
names(mpg)
```

```
##  [1] "manufacturer" "model"        "displ"        "year"
##  [5] "cyl"          "trans"        "drv"          "cty"
##  [9] "hwy"          "fl"           "class"
```

Question: Which variables in mpg are categorical? Which variables are continuous?

```
#edit me
```

Data Manipulations

Among the variables in mpg are:

`displ` – a car's engine size in litres.

`hwy` – a car's fuel efficiency on the highway, in miles per gallon (mpg). A car with a low fuel efficiency consumes more fuel than a car with a high fuel efficiency when they travel the same distance.

`class` – a class variable tells the class of each car

- Rename `displ` and `hwy` to `EngSize` and `FuelEff` (See `A Review` Step 2 for a hint how to do this.)

```
#edit me
```

```
colnames(mpg)[3] <- "EngSize"
colnames(mpg)[9] <- "FuelEff"
```

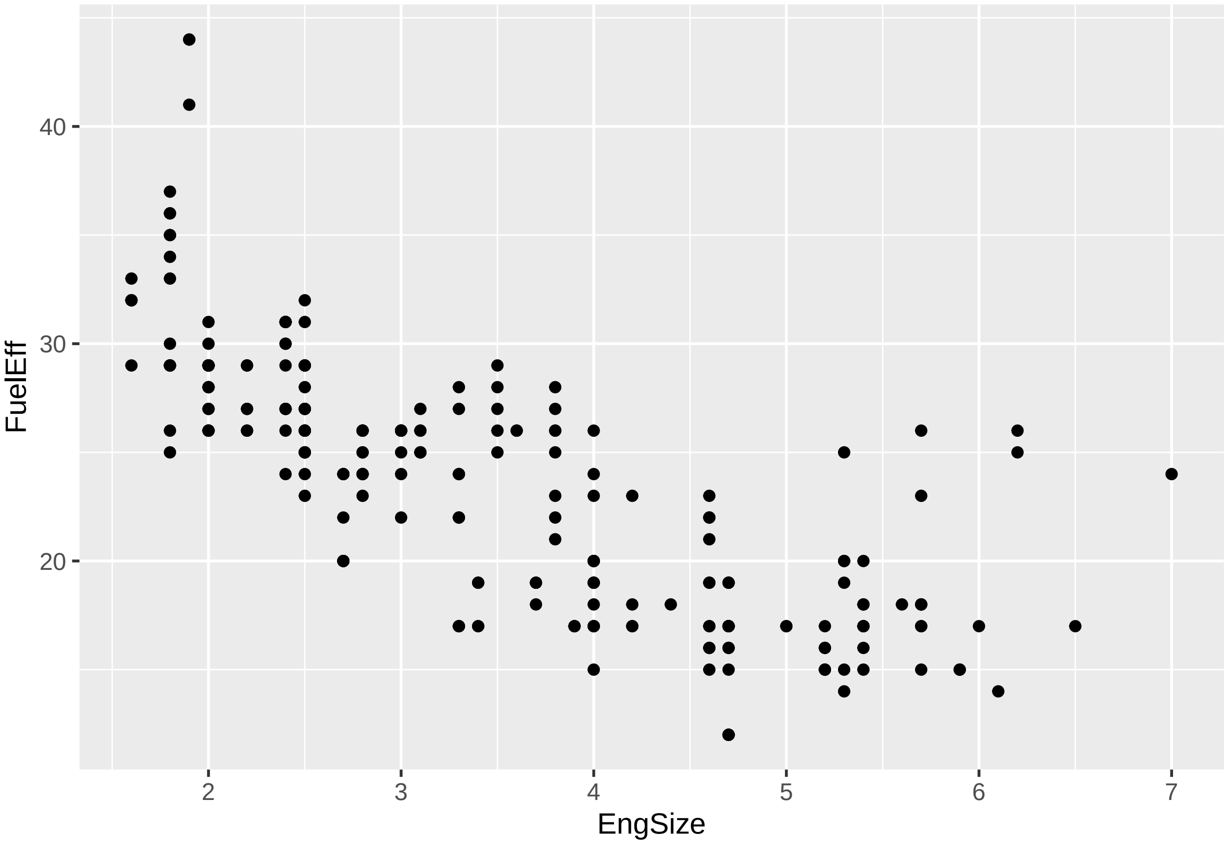
Visualize relationships

Note: (See `Advanced Graphics with ggplot2...` Steps 3, 4 and 5 for help)

- Use `ggplot` with `geom_point` to map the relationship between `EngSize` and `FuelEff`

