Describing distributions of data

Assignment Overview

There are a variety of conventional ways to visualize data - tables, histograms, bar graphs, etc. The purpose is always to examine the distribution of variables related to your research question. You will create a plot, follow up each graphic with a table of summary statistics (for quantitative variables) or frequency and proportion table (for categorical), and then a summary paragraph that brings it all together.

Do not worry about your data being "messy" at this point. Some of your graphics may look odd or "not right". You may have bars for NA values, or a huge spike of data at some odd value of 999. Address these oddities in your writeup. Fix them if you know how to. Data cleaning is the topic of week 3.

Instructions

- Use the template provided: [RMD] for R users, and [Word] for SPSS users. Rename this file to `univ_graphing_userid`
- Completely describe 2 categorical and 2 quantitative variables using
  - A table of summary statistics,
  - An appropriate plot with titles and axes labels,
  - A short paragraph description in full complete English sentences.
- Upload the final PDF to `03 Univariate Graphing` folder in Google Drive with the file name: `univ_graphing_userid.pdf`

To guide your description of this distribution try to include the following information:

- What is the trend in the data? What exactly does the chart show? (Use the chart title to help you answer this question)
- What are the axes and what are the units?
- Describe the shape:
  - Symmetry/Skewness - Is it symmetric, skewed right, or skewed left?
  - Modality - Is it uniform, unimodal, or bimodal?

- Describe the spread:
  - Variability - What is the approximate range of the data (x-axis)?
  - Does the variable have a lot of variability in the data (visually, are the participants responded to many different responses or mainly just one)?
- Describe the center: What is the mean/median/midpoint of the data? (Pick one or two). Don't
- Describe the outliers (note: there may not be any for every graph):
  - Are there any outliers for the variable?
  - If yes, are these true outliers or false (due to data management or input error) outliers?

Example

This example uses the `mpg` data set from the `ggplot2` package.

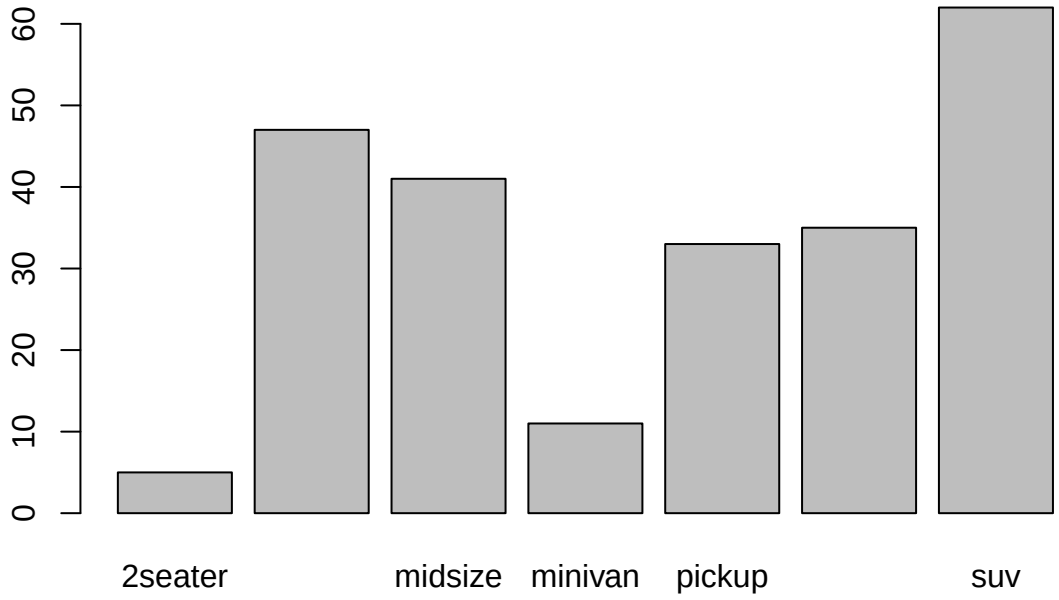
```
mpg <- ggplot2::mpg
```

Basic categorical

Draft style plot, direct computer output showing/copied. Poor grammar and/or sentence structure, no attempt at explaining what the variable means, extra unnecessary or incorrect information included. Typos.

class

```
library(descr)
freq(mpg$class)
```




```
## mpg$class
##            Frequency Percent
## 2seater            5   2.137
## compact           47  20.085
## midsize           41  17.521
## minivan           11   4.701
## pickup            33  14.103
## subcompact        35  14.957
## suv               62  26.496
## Total            234 100.000
```

