Descriptive Statistics

1. Overview of Last Week
2. Measurement
3. Descriptive Statistics
4. Readings
5. Wrap up

Measurement

- Social science is about developing and testing causal theories:
  - Does minimum wage change levels of employment?
  - Does outgroup contact influence views on immigration?
- Theories are made up of concepts:
  - Minimum wage, level of employment, outgroup contact, views on immigration.
  - We took these for granted when talking about causality.
- Important to consider how we measure these concepts.
  - Some more straightforward: what is your age?
  - Others more complicated: what does it mean to “be liberal”?
  - Have to create an operational definition of a concept to make it into a variable in our dataset.

Example

- Concept: presidential approval.
- Conceptual definition:
  - Extent to which US adults support the actions and policies of the current US president.
  - Operational definition:
- “On a scale from 1 to 5, where 1 is least supportive and 5 is more supportive, how much would you say you support the job that Donald Trump is doing as president?”

Measurement Error

- **Measurement error**: chance variation in our measurements.
  - individual measurement = exact value + chance error
  - chance errors tend to cancel out when we take averages.
- No matter how careful we are, a measurement could have always come out differently.
  - Panel study of 19,000 respondents: 20 reported being a citizen in 2010 and then a non-citizen in 2012.
  - Data entry errors.
- **Bias**: systematic errors for all units in the same direction.
  - individual measurement = exact value + bias + chance error.
  - “Did you vote?” ~ overreporting

Definitions

- A **variable** is a series of measurements about some concept.
- **Descriptive statistics** are numerical summaries of those measurements.
  - If we smart enough, we wouldn't need them: just look at the list of numbers and completely understand.
- Two salient features of a variable that we want to know:
  - **Central tendency**: where is the middle/typical/average value.
  - **Spread** around the center: are all the data close to the center or spread out?

Center of the Data

- "Center" of the data: Typical/average value
- **Mean**: sum of the values divided by the number od observations
- **Median**: the "middle" of a sorted list of numbers.
- Median more robust to outliers
  - Example 1: data = {0, 1, 2, 3, 5}, mean = 2.2, median = 2
  - Example 2: data = {0, 1, 2, 3, 100}, mean = 21.2, median = 2

Spread of the data

- Are the data close to the center?
- **Range**: [min(x), max(x]
- **Quantile** (quartile, quintile, percentile, etc):
  - 25th percentile = lower quartile (25% of the data below this value)
  - 50th percentile = median (50% of the data below this value)
  - 75th percentile = upper quartile (75% of the data below this value)
- **Interquartile range (IQR)**: a measure of variability
  - How spread out is the middle half of the data?
  - Is most of the data really close to the median or are the values spread out?
- One definition of outliers: over 1.5 × IQR above the upper quartile or below lower quartile.

Standard deviation

- **Standard deviation**: On average, how far away are data points from the mean?

$$\sqrt{\frac{\sum((x-\bar{x})^2)}{n-1}}$$

- Steps:

1. Subtract each data point by the mean
2. Square each resulting difference
3. Take the sum of these values
4. Divide by n-1
5. Take the square root

Variance

- **Variance** = standard deviation (squared)

$$\frac{\sum((x - \bar{x})^2)}{n - 1}$$

How large is large?

- Need a way to put any variable on common units.
- **z-score**:

$$\frac{x - \bar{x}}{\sigma}$$

- Interpretation:
  - Positive values above the mean, negative values below the mean
  - Units now on the scale of standard deviations away from the mean
  - Intuition: data more than 3 SDs away from mean are rare.

Shape

- *Skewness*
  - Positive/right skew
  - Symmetric
  - Negative/left skew
- *Kurtosis*: peakedness of a distribution

Skewness

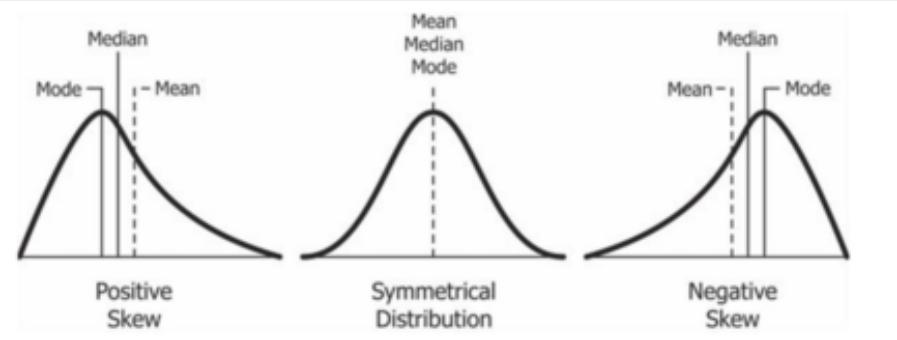


Figure 1: Skewness

Kurtosis

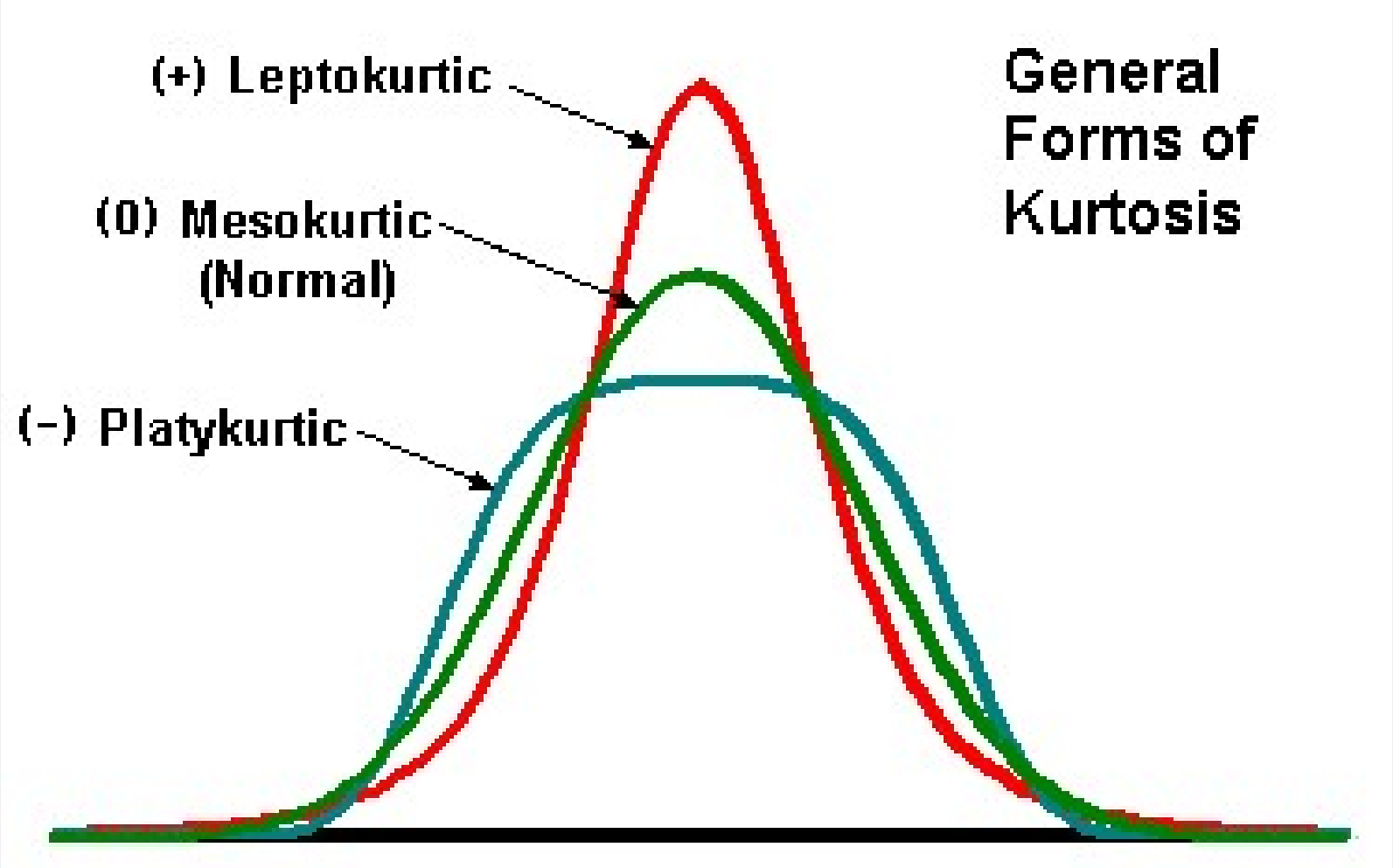


Figure 2: Kurtosis

Correlation

- How do variables move together on average?
- If I know one variable is big, does that tell me anything about how big the other variable is?
  - Positive correlation: when X is big, Y is also big
  - Negative correlation: when X is big, Y is small
  - High correlation: data cluster tightly around a line.

Whats is a strong correlation?

- In social sciences, usually we consider
  - lower than 0.3 = weak
  - between 0.3 and 0.5 = moderate
  - higher than 0.5 = strong
  - (but don't quote me on that!)

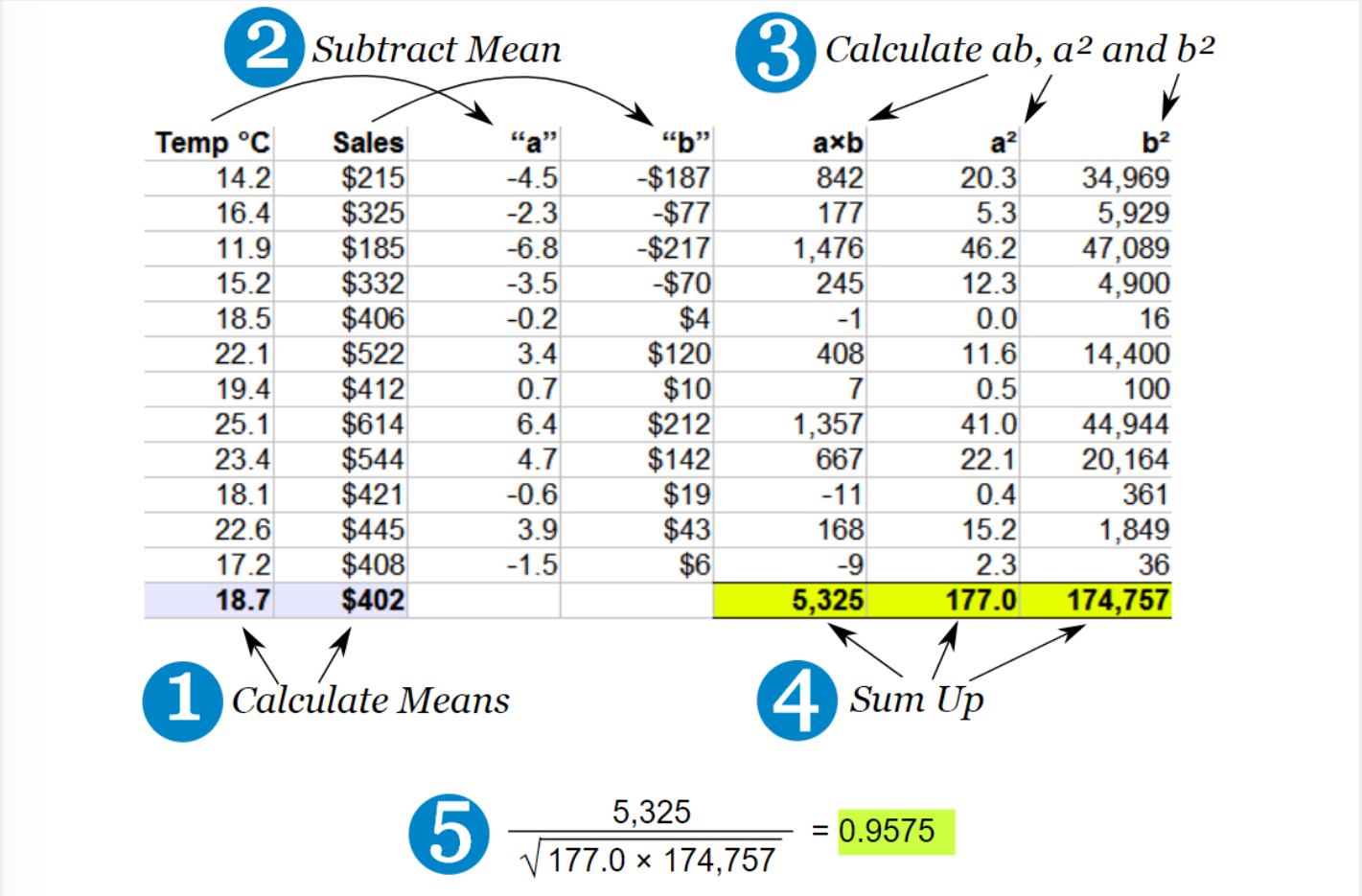
Relationship

- **Covariance**: provides a measure of the strength of the correlation between two or more sets of random variates.

