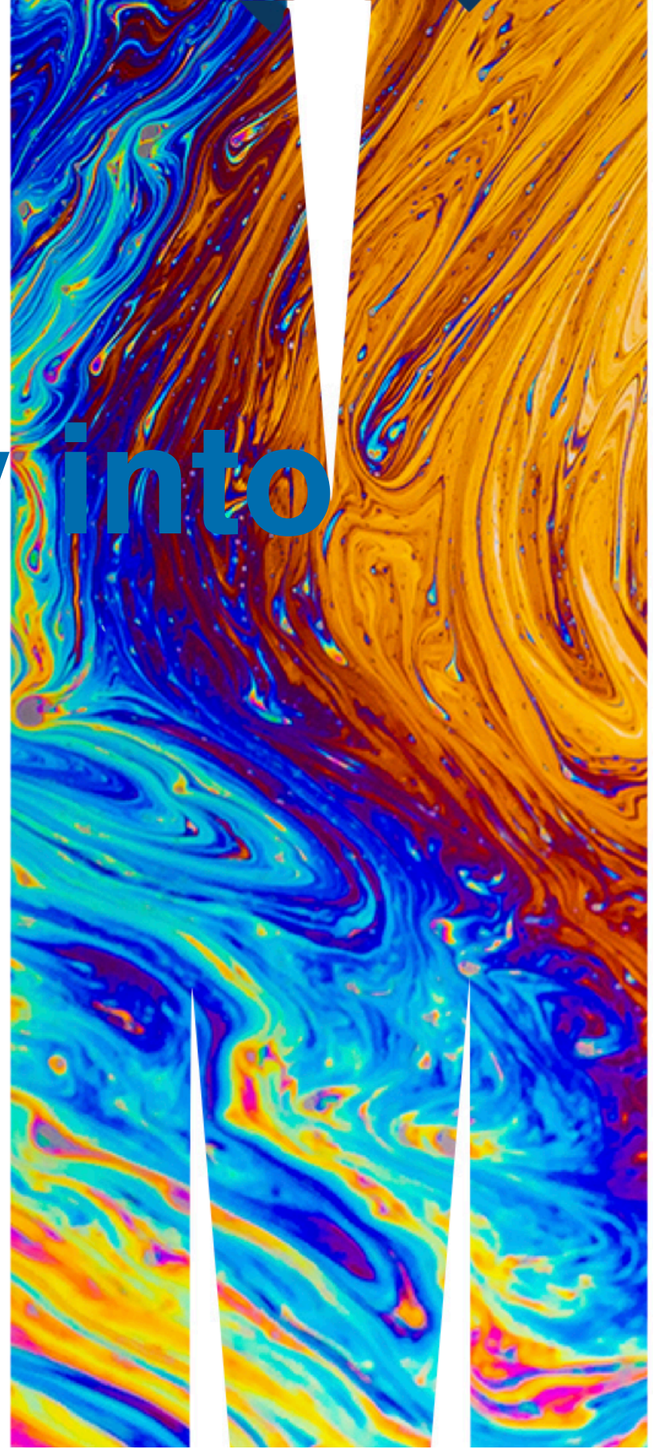
MONASH University

ETC5521: Diving Deeply into Data Exploration

Initial data analysis and model diagnostics



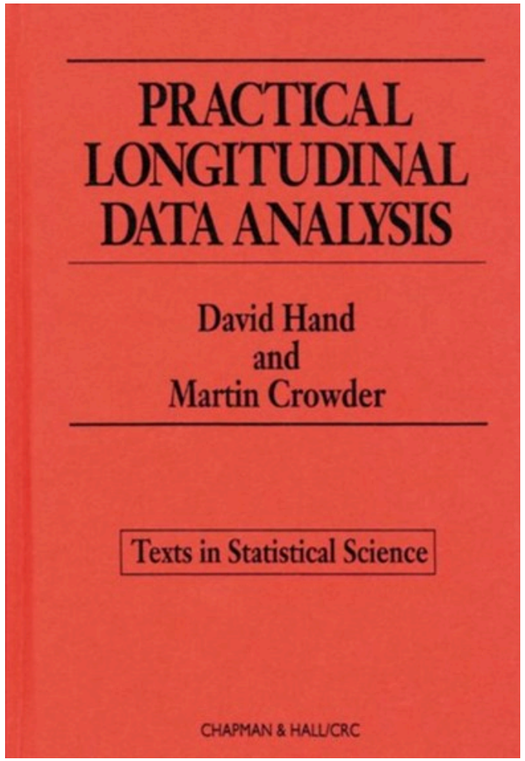
Professor Di Cook

Department of Econometrics and Business Statistics

The role of initial data analysis

The first thing to do with data is to look at them …. usually means tabulating and plotting the data in many different ways to see what's going on. With the wide availability of computer packages and graphics nowadays there is no excuse for ducking the labour of this preliminary phase, and it may save some **red faces** *later.*

Crowder, M. J. & Hand, D. J. (1990) "Analysis of Repeated Measures"



Initial Data Analysis and Confirmatory Analysis

> Prior to conducting a confirmatory data analysis, it is important to conduct an *initial data analysis (IDA)*.

- Confirmatory data analysis (CFA) is focused on statistical inference and includes procedures for:
  - hypothesis testing,
  - predictive modelling,
  - parameter estimation including uncertainty,
  - model selection.

- IDA includes:
  - describing the data and collection procedures
  - scrutinise data for errors, outliers, missing observations
  - check assumptions for confirmatory data analysis

IDA is sometimes called preliminary data analysis.

> IDA is related to exploratory data analysis (EDA) in the sense that it is primarily conducted graphically, and there are few formal tests available.

Taxonomies are useful but rarely perfect

Objectives of IDA?

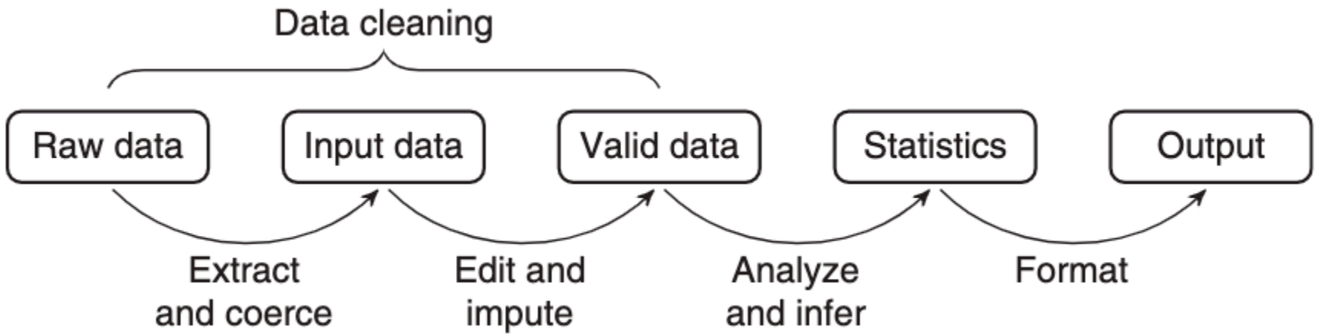
> The **main objective for IDA** is to intercept any problems in the data that might adversely affect the confirmatory data analysis.

- The role of **CFA** is to answer the intended question(s) that the data were collected for.
- ***IDA is often unreported*** in the data analysis reports or scientific papers, for various reasons. It might not have been done, or it may have been conducted but there was no space in the paper to report on it.

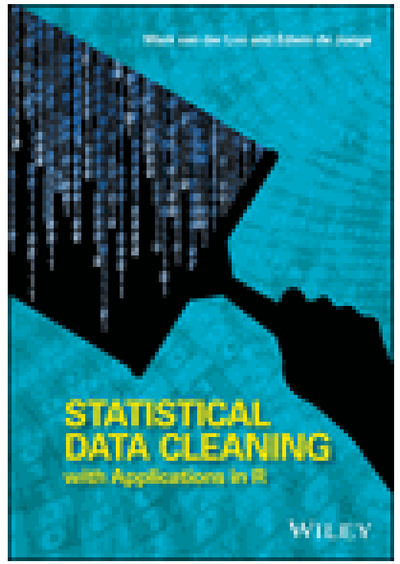
IDA in government statistics

The purpose of data cleaning is to bring data up to a level of quality such that it can reliably be used for the production of statistical models or statements.

