London Business School

# Session 1: Exploratory Data Analysis

Kostis Christodoulou
London Business School

London Business School

# Course Admin

London Business School

| | | Foundations: EDA and Intro to Data Science | Content | Example | Exercise | Assignment |
|---|---|---|---|---|---|---|
| 1 | 01 Sep | Lecture 1: Exploratory Data Analysis | | | | |
| 2 | 01 Sep | Workshop 1: Import, visualise, and manipulate data | | | | |
| 2 | 02 Sep | Github + portfolio website workshop | | | | |
| | 05 Sep | ★ **Homework 1 Due** | | | | |
| | | **Inferential Statistics** | **Content** | **Example** | **Exercise** | **Assignment** |
| 3 | 06 Sep | Lecture 2: Sampling and Probability Distributions | | | | |
| 4 | 06 Sep | Workshop 2: Confidence Intervals; reshape data | | | | |
| | 07 Sep | ★ **Homework 2 Due** | | | | |
| 5 | 08 Sep | Lecture 3: Hypothesis Testing; *there is only one test* | | | | |
| 6 | 08 Sep | Workshop 3: Hypothesis testing; A/B testing; simulating with `infer` | | | | |
| | 12 Sep | ★ **Homework 3 Due** | | | | |
| | | **Regression Modelling** | **Content** | **Example** | **Exercise** | **Assignment** |
| 7 | 13 Sep | Lecture 4: Introduction to regression models | | | | |
| 8 | 13 Sep | Workshop 4: Workshop on regression | | | | |
| 9 | 15 Sep | Lecture 5: Further regression topics; regression diagnostics | | | | |
| | 21 Sep | ★ **Final project due** | | | | |
| | 21 Sep | ★ **Portfolio website due** | | | | |
| 10 | 22 Sep | Group Case Presentations; course wrap up | | | | |
| | 23 Sep | ★ **Revision Session** | | | | |
| | 24 Sep | ★ **Final Exam** | | | | |

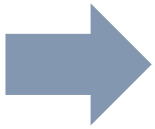
London Business School

# Course Material

- Lecture slides on Canvas
- Textbook ModernDive - https://www.moderndive.com
- Github repo: https://github.com/kostis-christodoulou/am01
- Course website: https://mam2022.netlify.app

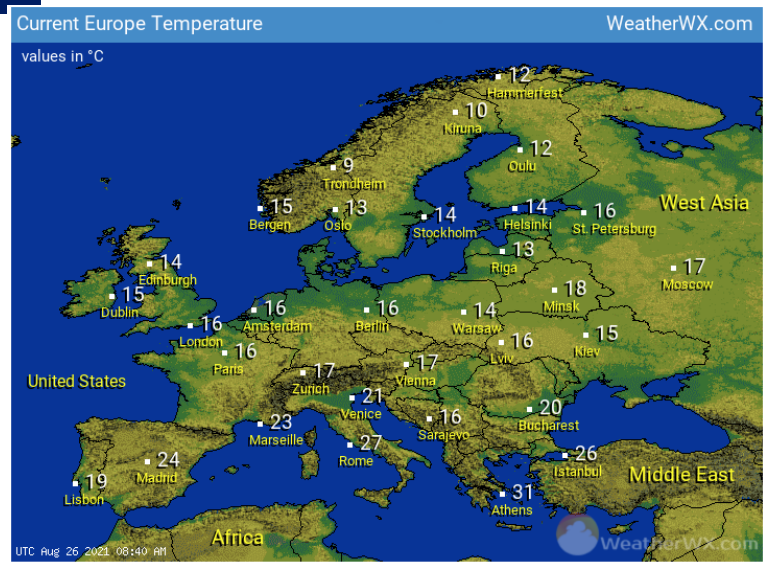
London Business School

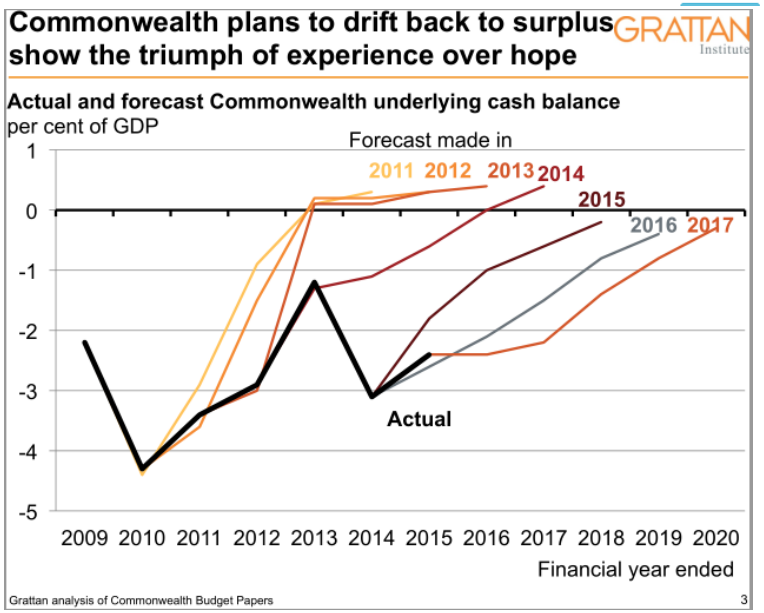
# Contents



- Descriptive Statistics
- Exploratory Data Analysis
- Normal Distribution
- Reproducibility and RMarkdown

# Uncertainty is challenging



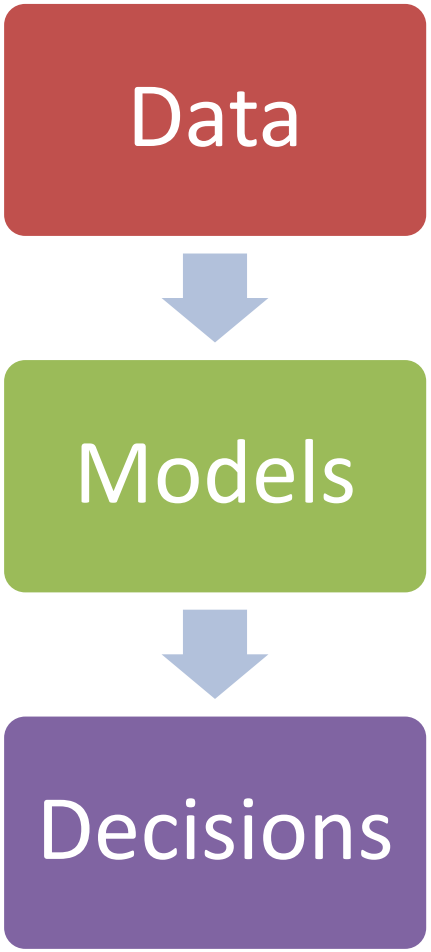

London Business School

"You are neither right nor wrong because the crowd disagrees with you. You are right because your *data* and *reasoning* are right."

-Warren Buffett

London Business School

# Handling uncertainty



# Why is statistics important?

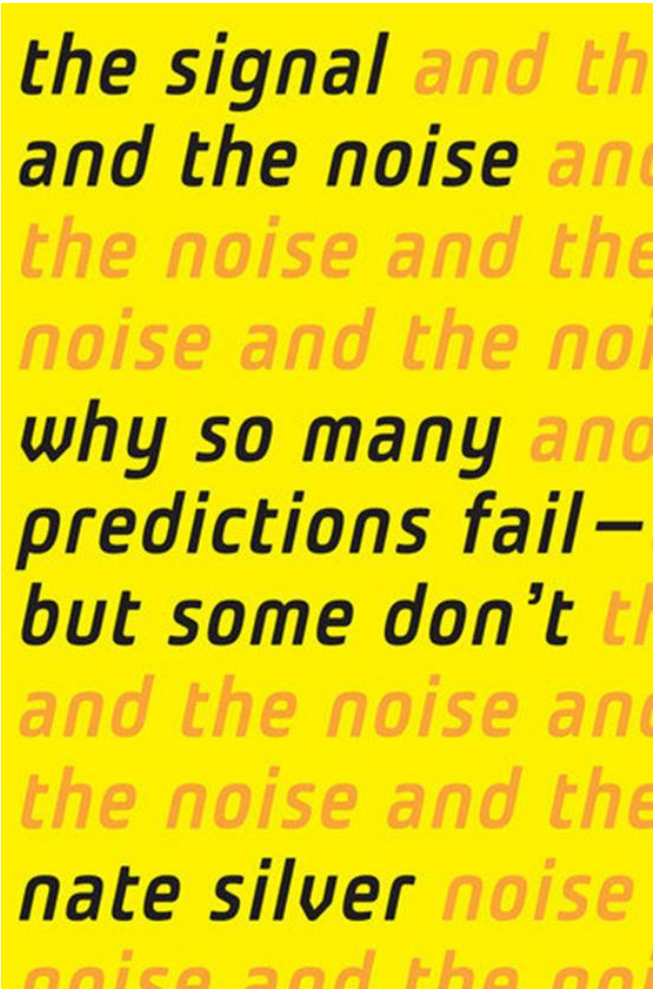
We use data to make decisions involving uncertainty, both at an individual and a public level. Many of the decisions affecting our lives have some statistical justification and statistics helps us understand variability.

Typically, data is collected by survey, experiment, or by an observational study. In many cases, the simple act of collecting and reporting data has a positive effect in organisations.

We build models and test how much of the variability can be explained by our '**model**' (systematic/model variance) versus the **noise,** the residual/random '**error**' (unsystematic/random variance)

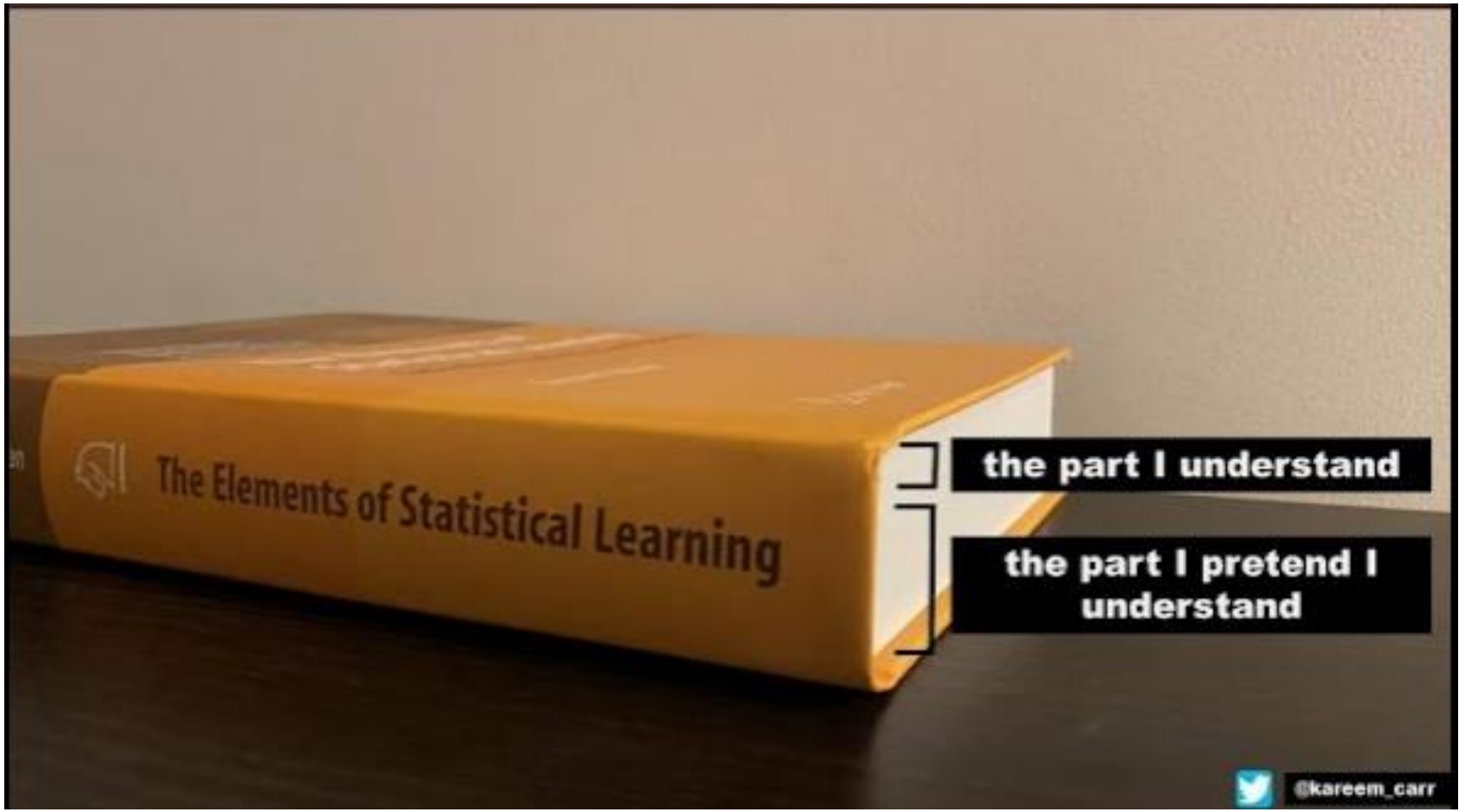
$$\mathbf{data = (\mathit{model}) + error_i}$$

The first, and easiest, model is the good old average, or arithmetic mean




*Spotting signals in the noise – that is statistics*
Stephen Ziliak

# It’s ok to be confused



London Business School

# What is statistics good for?

## Descriptive Statistics

- Collect
- Organize
- Summarize
- Display
- Analyze

### Typical Questions

- *How much time do MAMs sleep? Or exercise? Or spend online?*
- *Does height follow a symmetric distribution? What about the number of Facebook friends?*

## Inferential Statistics

- Predict and forecast values of population parameters
- Test hypotheses (draw conclusions) about values of population parameters
- Make decisions

### Typical Questions

- *Do women spend more than men on their haircut?*
- *Is there a relationship between statistics anxiety and academic motivation?*

*It is a capital mistake to theorize before you have the data.*
***Sherlock Holmes***

London Business School

# Probability distributions

- A **probability distribution** is an expression that describes the relative frequency (height of the curve) for every possible outcome, or a range of outcomes.

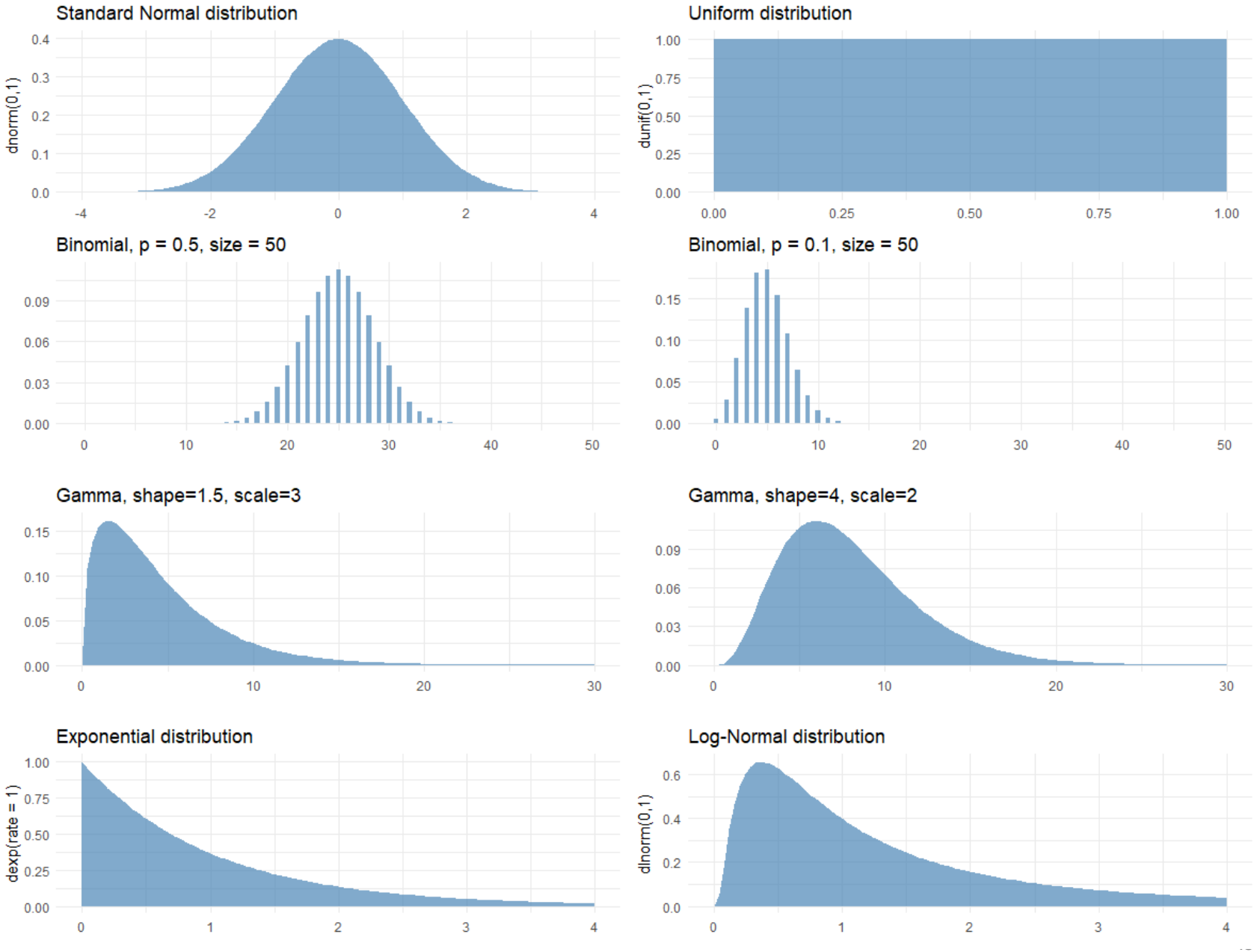
- A **probability model (e.g., the Normal distribution)** attempts to capture the essential structure of the real world by asking what the world might look like if an infinite number of observations were obtained. Nothing in the real world is exactly distributed as a probability model. However, a probability model often describes the world well enough to be useful in making decisions.

- The main purpose of a distribution is to indicate not ***accuracy***, but ***error***. If every estimate were precisely the correct measurement there would be no uncertainty; however, no single observation is a perfect example of generality.

$$\mathbf{data = (\mathit{model}) + error_i}$$

London Business School

# Probability distribution models




https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Distributions.html

London Business School

# Describing Probability Distributions

1. **Shape**: skewness, symmetry, modality

2. **Central Tendency:** an estimate of a ***typical*** observation in the distribution (mean, median, mode, etc.)

   $\boldsymbol{\mu}$: population mean, $\bar{x}$: sample mean

3. **Spread**: measure of variability in the distribution (standard deviation, IQR, range, etc.)

   $\boldsymbol{\sigma}$: population standard deviation, $s$: sample standard deviation

4. **Unusual observations:** observations that stand out from the rest of the data that may be suspected outliers

London Business School

# Handedness Histogram

Work in groups and draw your guess of the histogram of the handedness scores of the students in the class.