```
#edit me
```

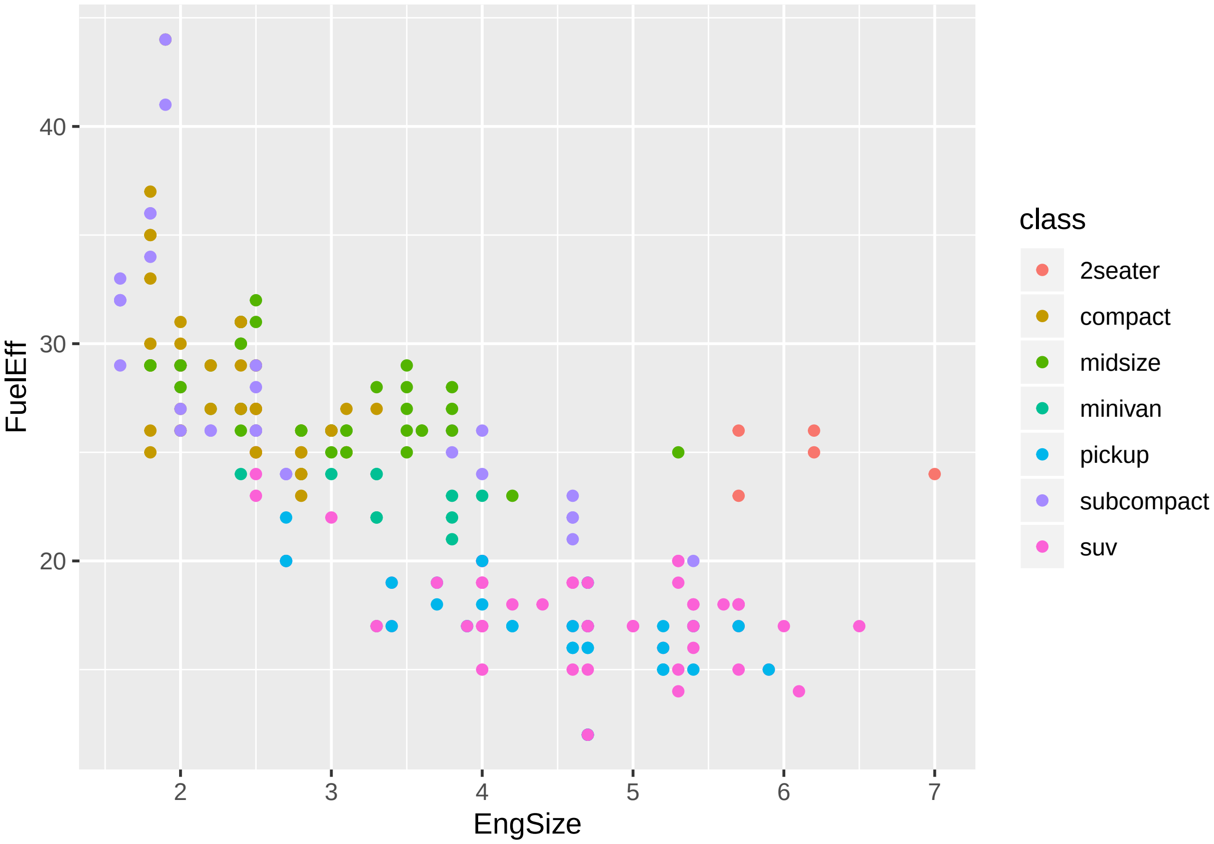
```
ggplot(data = mpg) +
geom_point(mapping = aes(x = EngSize, y = FuelEff))
```



- Add a `color` argument to better visualize your plot according to the `class` of the vehicle