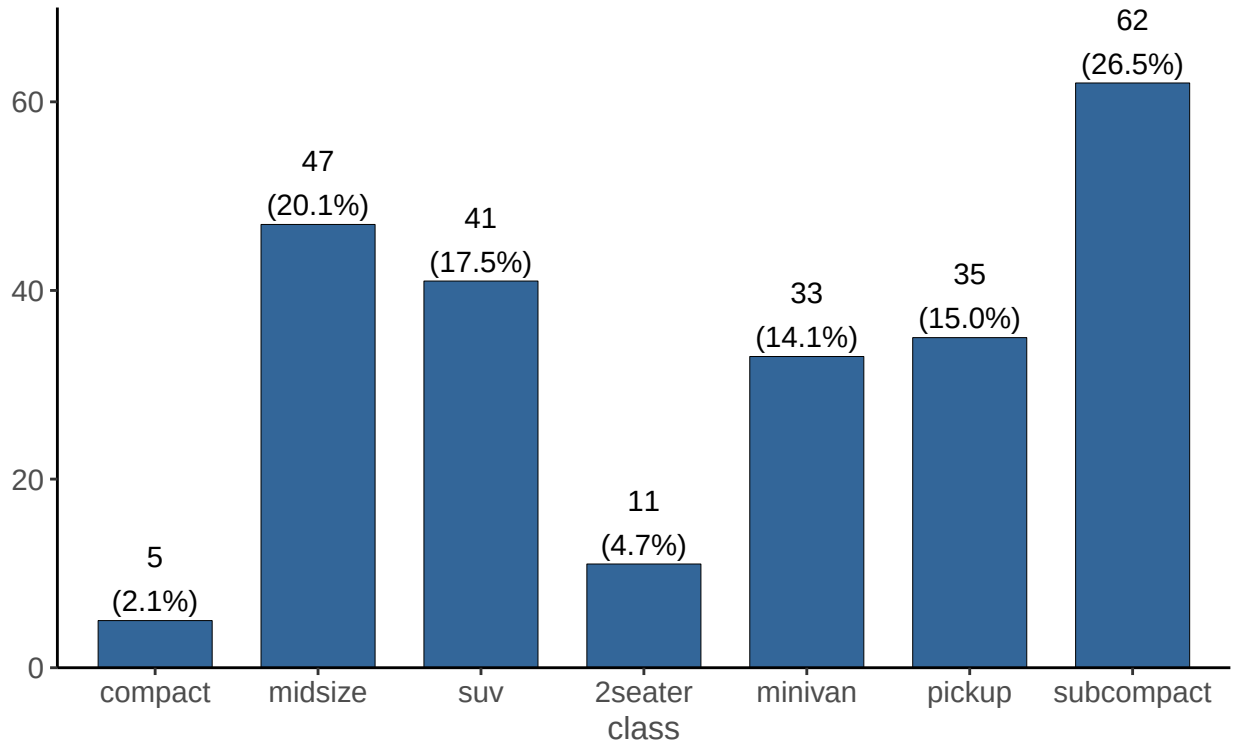
theres more suvs than compacts. 2% are 2seaters. there are 5 2seaters 47 cmpact 41 midize 11 minivans 33 pickups 35% subcompacts, 62 suv and 234 total cars.

Proficient categorical

Cleaned up plot, full English sentences, useful text formatting of variable names and levels. Explained what the variable was named and what it measured.

The `class` variable from the `mpg` data set is a catgorical variable that describes the type of vehicle being measured. Some levels of this categorical variable include *compact*, *pickup* and *suv*.