-1: Use left hand only

0: Completely ambidextrous

+1: Use right hand only

| Task | Left | Right |
|---|---|---|
| Writing | | |
| Drawing | | |
| Throwing | | |
| Scissors | | |
| Toothbrush | | |
| Knife (without fork) | | |
| Spoon | | |
| Broom (upper hand) | | |
| Striking match (hand that holds the match) | | |
| Opening box (hand that holds the lid) | | |
| Total | | |

Right − Left: Right + Left: $\frac{\text{Right} - \text{Left}}{\text{Right} + \text{Left}}$

London Business School

# The Average Student

## The average Early Careers student at LBS is…

- 173.82 cm tall
- Spends 38.90 hours online per week
- Exercises for 4.63 hours per week
- Sleeps for 7.70 hours
- Spent 35.19$ on last haircut
- Has 834.2 Facebook friends
- And is 0.487 male

*It is difficult to understand why statisticians commonly limit their inquiries to Averages, and do not revel in more comprehensive views. Their souls seem as dull to the charm of variety as that of the native of one of our flat English countries, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once*

Francis Galton (1822-1911)

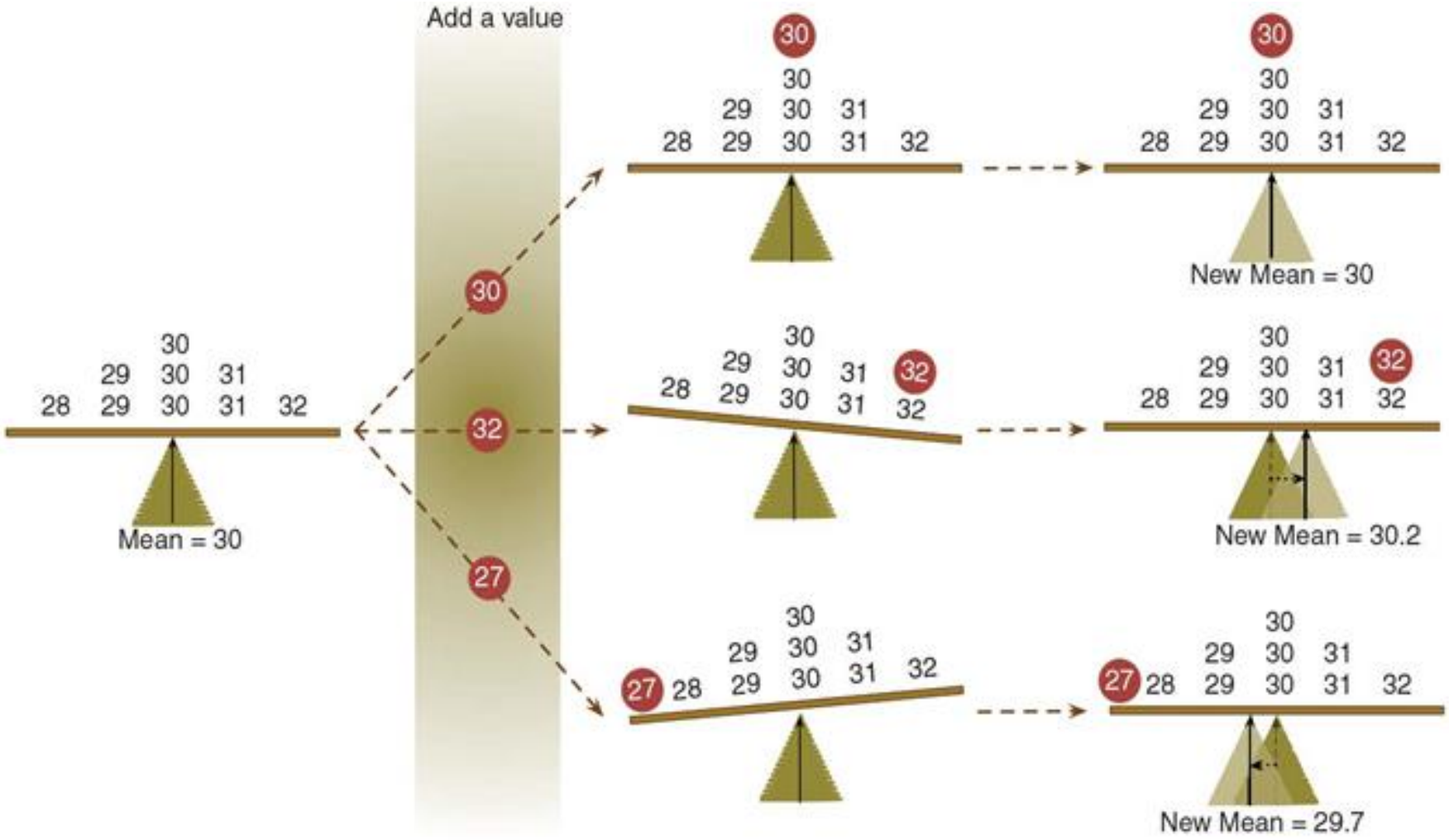
# Central Tendency and Variability

- Central tendency:
  - Mean: sum of all readings/total number = $\sum_{i=1}^{n} x_i$ /n
  - Median: middle reading when data is sorted in size order
  - Mode: the most frequent reading

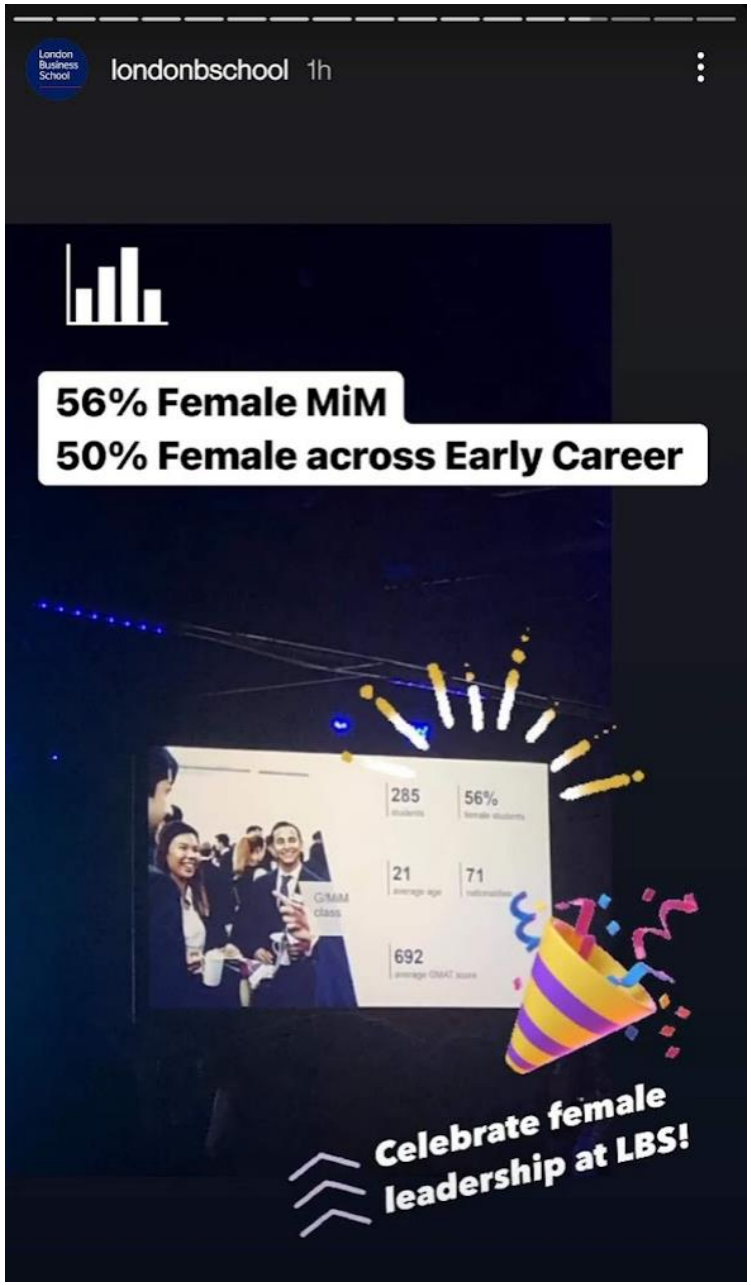
- Variability:
  - Mean absolute deviation (MAD) = $\frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n}$
  - Variance = $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$
  - Standard deviation = $\sqrt{\text{variance}}$
  - Coefficient of variation = $\frac{\text{SD}}{\text{mean}}$
  - Range = largest – smallest reading
  - Inter-quartile range (IQR) = upper quartile – lower quartile

*While nothing is more uncertain than a single life, nothing is more certain than the average duration of a thousand lives* -- Elizur Wright, 19th-century mathematician

# The Mean: Balancing point



# The Mean: Balancing point



London Business School

# The Mean: Balancing point

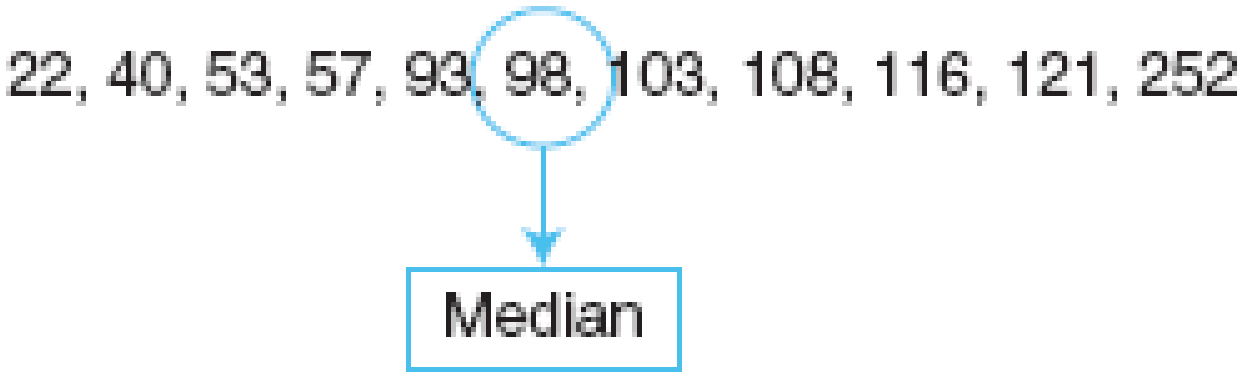
In January 1971, the Gallup poll asked: "A proposal has been made in Congress to require the US government to bring home all US troops before the end of this year. Would you be ***for*** or ***against*** this proposal?"

Guess the results, for respondents in each education category, and fill out this table (the two numbers should add up to 100%)

| | Grade school education | High school education | College education | Total adults |
|---|---|---|---|---|
| % ***for*** withdrawal of US troops (doves) | | | | 73% |
| % ***against*** withdrawal of US troops (hawks) | | | | 27% |
| Total | 100% | 100% | 100% | 100% |

# Central Tendency: The Median

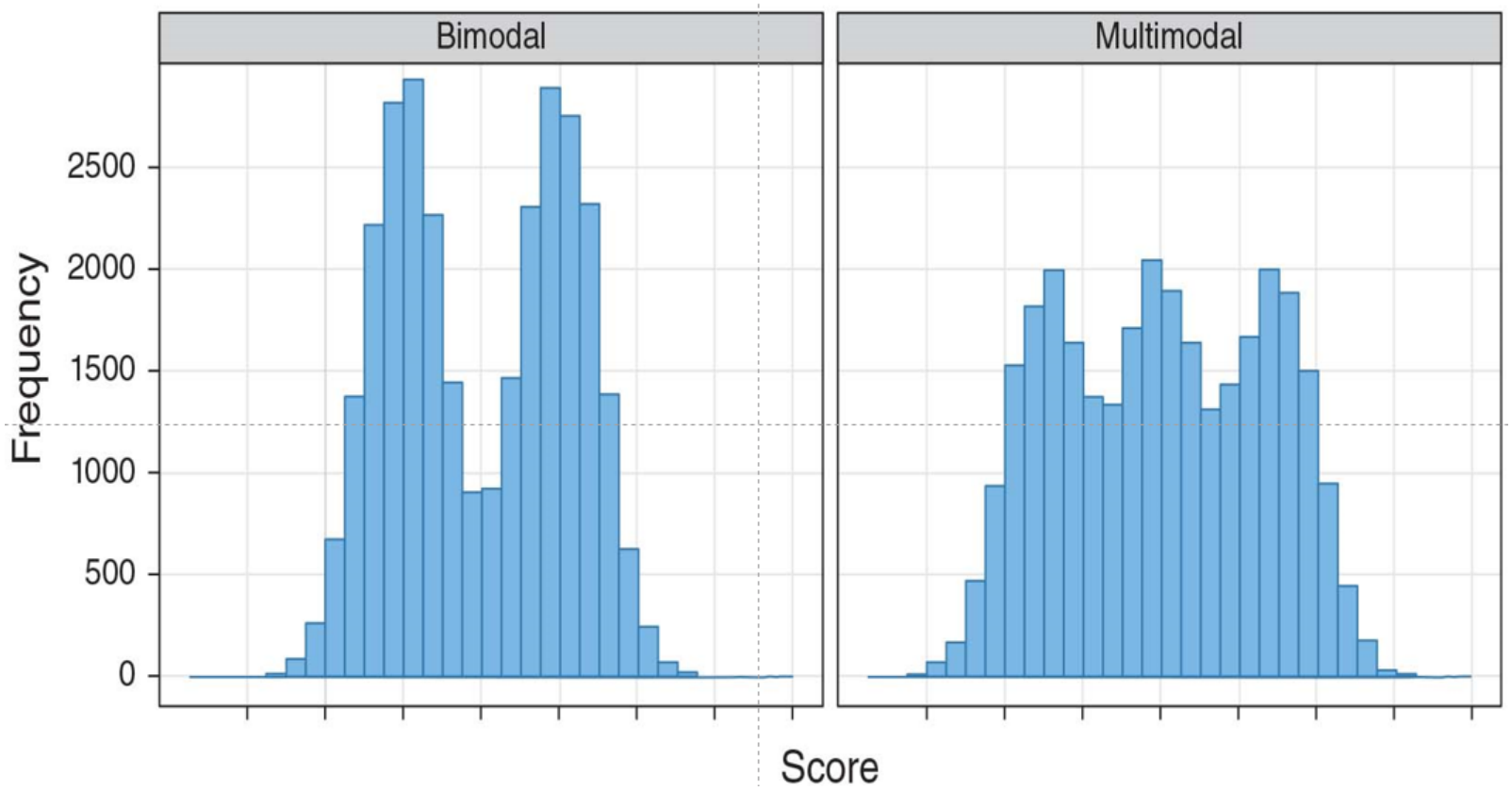
- Median
  - The middle score (50th percentile) when scores are ordered.
  - This second quartile 'slices' the data in half; 50% of the values are above it, 50% below it
- Example
  - Number of friends of 11 Facebook users.



London Business School

# Central tendency: The Mode

- Mode: The most frequent score
- Bimodal: Having two modes
- Multimodal: Having several modes



London Business School

# Skewness - Kurtosis

- Skewness
  - The symmetry of the distribution. Perfect symmetry has skew of 0
  - Positive skew (scores bunched at low values with the tail pointing to high values).
  - Negative skew (scores bunched at high values with the tail pointing to low values).

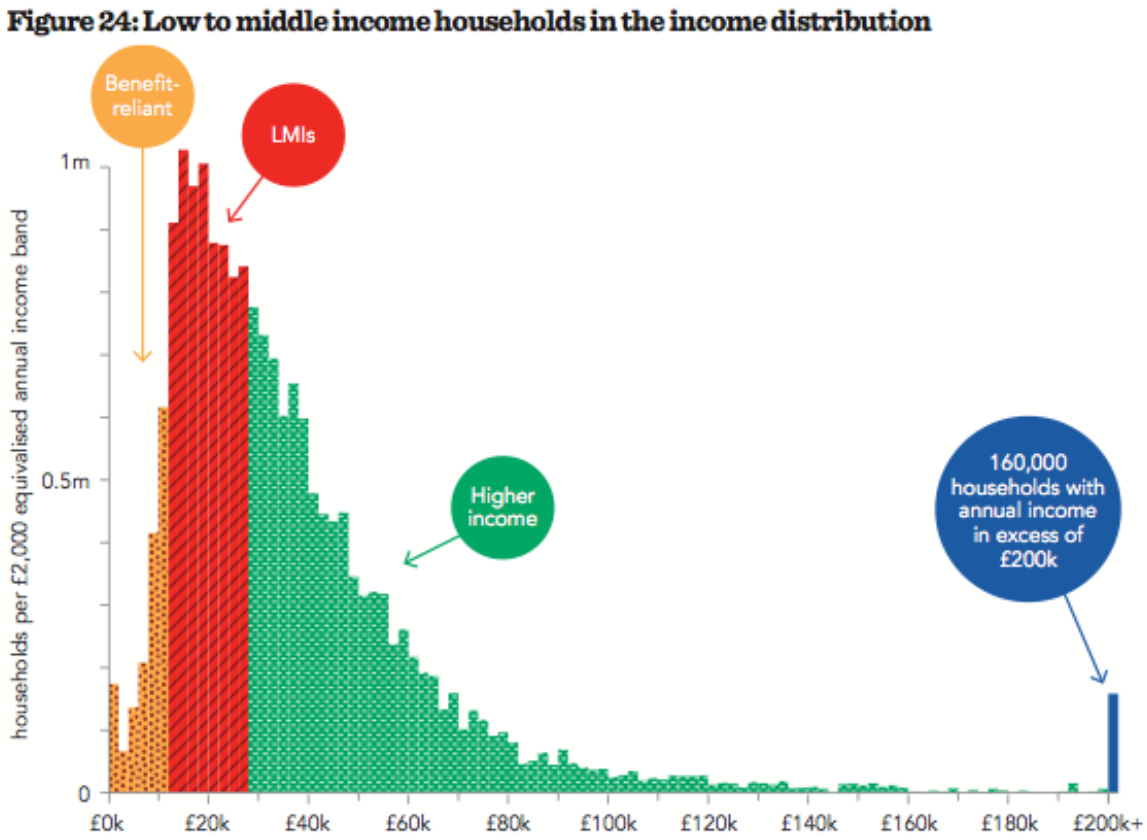
- Kurtosis
  - The ‘heaviness’ of the tails.
  - Leptokurtic = heavy tails.
  - Platykurtic = light tails.

# Example: Income distribution

Income is not symmetrically distributed - it is usually *skewed* to the right

What is an appropriate measure of central tendency for the figure?

What is the *average* income?

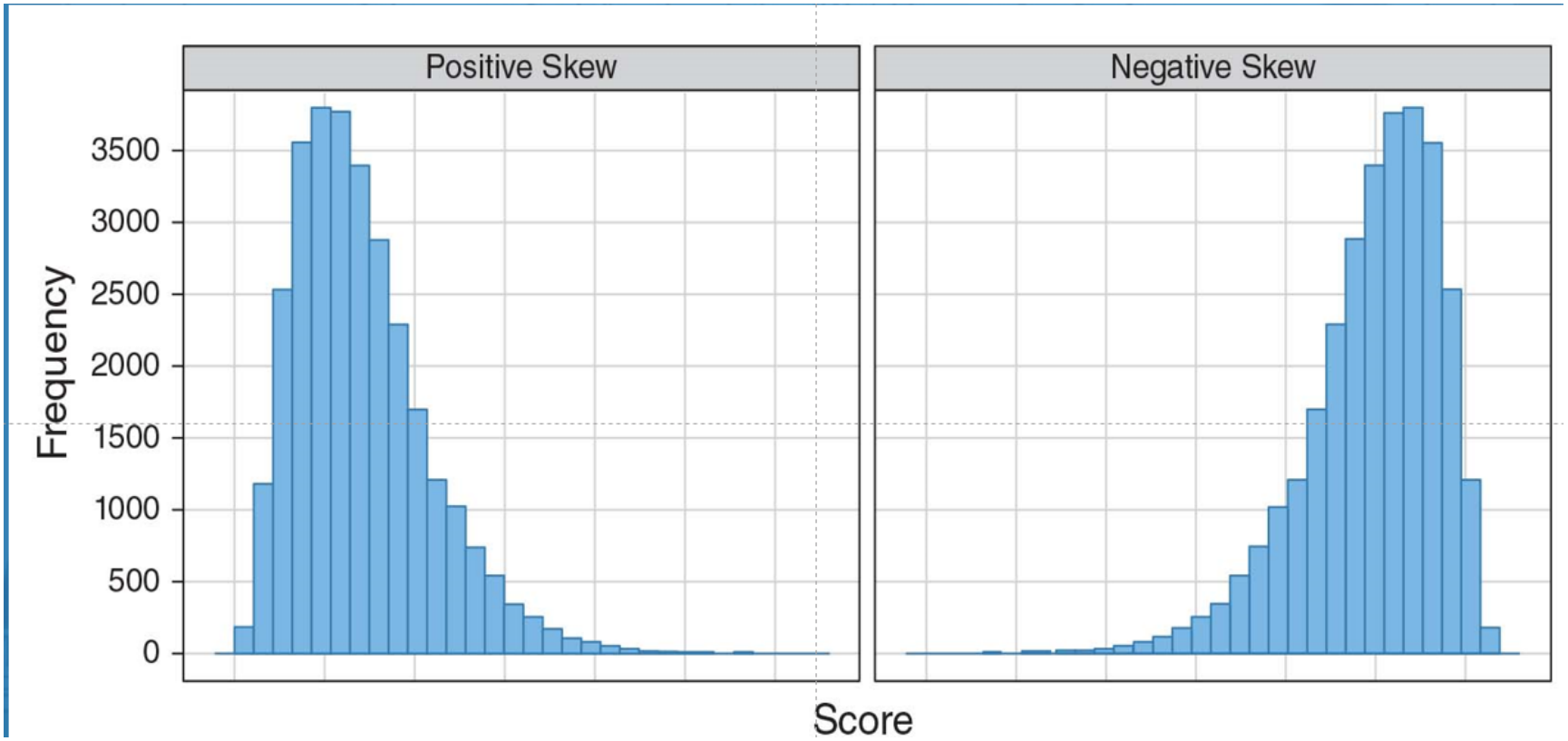



The median is a **robust statistic**, as extreme values do not affect it as much as they affect the mean

Source: http://ftalphaville.ft.com/2014/02/11/1768992/the-uks-squeezed-bottom-charted/

