Exploratory Data Analysis

Overview

Lecture outline

- Definitions
- Data types
- Steps in Exploratory Data Analysis (EDA)
    - General characteristics of the dataset
    - Descriptive statistics (univariate)
    - Correlation statistics (bivariate)
    - Exploratory visualisation - univariate and bivariate
    - Anomalies - outliers and inliers
    - Missing values
- EDA in real-life practice

Definitions

“ Exploratory data analysis can never be the whole story, but nothing else can serve as a foundation stone - as the first step.

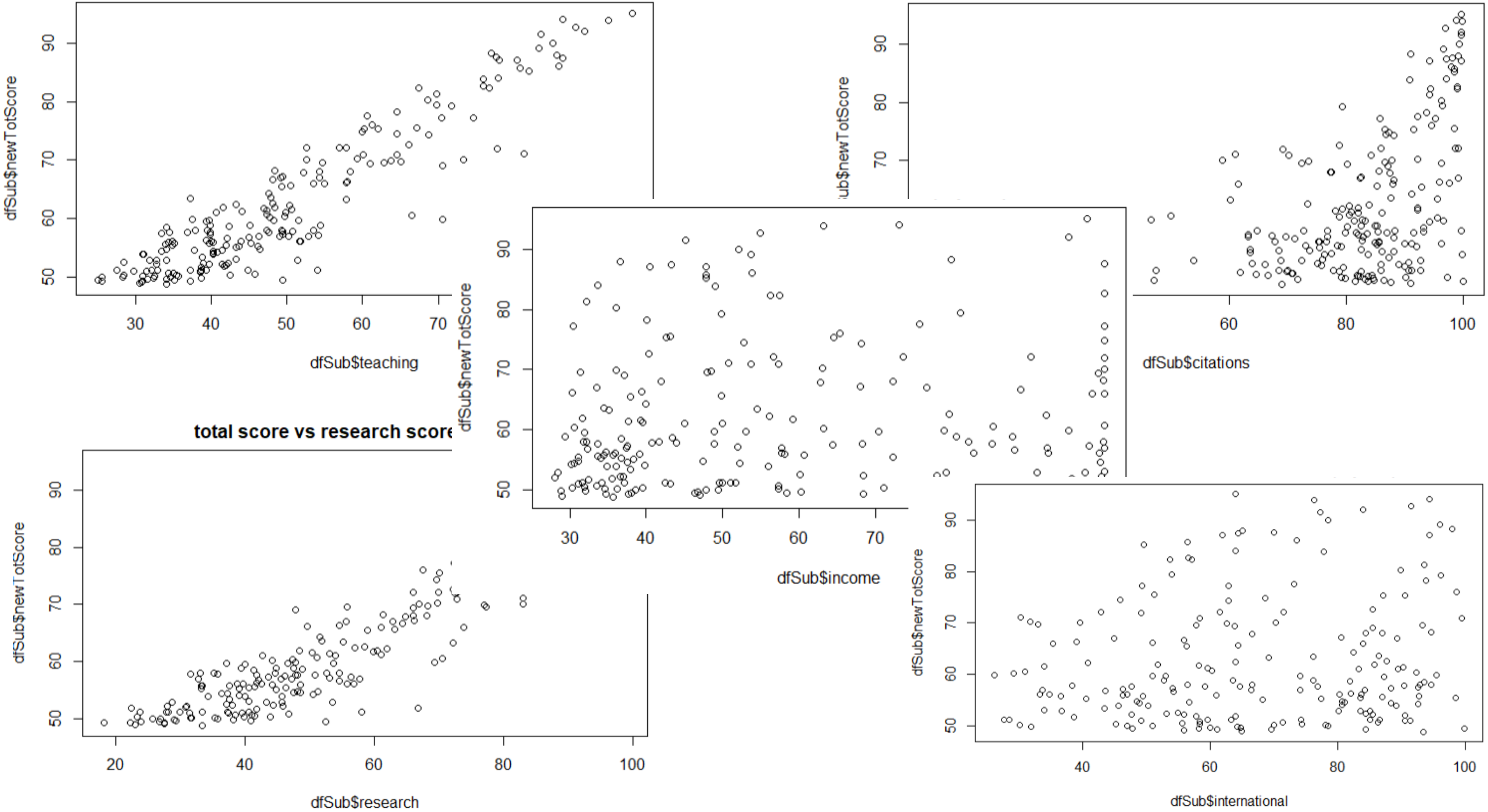
John Tukey, 1977, *Data Exploratory Analysis*, Addison-Wesley

“ Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there.