$$cov_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

- **Correlation** is the degree to which two or more quantities are linearly associated

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$



Temp °C	Sales	"a"	"b"	a×b	a²	b²
14.2	$215	-4.5	-$187	842	20.3	34,969
16.4	$325	-2.3	-$77	177	5.3	5,929
11.9	$185	-6.8	-$217	1,476	46.2	47,089
15.2	$332	-3.5	-$70	245	12.3	4,900
18.5	$406	-0.2	$4	-1	0.0	16
22.1	$522	3.4	$120	408	11.6	14,400
19.4	$412	0.7	$10	7	0.5	100
25.1	$614	6.4	$212	1,357	41.0	44,944
23.4	$544	4.7	$142	667	22.1	20,164
18.1	$421	-0.6	$19	-11	0.4	361
22.6	$445	3.9	$43	168	15.2	1,849
17.2	$408	-1.5	$6	-9	2.3	36
18.7	**$402**			**5,325**	**177.0**	**174,757**

Figure 3: Correlation by hand

Correlation

- Correlation is **Positive** when the values **increase** together
- Correlation is **Negative** when one value **decreases** as the other increases
- Correlation can have a value:
  - 1 is a perfect **positive** correlation
  - 0 is no correlation (the values don't seem linked at all)
  - -1 is a perfect **negative** correlation

Correlation

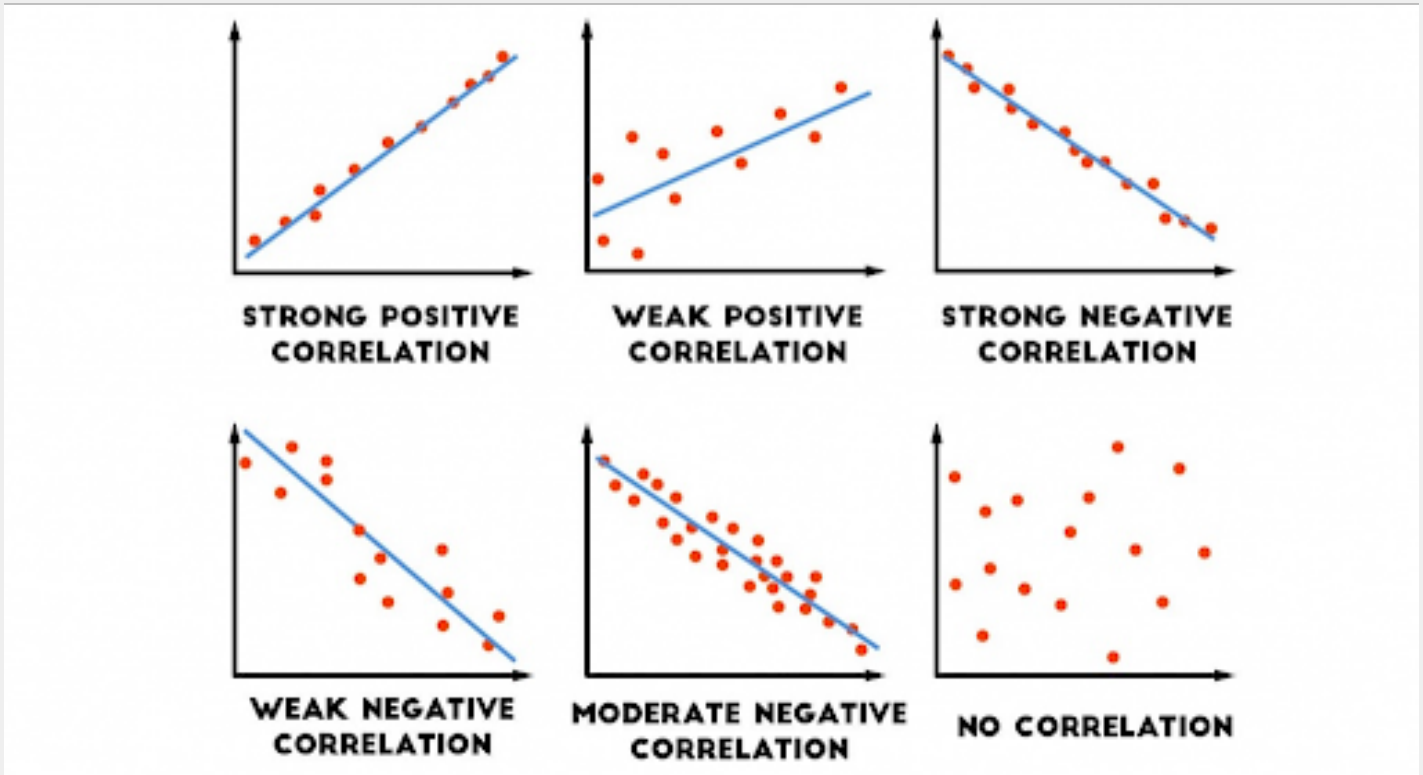


Figure 4: Correlation

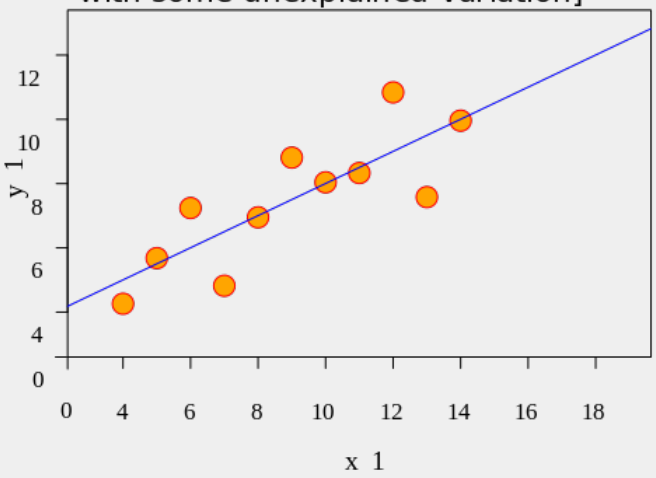
Properties of correlation coefficient

- Correlation measures **linear** association.
- Interpretation:
  - Correlation is between -1 and 1
  - Correlation of 0 means no linear association.
  - Positive correlations ~ positive associations.
  - Negative correlations ~ negative associations.
  - Closer to -1 or 1 means stronger association.
- Order doesn't matter: cor(x,y) = cor(y,x)
- Not affected by changes of scale:
  - Celsius vs. Fahreneheit; dollars vs. pesos; cm vs. in.

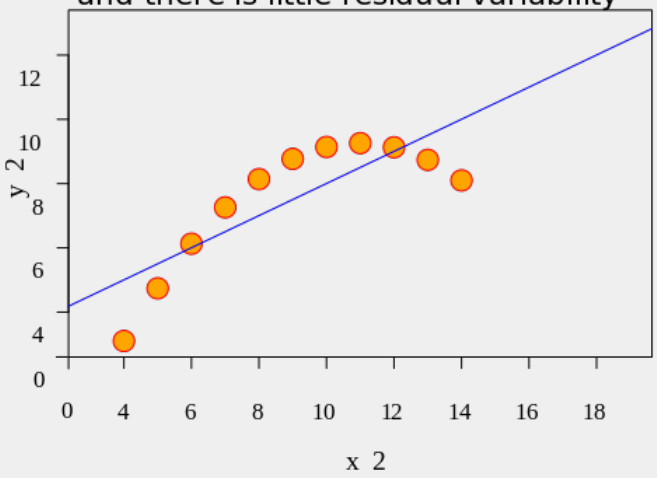
Correlation is not good at curves

Anscombe's Quartet

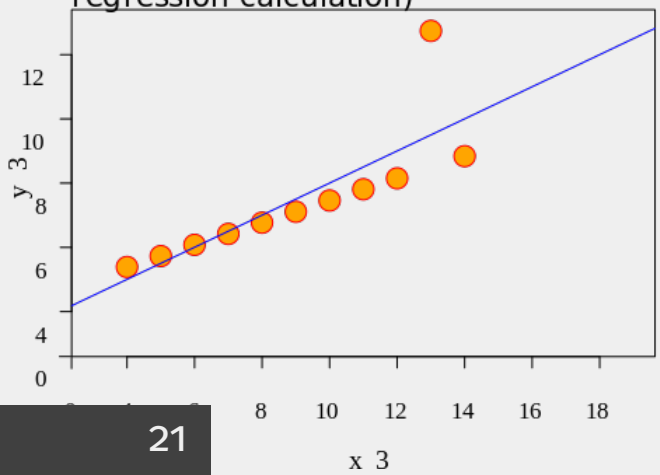
"what most people would see in their mind's eye [for a linear relationship with some unexplained variation]"



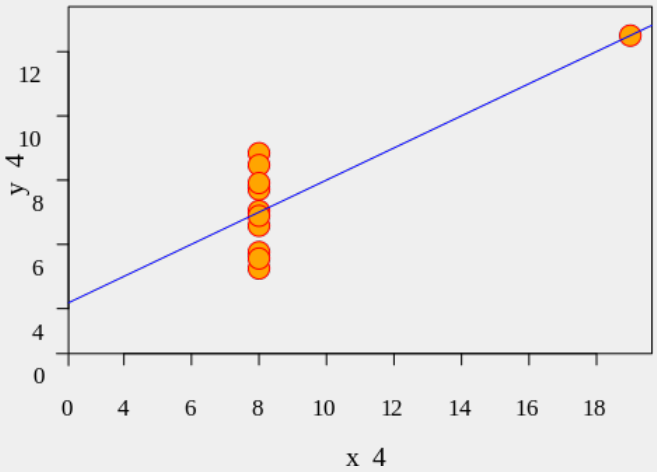
"y has a smooth curved relation with x, possibly quadratic, and there is little residual variability"



"all but one of the observations lie close to a straight line (not the one yielded by the standard regression calculation)"



"all the information about the slope of the regression line resides in one observation"



Correlation is not Causation

- What it really means is that a correlation does not prove one thing causes the other:
  - One thing might cause the other
  - The other might cause the first to happen
  - They may be linked by a different thing
  - Or it could be random chance!

Correlation is not causation

- Any correlation is potentially causal
  - X might cause Y
  - Y might cause X
  - X and Y might be caused by Z
  - X and Y might cause Z
  - There may be no causal relationship

Spurious Correlations

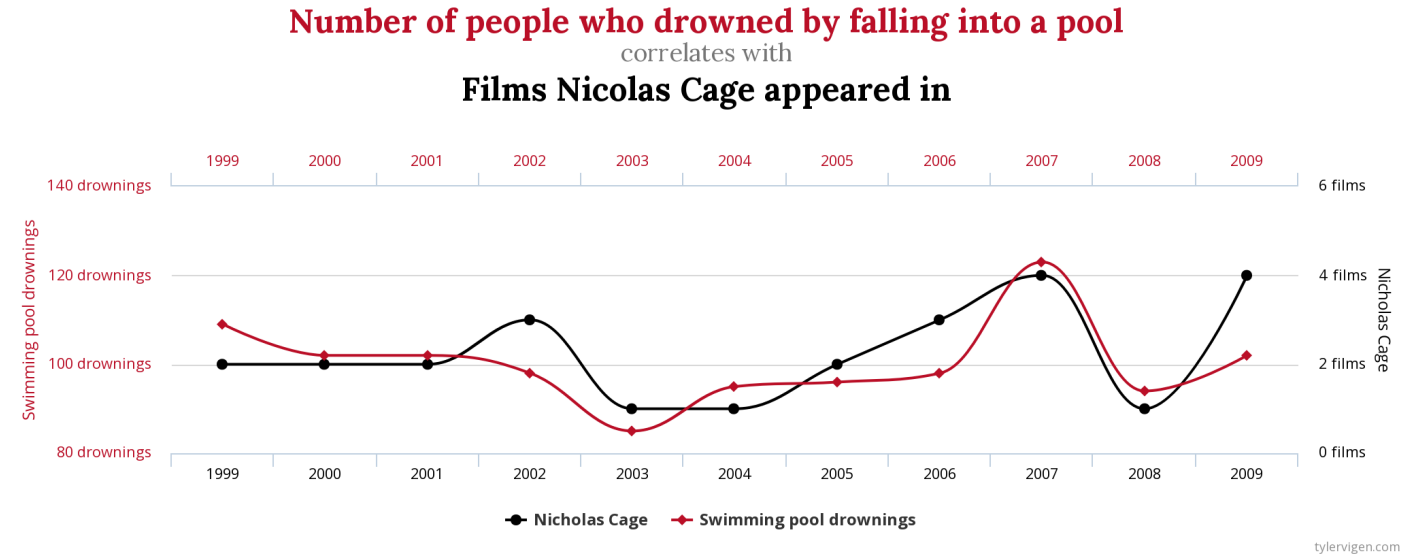


Figure 6: Correlation is NOT Causation

Naive Causal Inference

- Correlations are not necessarily causal
- Our mind thinks they are because humans are not very good at the kind of causal inference problems that social scientists care about
- Instead, we're good at understanding **physical causality**

Physical Causality

- Action and reaction
- Example:
  - Picture a ball resting on top of a hill
  - What happens if I push the ball?
- Features:
  - Observable
  - Single-case
  - Deterministic
  - Monocausal

Pre-Post Change Heuristic

- Our intuition about causation relies too heavily on simple comparisons of pre-post change in outcomes before and after something happens
- Why can this be wrong?