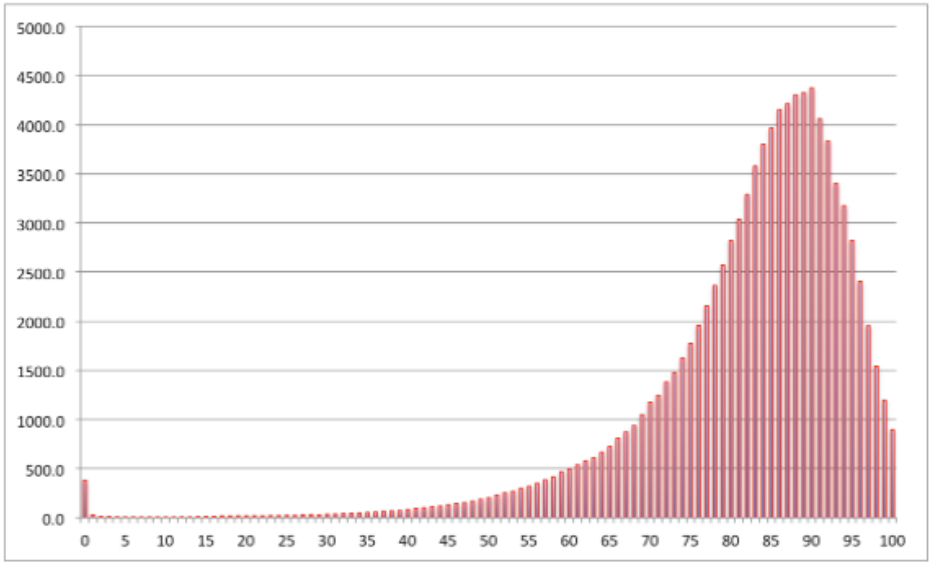
London Business School

# Skewness

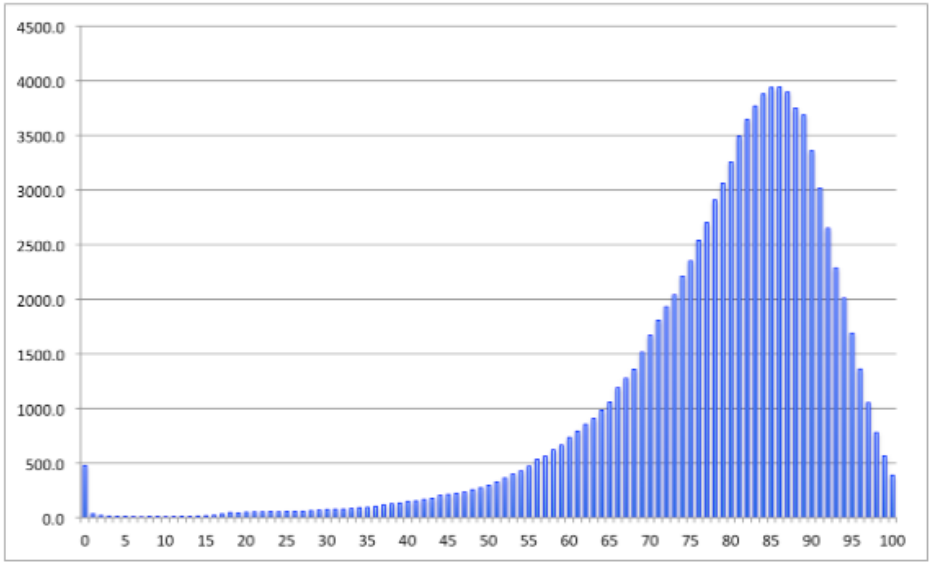



Positively skewed or skewed **right**

Negatively skewed or skewed **left**




*Numbers of women expected to die at each age, out of 100,000 born, assuming mortality rates stay the same as 2010-2012. The expectation is 83, median 86, the most likely value (mode) is 90.*




*Numbers of men expected to die at each age, out of 100,000 born, assuming mortality rates stay the same as 2010-2012. The expectation is 79, median 82, the most likely value (mode) is 86.*



Source: https://understandinguncertainty.org/why-life-expectancy-misleading-summary-survival

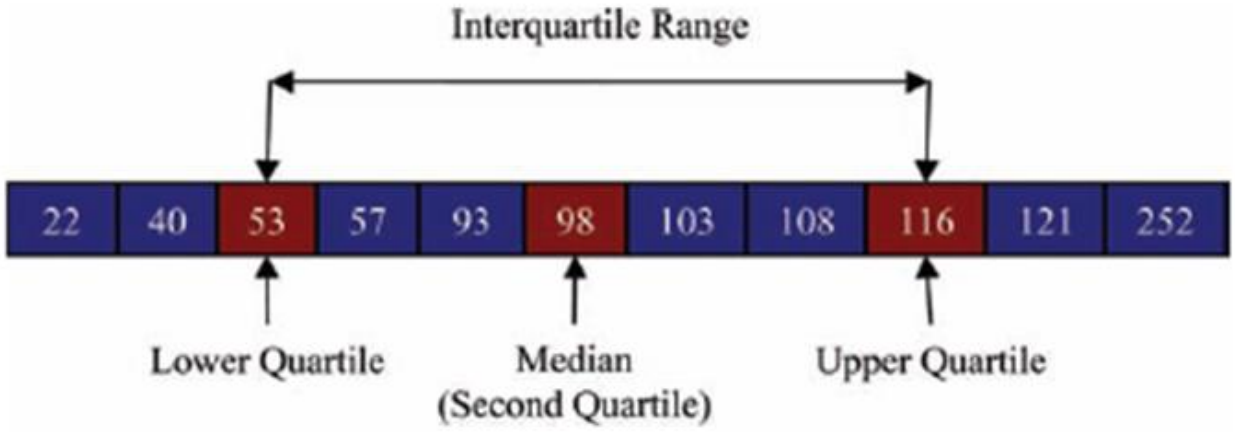
London Business School

# Variability: Range

- The Range
  - The smallest score subtracted from the largest
- Example
  - Number of friends of 11 Facebook users.
  - 22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252
  - Range = 252 – 22 = 230
  - Very biased by outliers

# Variability: Interquartile range (IQR)

- Quartiles
  - The three values that split the sorted data into four equal parts.
  - Second Quartile (Q2, 50$^{th}$ percentile) = median.
  - Lower quartile (Q1, 25$^{th}$ percentile) = median of lower half of the data
  - Upper quartile (Q3, 75$^{th}$ percentile) = median of upper half of the data



London Business School

# Variance and Standard Deviation

- The deviation is how far away each point is from the mean
- If we add up the deviations, we get zero
- **Variance**: The average sum of squared deviations
- **Standard Deviation:** the square root of variance

## Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x-\mu)^2}{N}$$

$$\sigma = \sqrt{\sigma^2}$$

Excel Functions
=VAR.P(range)
=STDEV.P(range)

## Sample Variance

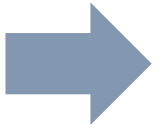
$$s^2 = \frac{\sum_{i=1}^{n}(x-\bar{x})^2}{(n-1)}$$

$$s = \sqrt{s^2}$$

Excel Functions
=VAR.S(range)
=STDEV.S(range)

London Business School

# Contents



- Descriptive Statistics
- Exploratory Data Analysis
- Normal Distribution
- Reproducibility and RMarkdown

London Business School

# *A picture is worth a thousand words*

Always plot the data. The main goals for statistical charts are to facilitate comparisons and to identify trends. Your goal is to transform data into information suited for human consumption. Visualization provides one of the most readable and compelling forms of information.

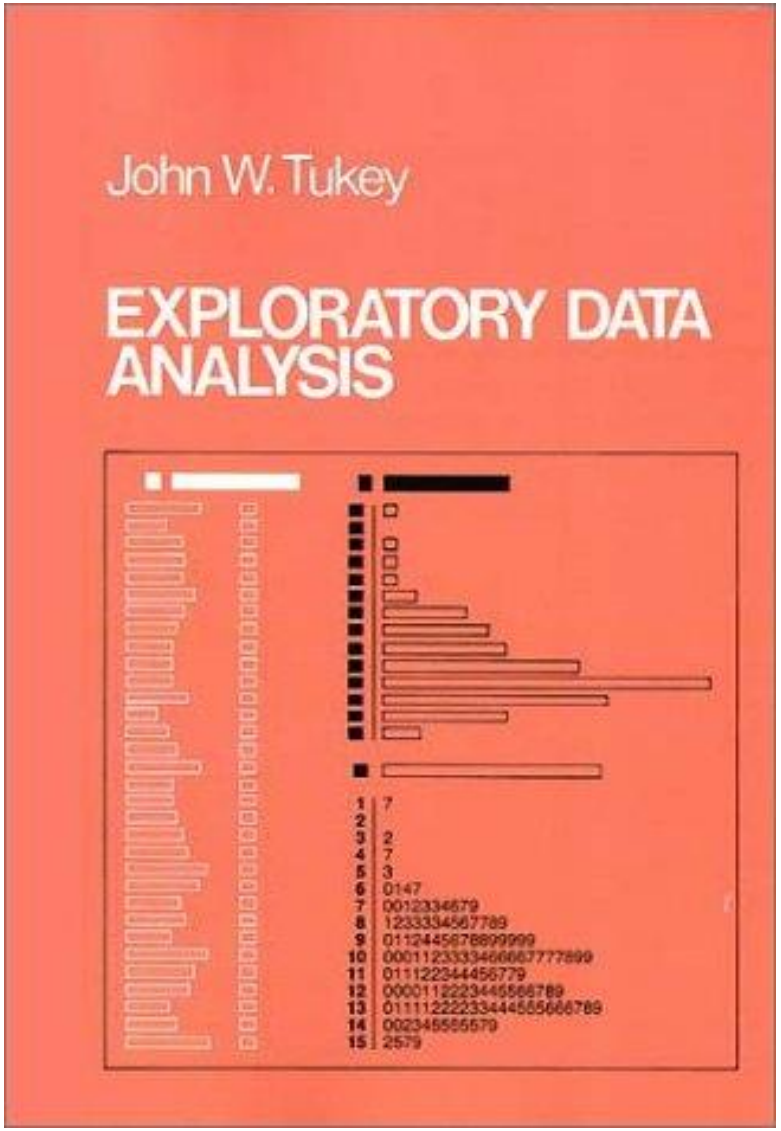
Replace (or complement) 'typical' tables of data or statistical results with charts that are more compelling and accessible.

Whenever possible, generate charts that overlay / juxtapose observed data and analytical results, e.g. the 'fit' of your sample data with the theoretical Normal distribution model.

Among the various charts you create, pick the most insightful ones. Do not use pie charts, ever!

London Business School

# Birth of EDA




**Exploratory Data Analysis (EDA) came of age when this book appeared in 1977**

Tukey believed that too much emphasis was placed on statistical hypothesis testis (confirmatory data analysis).

He thought that more emphasis needed to be placed on using data to suggest hypotheses to test

*The greatest value of a picture is when it forces us to notice what we never expected to see*

London Business School

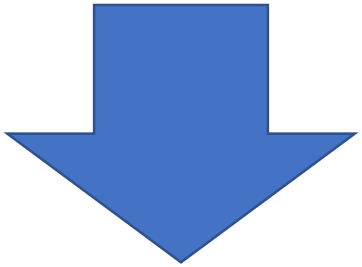
# Can we tell better stories?

**Contextual Understanding (hypotheses)**

- Market Dynamics
- Socio-Economic Status (SES)
- Culture

**Effective Data Analysis (Investigation)**

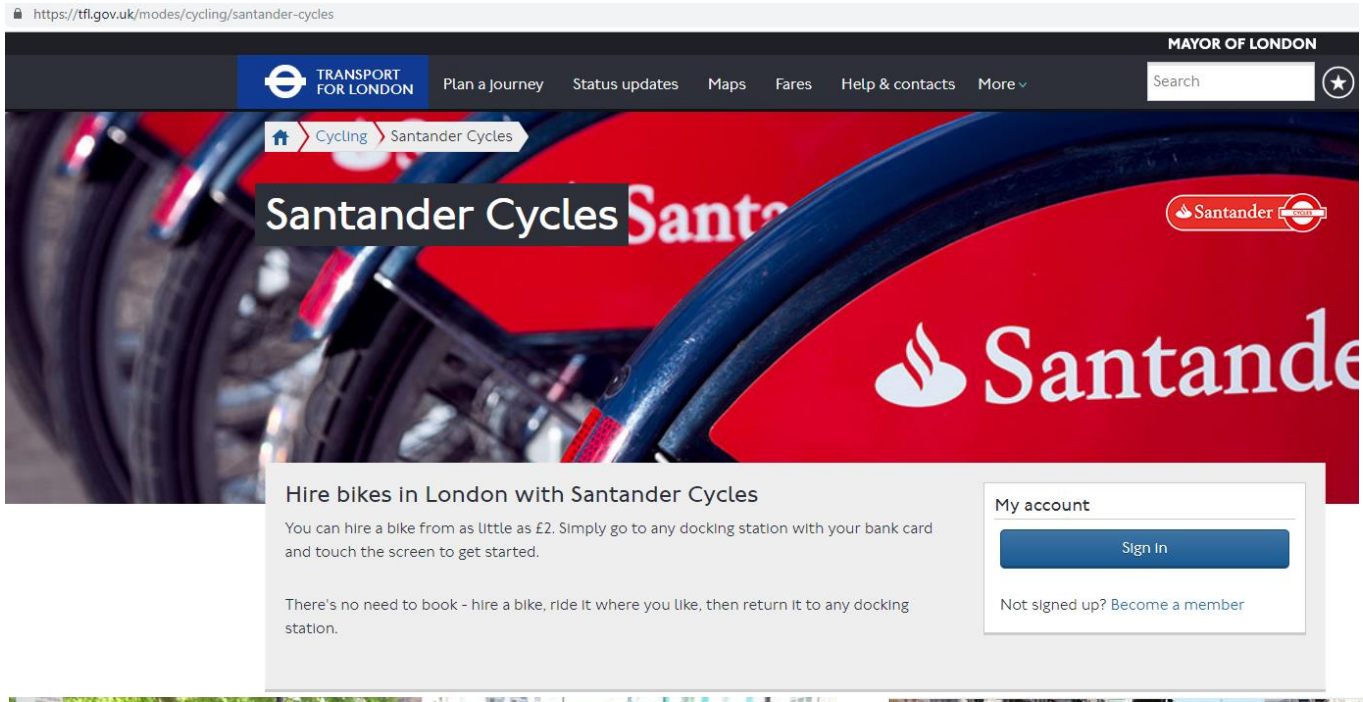
- Mindset
- Workflow
- Framework/Tools



**Actionable Insights**

*Data graphics are paragraphs about data and should be treated as such.*
*Words, graphics, and tables are different mechanisms with but a single purpose—the presentation of information. Why should the flow of information be broken up into different places on the page because the information is packaged one way or another?*
*The Visual Display of Quantitative Information*, Edward Tufte, Graphics Press, 1983, p.181

# London Bikes




London Business School

# London Bikes

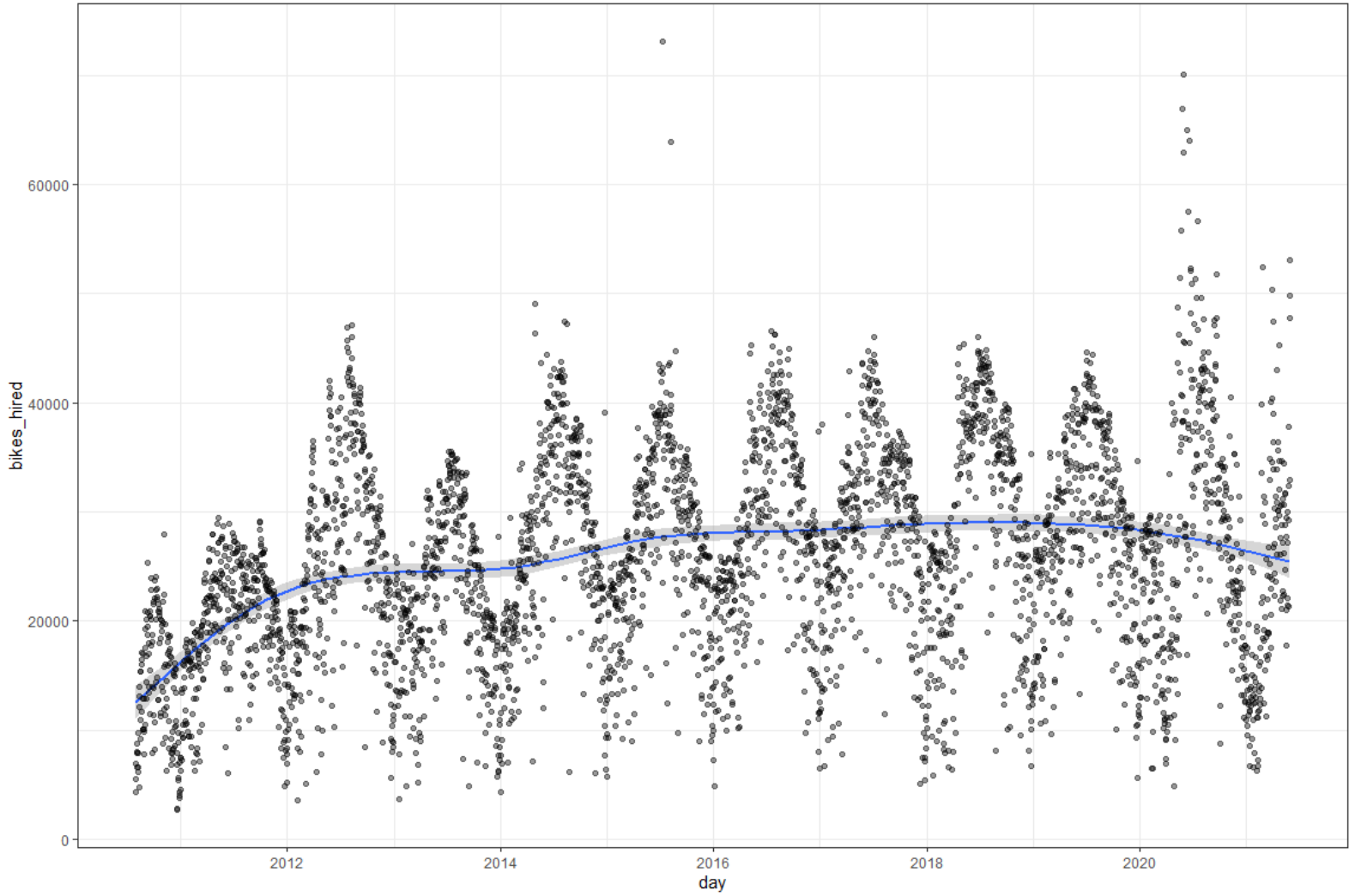
| | date | bikes_hired | season | avg_temp | avg_humidity | avg_pressure | avg_windspeed | rainfall_mm | rain | fog | thunderstorm | snow |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2011-01-01 | 4555 | 1 | 6 | 84 | 1025 | 10 | 0 | TRUE | FALSE | FALSE | FALSE |
| 2 | 2011-01-02 | 6250 | 1 | 3 | 79 | 1028 | 8 | 0 | FALSE | FALSE | FALSE | FALSE |
| 3 | 2011-01-03 | 7262 | 1 | 0 | 80 | 1024 | 6 | 0 | FALSE | FALSE | FALSE | TRUE |
| 4 | 2011-01-04 | 13430 | 1 | 3 | 87 | 1013 | 6 | 0 | TRUE | FALSE | FALSE | FALSE |
| 5 | 2011-01-05 | 13757 | 1 | 6 | 84 | 1000 | 19 | 0 | TRUE | FALSE | FALSE | FALSE |
| 6 | 2011-01-06 | 9595 | 1 | 4 | 92 | 996 | 5 | 0 | TRUE | TRUE | FALSE | TRUE |
| 7 | 2011-01-07 | 9294 | 1 | 6 | 92 | 999 | 11 | 0 | TRUE | TRUE | FALSE | FALSE |
| 8 | 2011-01-08 | 9338 | 1 | 6 | 82 | 997 | 23 | 0 | TRUE | FALSE | FALSE | FALSE |
| 9 | 2011-01-09 | 10558 | 1 | 3 | 79 | 1012 | 16 | 0 | FALSE | FALSE | FALSE | FALSE |

```
> glimpse(bike)
Observations: 3,103
Variables: 18
$ date          <date> 2011-01-01, 2011-01-02, 2011-01-03, 2011-01-04, 2011-01-05, 2011-01-06, 2011-01-07, 2011-01-...
$ bikes_hired   <dbl> 4555, 6250, 7262, 13430, 13757, 9595, 9294, 9338, 10558, 16058, 16412, 13894, 15911, 14834, 1...
$ season        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
$ avg_temp      <dbl> 6, 3, 0, 3, 6, 4, 6, 6, 3, 3, 6, 8, 12, 10, 9, 10, 7, 3, 3, 3, 1, 3, 7, 4, 4, 4, 2, 1, 1, 2, ...
$ avg_humidity  <dbl> 84, 79, 80, 87, 84, 92, 92, 82, 79, 87, 82, 89, 89, 87, 79, 83, 90, 85, 88, 86, 83, 86, 85, 8...
$ avg_pressure  <dbl> 1025, 1028, 1024, 1013, 1000, 996, 999, 997, 1012, 1011, 1006, 1011, 1009, 1010, 1015, 1016, ...
$ avg_windspeed <dbl> 10, 8, 6, 6, 19, 5, 11, 23, 16, 14, 16, 16, 23, 24, 24, 23, 8, 11, 8, 8, 8, 10, 13, 11, 14, 1...
$ rainfall_mm   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
$ rain          <lgl> TRUE, FALSE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE,...
$ fog           <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
$ thunderstorm  <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, TR...
$ snow          <lgl> FALSE, FALSE, TRUE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS...
$ year          <dbl> 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 201...
$ month         <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
$ month_name    <ord> Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan, Jan...
$ day           <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26...
$ day_of_week   <ord> Sat, Sun, Mon, Tue, Wed, Thu, Fri, Sat, Sun, Mon, Tue, Wed, Thu, Fri, Sat, Sun, Mon, Tue, Wed...
$ season_name   <fct> Winter, Winter, Winter, Winter, Winter, Winter, Winter, Winter, Winter, Winter, Winter, Winte...
```

# London Bikes through the years

This is a time series plot– On the x-axis, we have time

Which year has the lowest variability? The highest variability?