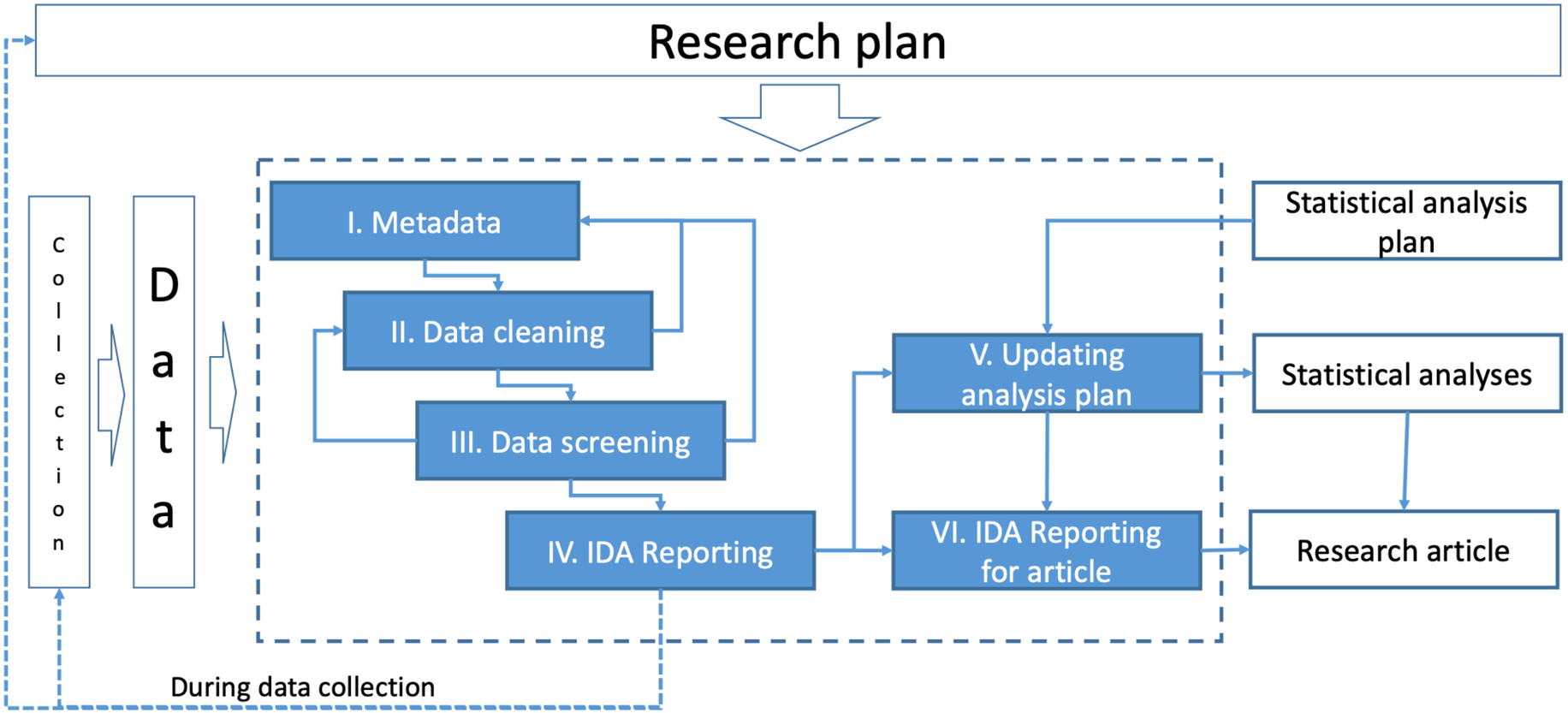
A statistical value chain is constructed by defining a number of meaningful intermediate data products, for which a chosen set of quality attributes are well described.



van der Loo & de Jonge (2018) Statistical Data Cleaning with Applications in R



IDA in health and medical data



Huebner et al (2018)'s six steps of IDA: (1) Metadata setup, (2) Data cleaning, (3) Data screening, (4) Initial reporting, (5) Refining and updating the analysis plan, (6) Reporting IDA in documentation.

Heed these words

> IDA prepares an analyst for CFA. One needs to be careful about NOT compromising the inference.

How do you compromise inference?

1. Change your inference or questions based on what you find in IDA.
2. Outlier removal or not.
3. Missing value imputation choices.
4. Treatment of zeros.
5. Handling of variable type, categorical temporal.
6. Lack of multivariate relationship checking, including subsets based on levels of categorical variables.
7. Choosing variables and observations.

How do you avoid these errors?

- Document ALL the IDA, using a reproducible analysis script.
- Pre-register your CFA plan, so that your CFA questions do not change.
- Decisions made on outlier removal, variable selection, recoding, sampling, handling of zeros have known affects on results, and are justifiable.

Insure yourself against accusations of data snooping, data dredging, data fishing.

Data screening

Data screening

> It's important to check how the data are understood by the computer.

that is, checking for *data type*:

- Was the date read in as character?
- Was a factor read in as numeric?

Also important for making inference is to know whether the data supports making broader conclusions. How was the data collected? Is it clear what the population of interest is, and that the data is a representative sample of this population?

Example: Checking the data type (1/2)

lecture3-example.xlsx

	A	B	C	D
1	id	date	loc	temp
2	1	3/1/10	New York	42
3	2	3/2/10	New York	41.4
4	3	3/3/10	New York	38.5
5	4	3/4/10	New York	41.1
6	5	3/5/10	New York	39.8

```r
library(readxl)
library(here)
df <- read_excel(here("data/lecture3-example.x
df
```

```
# A tibble: 5 × 4
     id date                loc       temp
  <dbl> <dttm>              <chr>    <dbl>
1     1 2010-01-03 00:00:00 New York  42
2     2 2010-02-03 00:00:00 New York  41.4
3     3 2010-03-03 00:00:00 New York  38.5
4     4 2010-04-03 00:00:00 New York  41.1
5     5 2010-05-03 00:00:00 New York  39.8
```

- What problems are there with the computer's interpretation of data type?
- What context specific issues indicate incorrect computer interpretation?

Example: Checking the data type (2/2)

```
library(lubridate)
df <- read_excel(here("data/lecture3-example.x
                  col_types = c("text",
                                "date",
                                "text",
                                "numeric"))

df |>
  mutate(id = as.factor(id),
         date = ydm(date)) |>
  mutate(
         day = day(date),
         month = month(date),
         year = year(date))
```

```
# A tibble: 5 × 7
  id    date       loc       temp   day month  year
  <fct> <date>     <chr>    <dbl> <int> <dbl> <dbl>
1 1     2010-03-01 New York  42       1     3  2010
2 2     2010-03-02 New York  41.4     2     3  2010
3 3     2010-03-03 New York  38.5     3     3  2010
4 4     2010-03-04 New York  41.1     4     3  2010
5 5     2010-03-05 New York  39.8     5     3  2010
```

- `id` is now a `factor` instead of `integer`
- `day`, `month` and `year` are now extracted from the `date`
- Is it okay now?

  - In the United States, it's common to use the date format MM/DD/YYYY (gasps) while the rest of the world commonly uses DD/MM/YYYY or better still YYYY/MM/DD.
  - It's highly probable that the dates are 1st-5th March and not 3rd of Jan-May.
  - You can validate interpretation of temperature using weather database.

Example: Specifying the data type with R

- You can robustify your workflow by ensuring you have a check for the expected data type in your code.

```
xlsx_df <- read_excel(here("data/lecture3-exam
                      col_types = c("text", "date",
  mutate(id = as.factor(id),
         date = as.character(date),
         date = as.Date(date, format = "%Y-%d-
```

- `read_csv` has a broader support for `col_types`

```
csv_df <- read_csv(here::here("data/lecture3-e
                   col_types = cols(
                        id = col_factor(),
                        date = col_date(format =
                        loc = col_character(),
                        temp = col_double()))
```

- The checks (or coercions) ensure that even if the data are updated, you can have some confidence that any data type error will be picked up before further analysis.

Example: Checking the data type with R

You can have a quick glimpse of the data type with:

```
1 dplyr::glimpse(xlsx_df)
```

```
Rows: 5
Columns: 4
$ id   <fct> 1, 2, 3, 4, 5
$ date <date> 2010-03-01, 2010-03-02, 2010-03-03, 2010-03-0…
$ loc  <chr> "New York", "New York", "New York", "New Yor…
$ temp <dbl> 42, 41, 38, 41, 40
```

```
1 dplyr::glimpse(csv_df)
```

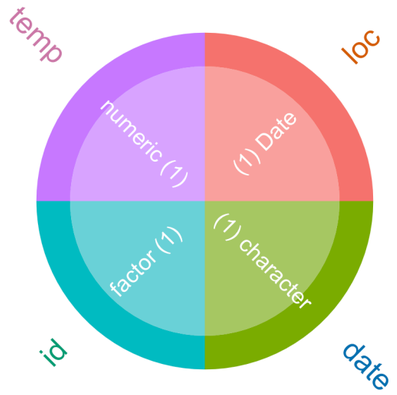
```
Rows: 5
Columns: 4
$ id   <fct> 1, 2, 3, 4, 5
$ date <date> 2010-03-01, 2010-03-02, 2010-03-03, 2010-03-0…
$ loc  <chr> "New York", "New York", "New York", "New Yor…
$ temp <dbl> 42, 41, 38, 41, 40
```

Example: Checking the data type visually

You can also visualise the data type with:

```
library(visdat)
vis_dat(xlsx_df)
```


```
library(inspectdf)
inspect_types(xlsx_df)  |>
  show_plot()
```



Data cleaning

Data cleaning (1/2)

Data quality checks should be one of the first steps in the data analysis to ***assess any problems with the data***.

These include using common or domain knowledge to check if the recorded data have sensible values.

- Are positive values, e.g. height and weight, recorded as positive values with a plausible range?
- If the data are counts, do the recorded values contain non-integer values?
- For compositional data, do the values add up to 100% (or 1)? If not, is that a measurement error or due to rounding? Or is another variable missing?
- Does the data contain only positives, ie disease occurrences, or warranty claims? If so, what would the no report group look like?

Data cleaning (2/2)

In addition, numerical or graphical summaries may reveal that there is unwanted structure in the data, for example,

- Does the treatment group have different demographic characteristics to the control group?
- Are the distributions similar between the or training and test sets?
- Are there sufficient measurements for each level of categorical variable, or across the range of numerical variables?

- Does the distribution of the data imply violations of assumptions for the CFA, such as
  - non-normality,
  - discrete rather real-valued, or
  - different variance in different domains?

Data scrutinizing is a process that you get better at with practice and have familiarity with the domain area.

Example: Checking the data quality

```
# A tibble: 9 × 4
  id    date       loc        temp
  <fct> <date>     <chr>     <dbl>
1 1     2010-03-01 New York   42
2 2     2010-03-02 New York   41.4
3 3     2010-03-03 New York   38.5
4 4     2010-03-04 New York   41.1
5 5     2010-03-05 New York   39.8
6 6     2020-03-01 Melbourne  30.6
7 7     2020-03-02 Melbourne  17.9
8 8     2020-03-03 Melbourne  18.6
9 9     2020-03-04 <NA>       21.3
```

- Numerical or graphical summaries or even just eye-balling the data helps to uncover some data quality issues.
- Any issues here?

- There's a missing value in `loc`.
- Temperature is in Farenheit for New York but Celsius in Melbourne (you can validate this again using external sources).