John Tukey, 1977, *Data Exploratory Analysis*, Addison-Wesley

“ The primary aim with exploratory data analysis is to examine the data for distribution, outliers and anomalies … hypothesis generation by visualising and understanding the data.

https://link.springer.com/chapter/10.1007/978-3-319-43742-2_15


dfSub$newTotScore
dfSub$teaching
dfSub$citations
total score vs research score
dfSub$research
dfSub$income
dfSub$international

Structured data vs unstructured data

Unstructured data: signals, images, text, graphs, sounds, etc.

Structured data - cross-sectional, panel, time series

- Data types: nominal, ordinal, interval, ratio, transaction, latitude/longitude, etc

Structured data types

Nominal - labels, mutually exclusive, no numerical significance, may or may not have orders.

What is your gender?
- M – Male
- F – Female

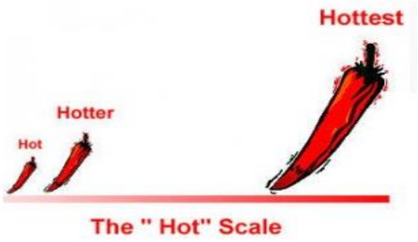
What is your hair color?
- 1 – Brown
- 2 – Black
- 3 – Blonde
- 4 – Gray
- 5 – Other

Where do you live?
- A – North of the equator
- B – South of the equator
- C – Neither: In the international space station

Ordinal - having order but the difference between variables not defined

How do you feel today?
- 1 – Very Unhappy
- 2 – Unhappy
- 3 – OK
- 4 – Happy
- 5 – Very Happy

How satisfied are you with our service?
- 1 – Very Unsatisfied
- 2 – Somewhat Unsatisfied
- 3 – Neutral
- 4 – Somewhat Satisfied
- 5 – Very Satisfied



Structured data types

Interval - having order, difference between variables defined, but don’t have a ‘true zero’, e.g. temperature, clock time.

For example, a glass of water with a temperature of 0 degree does not mean it has **no** temperature.

Ratio - like interval but with a ‘true zero’, e.g. income, age, years of education, weight.

EDA - General characteristics of the dataset

Assess the general characteristics of the dataset

- What kind of data structure is the dataset?
- How many records does this dataset contain?
- How many fields (variables) are there?
- What kind of variables are they?

EDA - General characteristics of the dataset

Example output from dataset in Bank.csv

	age	job	marital	education	default	balance	housing	loan	contact
0	59	admin.	married	secondary	no	2343	yes	no	unknown
1	56	admin.	married	secondary	no	45	no	no	unknown
2	41	technician	married	secondary	no	1270	yes	no	unknown
3	55	services	married	secondary	no	2476	yes	no	unknown
4	54	admin.	married	tertiary	no	184	no	no	unknown