London Business School

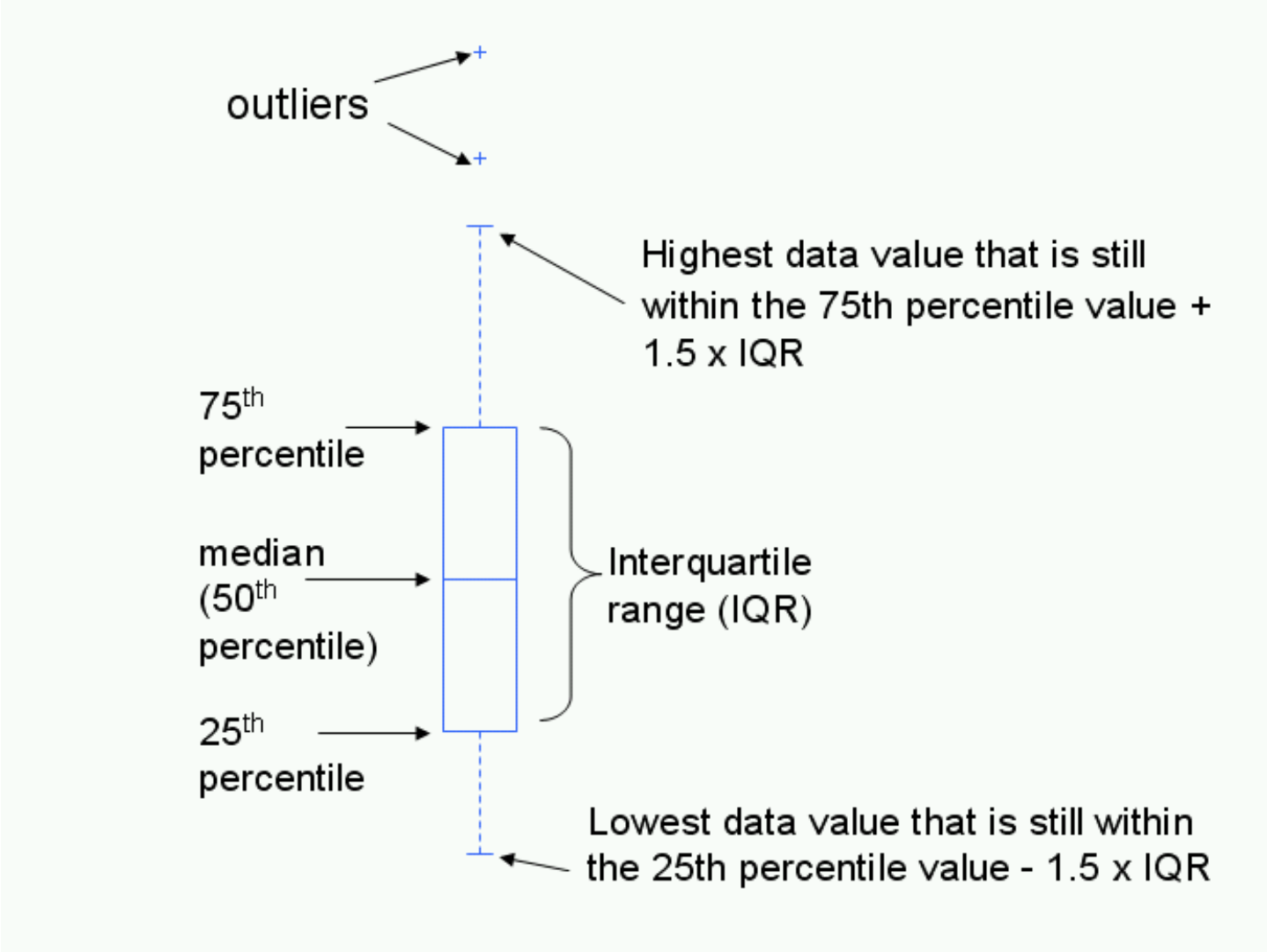
# London Bikes Summary Statistics

```
> favstats(bikes_hired ~ year, data=bike)
   year  min      Q1  median       Q3   max     mean        sd   n missing
1  2010 2764  9297.0 14010.0 18677.50 27964 14069.76  5616.909 155       0
2  2011 4555 16260.0 20264.0 23708.00 29417 19568.35  5497.469 365       0
3  2012 3531 19282.0 26178.5 32532.75 47102 26008.97  9429.395 366       0
4  2013 3728 17555.0 22021.0 27371.00 35580 22042.35  7276.488 365       0
5  2014 4327 20532.0 27676.0 34437.00 49025 27462.73  9065.086 365       0
6  2015 5779 22056.0 26618.0 32857.00 73094 27046.13  8547.890 365       0
7  2016 4894 22402.0 27881.5 35130.00 46625 28152.01  8850.958 366       0
8  2017 5143 24064.0 29490.0 34459.00 46035 28619.30  8376.552 365       0
9  2018 5859 21759.0 29190.0 37677.00 46084 28952.16 10174.346 365       0
10 2019 5649 24198.0 28927.0 34494.00 44668 28561.52  8087.904 365       0
11 2020 4871 20134.0 27500.0 36471.00 70074 28497.80 11580.091 366       0
12 2021 6252 14011.5 22318.0 30475.00 53069 23400.93 10440.581 151       0
> favstats(bikes_hired ~ day_of_week, data=bike)
  day_of_week  min       Q1  median       Q3   max     mean        sd   n missing
1         Sun 2764 13554.75 19984.0 29948.50 62905 21890.60 10735.171 566       0
2         Mon 3971 19975.75 25734.0 31640.25 67000 25870.88  8597.769 566       0
3         Tue 3763 21756.00 27528.0 34598.00 50325 27787.70  8686.706 565       0
4         Wed 4327 22120.00 27194.0 34122.00 52094 27763.93  8583.716 565       0
5         Thu 5649 21823.00 26826.0 34430.00 73094 27708.28  8840.338 565       0
6         Fri 5402 21017.00 26506.5 33297.00 50940 26779.87  8474.799 566       0
7         Sat 2805 15056.00 21392.0 30696.75 70074 23422.97 10939.834 566       0
> favstats(bikes_hired ~ month_name, data=bike)
   month_name   min       Q1  median       Q3   max     mean        sd   n missing
1         Jan  3728 13622.00 18206.0 23199.00 38042 18229.96  6109.965 341       0
2         Feb  3531 14808.00 19857.0 24225.50 52427 19653.52  6471.325 311       0
3         Mar  5062 17463.00 22914.0 26894.00 50325 22056.38  7135.274 341       0
4         Apr  4871 21192.50 26176.0 30826.50 49025 25926.52  7667.650 330       0
5         May 12049 24335.00 30145.0 36101.00 70074 30372.93  8578.489 341       0
6         Jun  6061 27190.50 33017.5 38616.25 65022 32703.72  8402.466 300       0
7         Jul  5564 30141.00 36375.5 41340.50 73094 35085.83  8265.741 312       0
8         Aug  4303 25270.00 33048.0 38278.00 63963 30987.29 10070.513 341       0
9         Sep  4826 24583.75 31898.5 36133.25 51750 30313.45  7997.329 330       0
10        Oct  7068 22886.00 27534.0 32588.00 39569 26905.27  6887.870 341       0
11        Nov  6030 18374.75 22534.5 27659.75 40467 22499.35  6367.520 330       0
12        Dec  2764 11584.00 16554.0 22939.00 39145 17006.94  7019.094 341       0
> favstats(bikes_hired ~ season_name, data=bike)
  season_name  min      Q1 median    Q3   max     mean       sd    n missing
1      Winter 2764 13339.0  18255 23415 52427 18255.82 6626.987  993       0
2      Spring 4871 20324.5  26065 31224 70074 26120.70 8525.566 1012       0
3      Summer 4303 27332.0  34240 39368 73094 32869.42 9143.360  953       0
4      Autumn 4826 21013.0  26888 32635 51750 26576.34 7787.347 1001       0
```

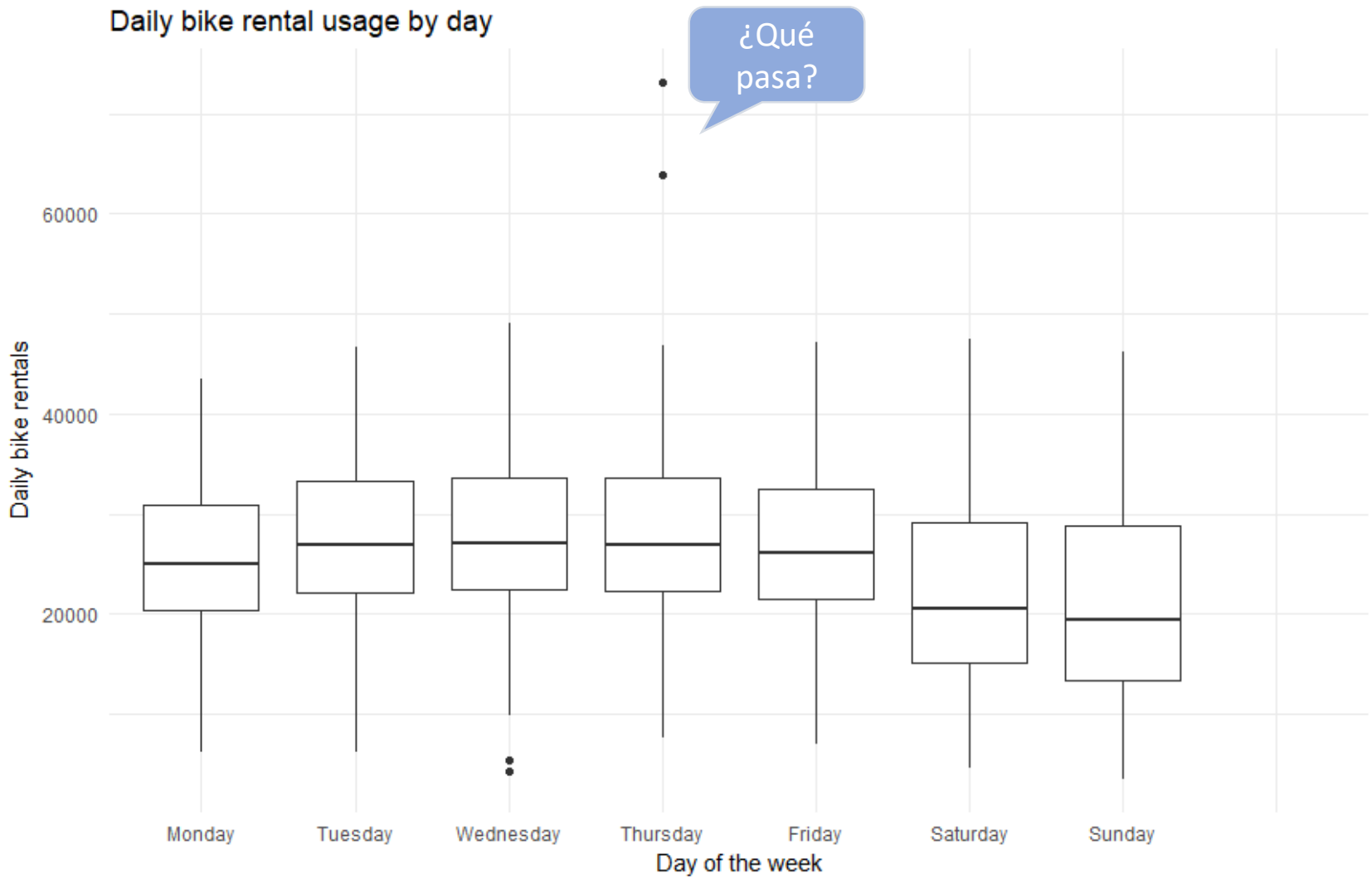
London Business School

# Histograms and Facetted Histograms

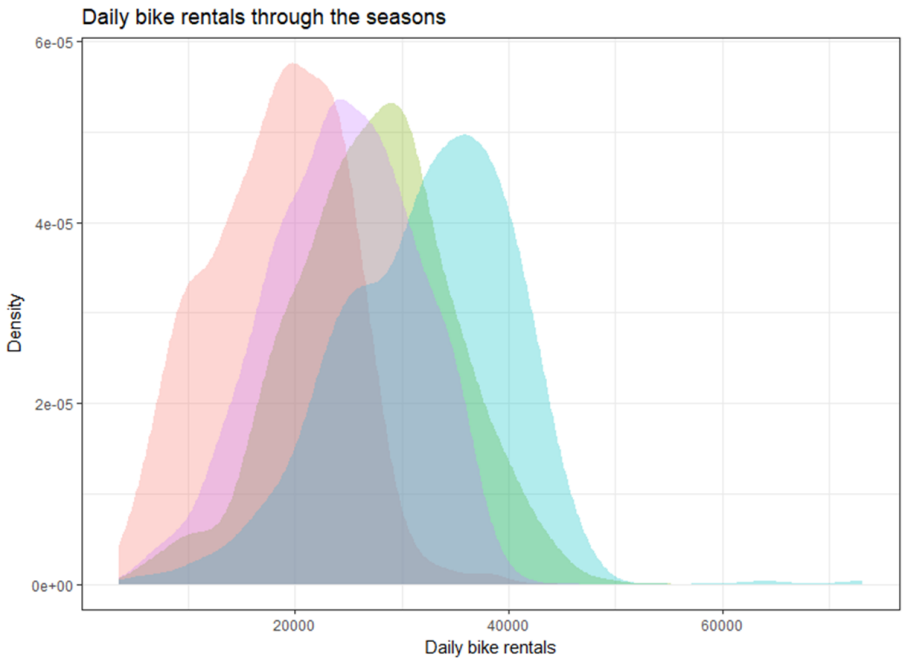
# Boxplots and Density Plots

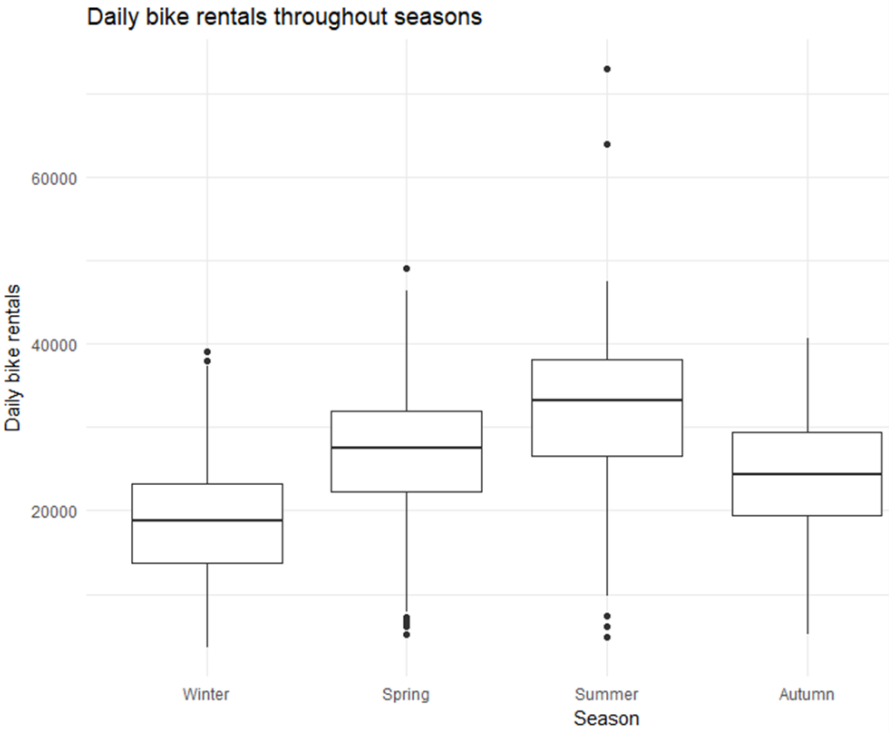


# Bikes rented up until 2018-12-31



# London City Bikes - Inference

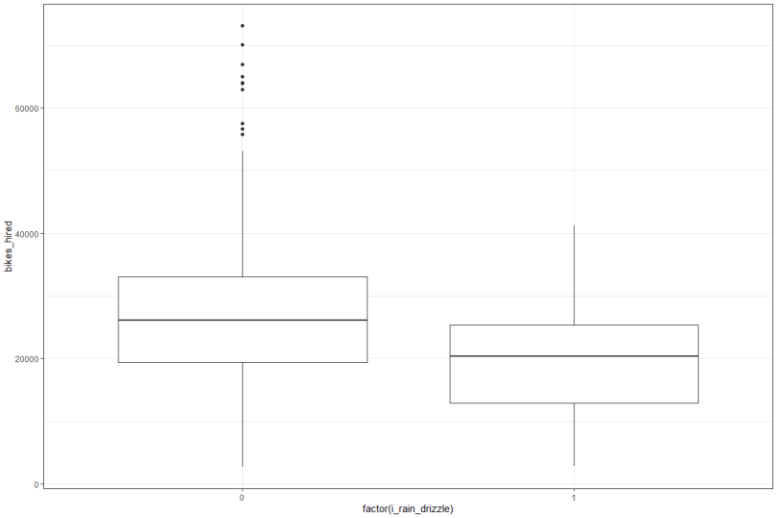


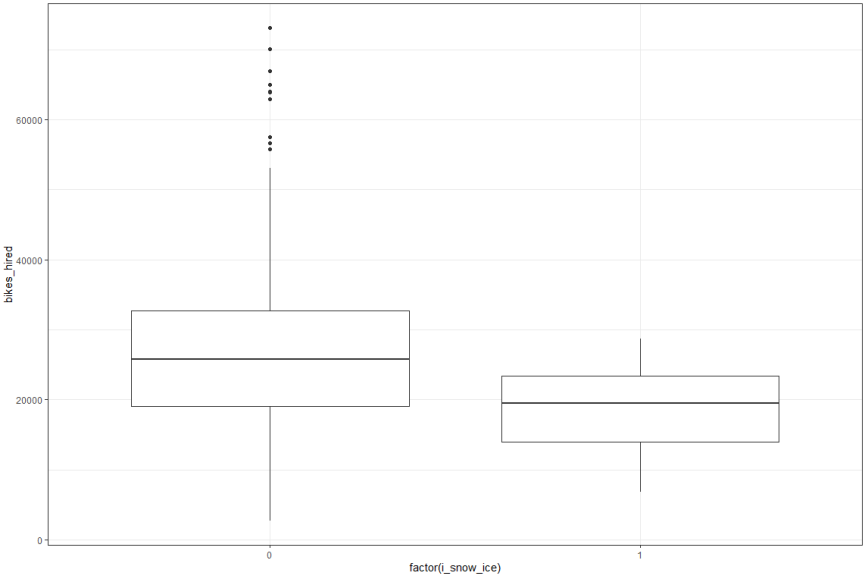



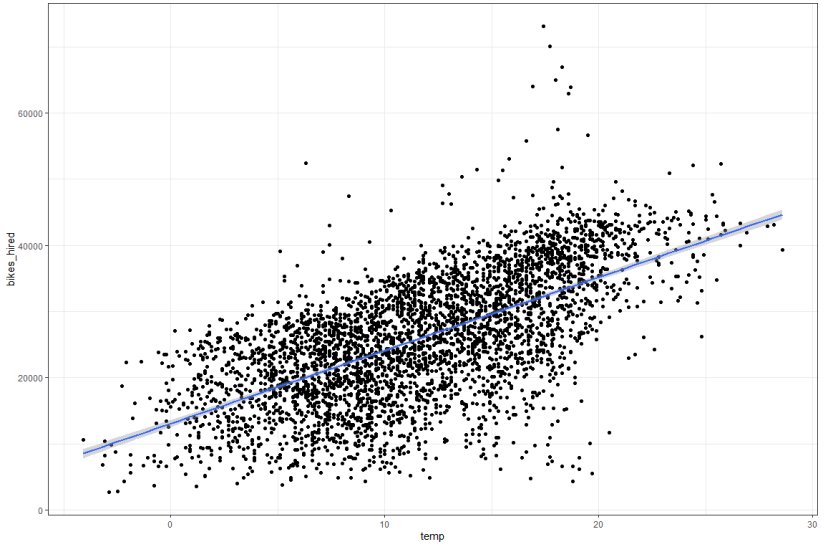

1. Let us compare Winter and Summer
   - Is there a difference in the **number** of bikes hired between Winter and Summer?
   - Is there a difference in the **average number** of bikes hired between Winter and Summer?