Case study: World development indicators (1/7)

```
options(width=80)
raw_dat <- read_csv(here("data/world-development-indicators.csv"),
                    na = "..", n_max = 11935)
glimpse(raw_dat)
```

```
Rows: 11,935
Columns: 54
$ `Country Name`  <chr> "Argentina", "Argentina", "Argentina", "Argentina", "A…
$ `Country Code`  <chr> "ARG", "ARG", "ARG", "ARG", "ARG", "ARG", "ARG", "ARG"…
$ `Series Name`   <chr> "Adolescent fertility rate (births per 1,000 women age…
$ `Series Code`   <chr> "SP.ADO.TFRT", "NV.AGR.TOTL.ZS", "ER.H2O.FWTL.ZS", "SH…
$ `1969 [YR1969]` <dbl> 6.4e+01, 9.2e+00, NA, NA, 3.3e+00, NA, 2.2e+01, NA, NA…
$ `1970 [YR1970]` <dbl> 6.5e+01, 9.6e+00, NA, NA, 3.5e+00, NA, 2.5e+01, NA, NA…
$ `1971 [YR1971]` <dbl> 6.7e+01, 1.1e+01, NA, NA, 3.7e+00, NA, 2.4e+01, 8.7e+0…
$ `1972 [YR1972]` <dbl> 6.8e+01, 1.1e+01, NA, NA, 3.6e+00, NA, 1.9e+01, 9.2e+0…
$ `1973 [YR1973]` <dbl> 7.1e+01, 1.2e+01, NA, NA, 3.7e+00, NA, 2.7e+01, 9.6e+0…
$ `1974 [YR1974]` <dbl> 7.5e+01, 1.0e+01, NA, NA, 3.7e+00, NA, 3.0e+01, 9.9e+0…
$ `1975 [YR1975]` <dbl> 7.8e+01, 6.6e+00, NA, NA, 3.6e+00, NA, 2.9e+01, 1.0e+0…
$ `1976 [YR1976]` <dbl> 8.1e+01, 8.2e+00, NA, NA, 3.8e+00, NA, 2.0e+01, 1.0e+0…
$ `1977 [YR1977]` <dbl> 8.4e+01, 8.1e+00, 9.5e+00, NA, 3.7e+00, NA, 2.6e+01, 1…
```

- What are the data types?
- What are the variables?
- What are the observations?
- Is the data in tidy form?

World Development Indicators (WDI), sourced from the World Bank Group (2019)

Case study: World development indicators (2/7)

```
country_code_df <- raw_dat  |>
  distinct(`Country Name`, `Country Code`)  |>
  rename_all(janitor::make_clean_names)  |>
  left_join(
    countrycode::codelist |> select(iso3c, region, continer
    by = c("country_code" = "iso3c")
  )  |>
  arrange(continent, region)
```

```
Rows: 217
Columns: 4
$ country_name <chr> "Algeria", "Djibouti", "Egypt, Arab Rep.", "Libya", "Moro…
$ country_code <chr> "DZA", "DJI", "EGY", "LBY", "MAR", "TUN", "AGO", "BEN", "…
$ region       <chr> "Middle East & North Africa", "Middle East & North Africa…
$ continent    <chr> "Africa", "Africa", "Africa", "Africa", "Africa", "Africa…
```

```
# A tibble: 6 × 2
  continent     n
  <chr>     <int>
1 Africa       54
2 Americas     46
3 Asia         50
4 Europe       46
5 Oceania      19
6 <NA>          2
```

```
# A tibble: 8 × 2
  region                         n
  <chr>                      <int>
1 East Asia & Pacific           37
2 Europe & Central Asia         56
3 Latin America & Caribbean     42
4 Middle East & North Africa    21
5 North America                  3
6 South Asia                     8
7 Sub-Saharan Africa            48
8 <NA>                           2
```

- How many countries are included
- How many continents, regions?
- Why are there NAs here?

```
country_code_df |> filter(is.na(continent))
```

```
# A tibble: 2 × 4
  country_name    country_code region continent
  <chr>           <chr>        <chr>  <chr>
1 Channel Islands CHI          <NA>   <NA>
2 Kosovo          XKX          <NA>   <NA>
```

Case study: World development indicators (3/7)

```
wdi_vars <- raw_dat  |>
  select(`Series Name`, `Series Code`) |>
  distinct() |>
  rename_all(janitor::make_clean_names)
```

series_name	series_code
Adolescent fertility rate (births per 1,000 women ages 15-19)	SP.ADO.TFRT
Agriculture, forestry, and fishing, value added (% of GDP)	NV.AGR.TOTL.ZS
Annual freshwater withdrawals, total (% of internal resources)	ER.H2O.FWTL.ZS
Births attended by skilled health staff (% of total)	SH.STA.BRTC.ZS
CO2 emissions (metric tons per capita)	EN.ATM.CO2E.PC
Contraceptive prevalence, any methods (% of women ages 15-49)	SP.DYN.CONU.ZS
Domestic credit provided by financial sector (% of GDP)	FS.AST.DOMS.GD.ZS

1–10 of 55 rows Previous **1** 2 3 4 5 6 Next

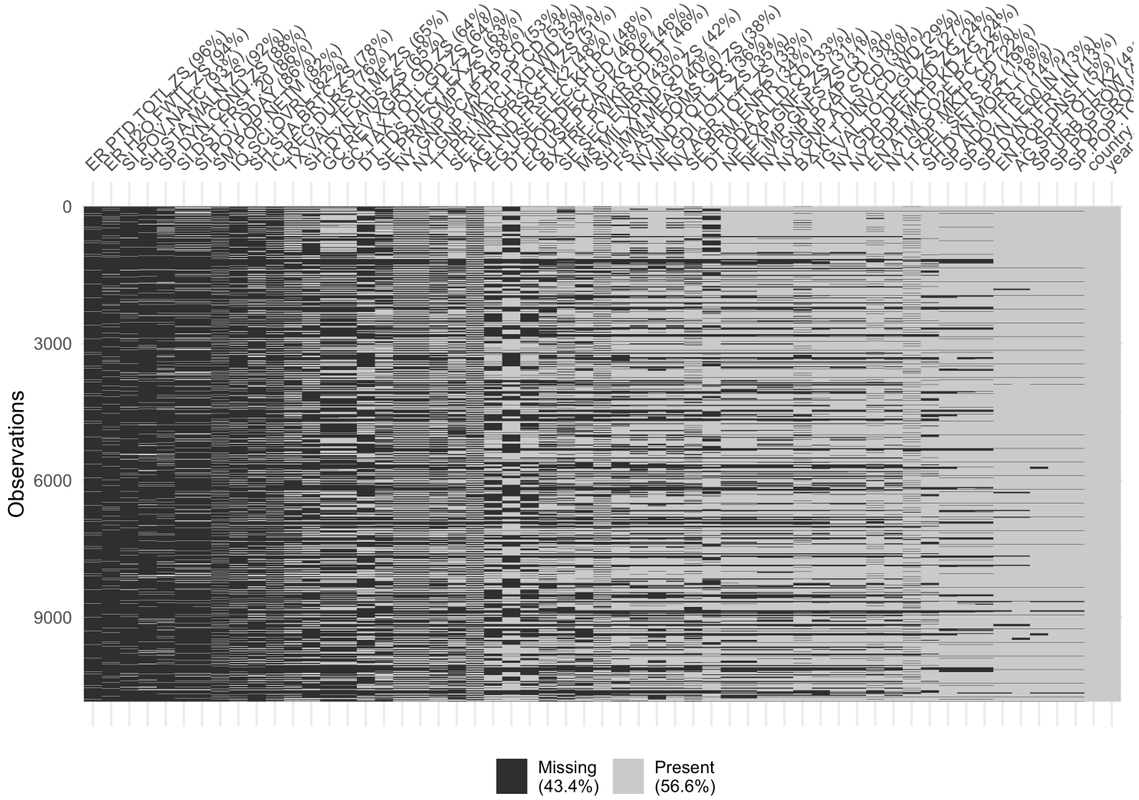
- Analysis will use the short name (`series_code`) for variables.
- Store full variable name (`series_name`) and short name (`series_code`) in a separate table.
- The `series_code` will be used as the key whenever the full name is needed.

Case study: World development indicators (4/7)

```
wdi <- raw_dat  |>
  select(`Country Code`, `Series Code`, `1969
  rename_all(janitor::make_clean_names) |>
  pivot_longer(x1969_yr1969:x2018_yr2018,
               names_to = "year",
               values_to = "value") |>
  mutate(year = as.numeric(str_sub(year, 2, 5)
  pivot_wider(names_from = series_code,
              values_from = value)

wdi2017 <- wdi  |> filter(year == 2017)
```

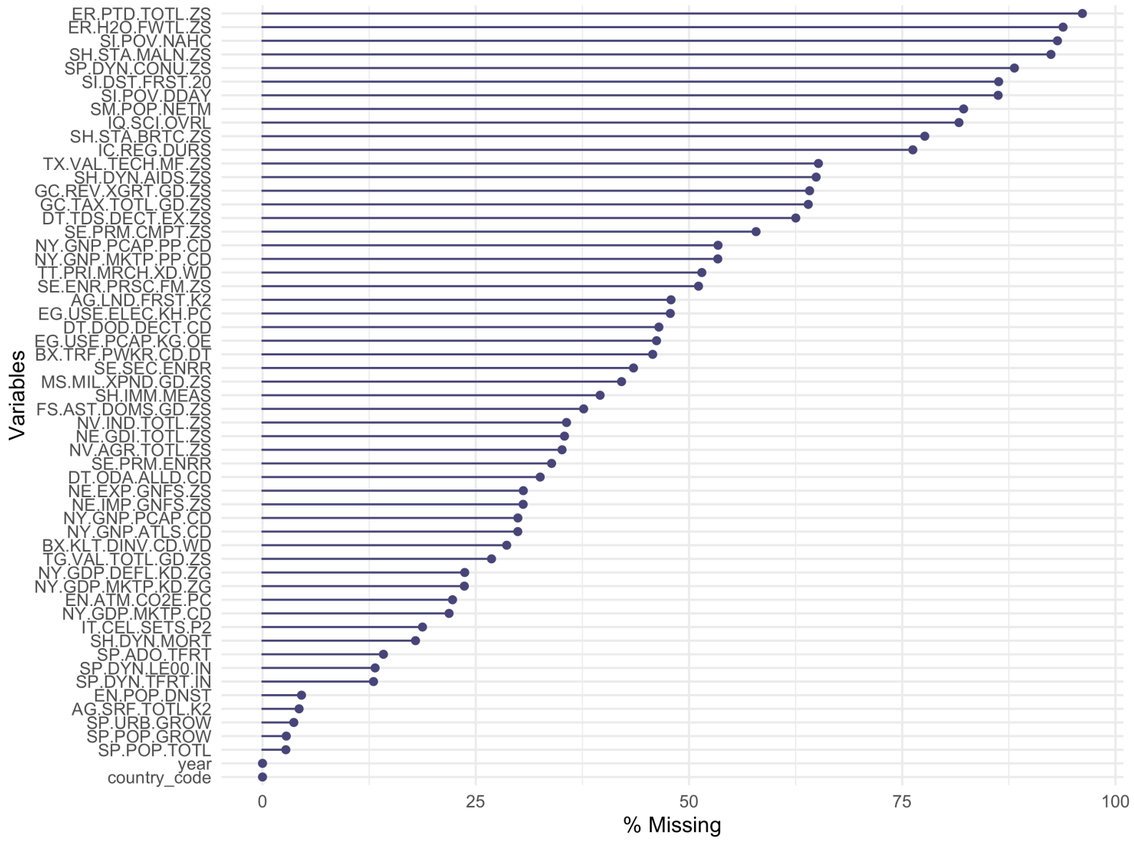
- Organise data into tidy form
- Check missing value distribution