Flaws in causal inference from pre-post comparisons

- Maturation or trends
- Regression to the mean
- Selection
- Simultaneous historical changes
- Instrumentation changes
- Monitoring changes behaviour

Maturation or trends

- Is a shift in an outcome before and after a policy change the impact of the policy or a small part of a longer time trend?
- Example:

Regression to the mean

- Is a shift in an outcome before and after a policy change the impact of the policy or a function of statistical variation?
- Example:

Selection

- Is a shift in an outcome before and after a policy the impact of the policy or the result of the policy being implemented when outcomes are extreme?

Simultaneous changes

- Is the shift in an outcome before and after a policy the impact of the policy or the result of a simultaneous historical shift?

Instrumentation changes

- Is the shift in an outcome before and after a policy the impact of the policy or a change in how the outcome is measured?

Monitoring changes behaviour

- Is the shift in an outcome before and after a policy the impact of the policy or a change in response to measuring the outcome per se?

Examples

- Age and conservatism
- GDP and democracy
- Personality traits and political ideologies
- Healthcare spending and happiness

In R...

- `mean()`
- `median()`, `min()`, `max()`, `quantile()`
- `var()`
- `sd()`
- `cov()`
- `cor()`

Studying Feelings Toward Democracy

- 2018 Pew Study
- 27 countries, ~ 30,000 respondents
- **Question**: How satisfied are you with the way democracy is working in our country – very satisfied, somewhat satisfied, not too satisfied, or not at all satisfied?
  1. very satisfied
  2. somewhat satisfied
  3. not too satisfied
  4. not at all satisfied
  5. don't know
  6. refused

- Load the data:

```
library(haven) # package to read the data
library(dplyr) # package for data manipulation
library(tidyverse) # package for 'tidy' data
pew2018 <- read_sav("Pew 2018.sav",
                    user_na=TRUE) %>%
                    as_factor()
pew2018 = pew2018 %>% select(COUNTRY, satisfied_democracy, age, sex, d_ptyid_us)
pew2018$partyid2 <- fct_collapse(pew2018$d_ptyid_us,
                         DK = c("No preference (DO NOT READ)",
                                "Other party (DO NOT READ)",
                                "Don't know (DO NOT READ)",
                                "Refused (DO NOT READ)"),
                         Rep = "Republican",
                         Ind = "Independent",
                         Dem = "Democrat")
```

Glimpsing at data

- `dim()`: Retrieve the dimension
- `names()`: Get the names
- `str()`: Display compactly the internal structure
- `glimpse()`: is the dplyr-version of str() showing values of each variable the whole sceen width, but does not display the number of levels and names of factor variables. But this feature of str() cannot be displayed completly with either many or long levels names.
- `View()`: With RStudio you can see and inspect the data set comfortably. The View() function invokes a spreadsheet-style data viewer.

Glimpsing at data

```r
dim(pew2018) # Dimensions
```

```
## [1] 30109     6
```

```r
names(pew2018) # Column names
```

```
## [1] "COUNTRY"             "satisfied_democracy" "age"
## [4] "sex"                 "d_ptyid_us"          "partyid2"
```

```r
glimpse(pew2018) # Structure of data
```

```
## Rows: 30,109
## Columns: 6
## $ COUNTRY             <fct> United States, United States, United States, Un...
## $ satisfied_democracy <fct> Not at all satisfied, Not too satisfied, Not to...
## $ age                 <fct> 60, 69, 71, 82, 46, 57, 64, 81, 84, Refused (DO...
## $ sex                 <fct> Male, Male, Male, Female, Male, Female, Female,...
## $ d_ptyid_us          <fct> Democrat, Independent, Democrat, Republican, De...
## $ partyid2            <fct> Dem, Ind, Dem, Rep, Dem, Dem, Dem, Dem, DK, DK,...
```

Contingency table

- The _table()_ function shows us how many respondents are in each category of a categorical variable:

```
table(pew2018$satisfied_democracy)
```

```
##
##          Very satisfied       Somewhat satisfied        Not too satisfied
##                    3403                    10402                     8801
##     Not at all satisfied Don't know (DO NOT READ)    Refused (DO NOT READ)
##                    6785                      662                       56
```

- We can use prop.table() to show what proportions of the data each response represents:

```
prop.table(table(pew2018$satisfied_democracy))
```

```
##
##          Very satisfied       Somewhat satisfied        Not too satisfied
##             0.113022684             0.345478096             0.292304627
##     Not at all satisfied Don't know (DO NOT READ)    Refused (DO NOT READ)
##             0.225347903             0.021986781             0.001859909
```

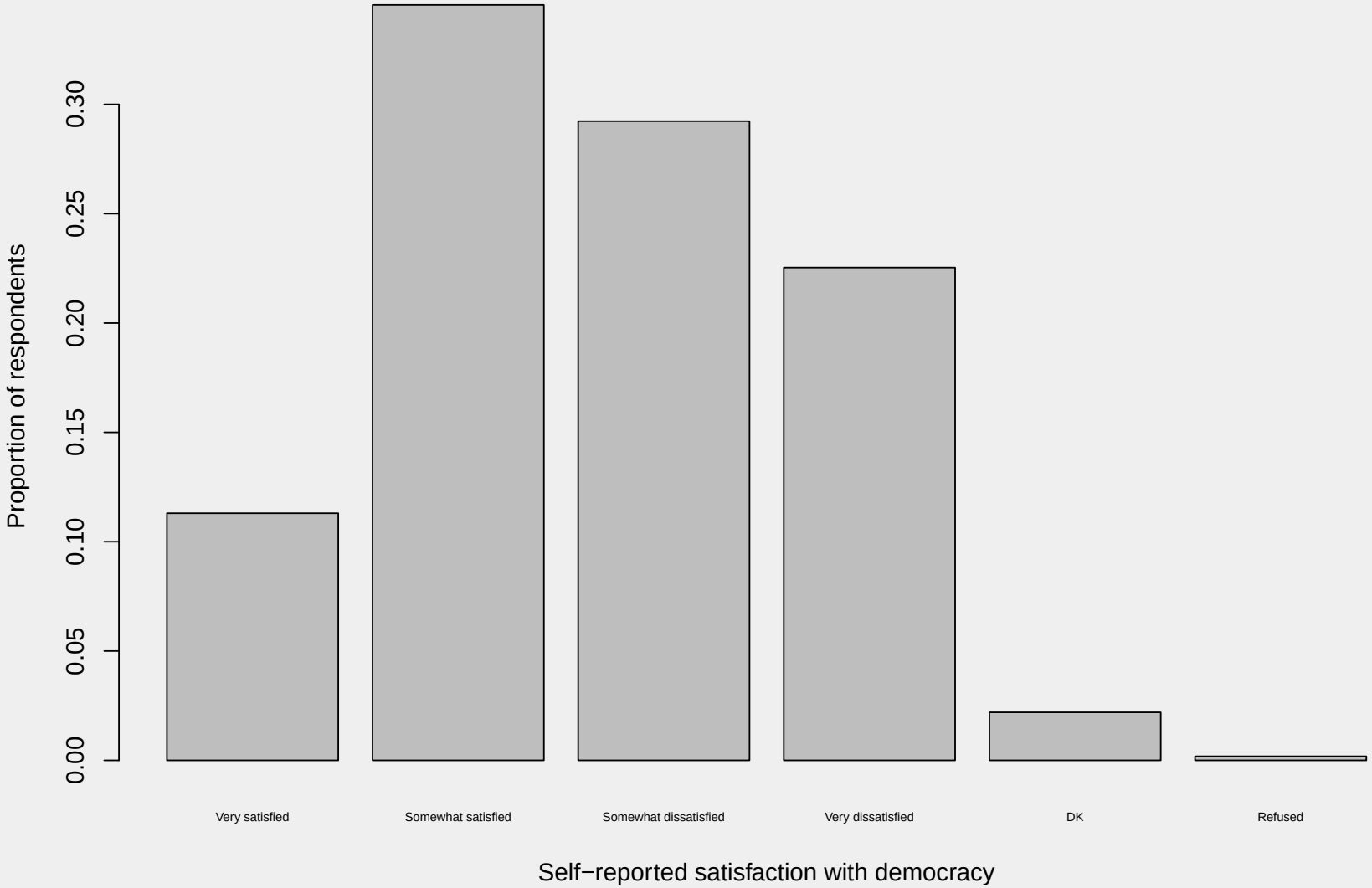
Barplot

- The `barplot()` function can help us visualize a contingency table:

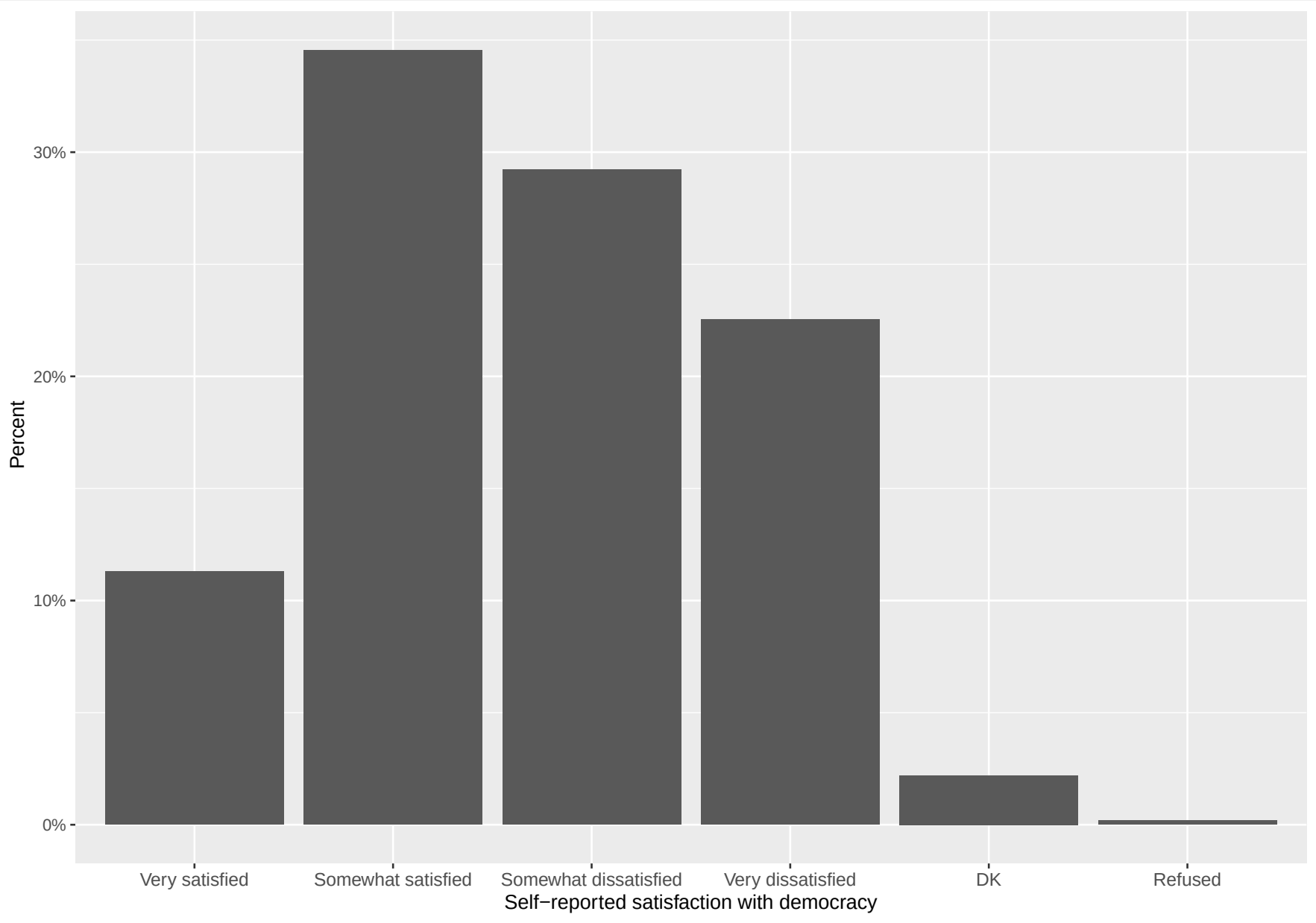
```
barplot(prop.table(table(pew2018$satisfied_democracy)),
xlab = "Self-reported satisfaction with democracy",
ylab = "Proportion of respondents")
```