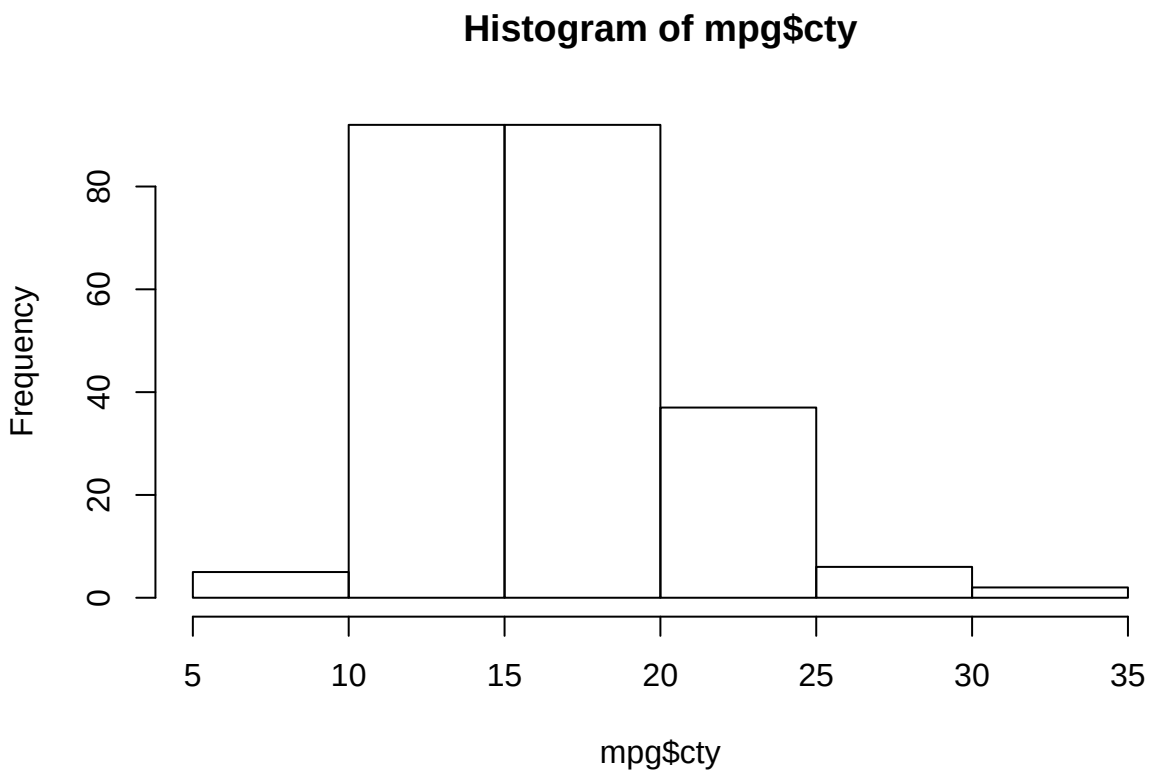
```
library(sjPlot); library(ggplot2)
set_theme(base = theme_classic())
sjp.frq(mpg$class)
```



Sub compact cars are the most frequently reported type of car, making up over one-quarter (26.5%) of the cars in this data set with n=62 cars represented. The least represented car is a compact car with n=5 (2.1%) records.

Basic quantitative

```
hist(mpg$cty)
```




```
summary(mpg$cty)
```

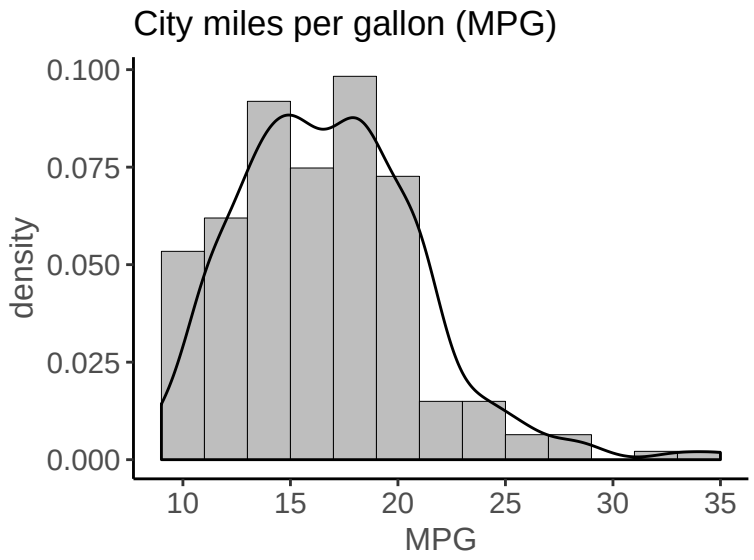
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   14.00   17.00   16.86   19.00   35.00
```

Proficient quantitative

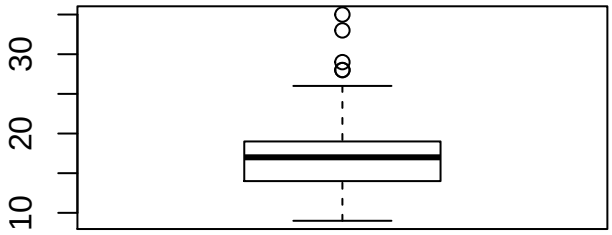
Overlaid a density curve on the histogram, also looked at a boxplot for outliers. Table of summary statistics present in a nicely formatted way, digits rounded appropriately. Plot cleaned up with appropriate axis and titles.

> The `cty` variable records the miles per gallon (mpg) achieved during city driving. This is a quantititative numeric variable.

```
ggplot(mpg, aes(x=cty)) + geom_histogram(aes(y=..density..), fill="grey", binwidth = 2) +
      geom_density()+ xlab("MPG") + ggtitle("City miles per gallon (MPG)")
```




```
boxplot(mpg$cty)
```




```
knitr::kable(t(c(summary(mpg$cty), sd=sd(mpg$cty))), digits=1)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
9	14	17	16.9	19	35	4.3

The MPG in the city ranges from 9 to 35, unimodal and is slightly skewed right with a mean of 16.9 close to the median of 17 and a standard deviation of 4.3mpg. The boxplot indicates that there are at least 4 upper end outliers achieving a city MPG of approximately over 28 mpg.