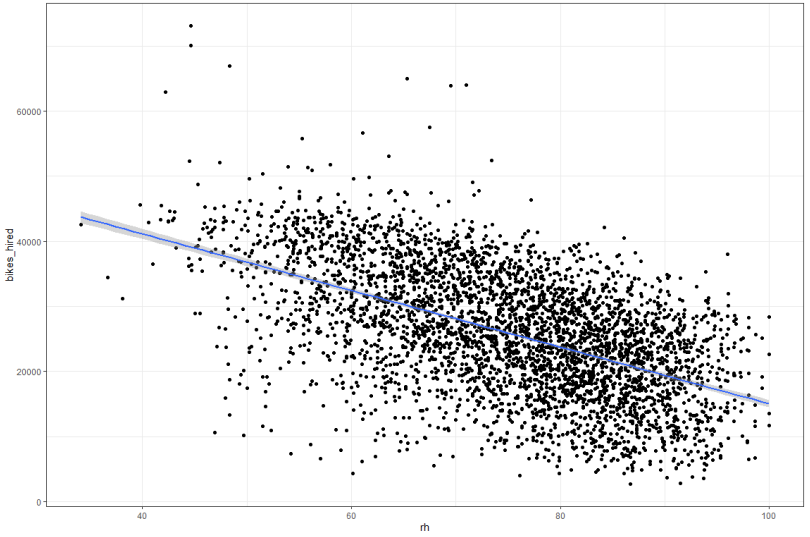
2. Let us compare Spring and Autumn
   - Is there a difference in the **number** of bikes hired between Spring and Autumn?
   - Is there a difference in the **average number** of bikes hired between Winter and Summer?

# Bikes Usage vs Weather



# Violent Crime in the US (2014)


FBI Home → UCR → UCR Data Online → Estimated Crime → State Level → One Year of Data Contact Us

**Results from state-level crime estimates database** Query date: August 07, 2019

Spreadsheet of this table (.csv file) | Spreadsheet help Revise this query | Get a different type of table

Definitions. Also see notes at the end of the page.

**Estimated crime in 2014**

| State | Population coverage | National or state crime: Violent crime: Violent crime total | Murder and nonnegligent manslaughter | Legacy rape[1] | Revised rape[2] | Robbery | Aggravated assault |
|---|---|---|---|---|---|---|---|
| Alabama | 4,849,377 | 20,727 | 276 | 1,436 | 2,005 | 4,701 | 13,745 |
| Alaska | 736,732 | 4,684 | 41 | 555 | 771 | 629 | 3,243 |
| Arizona | 6,731,484 | 26,916 | 319 | 2,464 | 3,378 | 6,249 | 16,970 |
| Arkansas | 2,966,369 | 14,243 | 165 | 1,182 | 1,763 | 2,050 | 10,265 |
| California | 38,802,500 | 153,709 | 1,699 | 8,398 | 11,527 | 48,680 | 91,803 |
| Colorado | 5,355,866 | 16,554 | 151 | 2,121 | 3,039 | 3,039 | 10,325 |
| Connecticut | 3,596,677 | 8,522 | 86 | 571 | 782 | 3,159 | 4,495 |
| Delaware | 935,614 | 4,576 | 54 | 249 | 386 | 1,269 | 2,867 |
| District of Columbia | 658,893 | 8,199 | 105 | 352 | 472 | 3,497 | 4,125 |
| Florida | 19,893,297 | 107,521 | 1,149 | 6,051 | 8,563 | 24,914 | 72,895 |
| Georgia | 10,097,343 | 38,097 | 580 | 2,159 | 3,048 | 12,417 | 22,052 |

| | state | population | violent_crime_total | murder_and_nonnegligent_manslaughter | legacy_rape_1 | revised_rape_2 | robbery | aggravated_assault |
|---|---|---|---|---|---|---|---|---|
| 1 | alabama | 4849377 | 20727 | 276 | 1436 | 2005 | 4701 | 13745 |
| 2 | alaska | 736732 | 4684 | 41 | 555 | 771 | 629 | 3243 |
| 3 | arizona | 6731484 | 26916 | 319 | 2464 | 3378 | 6249 | 16970 |
| 4 | arkansas | 2966369 | 14243 | 165 | 1182 | 1763 | 2050 | 10265 |
| 5 | california | 38802500 | 153709 | 1699 | 8398 | 11527 | 48680 | 91803 |
| 6 | colorado | 5355866 | 16554 | 151 | 2121 | 3039 | 3039 | 10325 |
| 7 | connecticut | 3596677 | 8522 | 86 | 571 | 782 | 3159 | 4495 |
| 8 | delaware | 935614 | 4576 | 54 | 249 | 386 | 1269 | 2867 |
| 9 | florida | 19893297 | 107521 | 1149 | 6051 | 8563 | 24914 | 72895 |
| 10 | georgia | 10097343 | 38097 | 580 | 2159 | 3048 | 12417 | 22052 |
| 11 | hawaii | 1419561 | 3680 | 26 | 314 | 445 | 1107 | 2102 |
| 12 | idaho | 1634464 | 3468 | 32 | 468 | 609 | 204 | 2623 |
| 13 | illinois | 12880580 | 47663 | 685 | 3081 | 4159 | 15299 | 27520 |
| 14 | indiana | 6596855 | 24099 | 330 | 1615 | 2186 | 6897 | 14686 |
| 15 | iowa | 3107126 | 8497 | 60 | 828 | 1128 | 1045 | 6264 |
| 16 | kansas | 2904021 | 10123 | 91 | 1075 | 1411 | 1362 | 7259 |
| 17 | kentucky | 4413457 | 9340 | 160 | 883 | 1440 | 3336 | 4404 |
| 18 | louisiana | 4649676 | 23934 | 477 | 992 | 1375 | 5695 | 16387 |
| 19 | maine | 1330089 | 1700 | 21 | 360 | 485 | 304 | 890 |
| 20 | maryland | 5976407 | 26661 | 365 | 1144 | 1619 | 9544 | 15133 |

https://ucrdatatool.gov/Search/Crime/State/OneYearofData.cfm

London Business School

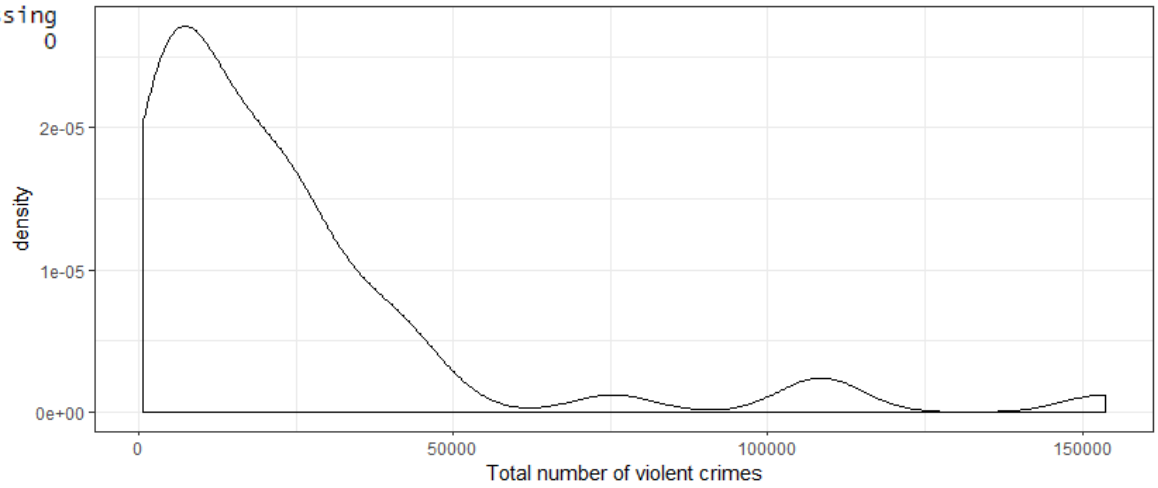
# Exploratory Data Analysis

```
> favstats(~violent_crime_total, data=us_crime)
 min      Q1 median       Q3    max     mean       sd  n missing
 622 5353.25  16042 26807.25 153709 23795.76 29960.49 50       0
```

```
#which are the top violent states?
us_crime %>%
  arrange(desc(violent_crime_total))
```

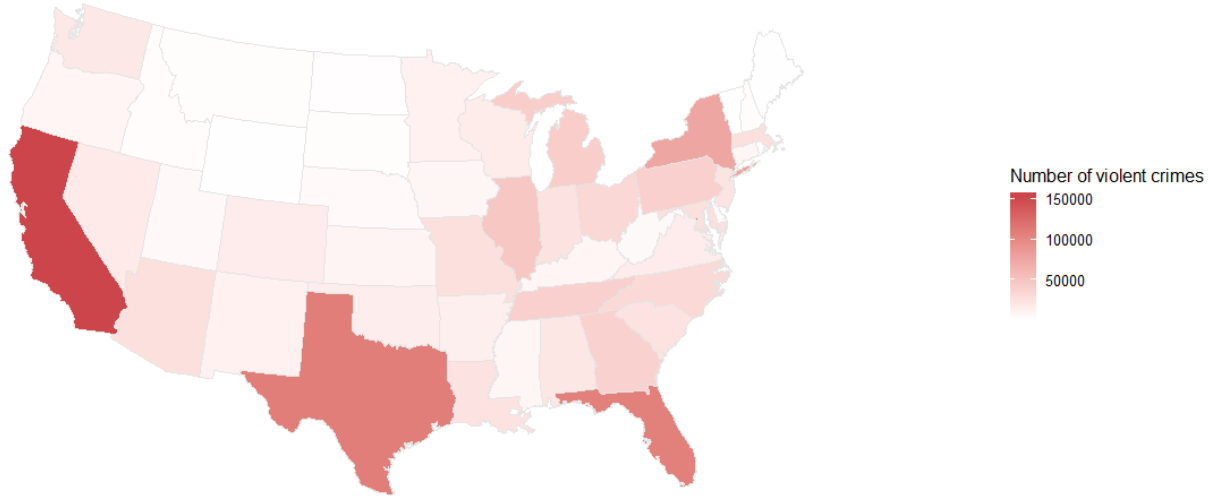
| | state | population | violent_crime_total |
|---|---|---|---|
| 51 | California | 38802500 | 153709 |
| 50 | Texas | 26956958 | 109414 |
| 49 | Florida | 19893297 | 107521 |
| 48 | New York | 19746227 | 75398 |
| 47 | Illinois | 12880580 | 47663 |
| 46 | Michigan | 9909877 | 42348 |
| 45 | Pennsylvania | 12787209 | 40164 |
| 44 | Tennessee | 6549352 | 39848 |
| 43 | Georgia | 10097343 | 38097 |
| 42 | Ohio | 11594163 | 33030 |
| 41 | North Carolina | 9943964 | 32767 |
| 40 | Arizona | 6731484 | 26916 |
| 39 | Missouri | 6063589 | 26856 |
| 38 | Maryland | 5976407 | 26661 |
| 37 | Massachusetts | 6745408 | 26399 |
| 36 | Indiana | 6596855 | 24099 |
| 35 | South Carolina | 4832482 | 24052 |
| 34 | Louisiana | 4649676 | 23934 |
| 33 | New Jersey | 8938175 | 23346 |

Crime in the USA (2014)




Source: FBI Uniform Crime Reporting, https://www.fbi.gov/services/cjis/ucr/

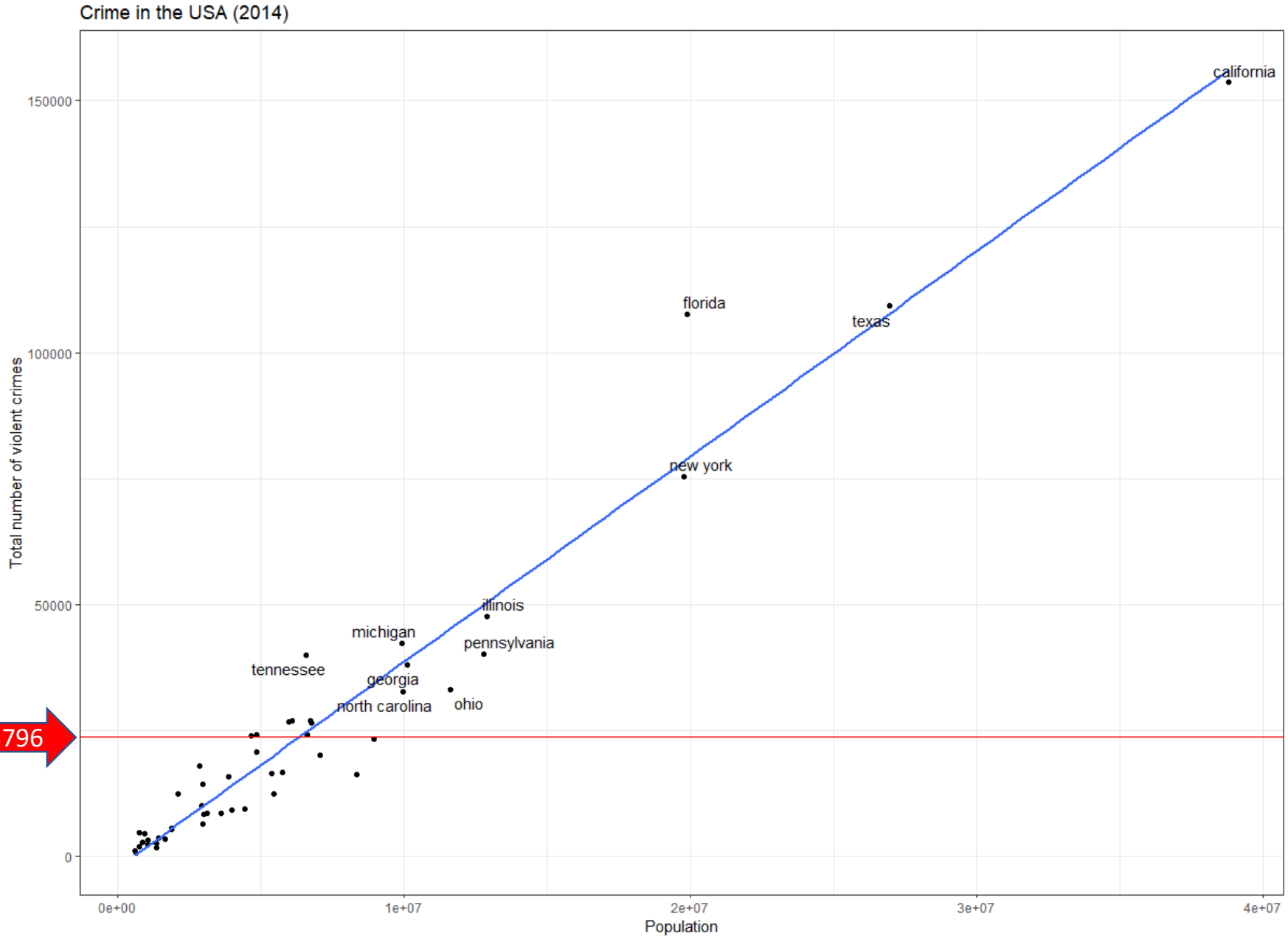
Crime in the US, 2014




Source: FBI Uniform Crime Reporting, https://www.fbi.gov/services/cjis/ucr/

London Business School

# Relationship with population?



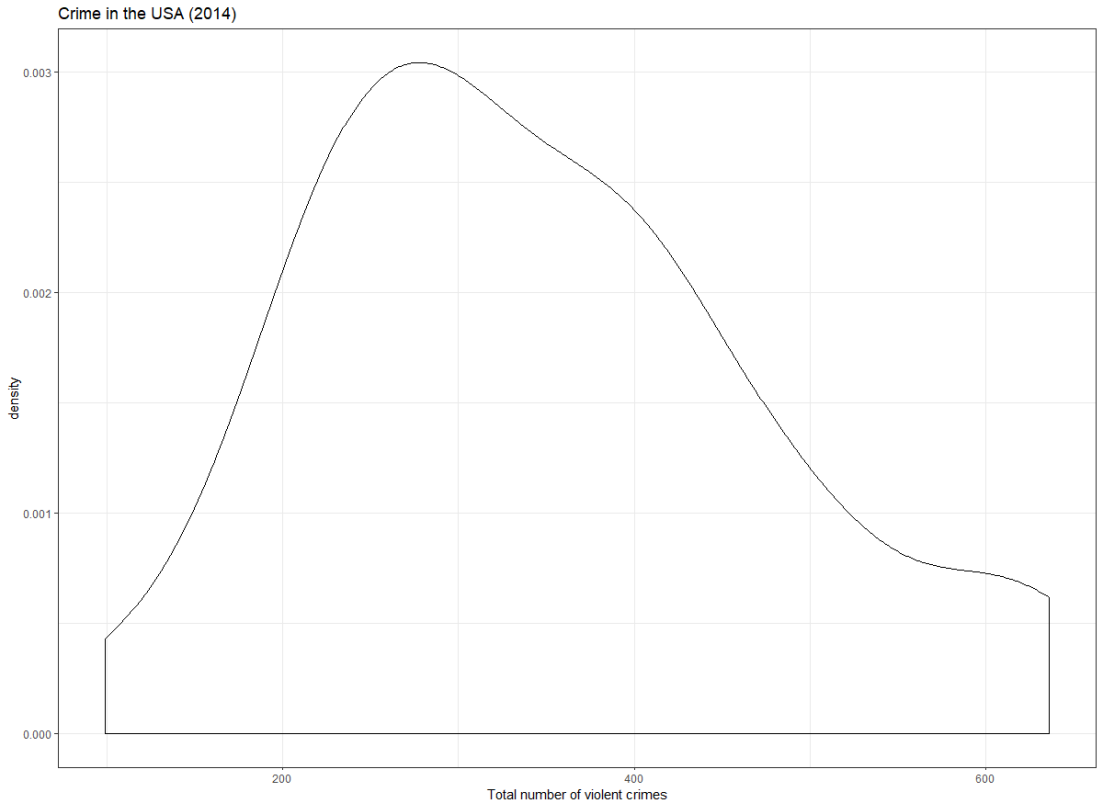
London Business School

# Define crime ***rate*** per 100K

```
#calculate violent per-capita violent crime rate (per 100K population)
us_crime <- us_crime %>%
  mutate(violent_crime_rate = 100000 * violent_crime_total / population)

us_crime %>%
  select(state, population, violent_crime_total, violent_crime_rate) %>%
  arrange(desc(violent_crime_rate))
```

```
> favstats(~violent_crime_rate, data=us_crime)
      min       Q1   median       Q3      max     mean       sd  n missing
 99.2719 259.7249 325.1066 421.9927 635.7807 346.8088 128.8192 50       0
```



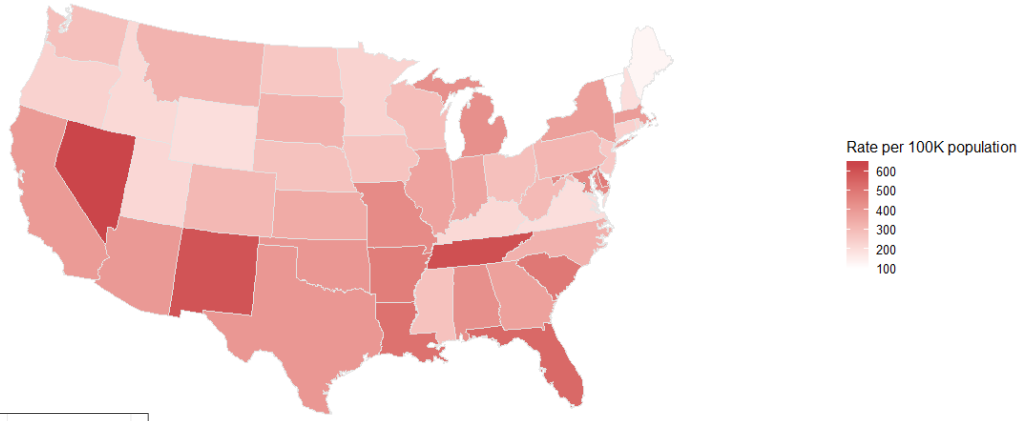

Source: FBI Uniform Crime Reporting, https://www.fbi.gov/services/cjis/ucr/

| | state | population | violent_crime_total | violent_crime_rate |
|---|---|---|---|---|
| 1 | alaska | 736732 | 4684 | 635.7807 |
| 2 | nevada | 2839099 | 18045 | 635.5890 |
| 3 | tennessee | 6549352 | 39848 | 608.4266 |
| 4 | new mexico | 2085572 | 12459 | 597.3901 |
| 5 | florida | 19893297 | 107521 | 540.4886 |
| 6 | louisiana | 4649676 | 23934 | 514.7455 |
| 7 | south carolina | 4832482 | 24052 | 497.7153 |
| 8 | delaware | 935614 | 4576 | 489.0906 |
| 9 | arkansas | 2966369 | 14243 | 480.1493 |
| 10 | maryland | 5976407 | 26661 | 446.1042 |
| 11 | missouri | 6063589 | 26856 | 442.9060 |
| 12 | alabama | 4849377 | 20727 | 427.4157 |
| 13 | michigan | 9909877 | 42348 | 427.3312 |
| 14 | oklahoma | 3878051 | 15744 | 405.9771 |
| 15 | texas | 26956958 | 109414 | 405.8841 |
| 16 | arizona | 6731484 | 26916 | 399.8524 |
| 17 | california | 38802500 | 153709 | 396.1317 |
| 18 | massachusetts | 6745408 | 26399 | 391.3625 |
| 19 | new york | 19746227 | 75398 | 381.8350 |
| 20 | georgia | 10097343 | 38097 | 377.2973 |
| 21 | illinois | 12880580 | 47663 | 370.0377 |
| 22 | indiana | 6596855 | 24099 | 365.3104 |
| 23 | kansas | 2904021 | 10123 | 348.5856 |
| 24 | north carolina | 9943964 | 32767 | 329.5165 |