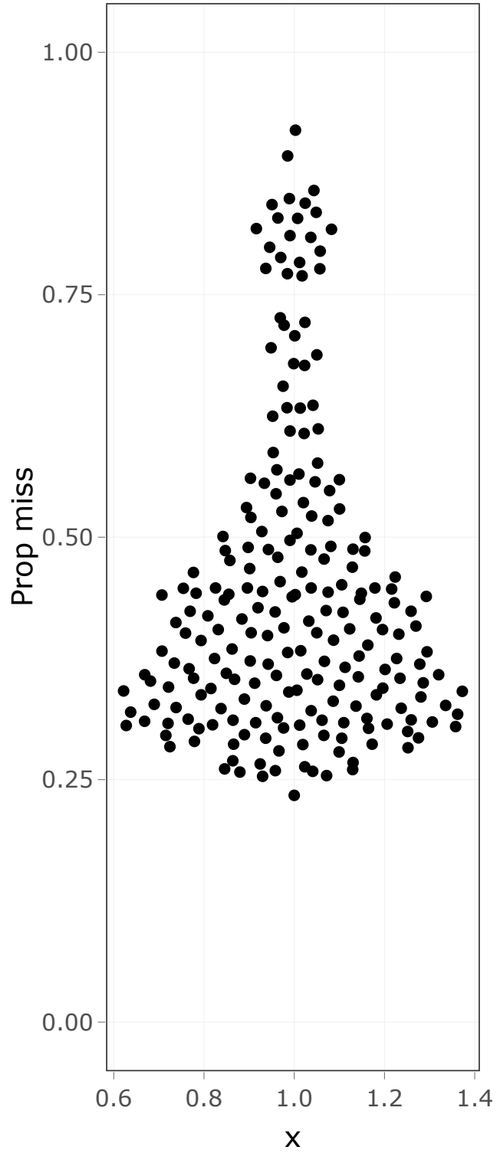


Case study: World development indicators (5/7)

Check missings by

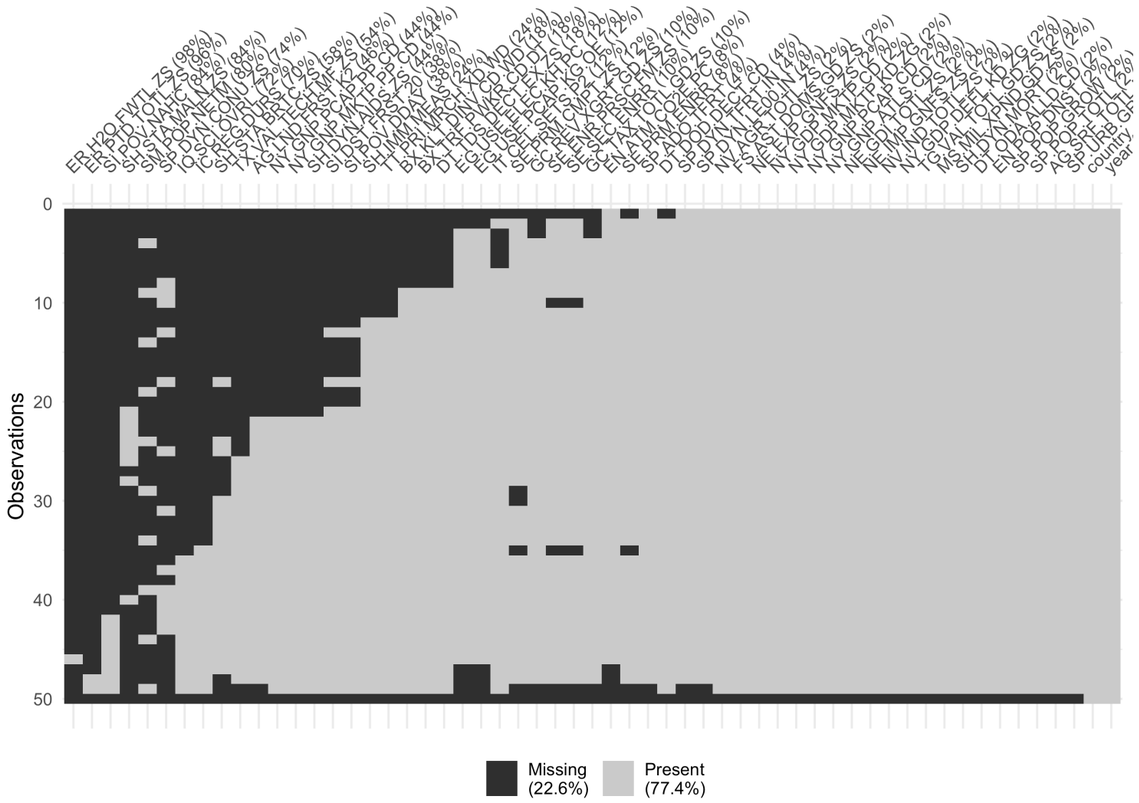
- variable
- country



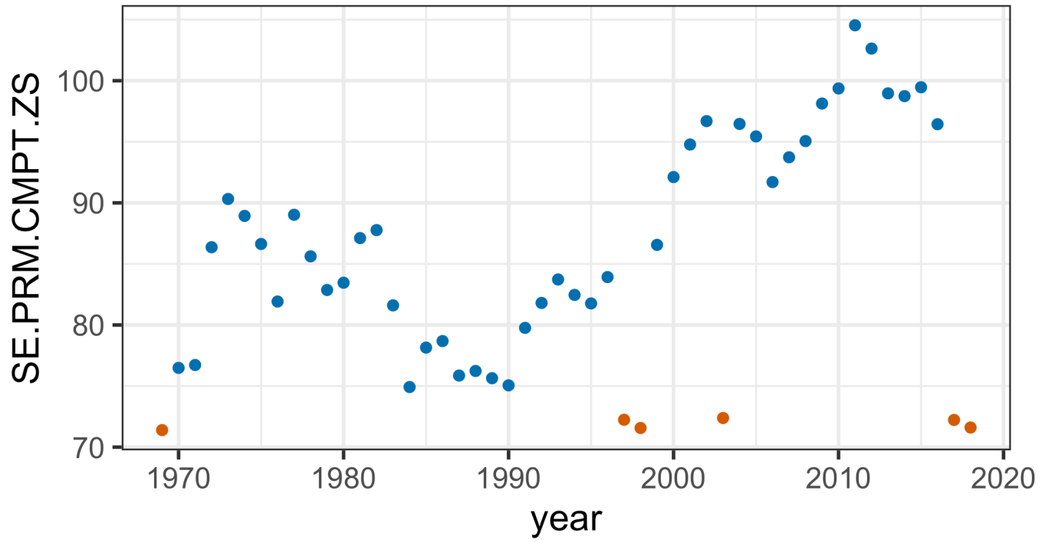


Case study: World development indicators (6/7)

Look at Costa Rica (CRI), most complete country



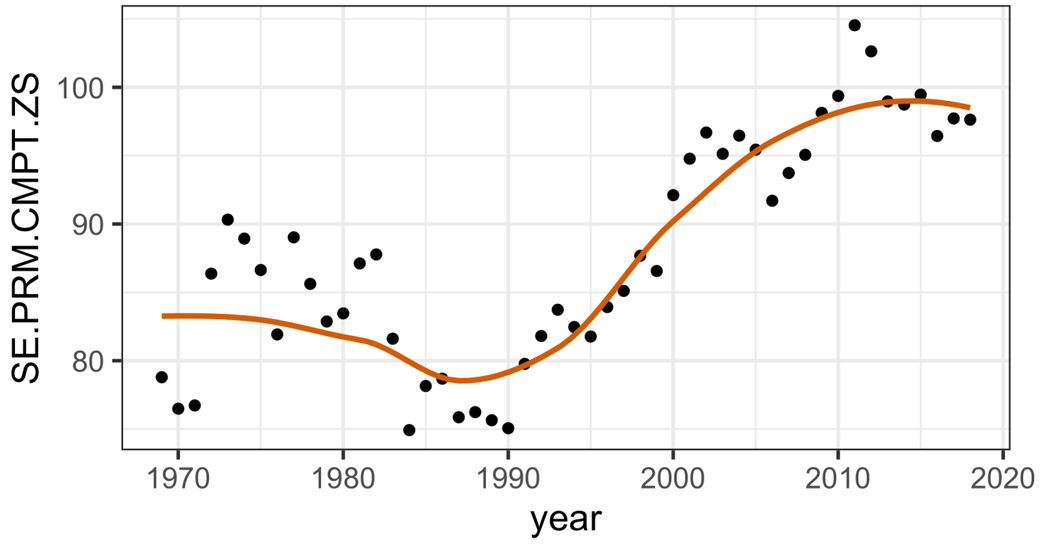
To illustrate imputation, we'll show one of the variables, that is relatively complete.



Impute a few temporal missings using nearest neighbours.