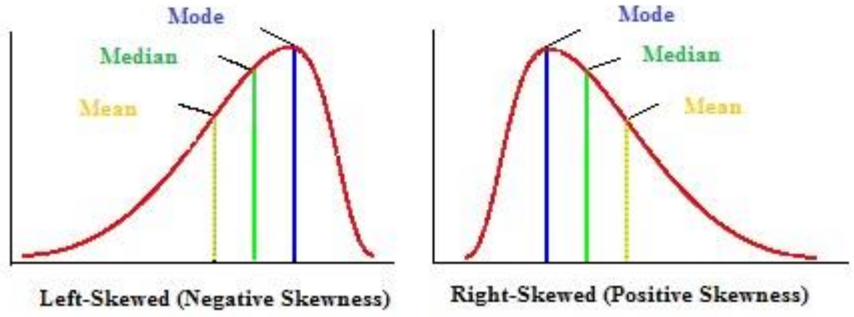
	day	month	duration	campaign	pdays	previous	poutcome	deposit
0	5	may	1042	1	-1	0	unknown	yes
1	5	may	1467	1	-1	0	unknown	yes
2	5	may	1389	1	-1	0	unknown	yes
3	5	may	579	1	-1	0	unknown	yes
4	5	may	673	2	-1	0	unknown	yes

```
RangeIndex: 11162 entries, 0 to 11161
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        11162 non-null  int64
 1   job        11162 non-null  object
 2   marital    11162 non-null  object
 3   education  11162 non-null  object
 4   default    11162 non-null  object
 5   balance    11162 non-null  int64
 6   housing    11162 non-null  object
 7   loan       11162 non-null  object
 8   contact    11162 non-null  object
 9   day        11162 non-null  int64
 10  month      11162 non-null  object
 11  duration   11162 non-null  int64
 12  campaign   11162 non-null  int64
 13  pdays      11162 non-null  int64
 14  previous   11162 non-null  int64
 15  poutcome   11162 non-null  object
 16  deposit    11162 non-null  object
dtypes: int64(7), object(10)
```

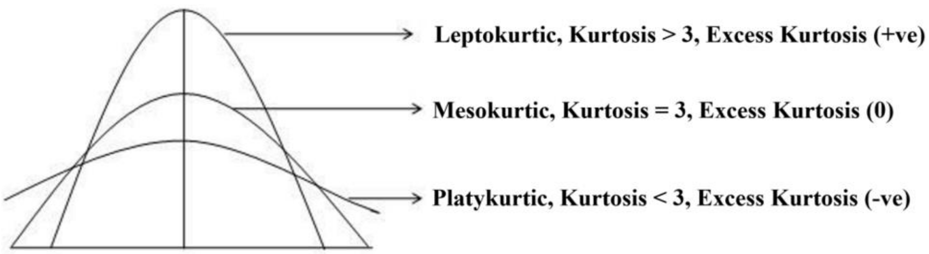
EDA - Descriptive statistics (univariate)

Numerical variables

- Measures of centre: mean, median, mode
- Measures of variability: range, standard deviation
- Measures of relative standings: quartiles, percentiles
- Measures of distribution: skewness and kurtosis



https://www.statisticshowto.com/probability-and-statistics/skewed-distribution/



https://towardsdatascience.com/skewness-kurtosis-simplified-1338e094fc85

EDA - Descriptive statistics (univariate)

Categorical variables

- Cardinality: number of unique values
- Unique counts: number of occurrences of each unique value

EDA - Descriptive statistics (univariate)

Example output from dataset in Bank.csv

	age	balance	day	duration	campaign
count	11162.000000	11162.000000	11162.000000	11162.000000	11162.000000
mean	41.231948	1528.538524	15.658036	371.993818	2.508421
std	11.913369	3225.413326	8.420740	347.128386	2.722077
min	18.000000	-6847.000000	1.000000	2.000000	1.000000
25%	32.000000	122.000000	8.000000	138.000000	1.000000
50%	39.000000	550.000000	15.000000	255.000000	2.000000
75%	49.000000	1708.000000	22.000000	496.000000	3.000000
max	95.000000	81204.000000	31.000000	3881.000000	63.000000

	pdays	previous
count	11162.000000	11162.000000
mean	51.330407	0.832557
std	108.758282	2.292007
min	-1.000000	0.000000
25%	-1.000000	0.000000
50%	-1.000000	0.000000
75%	20.750000	1.000000
max	854.000000	58.000000

management	2566
blue-collar	1944
technician	1823
admin.	1334
services	923
retired	778
self-employed	405
student	360
unemployed	357
entrepreneur	328
housemaid	274
unknown	70

Name: job, dtype: int64

married	6351
single	3518
divorced	1293

Name: marital, dtype: int64

secondary	5476
tertiary	3689
primary	1500
unknown	497

Name: education, dtype: int64

EDA - Correlation statistics (bivariate)