```
#edit me
```

```
ggplot(data = mpg) +
geom_point(mapping = aes(x = EngSize, y = FuelEff, color=class))
```

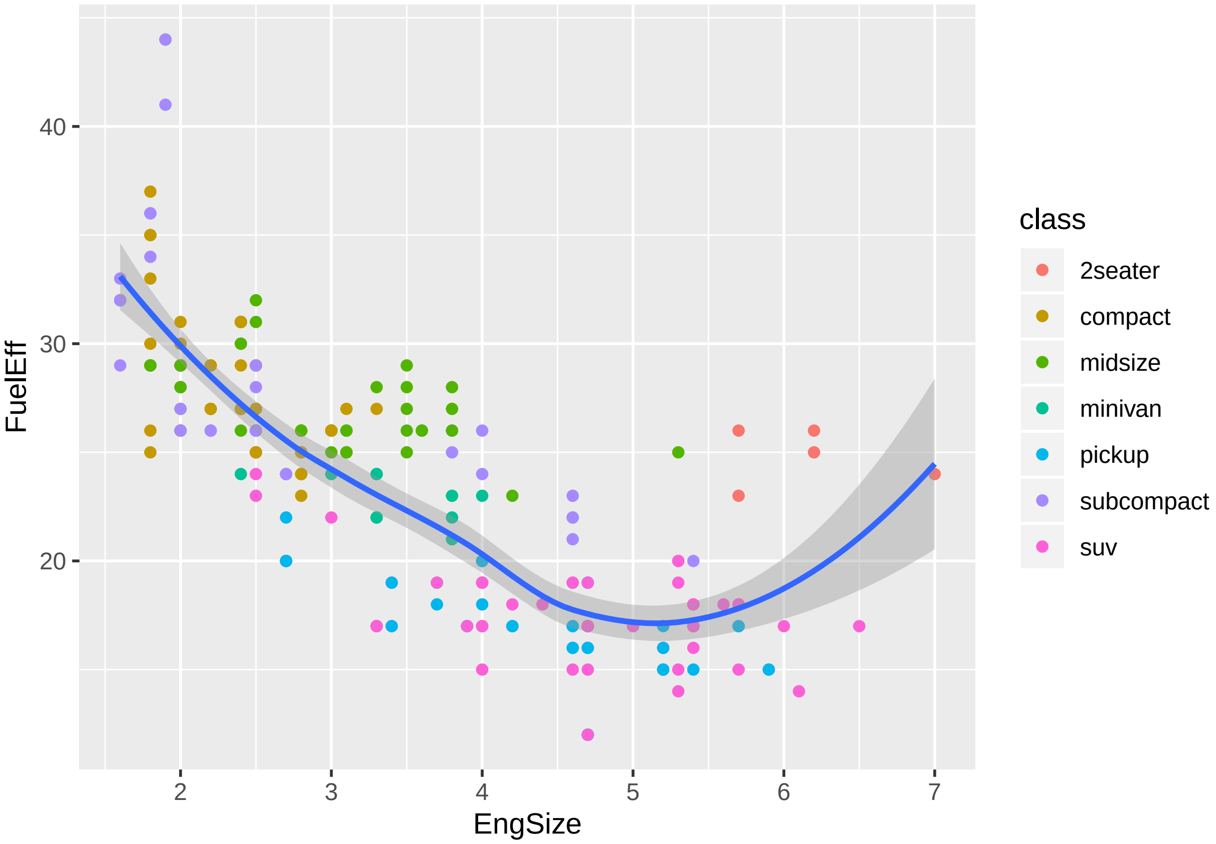


- Add a statistical overlay of using `geom_smooth`

```
#edit me
```

```
ggplot(data = mpg, mapping = aes(x = EngSize, y = FuelEff)) +
geom_point(mapping = aes(color = class)) +
geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Remove Outliers