Case study: World development indicators (6/7)

Missings imputed using `imputeTS` using the moving average method.



- Don't have to impute before scrutinizing data
- What are these numbers supposed to be?

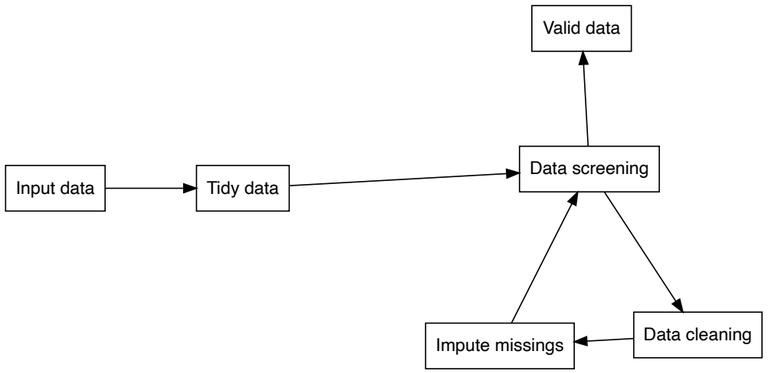
`SE.PRM.CMPT.ZS` is "Primary completion rate, total (% of relevant age group)"

Do we have any problems?

Yes. The explanation of the variable suggests the numbers should range between 0-100.

📋 Summary of the process

The steps we took roughly followed these:



At the end of this stage we would have:

- 3 tables of data: country name/code, variables name/key, time series of multiple variables for many countries
- What would you like to learn from this data? What sort of models might be fitted? What types of hypotheses might be tested?
- Have we done anything that might have compromised the later analysis?

Data collection

Case study: Employment Data in Australia (1/3)

Below is the data from ABS that shows the total number of people employed in a given month from February 1976 to December 2019 using the original time series.

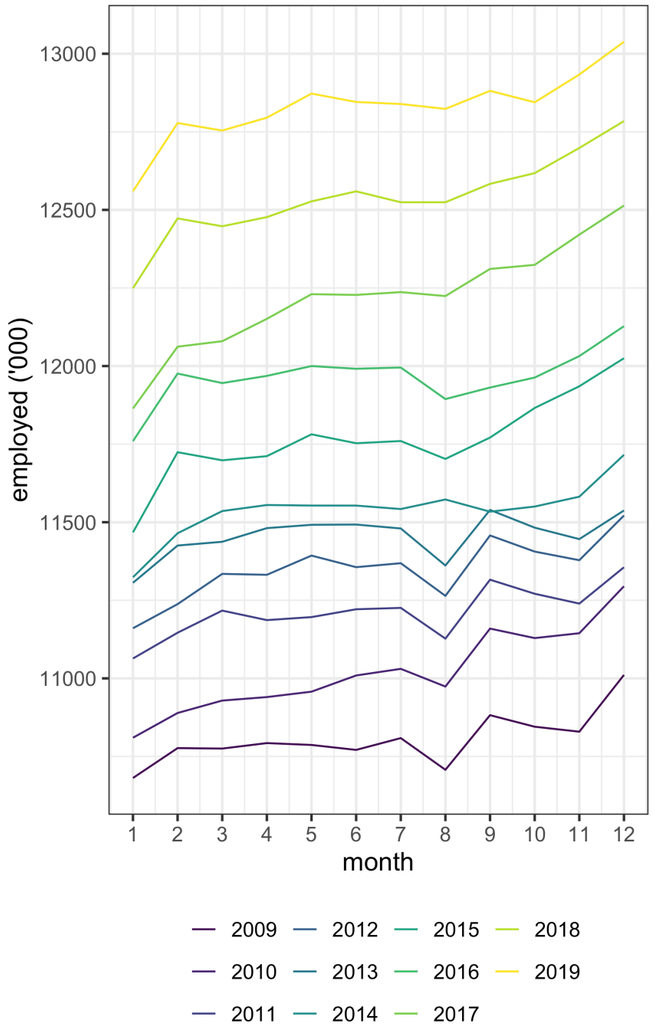
```
load(here("data/employed.rda"))
glimpse(employed)
```

```
Rows: 557
Columns: 4
$ date  <date> 1978-02-01, 1978-03-01, 1978-04-01, 1978-05-01, 1978-06-01, 197…
$ month <dbl> 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 1, 2, 3, 4, 5, 6, 7, 8, 9, 1…
$ year  <dbl> 1978, 1978, 1978, 1978, 1978, 1978, 1978, 1978, 1978, 1978, 1978…
$ value <dbl> 5986, 6041, 6054, 6038, 6031, 6036, 6005, 6024, 6046, 6034, 6125…
```

Australian Bureau of Statistics, Labour force, Australia, Table 01. Labour force status by Sex, Australia - Trend, Seasonally adjusted and Original

Case study: Employment Data in Australia (2/3)

Do you notice anything?

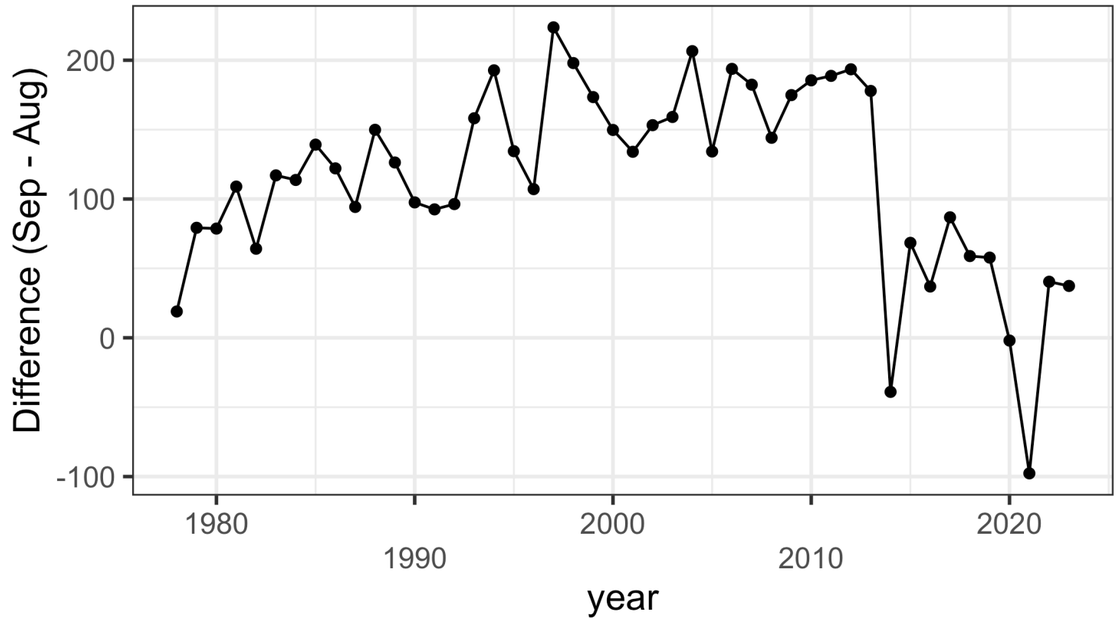


Why do you think the number of people employed is going up each year?

- Australian population is **25.39 million** in 2019
- 1.5% annual increase in population
- Vic population is 6.681 million (Sep 2020) - 26%
- NSW population is 8.166 (Sep 2020) - 32%

Case study: Employment Data in Australia (3/3)

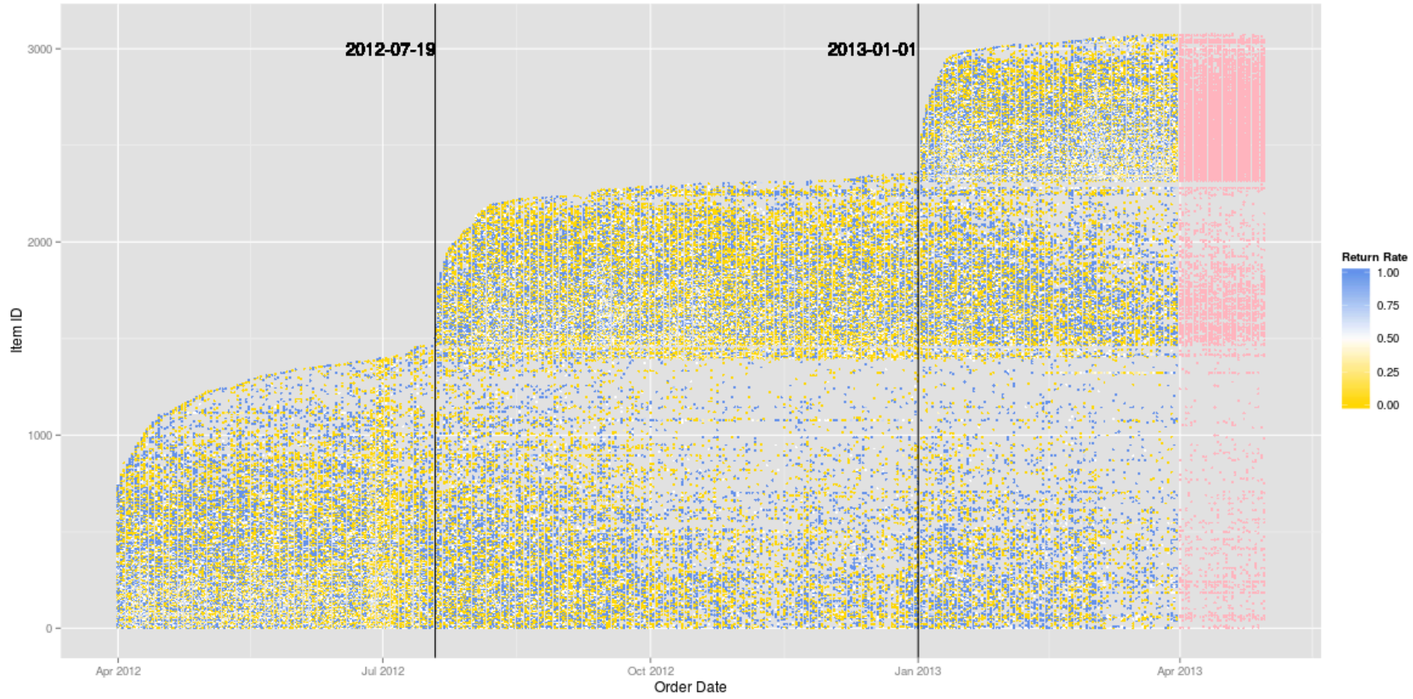
- There's a suspicious change in August numbers from 2014.