Qualitative analysis

Both categorical	Contingency table
Categorical (X) vs numerical (Y)	Descriptive statistics of Y for each value X

Quantitative analysis

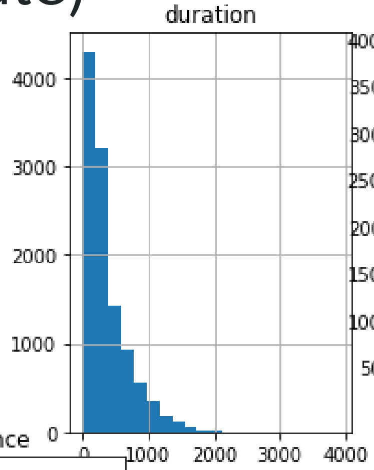
	Categorical	**Numerical**
Categorical	Chi-squared test	Student t-test, ANOVA, Logistic regression
Numerical	Student t-test, ANOVA, Logistic regression	Correlation, Linear regression

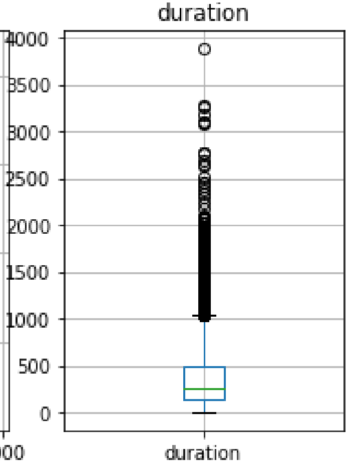
EDA - Exploratory visualisation (univariate)

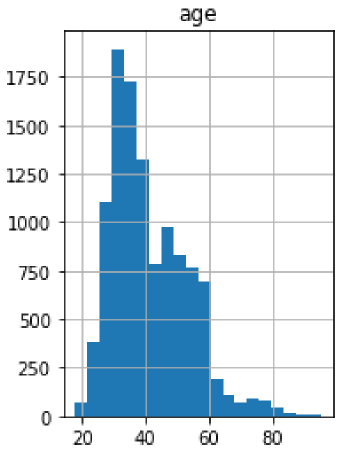
Numerical variables - histogram, boxplot

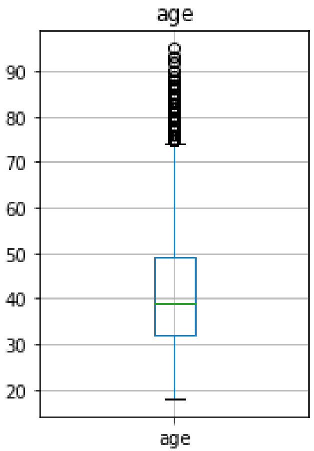
Freedman-Diaconis rule

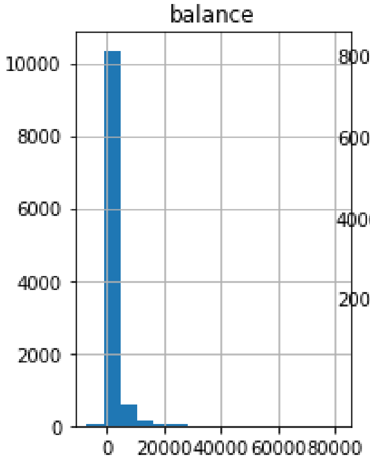
$$\text{Bin width} = 2\,\frac{\text{IQR}(x)}{\sqrt[3]{n}}$$