- Arguments:
  - First is the height each bar should take (we're using proportions in this case)
  - names are the labels for the each category
  - xlab, ylab are axis labels

Barplot


Proportion of respondents
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Very satisfied
Somewhat satisfied
Somewhat dissatisfied
Very dissatisfied
DK
Refused
Self−reported satisfaction with democracy

Bar plot (Using ggplot)



Histogram

- Visualize density of continuous/numeric variable.
- How to create a histogram by hand:
  1. create bins along the variable of interest
  2. count number of observations in each bin
  3. **density** = bin height
- In R, we use `hist()` with `freq = FALSE`:

```
hist(as.numeric(pew2018$age), freq = FALSE, ylim = c(0, 0
xlab = "Age", main = "Distribution of Respondent's Age")
```

- Other arguments:
  - `ylim` sets the range of the y-axis to show (if you don't set it, uses the range of the data).
  - `main` sets the title for the figure.

Histogram


Distribution of Respondent's Age
Density
0.00
0.01
0.02
0.03
0.04
0
20
40
60
80
Age

What is Density

- The areas of the blocks = proportion of observations in those blocks.
- area of the blocks sum to 1 (100%)
- Can lead to confusion: height of block can go above 1!

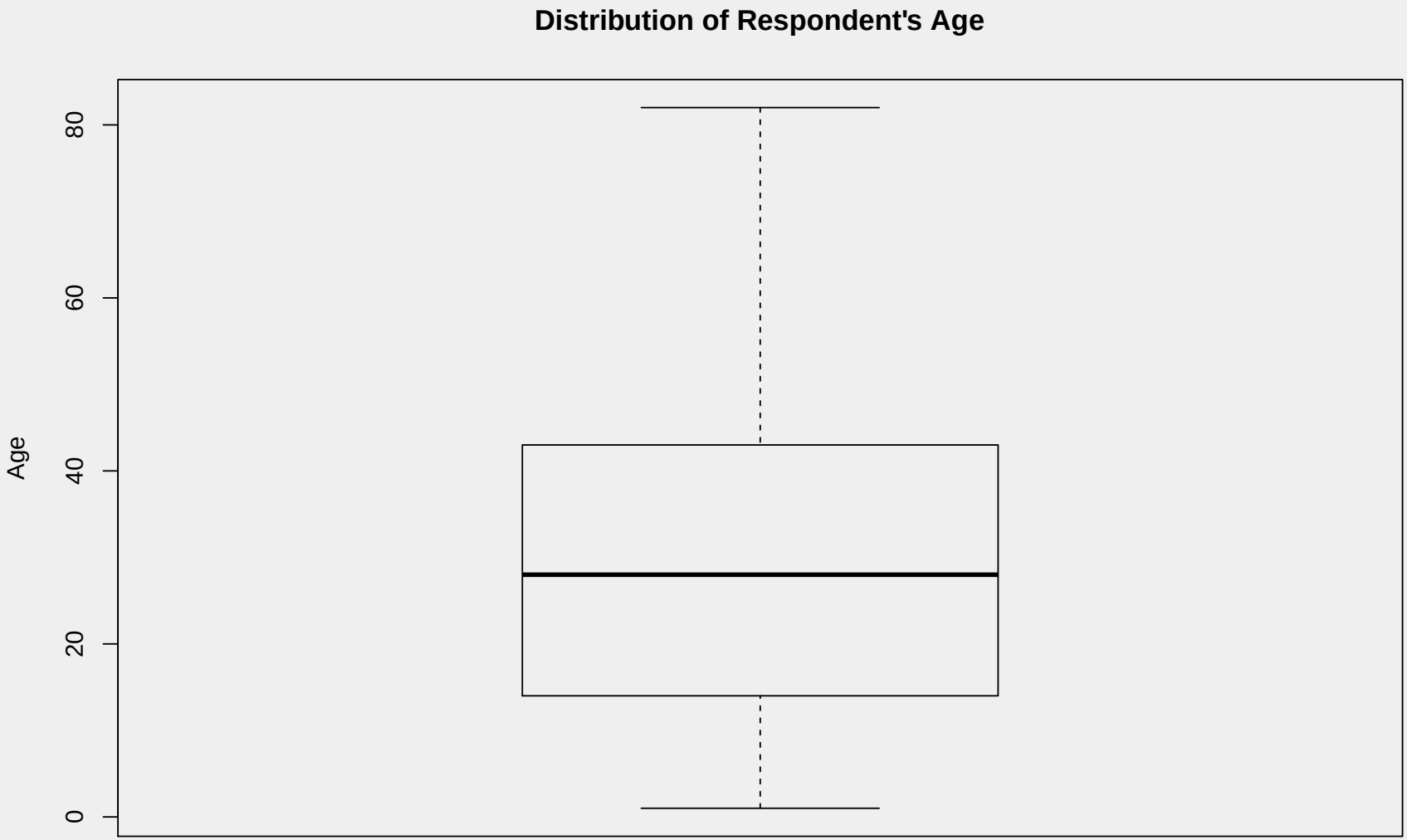
Boxplot

- A boxplot can characterize the distribution of continuous variables
- Use `boxplot()`:

```
boxplot(as.numeric(pew2018$age),
ylab = "Age",
main = "Distribution of Respondent's Age")
```

- "Box" represents range between lower and upper quartile.
- "Whiskers" represents either:
  - 1.5 × IQR or max/min of the data, whichever is tinyer.
  - Points beyond whiskers are outliers

Boxplot



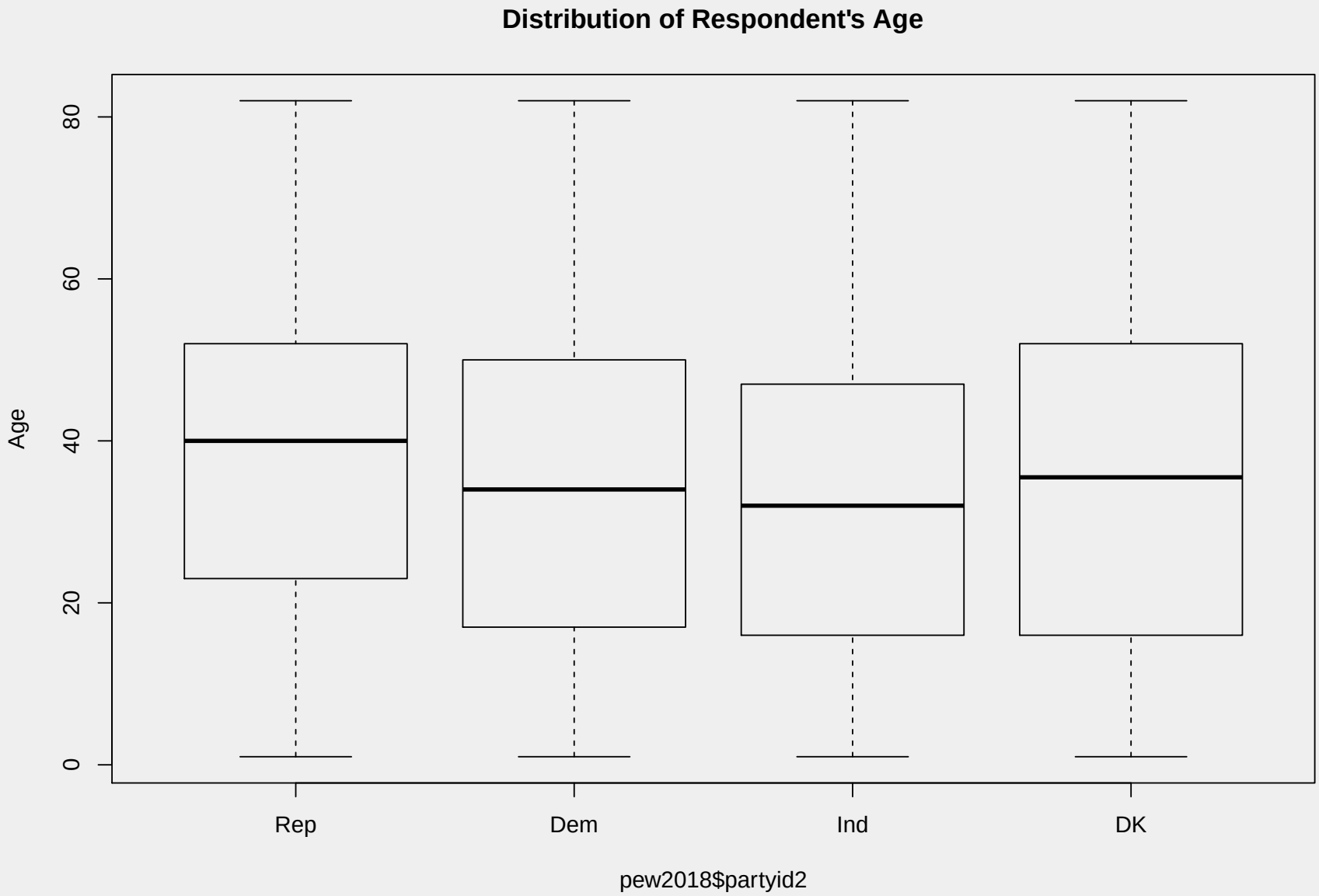
Comparing Distributions with the boxplot

- Useful for comparing a variable across groups:

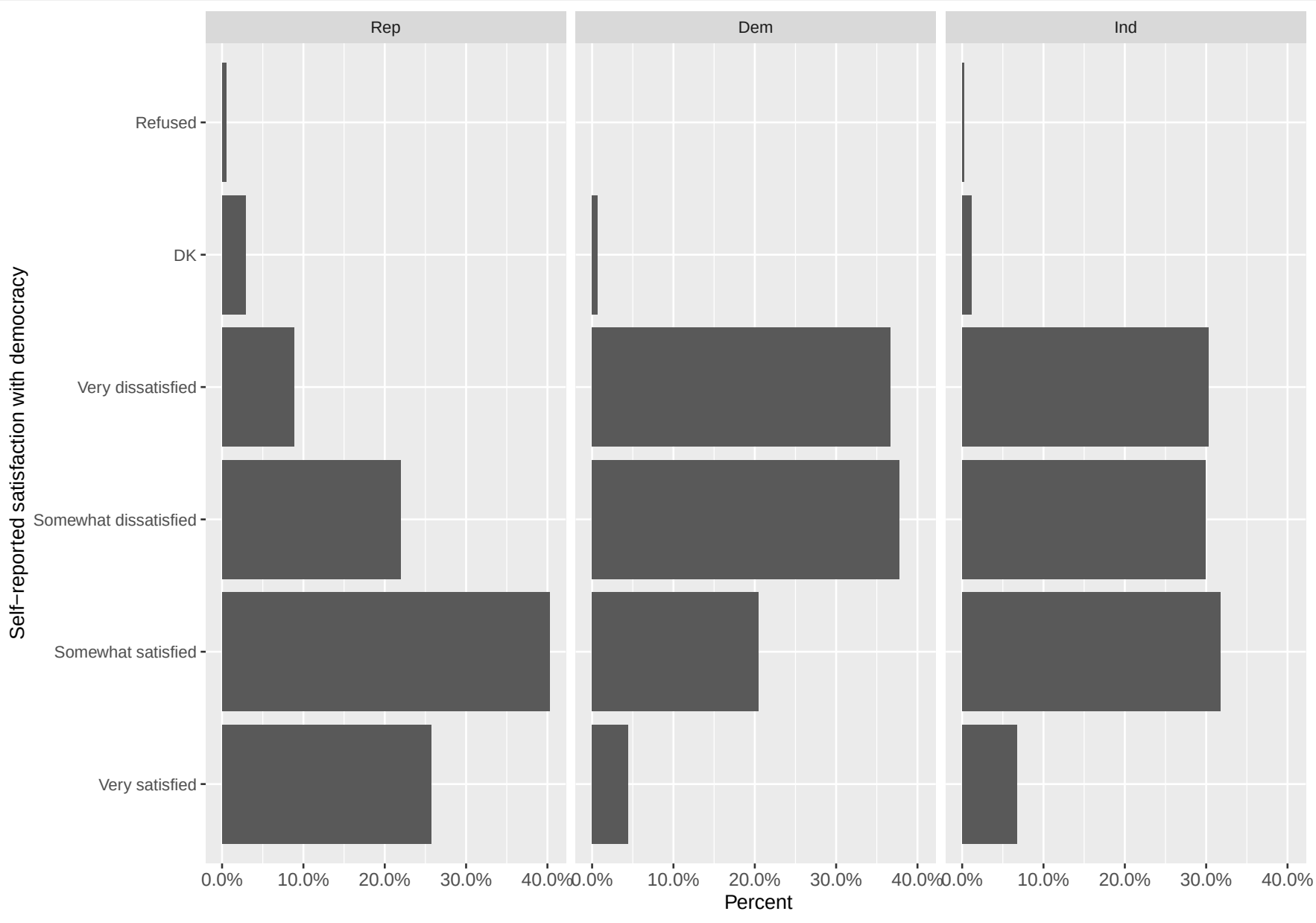
```
boxplot(as.numeric(pew2018$age) ~ pew2018$sex,
ylab = "Age",
main = "Distribution of Respondent's Age")
```

- First argument is called a formula, `y ~ x`:
  - `y` is the continuous variable whose distribution we want to explore.
  - `x` is the grouping variable.
  - When using a formula, we need to add a `data` argument

Comparing Distributions with the boxplot



Satisfaction with democracy by party



Satisfaction with democracy by party

- Why are Republicans more satified with democracy?

Correlations in R

- Use the `cor( )` function
- Missing values: set the use = ”pairwise” ~ available case analysis

```
# Read Happiness Data
happ2019 = read.csv("C:/Users/afisher/Documents/R Code/Resources/Data/Happiness/2019.csv")

# Structure of dataset
str(happ2019)
```

```
## 'data.frame':    156 obs. of  9 variables:
##  $ Overall.rank                : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Country.or.region           : Factor w/ 156 levels "Afghanistan",..: 44 37 106 58 99 134 
##  $ Score                       : num  7.77 7.6 7.55 7.49 7.49 ...
##  $ GDP.per.capita              : num  1.34 1.38 1.49 1.38 1.4 ...
##  $ Social.support              : num  1.59 1.57 1.58 1.62 1.52 ...
##  $ Healthy.life.expectancy     : num  0.986 0.996 1.028 1.026 0.999 ...
##  $ Freedom.to.make.life.choices: num  0.596 0.592 0.603 0.591 0.557 0.572 0.574 0.585 0.584 
##  $ Generosity                  : num  0.153 0.252 0.271 0.354 0.322 0.263 0.267 0.33 0.285 0
##  $ Perceptions.of.corruption   : num  0.393 0.41 0.341 0.118 0.298 0.343 0.373 0.38 0.308 0.
```

```
# Correlation
cor(happ2019$Score, happ2019$GDP.per.capita)
```

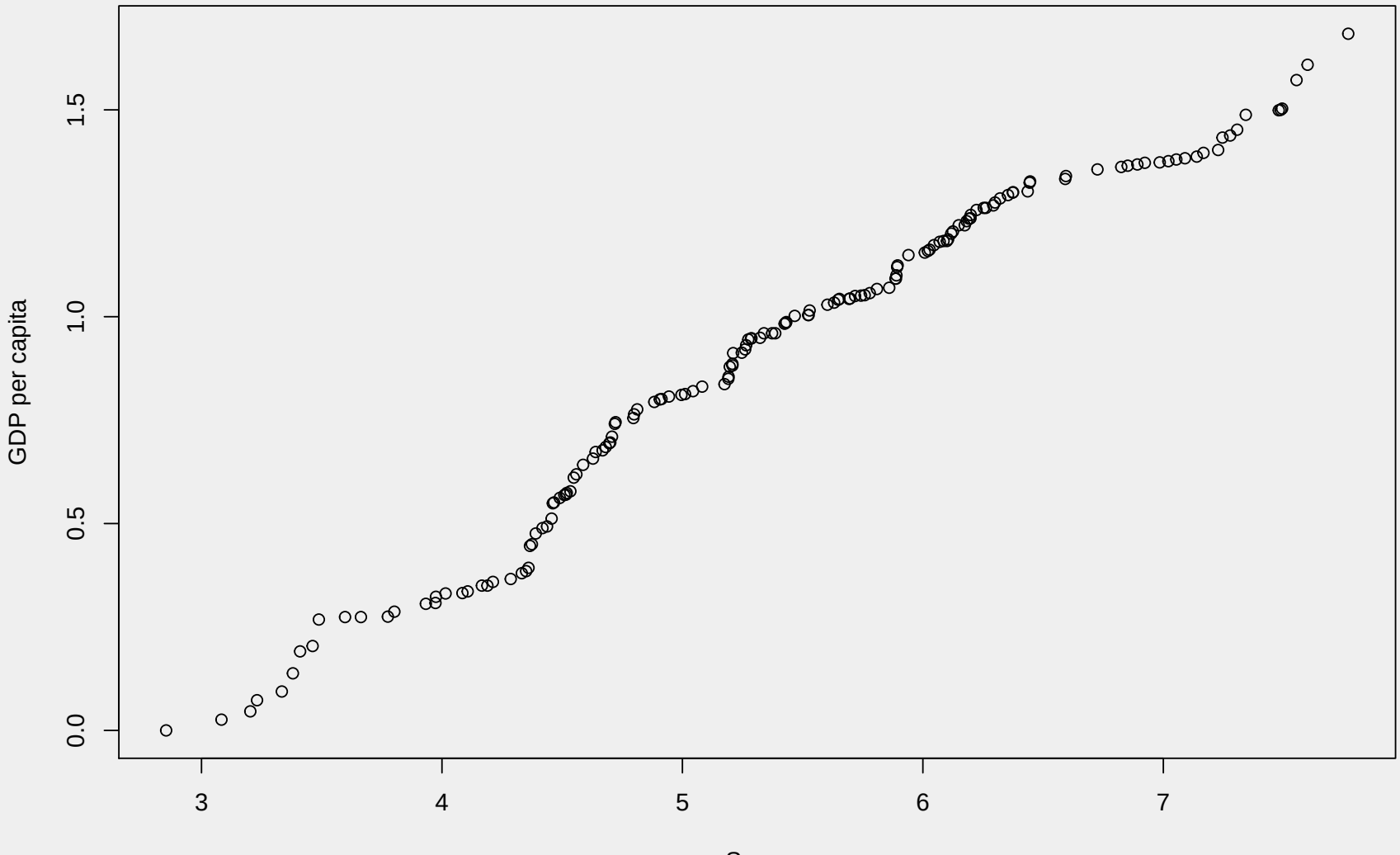
```
## [1] 0.7938829
```

QQ-plot example

- **Quantile-quantile plot (qq-plot)**: Plot the **quantiles** of each distribution against each other.
- Example points:
  - (min of X, min of Y)
  - (median of X, median of Y)
  - (25th percentile of X, 25th percentile of Y)
- 45 degree line indicates quality of the two distributions

QQ-PLOT EXAMPLE

```
qqplot(happ2019$Score, happ2019$GDP.per.capita, xlab = 'Score', ylab="GDP per capita")
```