- A potential explanation for this is that there was a *change in the survey from 2014*.

See discussion on this at Hyndsight blog (10 October 2014).

Case study: 2014 Data Mining Cup winners



Ugly plot of all observations provided in training sample, with response variable in colour, and test sample to predict.

What does this tell you about the test sample?

Case study: french fries/hot chips (1/2)

```
Rows: 696
Columns: 9
$ time      <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ treatment <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ subject   <fct> 3, 3, 10, 10, 15, 15, 16, 16, …
$ rep       <dbl> 1, 2, 1, 2, 1, 2, 1, 2, 1, 2, …
$ potato    <dbl> 2.9, 14.0, 11.0, 9.9, 1.2, 8.8…
$ buttery   <dbl> 0.0, 0.0, 6.4, 5.9, 0.1, 3.0, …
$ grassy    <dbl> 0.0, 0.0, 0.0, 2.9, 0.0, 3.6, …
$ rancid    <dbl> 0.0, 1.1, 0.0, 2.2, 1.1, 1.5, …
$ painty    <dbl> 5.5, 0.0, 0.0, 0.0, 5.1, 2.3, …
```

10 week sensory experiment, 12 individuals assessed taste of french fries on several scales (how potato-y, buttery, grassy, rancid, paint-y do they taste?), fried in one of 3 different oils, replicated twice.

- Is the design complete?
- Are replicates like each other?
- How do the ratings on the different scales differ?
- Are raters giving different scores on average?
- Do ratings change over the weeks?

Case study: french fries/hot chips (2/2)

- Is the design complete?

```
french_fries |> count(subject)
```

```
# A tibble: 12 × 2
   subject     n
   <fct>   <int>
 1 3          54
 2 10         60
 3 15         60
 4 16         60
 5 19         60
 6 31         54
 7 51         60
 8 52         60
 9 63         60
10 78         60
11 79         54
12 86         54
```

```
french_fries |> count(time)
```

```
# A tibble: 10 × 2
   time      n
   <fct> <int>
 1 1        72
 2 2        72
 3 3        72
 4 4        72
 5 5        72
 6 6        72
 7 7        72
 8 8        72
 9 9        60
10 10       60
```

```
french_fries |> count(treatment)
```

```
# A tibble: 3 × 2
  treatment     n
  <fct>     <int>
1 1           232
2 2           232
3 3           232
```

```
french_fries |> count(rep)
```

```
# A tibble: 2 × 2
    rep     n
  <dbl> <int>
1     1   348
2     2   348
```

Case study: Warranty claims

```
Rows: 4,561
Columns: 14
$ Region           <chr> "East", "West", "North …
$ State            <chr> "Delhi", "Gujarat", "We…
$ Area             <chr> "Urban", "Rural", "Urba…
$ City             <chr> "New Delhi", "Ahmedabad…
$ Consumer_profile <chr> "Personal", "Personal",…
$ TV_2001_Issue    <dbl> 1, 1, 0, 0, 0, 0, 1, 1,…
$ TV_2002_Issue    <dbl> 1, 1, 1, 0, 0, 0, 1, 1,…
$ TV_2003_Issue    <dbl> 1, 0, 1, 0, 0, 0, 1, 0,…
$ Claim_Value      <dbl> 25000, 4216, 4000, 5000…
$ Service_Centre   <dbl> 13, 10, 10, 12, 10, 10,…
$ Product_Age      <dbl> 60, 672, 275, 10, 4, 34…
$ Purchased_from   <chr> "Dealer", "Dealer", "De…
$ Call_details     <dbl> 1.3, 25.0, 11.0, 1.6, 0…
```

- `TV_2001_Issue`: failure of power supply
- `TV_2002_Issue`: failure of inverter
- `TV_2003_Issue`: failure of motherboard

- What is the population that this data is measuring?
- What is not measured?

```
# A tibble: 2 × 2
  City          n
  <chr>     <int>
1 Delhi       106
2 Bangalore   320
```

Can we say that Delhi has fewer problems with TVs than Bangalore?

Source: ExcelR Projects. (2019). Warranty Claims. Kaggle.

📋 Summary of checks for data collection

✅ Has the collection process been consistent?

✅ Does the set to be predicted match the training set?

✅ Is the experimental design correctly applied?

✅ Have treatments been appropriately *randomised* or assigned comprehensively across subjects?

✅ What is the population that the collected data describes?

✅ If the data is observational, can you group them into comparison sets?

Imputing missing values

Example 1: Olympic medals

	country	totalmedal
1	UnitedStates	104
2	China	88
3	Russia	82
4	GreatBritain	65
5	Germany	44
6	Japan	38
7	Australia	35
8	France	34
9	SouthKorea	28
10	Italy	28
11	Netherlands	20
12	Ukraine	20
13	Canada	18
14	Hungary	17

What is missing?

What is the correct average number of medals?

962/204 = 4.72

Working out what is missing can be hard!

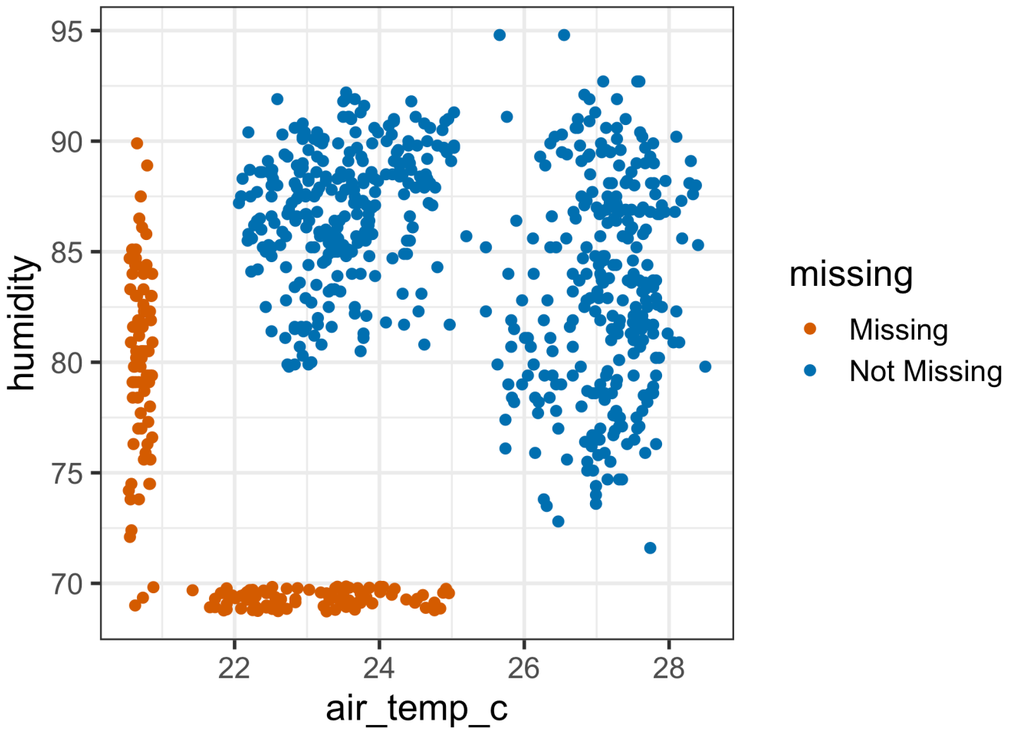
Is the average number of medals equal to 962/85 = 11.32?

Example 2: El Nino

Explore missings

- plotting on edge of plots, or
- using simple imputation like mean

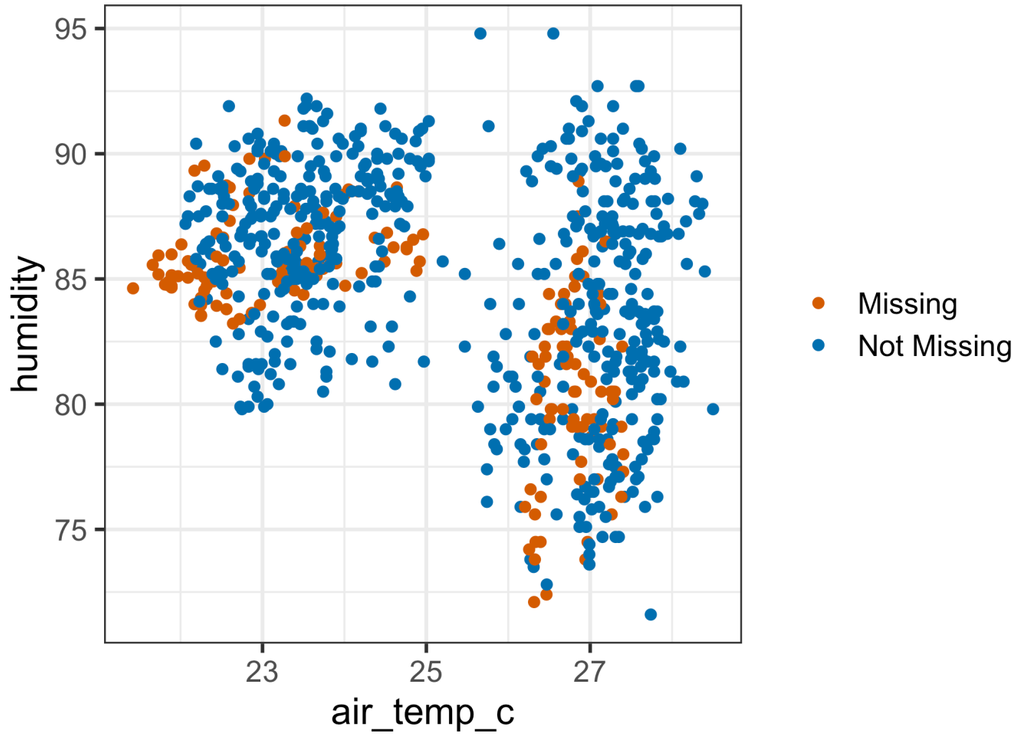
```
1 oceanbuoys |>
2   ggplot(aes(x=air_temp_c, y=humidity)) +
3   geom_miss_point()
```