# Visualise crime rate per 100K

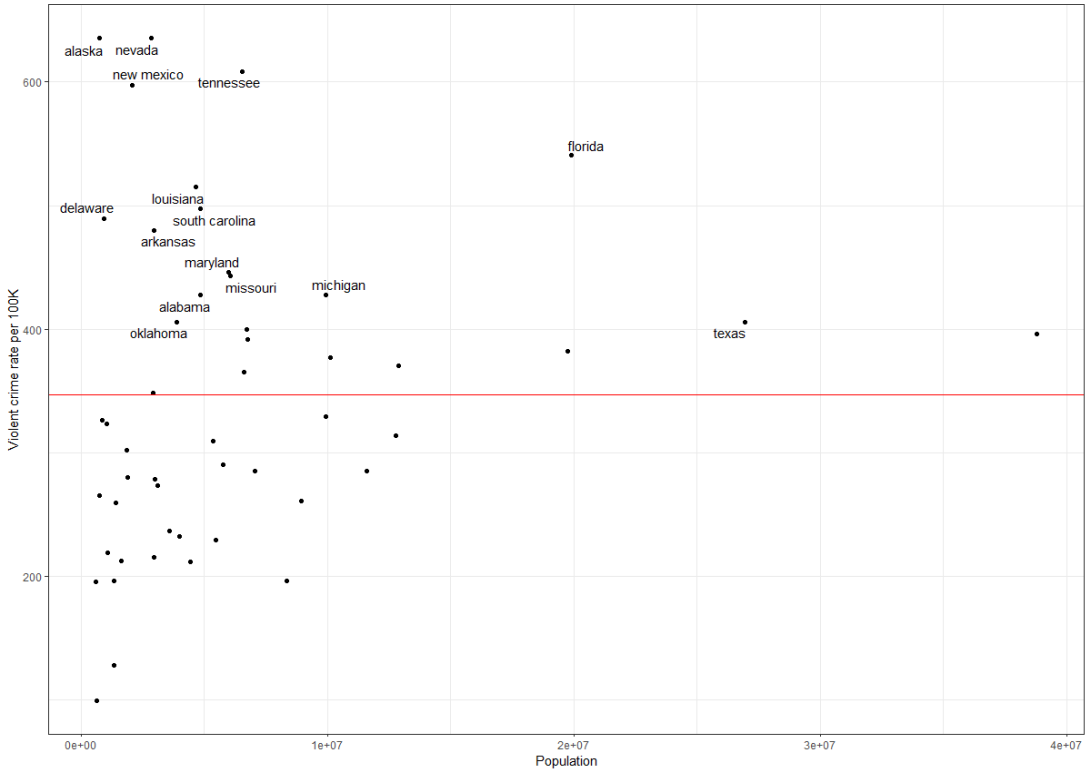
| | state | population | violent_crime_total | violent_crime_rate |
|---|---|---|---|---|
| 1 | alaska | 736732 | 4684 | 635.7807 |
| 2 | nevada | 2839099 | 18045 | 635.5890 |
| 3 | tennessee | 6549352 | 39848 | 608.4266 |
| 4 | new mexico | 2085572 | 12459 | 597.3901 |
| 5 | florida | 19893297 | 107521 | 540.4886 |
| 6 | louisiana | 4649676 | 23934 | 514.7455 |
| 7 | south carolina | 4832482 | 24052 | 497.7153 |
| 8 | delaware | 935614 | 4576 | 489.0906 |
| 9 | arkansas | 2966369 | 14243 | 480.1493 |

Crime in the US, 2014




Source: FBI Uniform Crime Reporting, https://www.fbi.gov/services/cjis/ucr/

Crime in the USA (2014)




Mean=347

Source: FBI Uniform Crime Reporting, https://www.fbi.gov/services/cjis/ucr/

London Business School

# Z-score

A ***Z score,*** or the number of standard deviations the observation falls above or below the mean, creates a common scale so you can assess data without worrying about the specific units in which it was measured.

$$Z = \frac{(observation - mean)}{SD}$$

Observations with a Z score > 2 or < -2, are usually considered unusual.

```
#calculate Z scores for violent per-capita violent crime rate (per 100K population)
us_crime <- us_crime %>%
  mutate(
  violent_crime_rate_Z =
    (violent_crime_rate - mean(violent_crime_rate)) /
    sd(violent_crime_rate)
)
```

| | state | population | violent_crime_total | violent_crime_rate | violent_crime_rate_Z |
|---|---|---|---|---|---|
| 1 | vermont | 626562 | 622 | 99.2719 | -1.92158395 |
| 2 | maine | 1330089 | 1700 | 127.8110 | -1.70004010 |
| 3 | wyoming | 584153 | 1142 | 195.4967 | -1.17460798 |
| 4 | new hampshire | 1326813 | 2602 | 196.1090 | -1.16985485 |
| 5 | virginia | 8326289 | 16340 | 196.2459 | -1.16879236 |
| 6 | kentucky | 4413457 | 9340 | 211.6255 | -1.04940331 |
| 7 | idaho | 1634464 | 3468 | 212.1797 | -1.04510146 |
| 8 | utah | 2942902 | 6346 | 215.6375 | -1.01825888 |
| 9 | rhode island | 1055173 | 2313 | 219.2058 | -0.99055903 |
| 10 | minnesota | 5457173 | 12505 | 229.1479 | -0.91337962 |

| | state | population | violent_crime_total | violent_crime_rate | violent_crime_rate_Z |
|---|---|---|---|---|---|
| 1 | alaska | 736732 | 4684 | 635.7807 | 2.24323717 |
| 2 | nevada | 2839099 | 18045 | 635.5890 | 2.24174864 |
| 3 | tennessee | 6549352 | 39848 | 608.4266 | 2.03089213 |
| 4 | new mexico | 2085572 | 12459 | 597.3901 | 1.94521749 |
| 5 | florida | 19893297 | 107521 | 540.4886 | 1.50350156 |
| 6 | louisiana | 4649676 | 23934 | 514.7455 | 1.30366296 |
| 7 | south carolina | 4832482 | 24052 | 497.7153 | 1.17145989 |
| 8 | delaware | 935614 | 4576 | 489.0906 | 1.10450816 |
| 9 | arkansas | 2966369 | 14243 | 480.1493 | 1.03509857 |
| 10 | maryland | 5976407 | 26661 | 446.1042 | 0.77081226 |

London Business School

# Why you should always visualize your data

- *bakers_dozen* is a dataframe with 13 datasets, each with n=142 datapoints, and all datasets have the same mean and standard deviation for both x and y.

| id | n | mean_x | mean_y | sd_x | sd_y | correlation |
|---|---|---|---|---|---|---|
| 1 | 142 | 54.3 | 47.8 | 16.8 | 26.9 | -0.064 |
| 2 | 142 | 54.3 | 47.8 | 16.8 | 26.9 | -0.069 |
| 3 | 142 | 54.3 | 47.8 | 16.8 | 26.9 | -0.068 |
| 4 | 142 | 54.3 | 47.8 | 16.8 | 26.9 | -0.064 |
| 5 | 142 | 54.3 | 47.8 | 16.8 | 26.9 | -0.060 |
| 6 | 142 | 54.3 | 47.8 | 16.8 | 26.9 | -0.062 |
| 7 | 142 | 54.3 | 47.8 | 16.8 | 26.9 | -0.069 |
| 8 | 142 | 54.3 | 47.8 | 16.8 | 26.9 | -0.069 |
| 9 | 142 | 54.3 | 47.8 | 16.8 | 26.9 | -0.069 |
| 10 | 142 | 54.3 | 47.8 | 16.8 | 26.9 | -0.063 |
| 11 | 142 | 54.3 | 47.8 | 16.8 | 26.9 | -0.069 |
| 12 | 142 | 54.3 | 47.8 | 16.8 | 26.9 | -0.067 |
| 13 | 142 | 54.3 | 47.8 | 16.8 | 26.9 | -0.066 |

- Would you expect any differences among the 13 datasets? After all, they all have the same mean, SD, and correlation(x, y).

London Business School

# Statistics on youtube

In early 2011, the BBC aired a documentary "The Joy of Stats" with Hans Rosling http://www.bbc.co.uk/programmes/b00wgq0l

You may want to watch the following:

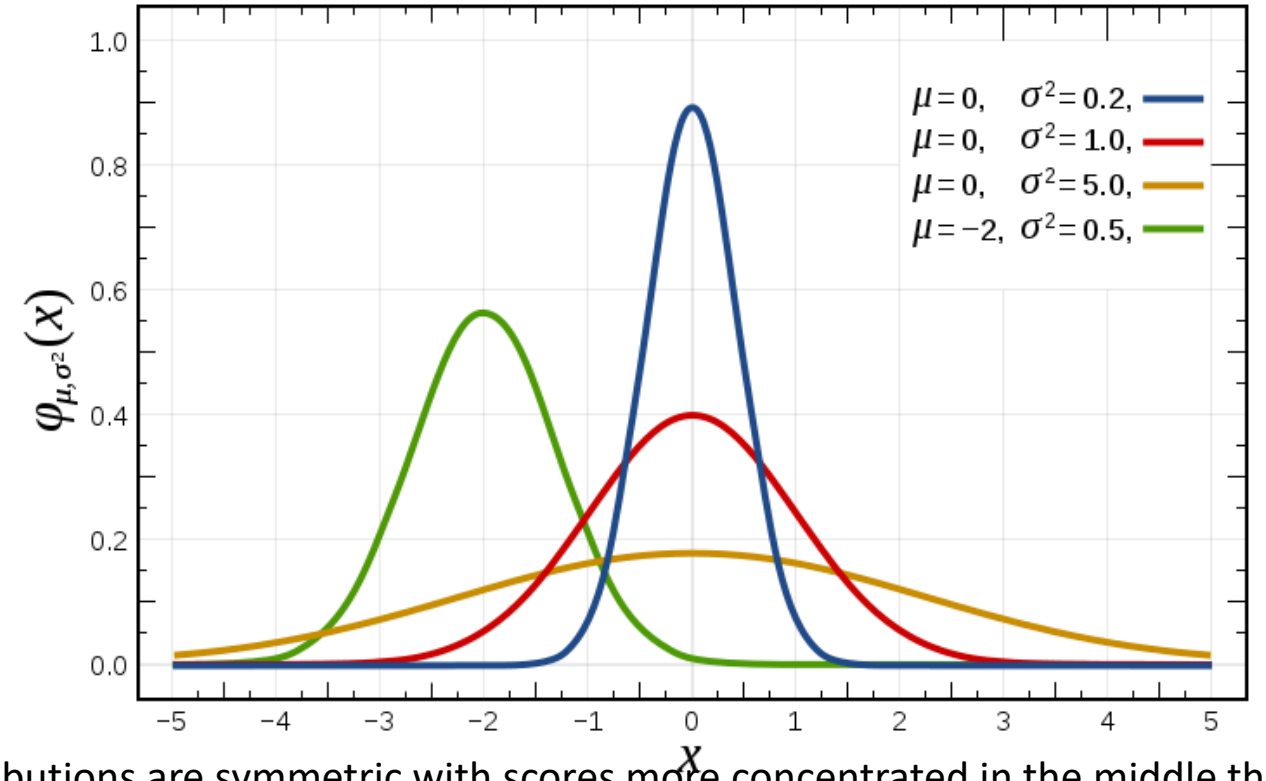
- 200 hundred years in four minutes: http://www.youtube.com/watch?v=jbkSRLYSojo
- Averages: http://www.youtube.com/watch?v=hUGUWr-TjR8
- Correlation: http://www.youtube.com/watch?v=6RzDMEW5omc
- Communicating findings: http://www.youtube.com/watch?v=yhX0OR1_Vfc
- Crime-spotting: http://www.youtube.com/watch?v=en2ix9f8ceM
- Google's automatic translation http://www.youtube.com/watch?v=AEac-jP5Eho

London Business School

# Contents

- Descriptive Statistics
- Exploratory Data Analysis
- Normal Distribution
- Reproducibility and RMarkdown

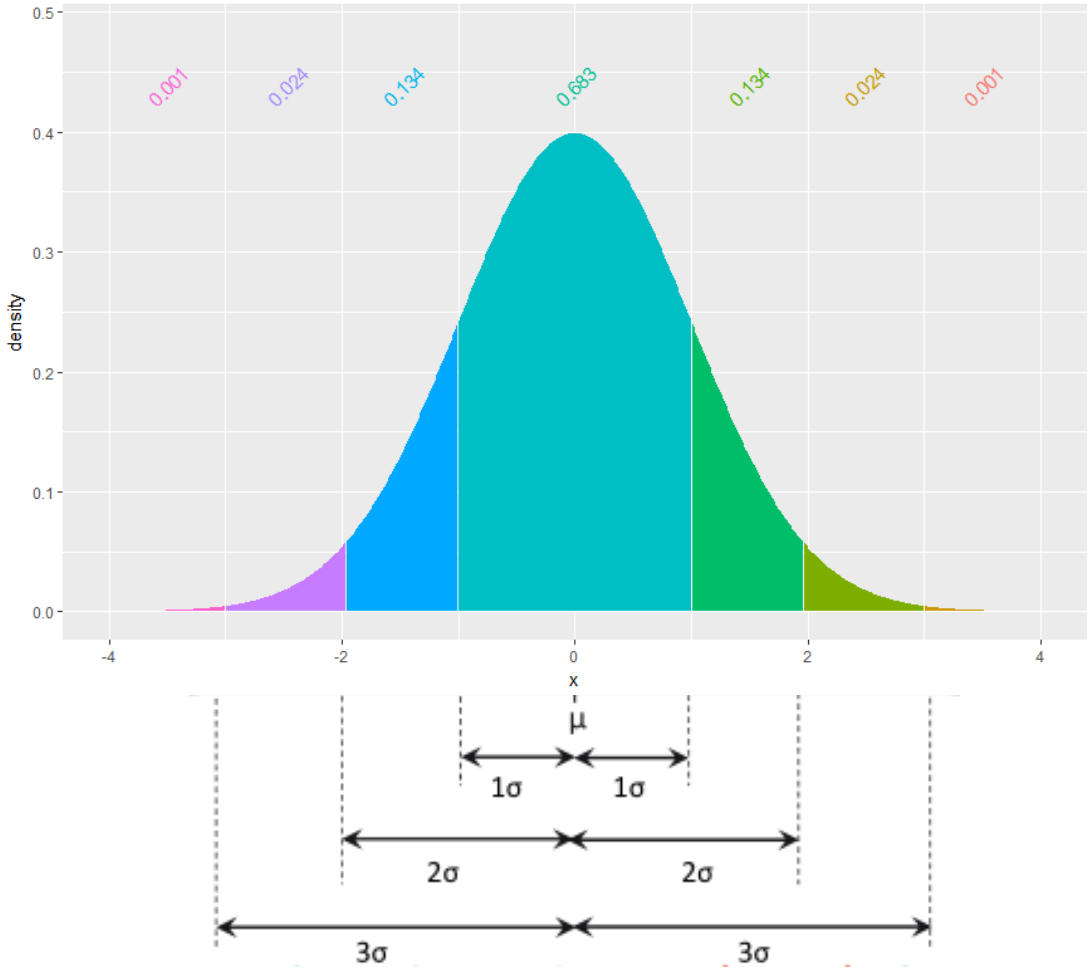
# Properties of the Normal Distribution




Normal distributions are symmetric with scores more concentrated in the middle than in the tails. The height of the distribution can be specified mathematically in terms of two parameters: the mean  and the standard deviation
The main difference in example above is how spread out they are.

68% chance of  being within (+/-) one standard-deviation of the mean
95% chance of being within  (+/-) two standard-deviations of the mean
99.7% chance of being within (+/-) three  standard-deviations from the mean

# Properties of the Normal Distribution




**95% chance of being within (+/-) 1.96 standard-deviations of the mean**

A ***Z score,*** or the number of standard deviations the observation falls above or below the mean, creates a common scale so you can assess data without worrying about the specific units in which it was measured.

$$Z = \frac{(observation - mean)}{SD}$$

Z distribution (also called the standardi**Z**ed normal distribution, is a special case of the normal distribution where μ = 0 and σ = 1
$Z \sim N(\mu = 0, \sigma = 1)$

Observations with a Z score > 2 or < -2, are usually considered unusual.

# Normal distribution and hiring policy

ZTel's personnel department is reconsidering their hiring policy.

- Currently all applicants take a standardized test.
- Test scores are normally distributed with a **mean= 525** and **SD= 55**.

- People are automatically classified as
  - Accepted, if their exam score > = 600
  - Rejected, if their exam score < = 425

- The rest go through a second phase where they assess their previous job experience, special talents and other factors as hiring criteria.

- ZTel's personnel manager want to calculate the percentage of applicants who are automatically accepted and rejected, given the current policy.

- She also wants to know how to change the standards in order to automatically reject 10% of all applicants and automatically accept 15% of all applicants.

London Business School

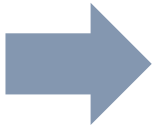
# Normal distribution caution

- One must be cautious not to model *every* continuous random variable with a normal distribution.
- This can be dangerous for at least two reasons.
  - First, not all random variables have a symmetric distribution. Some are skewed left or right and the normal distribution can be a poor approximation.
  - Second, many random variables in real applications must be nonnegative, and the normal distribution allows the possibility of negative values.

London Business School

# Contents

- Descriptive Statistics
- Exploratory Data Analysis
- Normal Distribution
- Reproducibility and RMarkdown



London Business School

# Literate Programming and Rmarkdown (Rmd)

*"You are always working with at least one collaborator: Future you." -Hadley Wickham*

Literate programming is a technique that mixes text and chunks of code together. This makes documentation of code easier, and/or the production of a full written report that incorporates the code and results used to create it.

*Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to humans what we want the computer to do. - Donald E. Knuth, Literate Programming, 1984*

**Rmarkdown**: A powerful package for literate programming, reproducible analysis, and document generation, which can (often working with *knitr*):
Combine R code and Markdown syntax
Produce documents in PDF , Microsoft Word and various types of HTML documents
In HTML format, it can incorporate "extras" like interactive graphics

**Markdown** is a lightweight markup language that allows you to quickly write and format text, which is then converted to different types of output.

https://www.markdowntutorial.com/

# Anatomy of Rmarkdown documents

Anatomy of an RMarkdown Document

1. YAML (front matter)
    - YAML = YAML Ain't No Markup Language (not helpful)
    - aka "R Markdown front-matter"
    - You can only have one YAML front-matter for each .Rmd file
    - YAML is the "metadata" - the data about your document - that goes at the top of your file and tells RMarkdown things like titles and document options.

2. Markdown text

3. R code
    - To insert a new chunk of R code, type **Ctrl+Alt+I** (Windows) or **Cmd+Alt+I** (Mac)

4. Repeat 2 & 3 as needed

https://rmarkdown.rstudio.com

# Putting the R in Rmarkdown

RMarkdown documents are Markdown-formatted text, with chunks of R code placed where you need them. You designate an R code chunk using three backticks ``` and braces { }:

```
```{r chunk_name, [options]}
# R code here
```
```

Options include various things that you want to control, including things like

- Figure height and width
- Whether to echo (show) or evaluate (run) that code chunk in the final document
- Whether to show warnings and messages

- **Be careful to end a chunk of code with three braces ```. If you accidentally delete them, the Rmd will not knit.**

# Putting the R in Rmarkdown

Our HTML document has some great features, but it's rather plain.

We can use RMarkdown's built-in themes, which are based on CSS, to make it pretty.