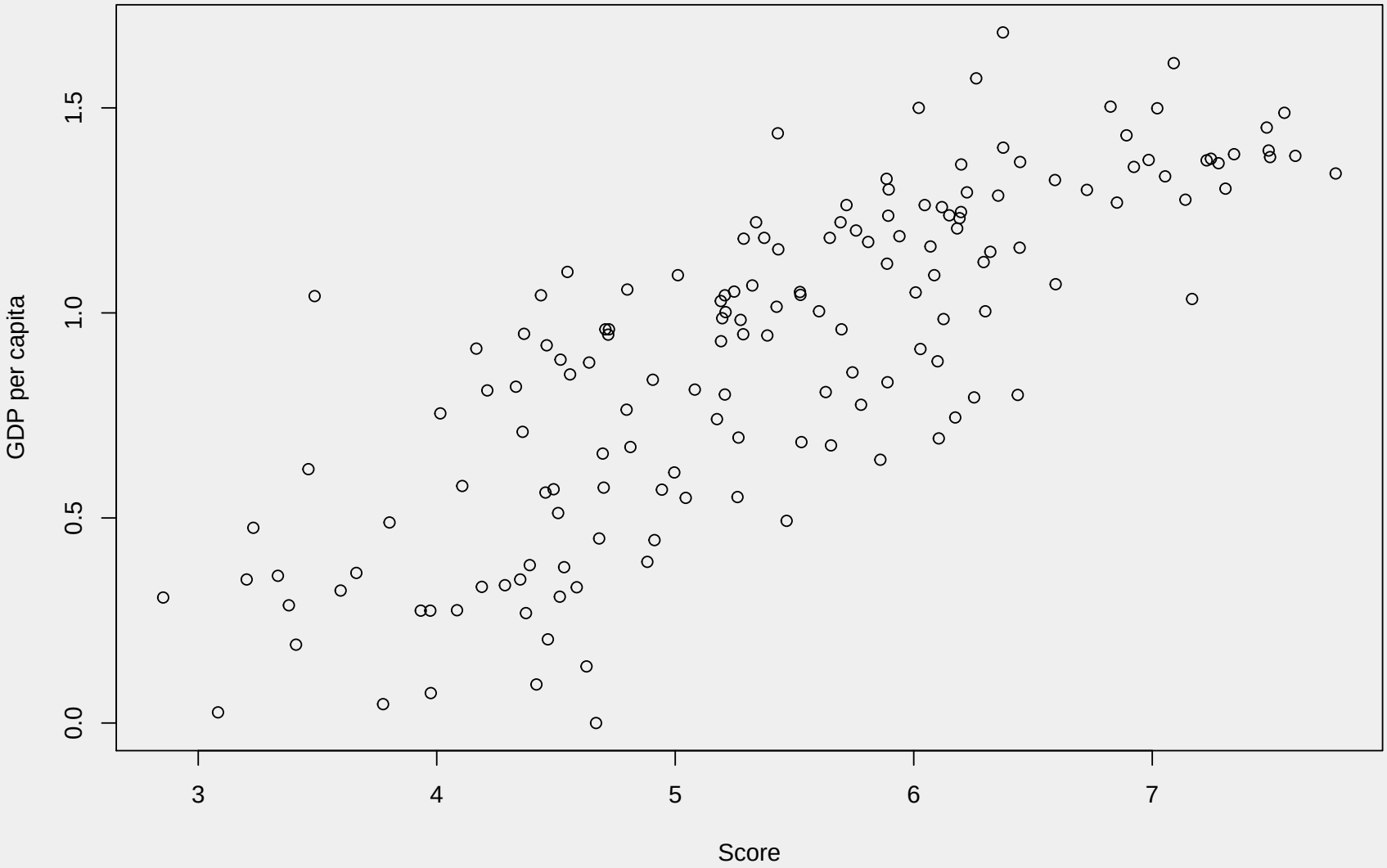


Scatterplot

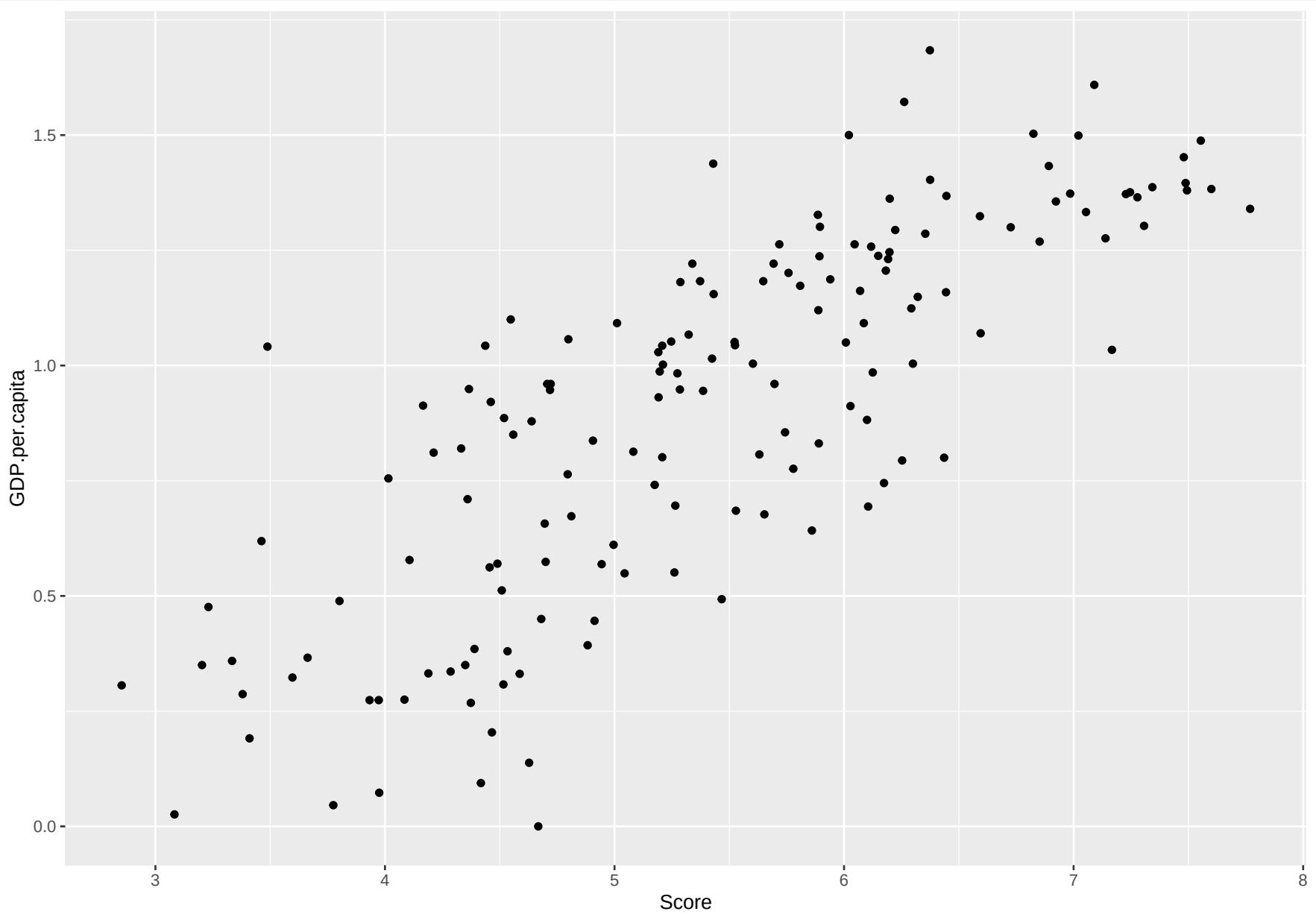
```r
## Base R
plot(happ2019$Score, happ2019$GDP.per.capita, xlab = 'Score', yl

## ggplot
ggplot(happ2019) +
  aes(x=Score, y=GDP.per.capita)+
  geom_point()+
  theme_classic()
```

Base R - Scatterplot



ggplot - Scatterplot



Tables and Missing Data in R

Tables in R

```
# Read File
afghan <- read.csv("C:/Users/afisher/Documents/R Code/qss/MEASUR

# Column names
names(afghan)
```

```
##  [1] "province"              "district"              "village.id"
##  [4] "age"                   "educ.years"            "employed"
##  [7] "income"                "violent.exp.ISAF"      "violent.exp
## [10] "list.group"            "list.response"
```

```
# Tables
table(ISAF = afghan$violent.exp.ISAF,
      Taliban = afghan$violent.exp.taliban)
```

```
##     Taliban
## ISAF    0    1
##    0 1330  354
##    1  475  526
```

Table in R: prop.table()

We can also use prop.table() to see the proportion of cases in each cell

We have to include table() within parentheses too:

```
prop.table(table(ISAF = afghan$violent.exp.ISAF,
                 Taliban = afghan$violent.exp.taliban))
```

```
##     Taliban
## ISAF         0         1
##    0 0.4953445 0.1318436
##    1 0.1769088 0.1959032
```

Round Function

- Since we're already using nested functions, we can also use round() to round the values in each cell
- Notice the , 2 in the code below. It indicates that we will round the numbers up to two significant digits

```
round(prop.table(table(ISAF = afghan$violent.exp.ISAF,
                       Taliban = afghan$violent.exp.taliban)), 2
```

```
##      Taliban
## ISAF    0    1
##    0 0.50 0.13
##    1 0.18 0.20
```

Missing Data

- Not all individuals answer to surveys
- Two types of non-response:
  - Individual non-response
  - Item non-response
- Both tend to bias the results
- So it is very important that we know where (and think about why) we see gaps in our data

Missing Data in R

- R has a special code for missing data, NA
- Since NA is only used for missing observations, we can count their numbers with `is.na()`

```
head(afghan$income, 10)
```

```
##  [1] 2,001-10,000  2,001-10,000  2,001-10,000  2,001-10,000
##  [6] <NA>          10,001-20,000 2,001-10,000  2,001-10,000
## 5 Levels: 10,001-20,000 2,001-10,000 20,001-30,000 ... over 3
```

```
# number of missings
sum(is.na(afghan$income))
```

```
## [1] 154
```

```
# proportion of missings
round(mean(is.na(afghan$income)), 2)
```

```
## [1] 0.06
```

Missing Data

- Some R function don't work if there's missing data
- We add `na.rm = TRUE` to the code

```
# Victims of Taliban violence
sum(is.na(afghan$violent.exp.taliban))
```

```
## [1] 54
```

```
# Mean violance by taliban
mean(afghan$violent.exp.taliban)
```

```
## [1] NA
```

```
# Mean violance by taliban
round(mean(is.na(afghan$income)), 2)
```

```
## [1] 0.06
```

- Why do you think we have missing data here?

Missing Data in R

- You can also visualise the number of missing observations with `summary()`

```
summary(afghan$violent.exp.taliban)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.3289  1.0000  1.0000
```

```
sum(is.na(afghan$violent.exp.taliban))
```

```
## [1] 54
```

Data Visualization

Bar Plots

- Bar plots are used to visualise **factor/character variables**
- Proportion of observations in each category as the height of each bar
- Options:
  - `main = "Title"`
  - `xlab = "X label"`
  - `ylab = "Y label"`
  - `xlim = c(number, number)` limits for the x variable
  - `ylim = c(number, number)` limits for the y variable
  - `names.arg = c("Bars labels")` - in the same order of the variable
  - `horiz = TRUE` for horizontal plots
  - `cols` = “colour name” bar colour (see:)
- You can use `barplot()` with `prop.table()` instead of pie charts

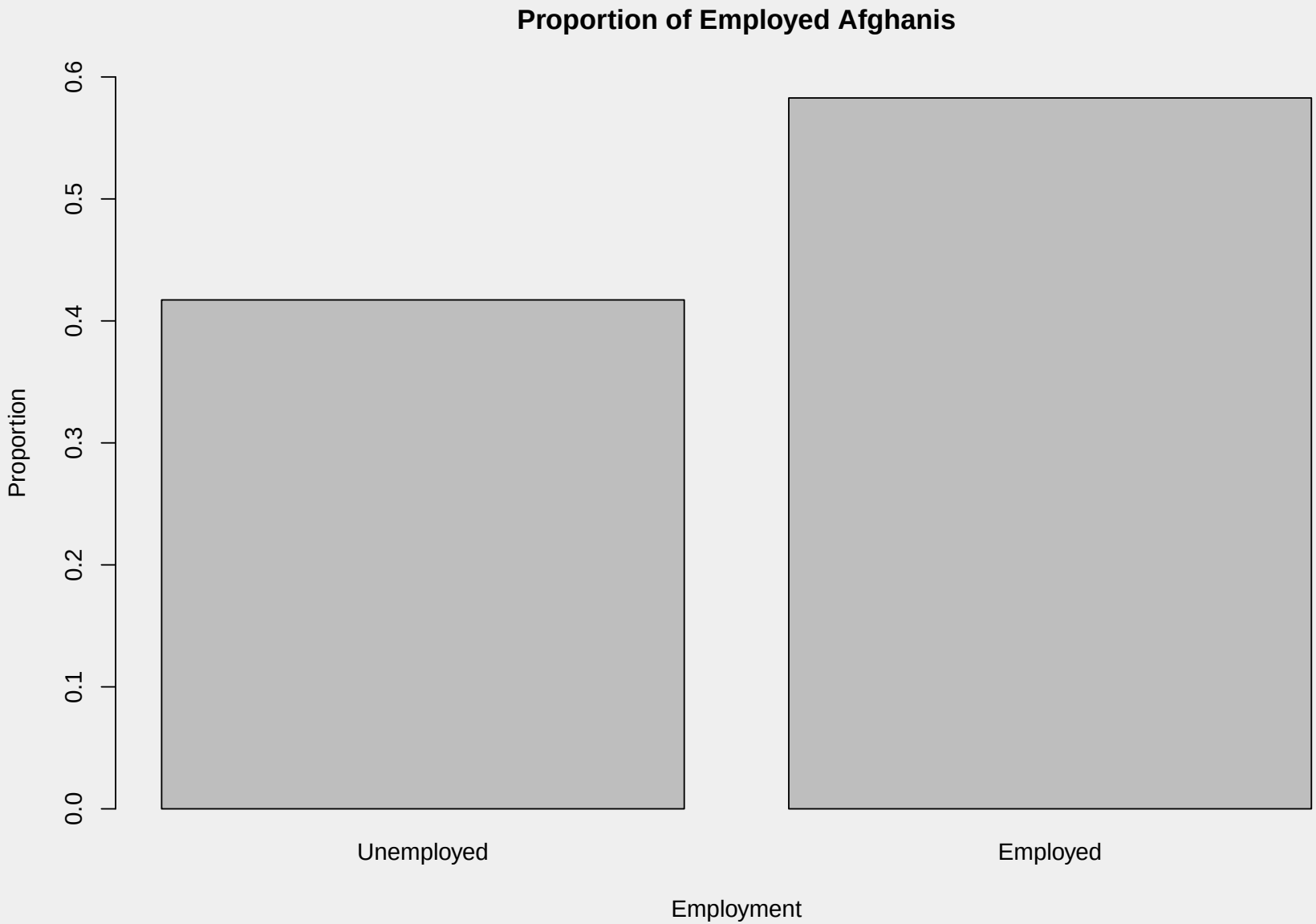
Barplots

```
employed.ptable <- prop.table(table(afghan$employed))
employed.ptable
```

```
        0         1
0.4172113 0.5827887
```

```
barplot(employed.ptable,
        names.arg = c("Unemployed", "Employed"),
        main = "Proportion of Employed Afghanis",
        xlab = "Employment",
        ylab = "Proportion",
        ylim = c(0, 0.6))
```