- Remove 2seater cars from the data using `subset` and the argument `class != '2seater' (see Note above in`Statistical Overlay‘ section)

```
#edit me
```

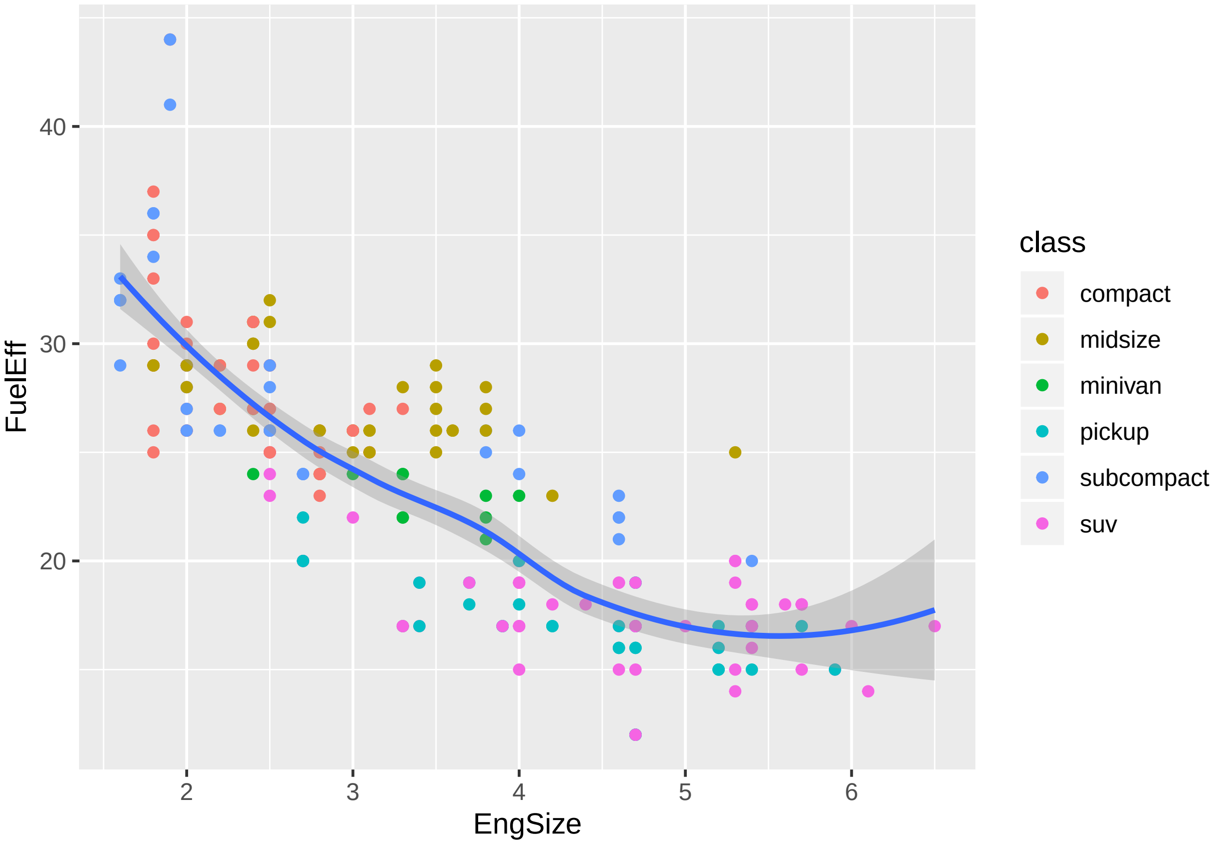
```
mpg<-subset(mpg, subset= class != "2seater")
```

- Re-plot the data again using `geom_point` and `geom_smooth`

```
#edit me
```

```
ggplot(data = mpg, mapping = aes(x = EngSize, y = FuelEff)) +
geom_point(mapping = aes(color = class)) +
geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



-Question: What do you notice happened to the regression line after removing the 2seater class from the data?

```
#edit me
```