Impute and check

- Impute using regression or simulation
- Check distribution relative to complete cases

▶ Code



Validators

Automating some checks

Case study: Dutch supermarket revenue and cost (1/3)

- Data contains the revenue and cost (in Euros) for 60 supermarkets
- Data has been anonymised and distorted

```
data("SBS2000", package = "validate")
dplyr::glimpse(SBS2000)
```

```
Rows: 60
Columns: 11
$ id           <fct> RET01, RET02, RET03, RET04, …
$ size         <fct> sc0, sc3, sc3, sc3, sc3, sc0…
$ incl.prob    <dbl> 0.02, 0.14, 0.14, 0.14, 0.14…
$ staff        <int> 75, 9, NA, NA, NA, 1, 5, 3, …
$ turnover     <int> NA, 1607, 6886, 3861, NA, 25…
$ other.rev    <int> NA, NA, -33, 13, 37, NA, NA,…
$ total.rev    <int> 1130, 1607, 6919, 3874, 5602…
$ staff.costs <int> NA, 131, 324, 290, 314, NA, …
$ total.costs <int> 18915, 1544, 6493, 3600, 553…
$ profit       <int> 20045, 63, 426, 274, 72, 3, …
$ vat          <int> NA, NA, NA, NA, NA, NA, 1346…
```

Case study: Dutch supermarket revenue and cost (2/3)

- Checking for completeness of records

```
library(validate)
rules <- validator(
          is_complete(id),
          is_complete(id, turnover),
          is_complete(id, turnover, profit))
out <- confront(SBS2000, rules)
summary(out)
```

```
  name items passes fails nNA error warning
1   V1    60     60     0   0 FALSE   FALSE
2   V2    60     56     4   0 FALSE   FALSE
3   V3    60     52     8   0 FALSE   FALSE
                      expression
1                 is_complete(id)
2         is_complete(id, turnover)
3 is_complete(id, turnover, profit)
```

Case study: Dutch supermarket revenue and cost (3/3)

- Sanity check derived variables

```
library(validate)
rules <- validator(
    total.rev - profit == total.costs,
    turnover + other.rev == total.rev,
    profit <= 0.6 * total.rev
)
out <- confront(SBS2000, rules)
summary(out)
```

```
  name items passes fails nNA error warning
1   V1    60     39    14   7 FALSE   FALSE
2   V2    60     19     4  37 FALSE   FALSE
3   V3    60     49     6   5 FALSE   FALSE
                                    expression
1 abs(total.rev - profit - total.costs) <= 1e-08
2 abs(turnover + other.rev - total.rev) <= 1e-08
3              profit - 0.6 * total.rev <= 1e-08
```

IDA for hypothesis testing

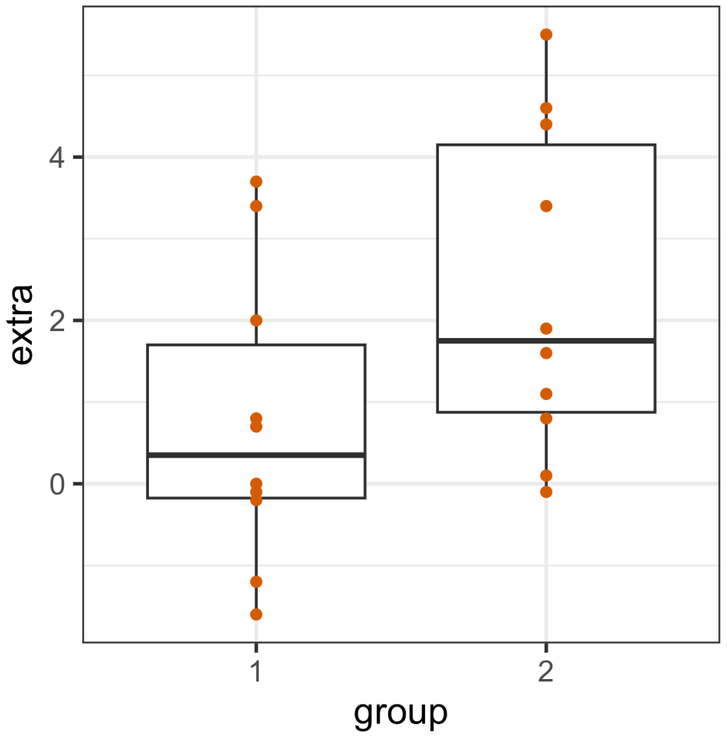
Hypothesis testing (1/3)

- State the hypothesis (pair), e.g. $\mathrm{H_o} : \mu_1 = \mu_2$ vs $\mathrm{H_a} : \mu_1 < \mu_2$.
- Test statistic depends on assumption about the distribution, e.g.
  - t-test will assume that distributions are *normal*, or small departures from if we have a large sample.
  - two-sample might assume both groups have the *same variance*

- Steps to complete:
  - Compute the test statistic
  - Measure it against a standard distribution
  - If it is extreme, p-value is small, decision is to reject $\mathrm{H_o}$
  - p-value is the probability of observing a value as large as this, or large, assuming $\mathrm{H_o}$ is true.

Example 1: Checking variance and distribution (2/3)

▶ Code



Few observations. Nothing strongly suggests violation of normality and spread of points is similar for each group.

```
tt <- with(sleep,
     t.test(extra[group == 1],
            extra[group == 2],
            paired = TRUE))
tt
```

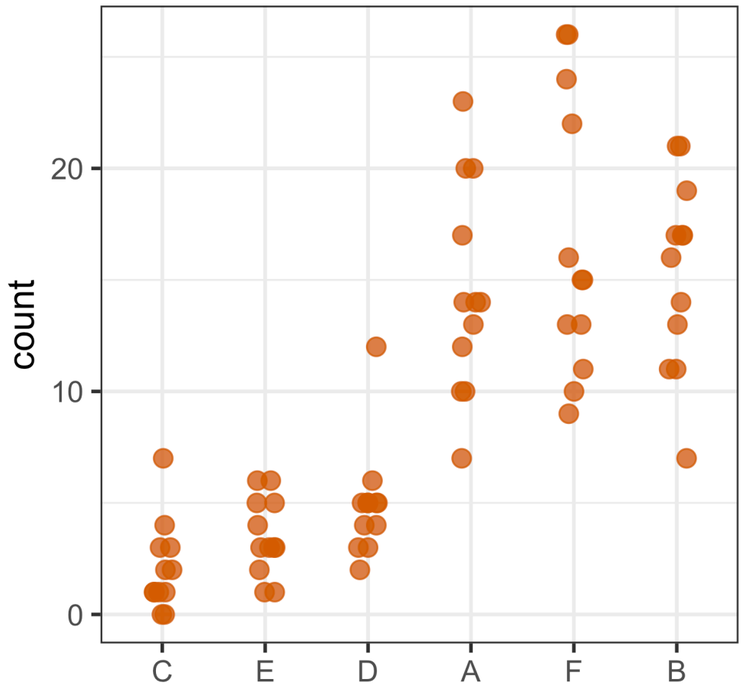
```
    Paired t-test

data:  extra[group == 1] and extra[group == 2]
t = -4, df = 9, p-value = 0.003
alternative hypothesis: true mean difference is not equal
to 0
95 percent confidence interval:
 -2.5 -0.7
sample estimates:
mean difference
            -1.6
```

Cushny, A. R. and Peebles, A. R. (1905) The action of optical isomers: II hyoscines. The Journal of Physiology 32, 501–510.

Example 2: Checking distribution and variance (3/3)