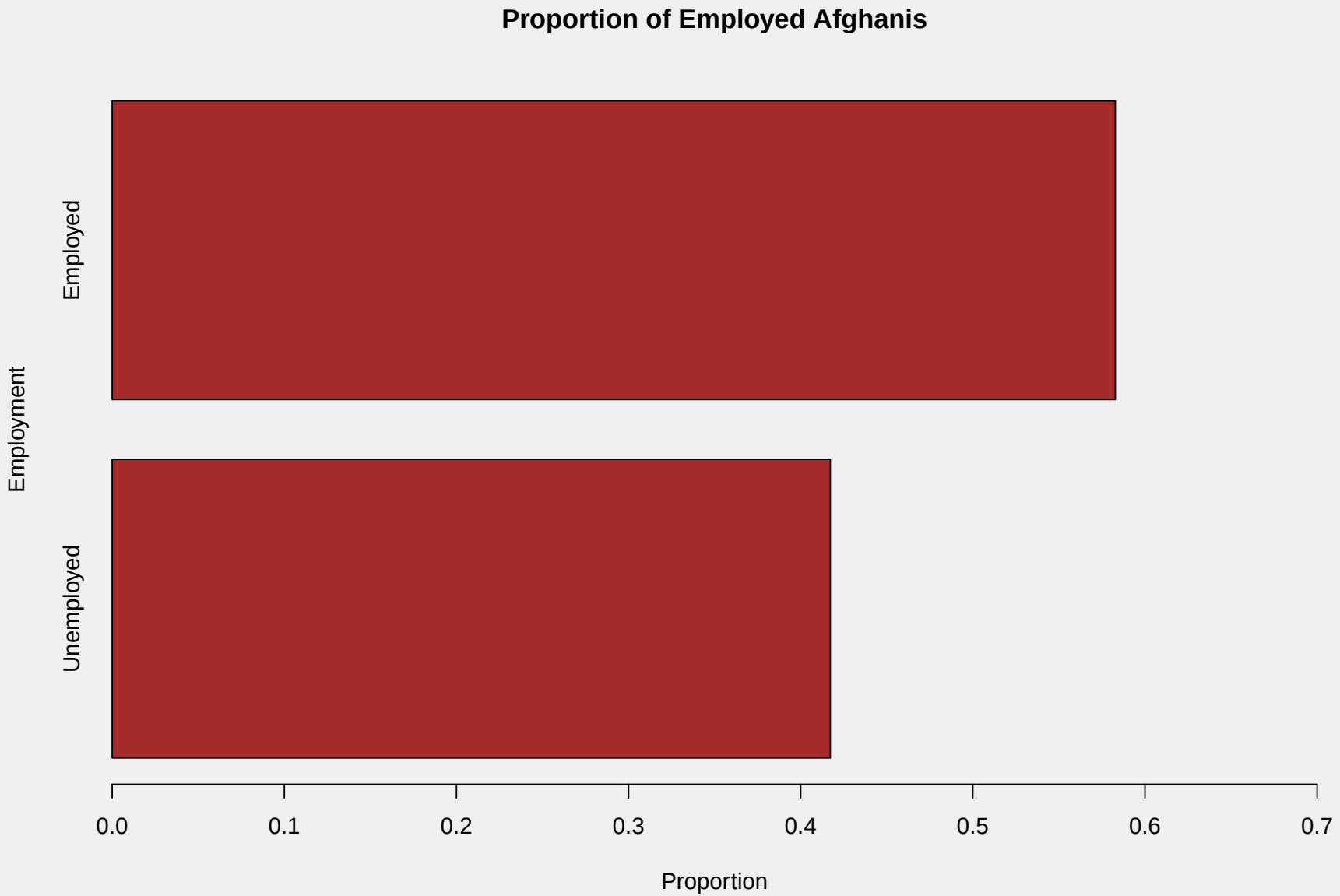
Barplots


Proportion of Employed Afghanis
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Unemployed
Employed
Employment

Horizontal Bar plot

```r
barplot(employed.ptable,
        names.arg = c("Unemployed", "Employed"), # 0 and 1, resp
        main = "Proportion of Employed Afghanis",
        ylab = "Employment", # change the axes
        xlab = "Proportion",
        xlim = c(0, 0.7), # now it's xlim
        horiz = TRUE,     # because the plot is horizontal
        col = "brown")
```

Horizontal Bar plot



Histogram

- We use histograms to display the distribution of a numeric variable
- Numeric variables are binned into groups
  - Histograms shows the density of each bin
  - Important: Height is share of observations in bin divided by bin size
- We care less about the density of each bin than about the distribution of the variable as a whole
- Area of each bar is the share of observations that fall into that bin
- Area of all bins sum to one

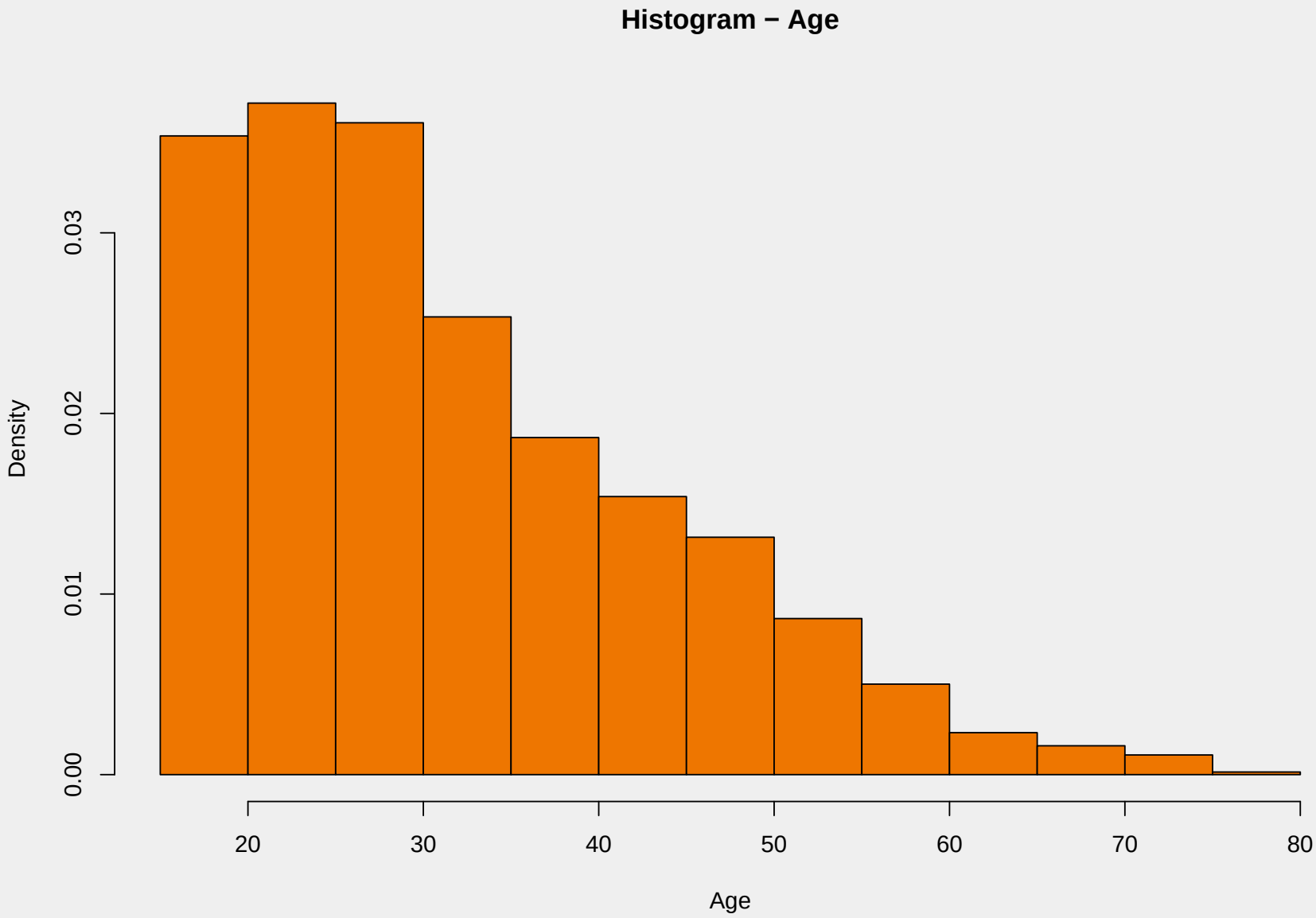
Histogram

- Many options are similar to those of `barplot()`: `main`, `xlab`, `ylim`, `col`
- We can also add `freq = FALSE` to show the density of each histograms
- `breaks` = changes the size of the bins
- Densities are useful to compare different distributions
- *Densities are not percentages*: “percentage per horizontal unit”

Histogram in R

```r
hist(afghan$age,
     main = "Histogram - Age",
     xlab = "Age",
     xlim = c(0, 0.04),
     freq = FALSE,
     col = "darkorange2")
```

Histogram in R

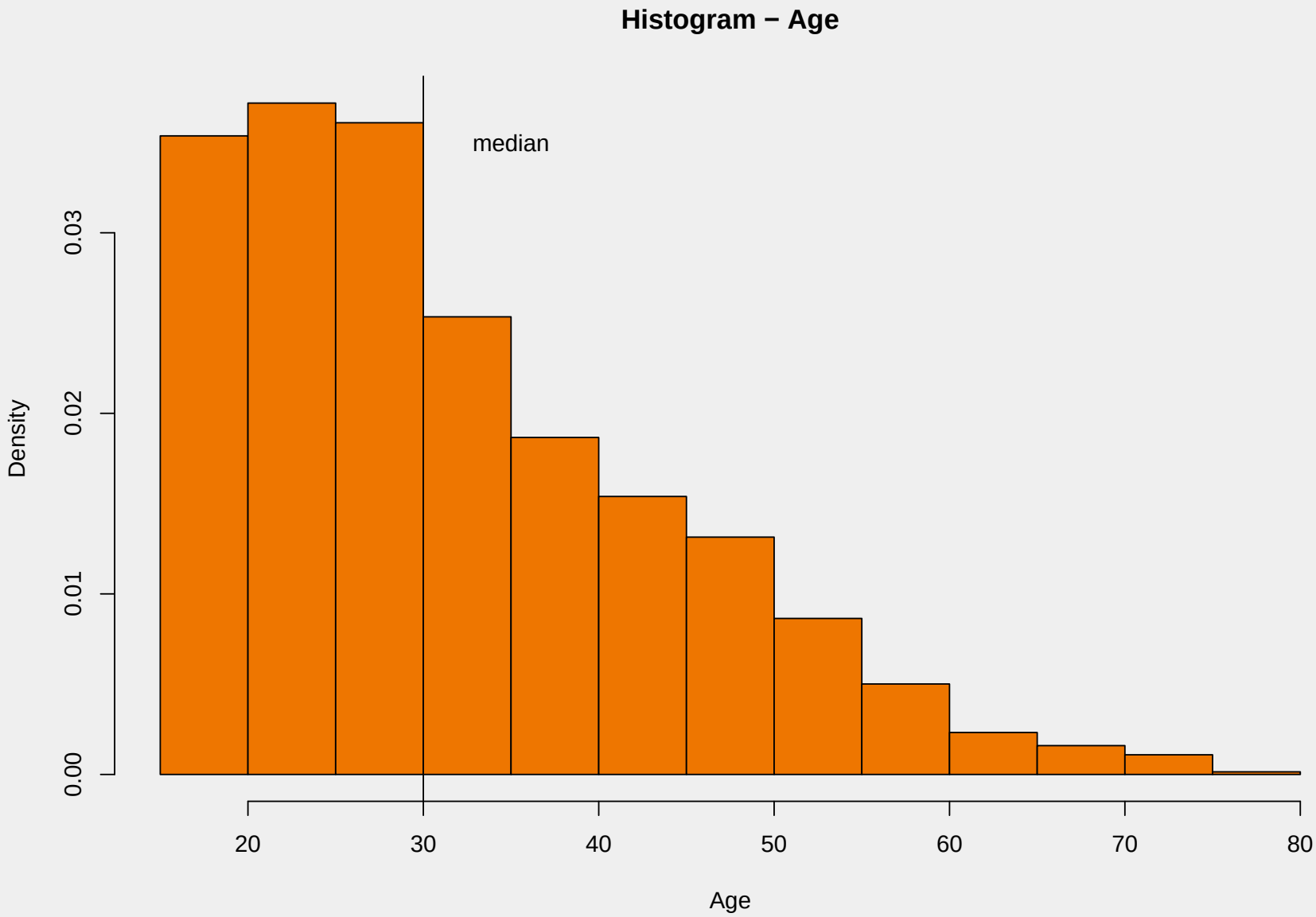


Histogram in R

- We can also add text and fitted lines to all R plots
- Use `text()` and `abline()` after `hist()`

```
hist(afghan$age,
     main = "Histogram - Age",
     xlab = "Age",
     xlim = c(0, 0.04),
     freq = FALSE,
     col = "darkorange2")
## add a text label at (x, y) = (35, 0.35)
text(x = 35, y = 0.035, "median")
## add a vertical line representing median
abline(v = median(afghan$age))
```