From the RMarkdown HTML documentation https://bookdown.org/yihui/rmarkdown/html-document.html#appearance-and-style

Table of Contents:

- **toc:** include TOC, only at top by default
- **toc_float:** make the TOC "float" on the side
- **toc_depth:** how many section levels should your TOC show?

```
---
title: "Session1_Work"
author: "Your Name Goes Here"
date: "Date goes here"
output:
  html_document:
    theme: flatly
    highlight: zenburn
    toc: yes
    toc_float: yes
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(
  message = FALSE,
  warning = FALSE,
  tidy=FALSE,     # display code as typed
  size="small")   # slightly smaller font for code
```
```

### 3.1.4 Appearance and style

There are several options that control the appearance of HTML documents:

- `theme` specifies the Bootstrap theme to use for the page (themes are drawn from the Bootswatch theme library). Valid themes include default, cerulean, journal, flatly, darkly, readable, spacelab, united, cosmo, lumen, paper, sandstone, simplex, and yeti. Pass `null` for no theme (in this case you can use the `css` parameter to add your own styles).
- `highlight` specifies the syntax highlighting style. Supported styles include `default`, `tango`, `pygments`, `kate`, `monochrome`, `espresso`, `zenburn`, `haddock`, and `textmate`. Pass `null` to prevent syntax highlighting.
- `smart` indicates whether to produce typographically correct output, converting straight quotes to curly quotes, `---` to em-dashes, `--` to en-dashes, and `...` to ellipses. Note that `smart` is enabled by default.

# Your first and second code chunks

In the first R code chunk, we typically include any knitr options we want to use throughout the document. More options can be found here https://yihui.name/knitr/options/

In the second chunk, we typically load packages needed for our analysis.

To insert a new chunk of R code, type **Ctrl+Alt+I** (Windows) or **Cmd+Alt+I** (Mac)

````
---
title: "Session1_Work"
author: "Your Name Goes Here"
date: "Date goes here"
output:
  html_document:
    theme: flatly
    highlight: zenburn
    toc: yes
    toc_float: yes
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(
  message = FALSE,
  warning = FALSE,
  tidy=FALSE,     # display code as typed
  size="small")   # slightly smaller font for code
```

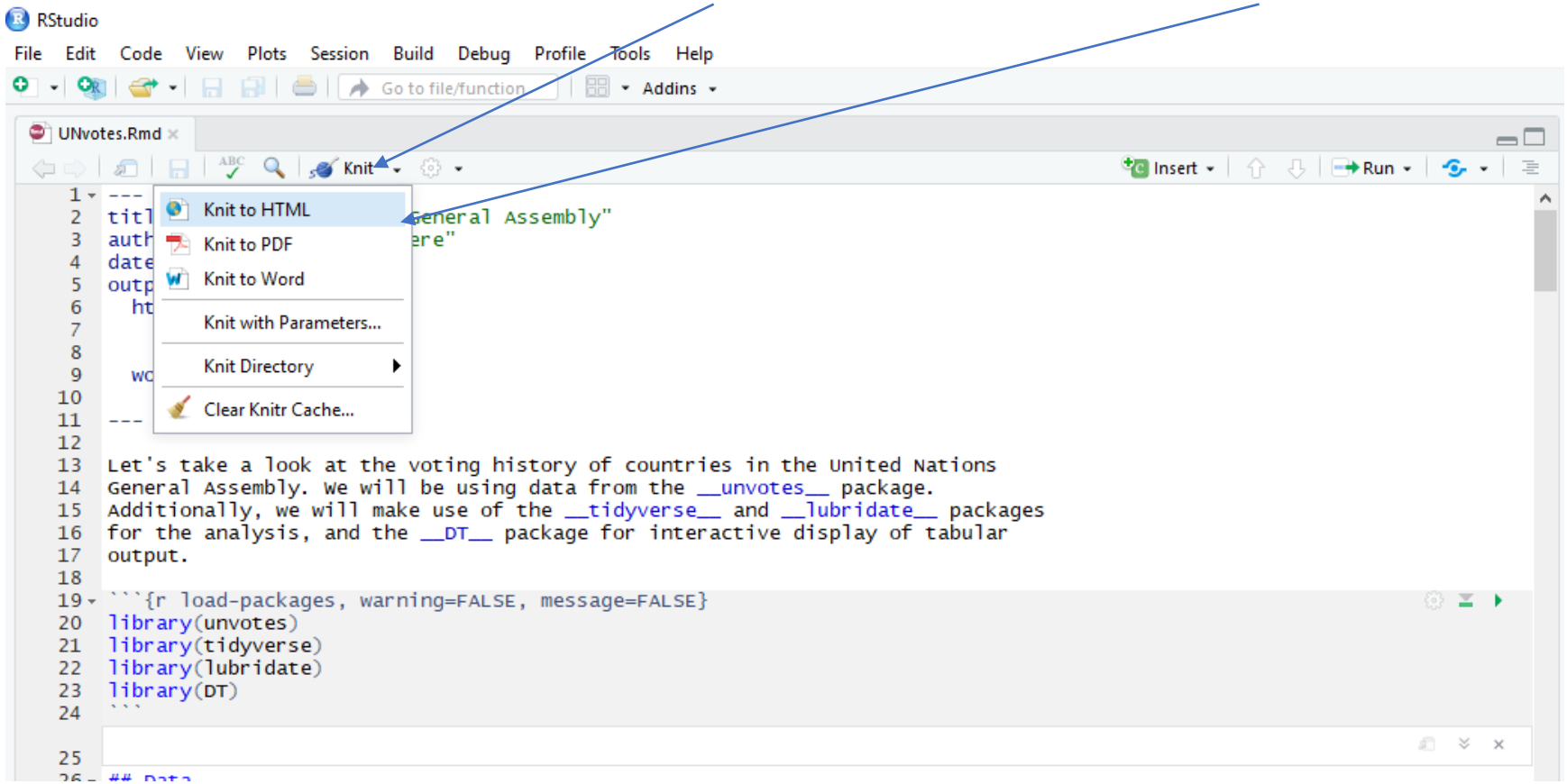

```{r load-packages, include=FALSE}
library(knitr)
library(tidyverse)
library(mosaic)
```
````

London Business School

# *UNVotes.Rmd*

Download the R Markdown file *UNvotes.Rmd* that will help us analyse voting in the United Nations (UN) General Assembly. It is an R Markdown Document (RMD) that opens in RStudio.

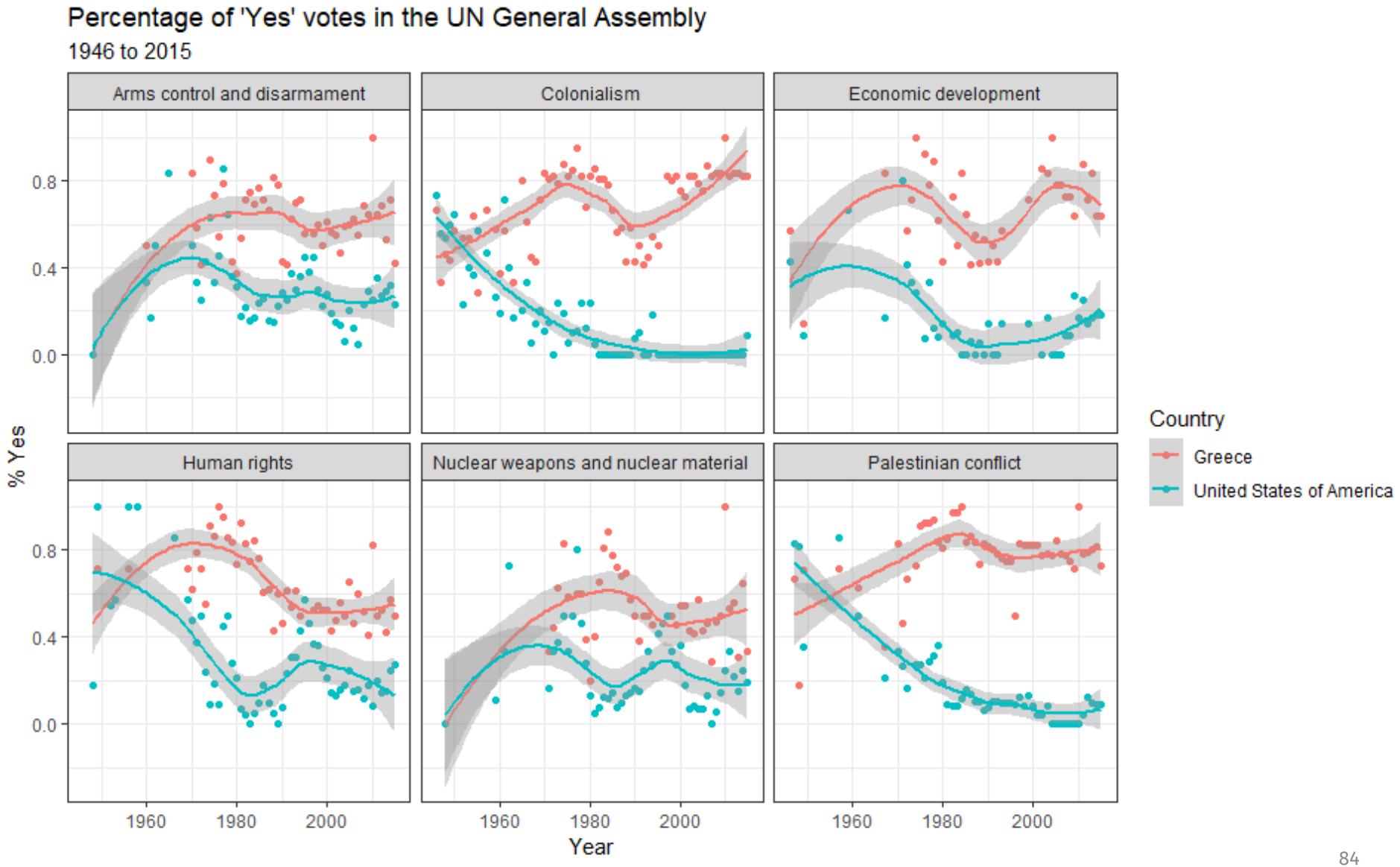
Knit the document by clicking on the *Knit* button, or by selecting *Knit to HTML*




Discuss the results with your neighbour. Then change countries plotted (line 100) and knit again

London Business School

# *UNVotes.Rmd*



London Business School

# R Markdown Cheat Sheet

## R Markdown Cheat Sheet

learn more at rmarkdown.rstudio.com

RStudio

### .Rmd files

An R Markdown (.Rmd) file is a record of your research. It contains the code that a scientist needs to reproduce your work along with the narration that a reader needs to understand your work.

### Reproducible Research

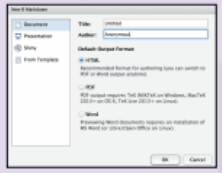
At the click of a button, or the type of a command, you can rerun the code in an R Markdown file to reproduce your work and export the results as a finished report.

### Dynamic Documents

You can choose to export the finished report as a html, pdf, MS Word, ODT, RTF, or markdown document; or as a html or pdf based slide show.

## Workflow

**1 Open a new .Rmd file** at File ▸ New File ▸ R Markdown. Use the wizard that opens to pre-populate the file with a template



**2 Write document** by editing template

**3 Knit document to create report** Use knit button or **render()** to knit

**4 Preview Output** in IDE window

**5 Publish** (optional) to web or server

Synch publish button to accounts at
- rpubs.com,
- shinyapps.io
- RStudio Connect

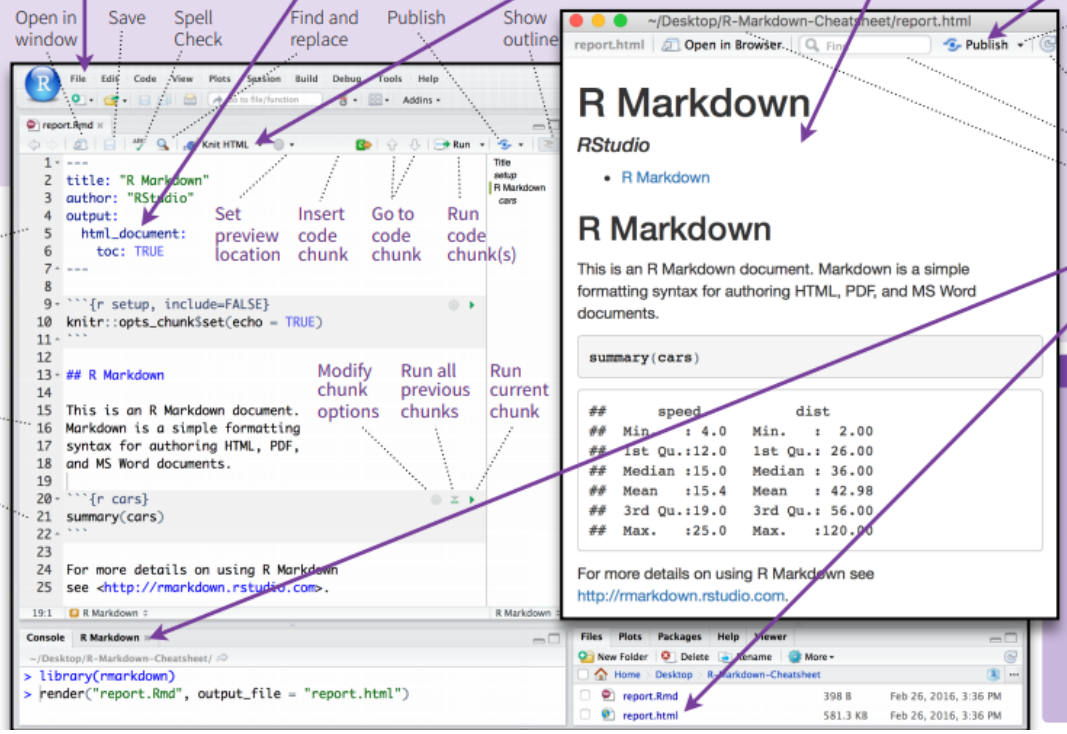
Reload document

Find in document

File path to output document

**6 Examine build log** in R Markdown console

**7 Use output file** that is saved alongside .Rmd




### .Rmd structure

**YAML Header**
Optional section of render (e.g. pandoc) options written as key:value pairs (YAML).
- At start of file
- Between lines of ---

**Text**
Narration formatted with markdown, mixed with:

**Code chunks**
Chunks of embedded code. Each chunk:
- Begins with ```{r}
- ends with ```

R Markdown will run the code and append the results to the doc.

It will use the location of the .Rmd file as the **working directory**

### render()

Use rmarkdown::**render()** to render/knit at cmd line.
Important args:

**input** - file to render

**output_format**

**output_options** - List of render options (as in YAML)

**output_file**

**output_dir**

**params** - list of params to use

**envir** - environment to evaluate code chunks in

**encoding** - of input file

## Interactive Documents

Turn your report into an interactive Shiny document in 4 steps

1 Add **runtime: shiny** to the YAML header.

2 Call Shiny **input** functions to embed input objects.

3 Call Shiny **render** functions to embed reactive output.

4 Render with rmarkdown::**run** or click **Run Document** in RStudio IDE

```
---
output: html_document
runtime: shiny
---

```{r, echo = FALSE}
numericInput("n",
  "How many cars?", 5)

renderTable({
  head(cars, input$n)
})
```
```

How many cars? 5

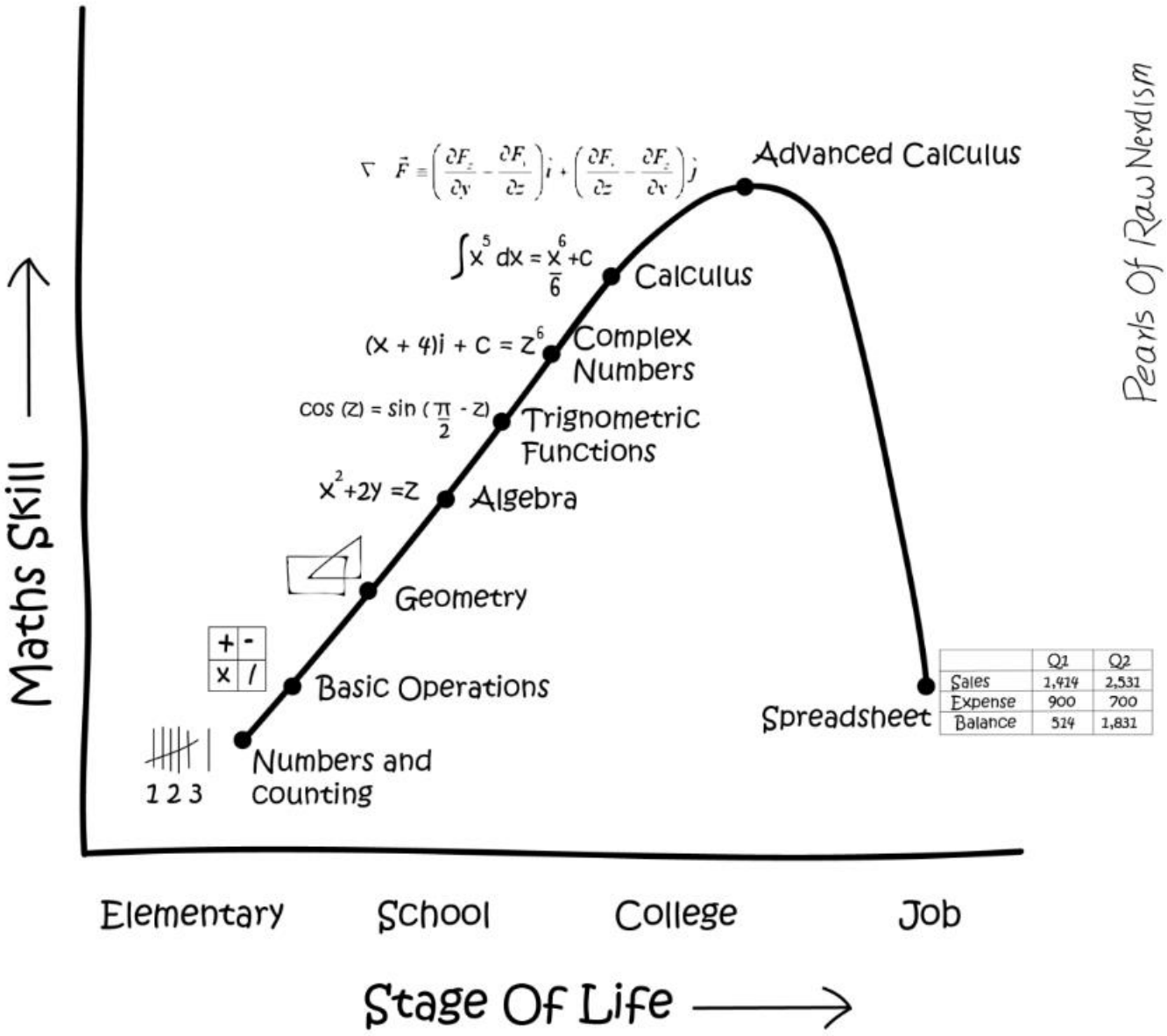
| | speed | dist |
|---|---|---|
| 1 | 4.00 | 2.00 |
| 2 | 4.00 | 10.00 |
| 3 | 7.00 | 4.00 |
| 4 | 7.00 | 22.00 |
| 5 | 8.00 | 16.00 |

Embed a complete app into your document with shiny::**shinyAppDir()**

* *Your report will rendered as a Shiny app, which means you must choose an html output format, like* ***html_document****, and serve it with an active R Session.*

https://www.rstudio.com/wp-content/uploads/2016/03/rmarkdown-cheatsheet-2.0.pdf

# Excel is ok, but....




http://pearlsofrawnerdism.com/life-and-maths/

London Business School

# Austerity and Excel

Growth in a Time of Debt
Carmen M. Reinhart and Kenneth S. Rogoff
NBER Working Paper No. 15639
January 2010, Revised January 2010
JEL No. E2,E3,E6,F3,F4,N10


**ABSTRACT**

We study economic growth and inflation at different levels of government and external debt. Our analysis is based on new data on forty-four countries spanning about two hundred years. The dataset incorporates over 3,700 annual observations covering a wide range of political systems, institutions, exchange rate arrangements, and historic circumstances. Our main findings are: First, the relationship between government debt and real GDP growth is weak for debt/GDP ratios below a threshold of 90 percent of GDP. Above 90 percent, median growth rates fall by one percent, and average growth falls considerably more. We find that the threshold for public debt is similar in advanced and emerging economies. Second, emerging markets face lower thresholds for external debt (public and private)—which is usually denominated in a foreign currency. When external debt reaches 60 percent of GDP, annual growth declines by about two percent; for higher levels, growth rates are roughly cut in half. Third, there is no apparent contemporaneous link between inflation and public debt levels for the advanced countries as a group (some countries, such as the United States, have experienced higher inflation when debt/GDP is high). The story is entirely different for emerging markets, where inflation rises sharply as debt increases.


**Debt:GDP = 90%+ → −1.0% growth**

https://www.bloomberg.com/news/articles/2013-04-18/faq-reinhart-rogoff-and-the-excel-error-that-changed-history
https://www.fa-mag.com/news/reinhart-rogoff-paper-cited-by-ryan-faulted-by-umass-economists-13993.html

# Austerity and Excel

Finally, Ms. Reinhart and Mr. Rogoff allowed researchers at the University of Massachusetts to look at their original spreadsheet — and the mystery of the irreproducible results was solved. First, they omitted some data; second, they used unusual and highly questionable statistical procedures; and finally, yes, they made an Excel coding error. Correct these oddities and errors, and you get what other researchers have found: some correlation between high debt and slow growth, with no indication of which is causing which, but no sign at all of that 90 percent "threshold."

But no, he was correct - he'd spotted a basic error in the spreadsheet. The Harvard professors had accidentally only included 15 of the 20 countries under analysis in their key calculation (of average GDP growth in countries with high public debt).

Australia, Austria, Belgium, Canada and Denmark were missing.

Oops.

> **New Zealand's single year, 1951, at -8% growth is held up with the same weight as Britain's nearly 20 years in the high public debt category at 2.5% growth**
> Prof Michael Ash

Herndon and his professors found other issues with Growth in a Time of Debt, which had an even bigger impact on the famous result. The first was the fact that for some countries, some data was missing altogether.