▶ Code




```
fm1 <- aov(count ~ spray, data = InsectSprays)
summary(fm1)
```

```
            Df Sum Sq Mean Sq F value Pr(>F)
spray        5   2669     534    34.7 <2e-16 ***
Residuals   66   1015      15
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What hypothesis being tested? What would the decision be?

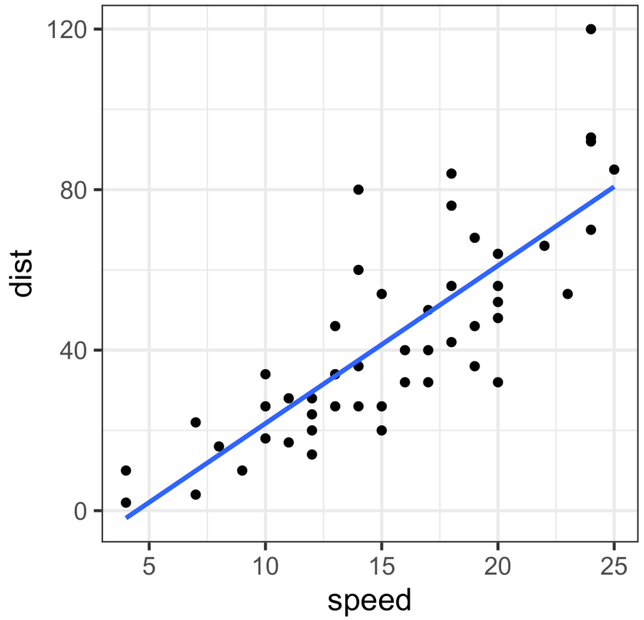
Why does equal variance matter in this test?

Is it plausible that the samples are from a normal population? Do they have equal variance?

IDA for inferential modeling

Linear models (1/3)

▶ Code



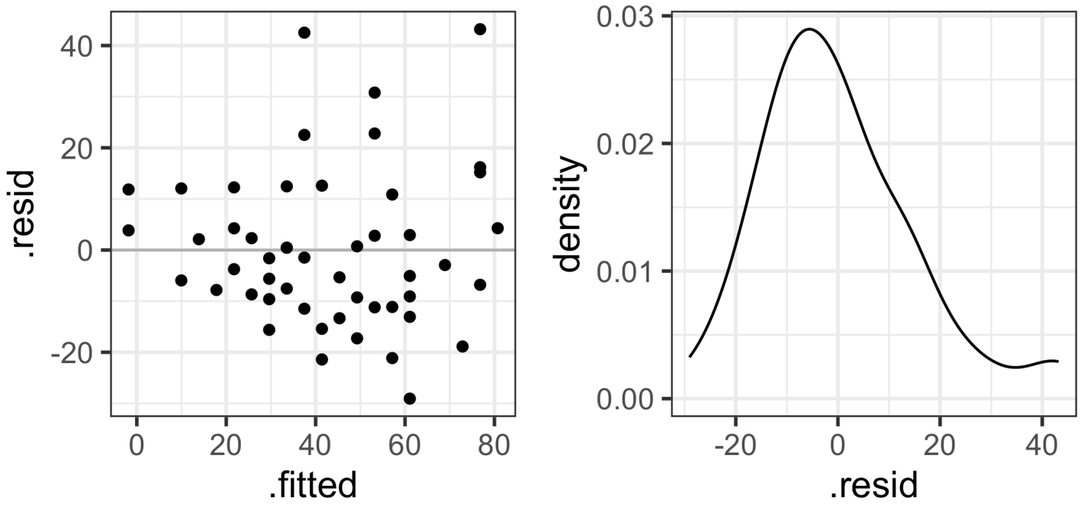
$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Assumptions:

- Form is linear
- Error is normally distributed around 0

Check using residual plots

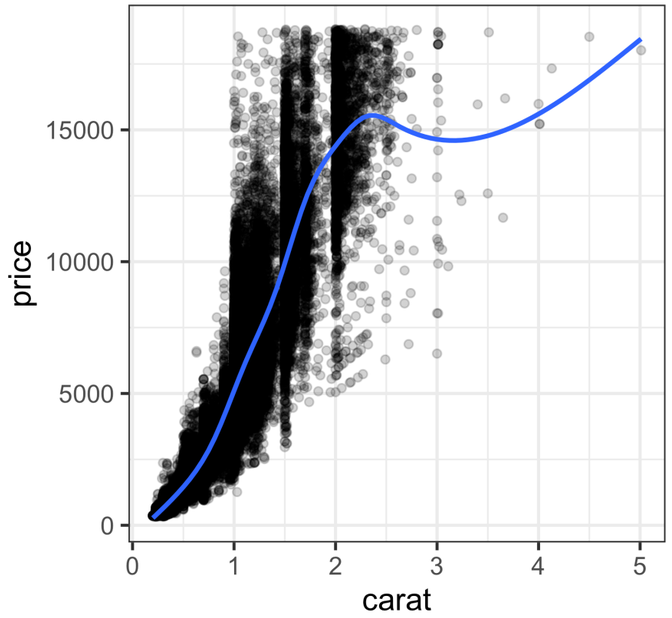
▶ Code



Linear models (2/3)

Data and loess smoother

▶ Code

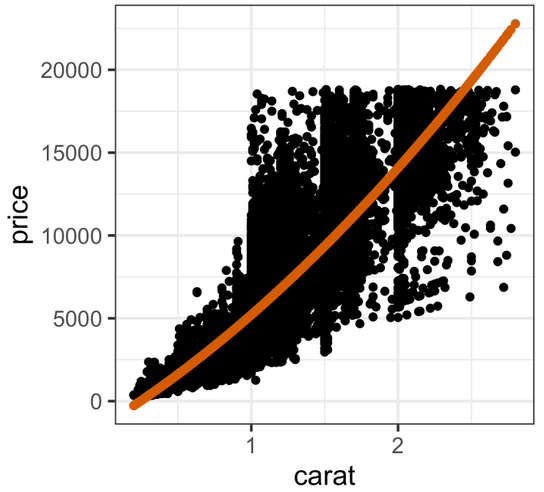


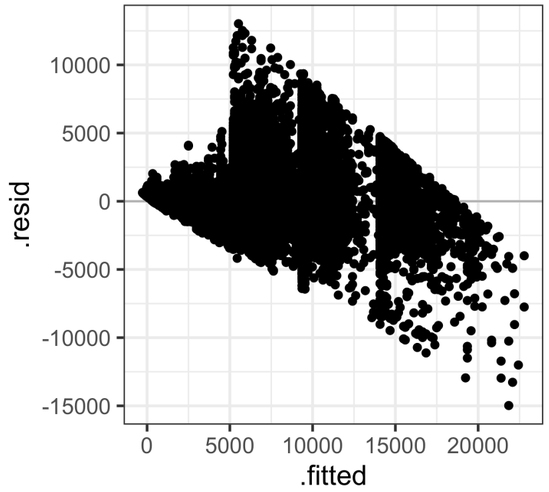
- Form is not linear!
- Also, insufficient data on large diamonds.

Fix 1: fit polynomial form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i\,.$$

▶ Code



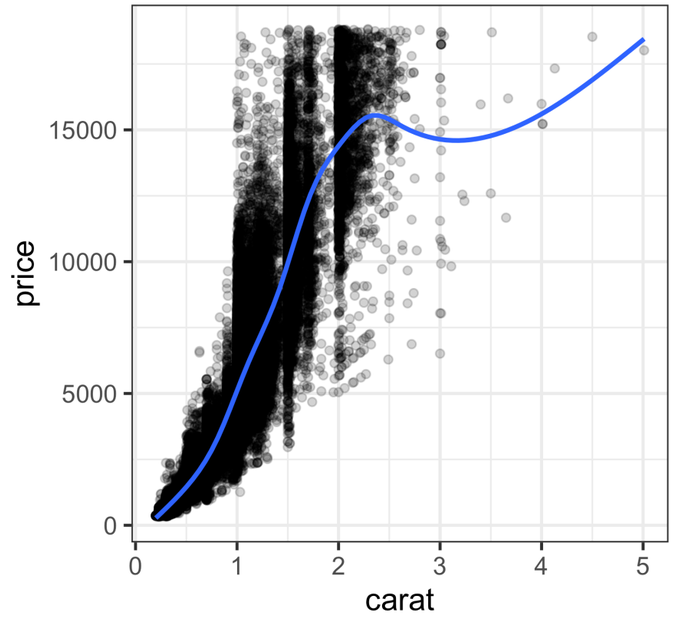


Form is not quadratic, continue to explore additional polynomial terms.

Linear models (3/3)

Data and loess smoother

▶ Code

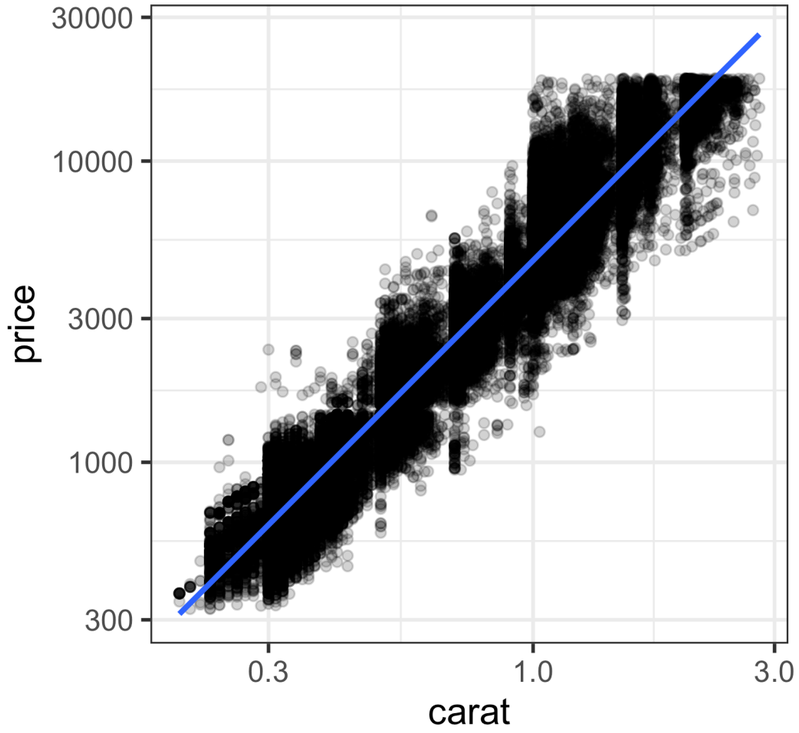


- Form is not linear!
- Also, insufficient data on large diamonds.

Fix 2: linearise

The log transformation of both variables linearises the relationship, so that a simple linear model can be used, and can correct heteroskedasticity.

▶ Code



Cautions

- Notice that there was ***no formal statistical inference*** when trying to determine an appropriate model form.

> Discarded models are hardly ever reported. Consequently, majority of reported statistics give a distorted view and it's important to remind yourself what might ***not*** be reported.

Summary

- IDA is a model-focused exploration to support a CFA with:
  - data description and collection
  - data quality checking, and
  - checking assumptions
  - model fit without any formal statistical inference.
- IDA is part of EDA, even when no CFA is planned.
- IDA may never see the limelight BUT it forms the foundation that the main analysis is built upon. Document it! Do it well!

> *The Census Bureau tabulates same-sex couples in both the American Community Survey (ACS) and the Decennial Census. Two questions are used to identify same-sex couples: relationship and sex. The agency follows edit rules that are used to change data values for seemingly contradictory answers. The edit rules for combining information from relationship and sex have evolved since the category of unmarried partner was added in 1990. In that census, if a household consisted of a married couple and both spouses reported the same sex, the relationship category remained husband or wife, but the sex of the partner who reported being a spouse to the householder was changed.* Humans all the way down

Human actions are ubiquitous in every part of data analysis! The most objective methods often have had subjective actions before and after.

Further reading

- Huebner et al (2018) A Contemporary Conceptual Framework for Initial Data Analysis
- Huebner et al (2020) Hidden analyses
- Chatfield (1985) The Initial Examination of Data. *Journal of the Royal Statistical Society. Series A (General)* **148**
- Cox & Snell (1981) Applied Statistics. *London: Chapman and Hall.*
- van der Loo and de Jonge (2018). Statistical Data Cleaning with Applications in R. John Wiley and Sons Ltd.
- Hyndman (2014) Explaining the ABS unemployment fluctuations