Histogram in R



Box Plots

- Like histograms, box plots also display the distribution of a numeric variable
- Box plots show the median, quartiles, and IQR
- Useful to compare different distributions side-by-side
- It is also useful to identify outliers, that is, data points that are above 1.5 times the **interquartile range (IQR)**
- `boxplot()`

Box Plots

- `boxplot()` also has a series of optional arguments:
  - `main`, `ylab`, `ylim`, `col`
  - `formula = y ~ group`, `y` is numeric variable and `group` is a factor

Quantiles

```
median(afghan$age)
```

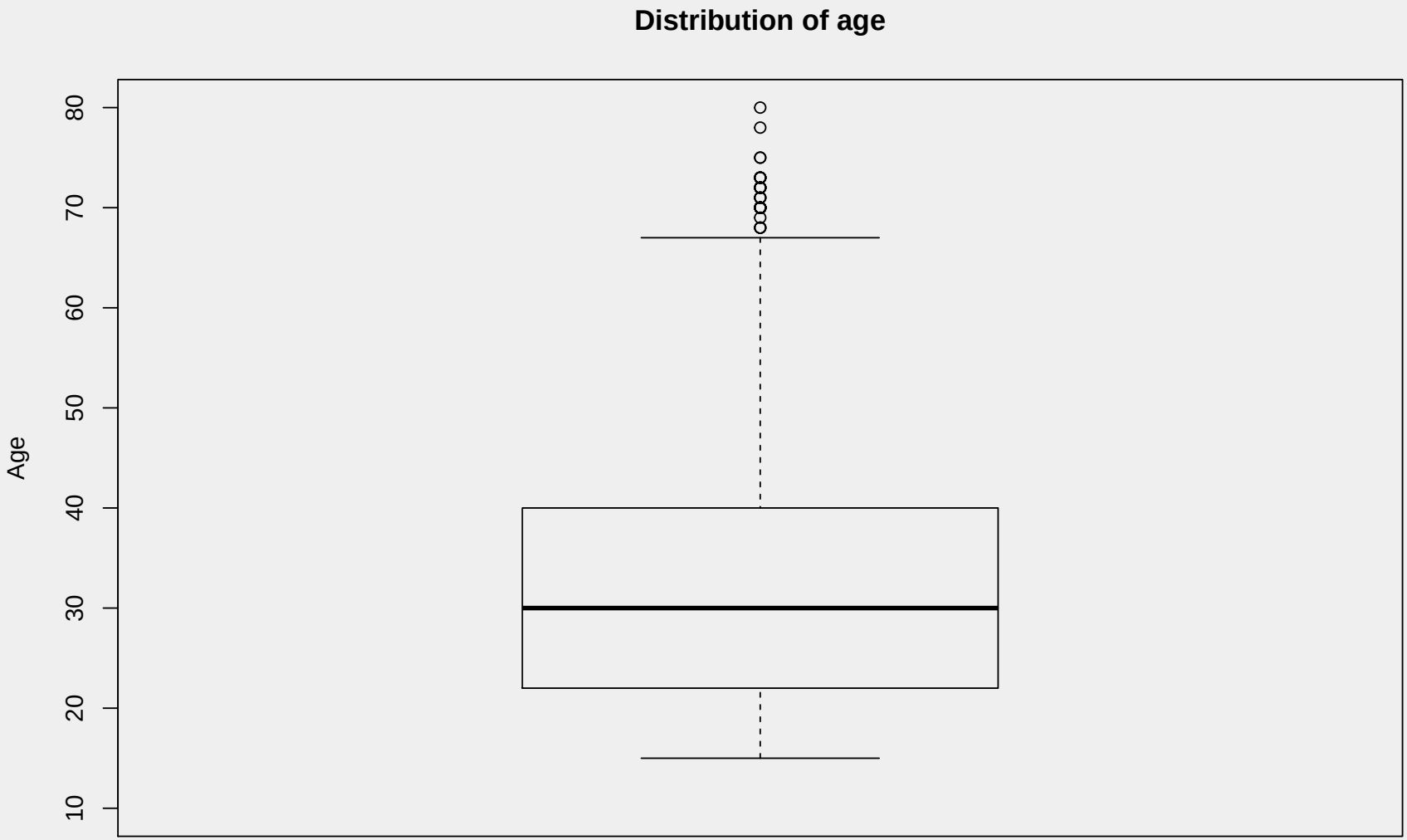
```
## [1] 30
```

```
quantile(afghan$age, probs = c(0, .25, .5, .75, 1))
```

```
##   0%  25%  50%  75% 100%
##   15   22   30   40   80
```

```
boxplot(afghan$age, main = "Distribution of age", ylab = "Age",
```

Box Plots



Box Plot

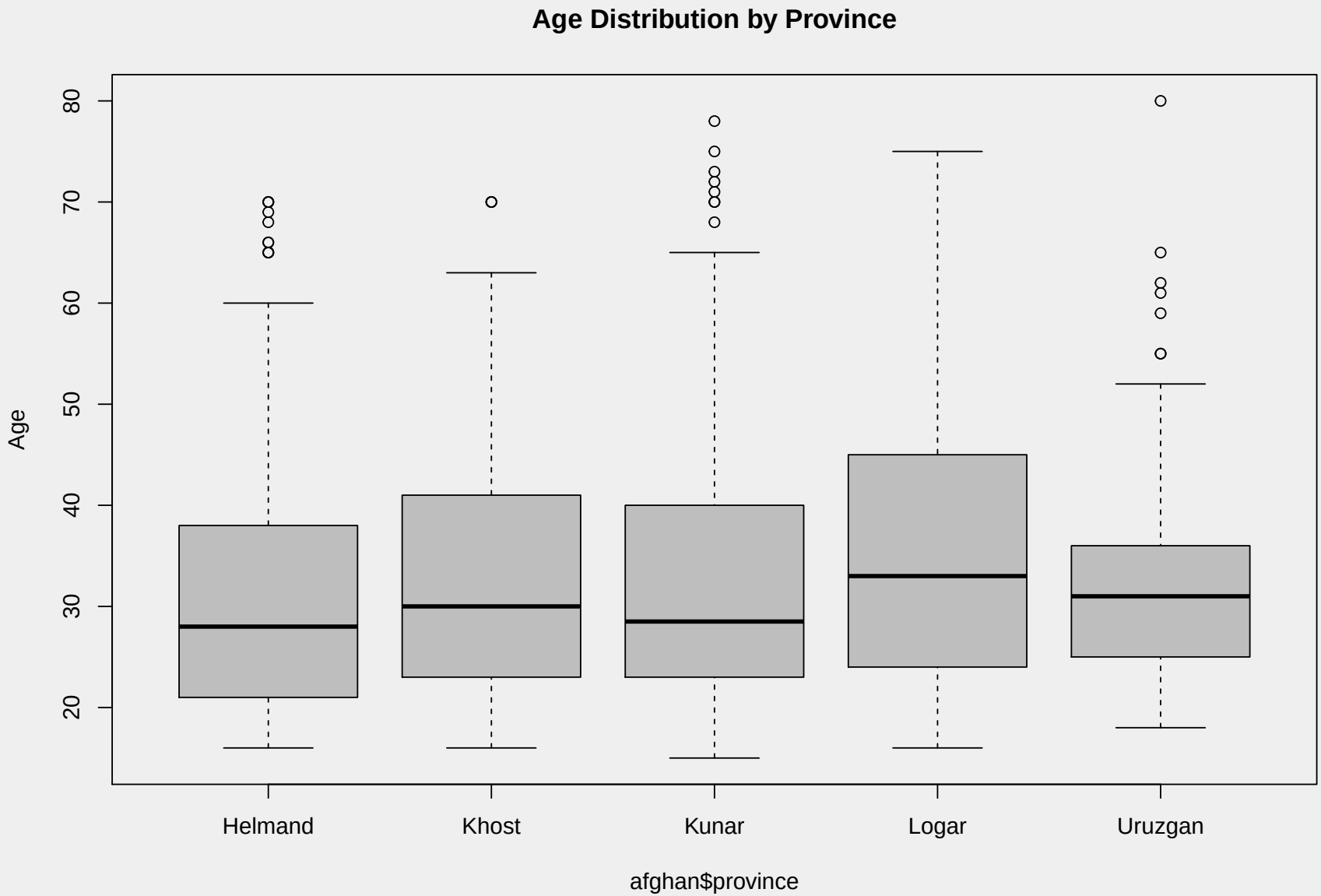
- But box plots provide an easy way to compare multiple observations at a time
- Similar to `tapply()`

```
tapply(afghan$age, afghan$province, median, na.rm = TRUE)
```

```
## Helmand   Khost   Kunar   Logar Uruzgan
##    28.0    30.0    28.5    33.0    31.0
```

```
boxplot(afghan$age ~ afghan$province,
        main = "Age Distribution by Province", ylab = "Age", col
```

Boxplot by group



Survey Sampling

Survey Sampling

		DATES	POLLSTER	SAMPLE	RESULT				NET RESULT
Presidential approval		APR 28-30, 2020	C+ Rasmussen Reports/Pulse Opinion Research	1,500 LV	Approve	45%	53%	Disapprove	Disapprove +8
		APR 27-30, 2020	Global Strategy Group/GBAO	1,008 RV	Approve	45%	53%	Disapprove	Disapprove +8
		APR 26-29, 2020	A/B IBD/TIPP	1,225 A	Approve	44%	44%	Disapprove	EVEN
President: general election		APR 26-29, 2020	A/B IBD/TIPP	948 RV	Biden	43%	43%	Trump	EVEN
	N.C.	APR 27-28, 2020	Meredith College	604 RV	Biden	47%	40%	Trump	Biden +7
	Ga.	APR 25-27, 2020	A/B Cygnal*	591 LV	Biden	44%	45%	Trump	Trump +1

- Most polls only interview several hundred voters
- Goal: infer what 200 million voters are thinking

The 1936 Literary Digest Poll

- Mail questionnaire to 10 million people
- Final sample size: over 2.3 million returned
- Addresses came from phone books and club memberships
- The young Gallup used 50,000 respondents
- Predicted FDR would get 43% of the vote… actually recieved 62%

WTF went wrong?

- Biased sample:
  - slanted toward middle- and upper-class voters, and by default to exclude lower-income voters
  - people who respond to surveys are different from people who don’t
- Two morals of the story:
  - A badly chosen big sample is muc worse than a well-chosen small sample
  - Watch out for selection bias and nonresponse bias.

Quota Sampling vs. Random Sampling

- **Quota sampling:** Sample certain groups until quota is filled
  - Problem: Unobservables can bias the results
- **Random sampling:** Random draws without replacement from the population
  - Everybody has the same chance of being in the sample
  - Problem: none, sample is unbiased (in theory…)!

Random Sampling

- Not every single sample will match all characteristics of the population exactly
- But as the number of samples gets larger (say 1000 samples of 1000 respondents), on average the samples would be representative
- Polls are associated with uncertainty: plus or minus a number
- But getting a random sample is hard

Why is this hard

- Problems of telephone survey
  - Random digit dialing from phone book
  - Wealthy individuals have higher chances of being called
  - Caller ID screening (unit non-response)
- Problems of internet survey
  - Non-probability sampling
  - Cheap but non-representative
  - Young, urban, rich groups are overrepresented
  - Requires statistical corrections (usually weights)

Social Desirability Bias

- Respondents sometimes do not state their true preferences
- Examples: support for drug use, abortion, etc
- Under- or overestimation of true proportion

List Experiments

- List experiments can minimise the problem
- Grant anonymity to respondents
- Control group sees a list of statements
- Treatment group sees the same list plus a sensitive item
- Assuming that respondents don’t lie and that both groups would answer the same number of non-sensitive items, we can infer their true preferences

List Experiments

- https://statmodeling.stat.columbia.edu/2014/04/23/thinking-list-experiment-heres-list-reasons-think/

```
<iframe width="560" height="315" src="http://www.you
```