*"I clicked on cell L51, and saw that they had only averaged rows 30 through 44, instead of rows 30 through 49."*

https://www.nytimes.com/2013/04/19/opinion/krugman-the-excel-depression.html
http://www.bbc.com/news/magazine-22223190

# Austerity and Excel

Table 1. Real GDP Growth as the Level of Government Debt Varies:
Selected Advanced Economies, 1790-2009
(annual percent change)

| Country | Period | Central (Federal) government debt/ GDP | | | |
|---|---|---|---|---|---|
| | | Below 30 percent | 30 to 60 percent | 60 to 90 percent | 90 percent and above |
| Australia | 1902-2009 | 3.1 | 4.1 | 2.3 | 4.6 |
| Austria | 1880-2009 | 4.3 | 3.0 | 2.3 | n.a. |
| Belgium | 1835-2009 | 3.0 | 2.6 | 2.1 | 3.3 |
| Canada | 1925-2009 | 2.0 | 4.5 | 3.0 | 2.2 |
| Denmark | 1880-2009 | 3.1 | 1.7 | 2.4 | n.a. |
| Finland | 1913-2009 | 3.2 | 3.0 | 4.3 | 1.9 |
| France | 1880-2009 | 4.9 | 2.7 | 2.8 | 2.3 |
| Germany | 1880-2009 | 3.6 | 0.9 | n.a. | n.a. |
| Japan | 1885-2009 | 4.9 | 3.7 | 3.9 | 0.7 |
| Netherlands | 1880-2009 | 4.0 | 2.8 | 2.4 | 2.0 |
| New Zealand | 1932-2009 | 2.5 | 2.9 | 3.9 | **3.6** |
| Norway | 1880-2009 | 2.9 | 4.4 | n.a. | n.a. |
| Portugal | 1851-2009 | 4.8 | 2.5 | 1.4 | n.a. |
| Spain | 1850-2009 | ***1.6*** | 3.3 | ***1.3*** | 2.2 |
| Sweden | 1880-2009 | 2.9 | 2.9 | 2.7 | n.a. |
| United Kingdom | 1830-2009 | 2.5 | 2.2 | 2.1 | 1.8 |
| United States | 1790-2009 | 4.0 | 3.4 | 3.3 | ***-1.8*** |
| Average | | **3.7** | **3.0** | **3.4** | **1.7** |
| Median | | **3.9** | **3.1** | **2.8** | **1.9** |
| Number of observations = **2,317** | | 866 | 654 | 445 | 352 |

**Debt:GDP = 90%+ → 2.2% growth**

https://politicalsciencereplication.wordpress.com/2013/04/19/what-reinhart-rogoff-means-for-the-replication-debate

London Business School

# Excel and PHE

bbc.co.uk/news/uk-54412581

BBC | Sign in | Home | News | Sport | Weather | iPlayer

NEWS

Home | Coronavirus | Brexit | UK | World | Business | Politics | Tech | Science | Health | Family & Education

UK | England | N. Ireland | Scotland | Alba | Wales | Cymru | Isle of Man | Guernsey | Jersey | Local News

## Covid: 16,000 coronavirus cases missed in daily figures after IT error

5 October 2020

Coronavirus pandemic


**A technical glitch that meant nearly 16,000 cases of coronavirus went unreported has delayed efforts to trace contacts of people who tested positive.**

Public Health England said 15,841 cases between 25 September and 2 October were left out of the UK daily case figures.

bbc.co.uk/news/technology-54423988

Home | Coronavirus | Brexit | UK | World | Business | Politics | Tech | Science | Health | Family & Education

Technology

## Excel: Why using Microsoft's tool caused Covid-19 results to be lost

By Leo Kelion
Technology desk editor

5 October 2020

Coronavirus pandemic

**The badly thought-out use of Microsoft's Excel software was the reason nearly 16,000 coronavirus cases went unreported in England.**

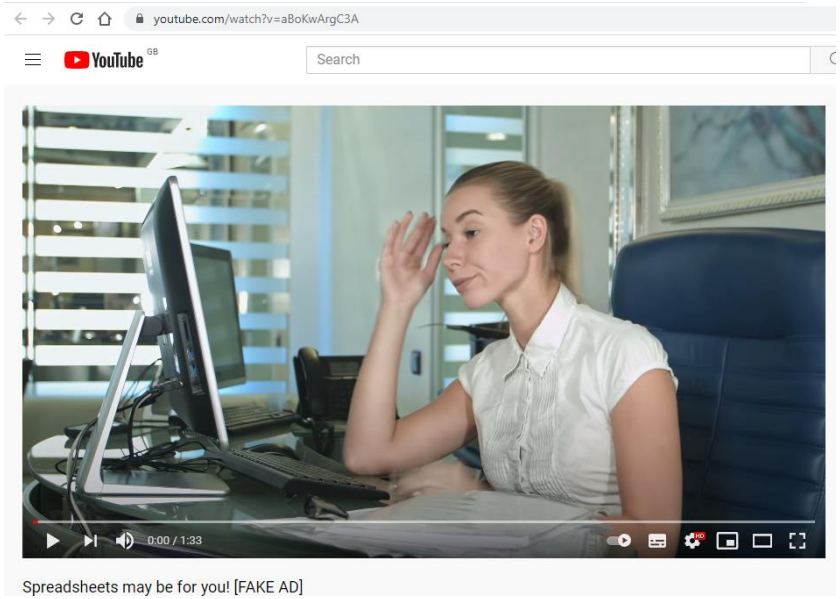
And it appears that Public Health England (PHE) was to blame, rather than a third-party contractor.

The issue was caused by the way the agency brought together logs produced by commercial firms paid to analyse swab tests of the public, to discover who has the virus.

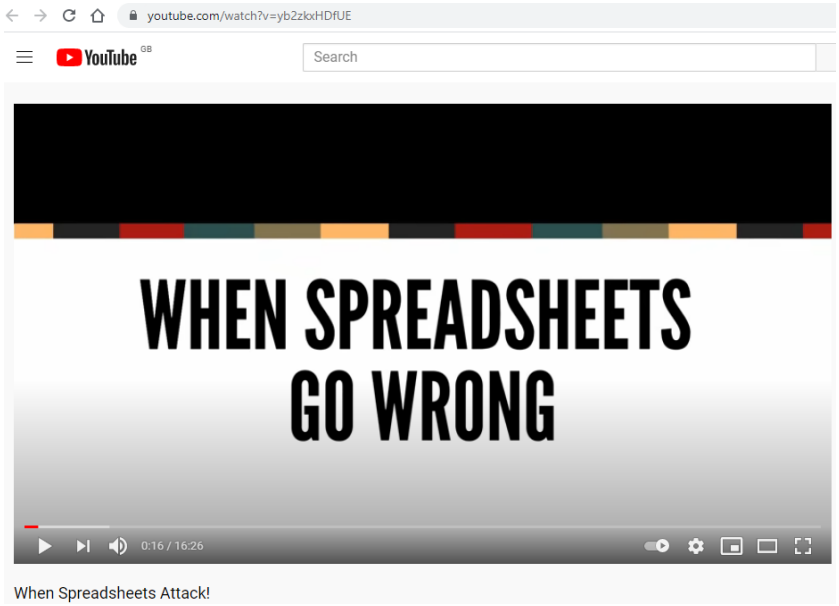
They filed their results in the form of text-based lists - known as CSV files - without issue.

PHE had set up an automatic process to pull this data together into Excel templates so that it could then be uploaded to a central system and made available to the NHS Test and Trace team, as well as other government computer dashboards.

https://www.bbc.co.uk/news/technology-54423988
https://www.youtube.com/watch?v=zUp8pkoeMss




https://www.youtube.com/watch?v=aBoKwArgC3A




https://www.youtube.com/watch?v=yb2zkxHDfUE

London Business School

# Reproducibility Guidelines

1. Don’t touch the raw data
   - If you do, explain what you did and why

2. Use self-documenting code: R Markdown

3. Ensure code is reproducible: R Markdown

4. Use open formats: Use .csv, not .xlsx

London Business School

# R Markdown in real life

## How R Helps Airbnb Make the Most of Its Data

Ricardo Bion
Airbnb
Ricardo.Bion@airbnb.com
and
Robert Chang
Airbnb
Robert.Chang@airbnb.com
and
Jason Goodman
Airbnb
Jason.Goodman@airbnb.com

August 23, 2017

**Abstract**

At Airbnb, R has been amongst the most popular tools for doing data science in many different contexts, including generating product insights, interpreting experiments, and building predictive models. Airbnb supports R usage by creating internal R tools and by creating a community of R users. At the end of the post, the authors provide some specific advice for practitioners who wish to incorporate R into their day-to-day workflow.

### 3.1.2 Data Visualization

We use ggplot2 as our main package to create ad-hoc exploratory graphics as well as polished-looking customized visualizations. When combined with tools to clean and transform data, ggplot2 allows analysts to quickly translate insights into high quality, compelling visualizations. In addition to the static graphics of `ggplot2`, we often make interactive visualizations or dashboards using R packages such as `plotly` (Sievert et al. 2017), `leaflet` (Cheng et al. 2017), `dygraphs` (Vanderkam et al. 2017), `DiagrammeR` (Sveidqvist et al. 2017), and `shiny` (Chang et al. 2017).

### 3.1.3 Reproducible Research

At Airbnb, all R analyses are documented in `rmarkdown`, where code and visualizations are combined within a single written report. Posts are carefully reviewed by experts in the content area and techniques used, both in terms of methodologies and code style, before publishing and sharing with the business partners. The peer review process is

https://dataingovernment.blog.gov.uk/2017/03/27/reproducible-analytical-pipeline/

Blog

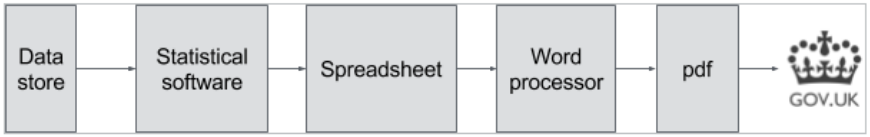
## Data in government

Organisations: Civil Service

## Reproducible Analytical Pipelines

Matt Upson, 27 March 2017 - Data science

Producing official statistics for publications is a key function of many teams across government. It's a time consuming and meticulous process to ensure that statistics are accurate and timely. With open source software becoming more widely used, there's now a range of tools and techniques that can be used to reduce production time, whilst maintaining and even improving the quality of the publications. This post is about these techniques: what they are, and how we can use them.





https://peerj.com/preprints/3182.pdf
https://gdsdata.blog.gov.uk/2017/03/27/reproducible-analytical-pipeline

# R Markdown in real life

https://www.cultureofinsight.com/blog/2018/10/22/2018-08-20-automated-data-reports-with-r/

HOME PORTFOLIO TRAINING BLOG

## AUTOMATED DATA REPORTS WITH R

BY PAUL CAMPBELL | OCTOBER 22, 2018

A lot of data analysts will find themselves doing repetitive manual tasks on a data set every day/week/month in Excel, then copying and pasting their updated pivot tables and charts into Word or PowerPoint reports for their stakeholders. If this sounds like your job description, you may want to consider switching to a programming language like R.

Writing scripts will allow you to automate the majority of these processes; from importing your data all the way through to emailing your boss the final report. They'll never know you were actually in the pub the whole time.

### Automation

Automation may sound like a scary word to any human being with a job that would like to keep it, but learning to automate some of your most common data tasks can be seriously beneficial to both your organisation and your own job security! Some of the benefits of an automated reporting workflow over a manual one include:

1. **It saves you time.** Most people will feel like they don't have the time they need to fulfill all that is asked of them at work. So if you can cut out the time taken on manual data processing and focus more analysis and insight, that can only be a good thing.
2. **It reduces errors.** When your reporting relies on manual data entry and formula with hard-coded cell references, one typo or out-of-place number can lead to results that are way off the mark. Automating the process with a script will remove the possibility of human error completely.
3. **It expands your data visualisation options.** Using an open-source software like R will allow to draw on a vast array of tools and charting libraries not available in proprietary software. For example, HTML reports with the rmarkdown package can include interactive charts, maps and tables that utilise the latest web-technologies - more

https://www.cultureofinsight.com/blog/2018/10/22/2018-08-20-automated-data-reports-with-r

London Business School

# Session Summary

We covered

- Exploratory Data Analysis
  - Shape of the distribution
  - Measures of centre and spread
- Review of the Normal Distribution
- Reproducibility and Rmarkdown
- Session 2, Workshop 1 later today

London Business School

# To do **NOW**

To do before Thursday 02 Sep

- **Individual portfolio website**
  - Make sure you sign up for Github+Netlify
  - *Happy Git with R* *https://happygitwithr.com/*
    - Chapter 6: Install Git
    - Chapter 7: Introduce yourself to Git
  - Open a pull request to add your Github username to the class list as explained in https://github.com/kostis-christodoulou/github-practice-am01-2021

London Business School

# To do

To do before Monday 06 Sep

- **Workshop: Homework1.Rmd**
  - Group submissions: you can ask for help, but all team members should understand the code
  - As usual, please remove any introductory text I have provided
  - Comment your code, as this helps you and others understand your work
  - Tell a short story  of what you can infer from your Exploratory Data Analyses. Don't just explain what's in the graph/summary tables, but seek the “so-what?” of your data
  - Upload your knitted HTML file on Canvas

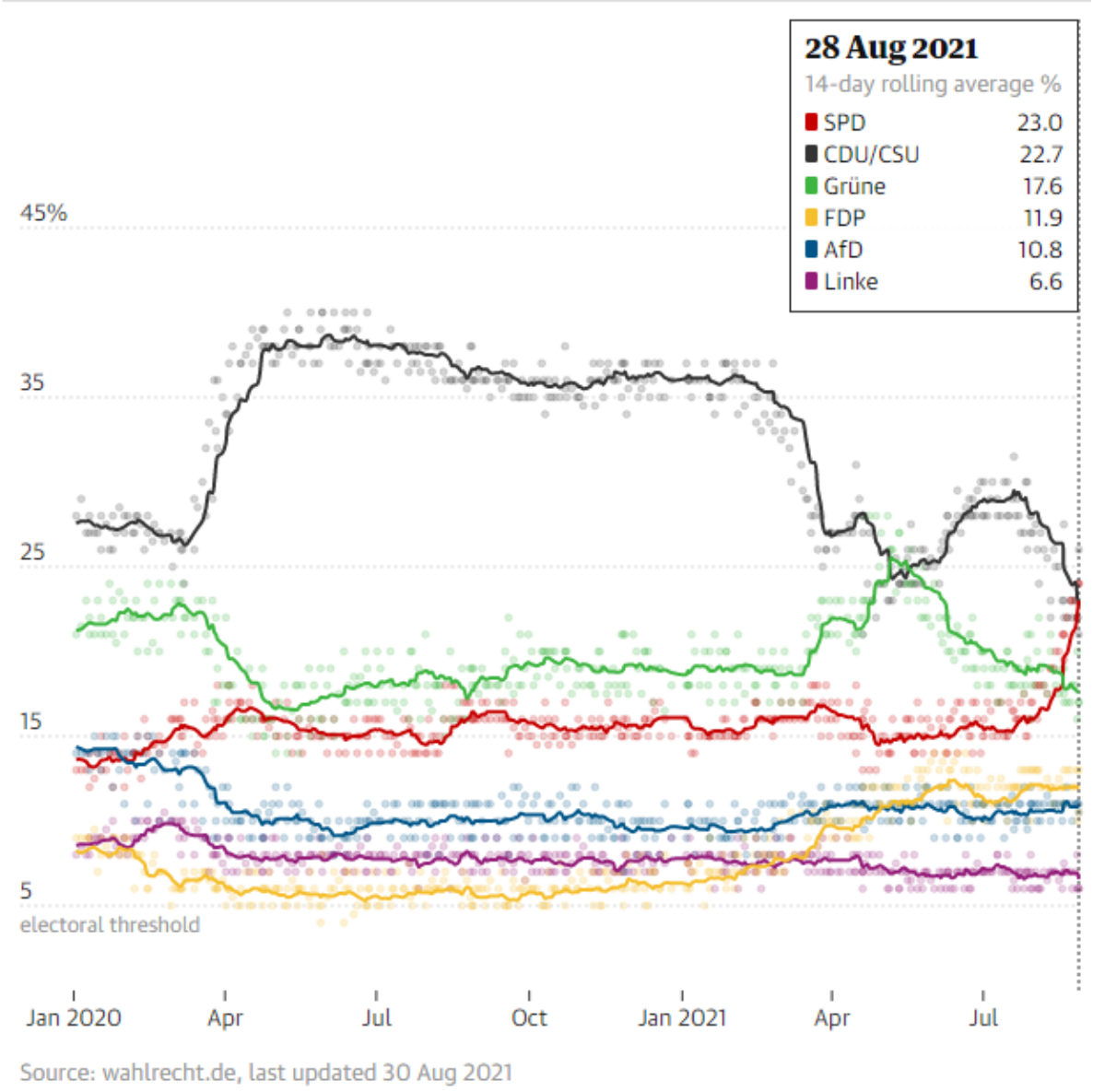
**Supporting material**

- Course website contains worked examples and interactive, online exercises

London Business School

# *And finally…*

# 2021 German Federal election



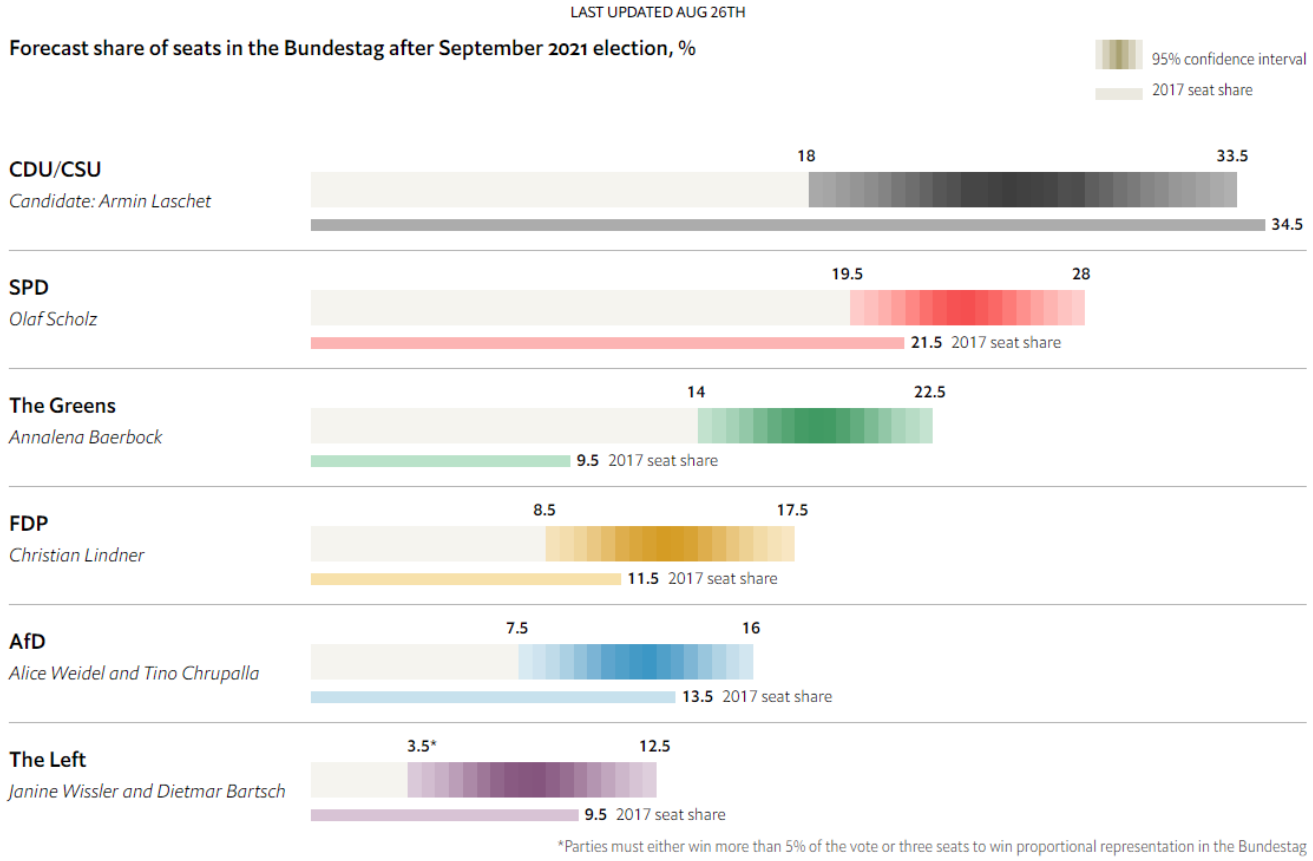

https://www.theguardian.com/world/2021/aug/20/german-election-poll-tracker-who-will-be-the-next-chancellor

# 2021 German Federal election

German election 2021

## Who will succeed Angela Merkel?

Our forecast shows who might be next into the chancellery




https://www.economist.com/graphic-detail/who-will-succeed-angela-merkel/