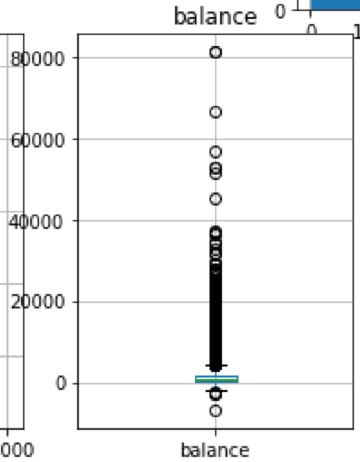






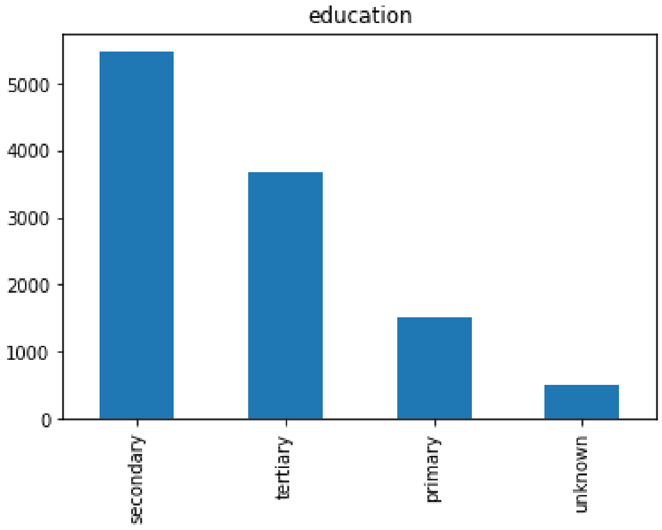


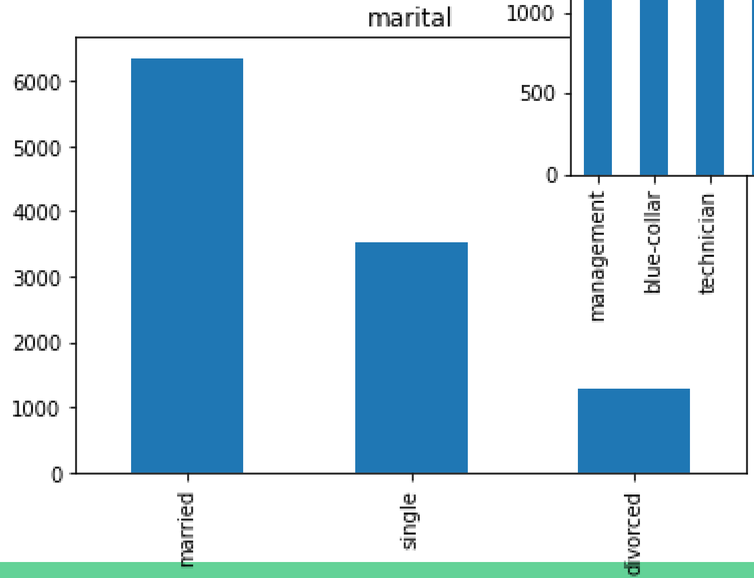


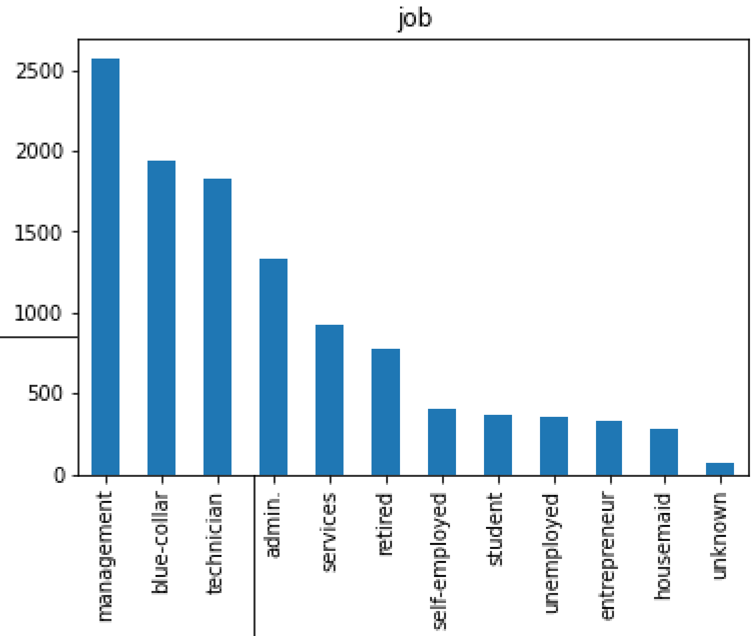


EDA - Exploratory visualisation (1 dimensional)

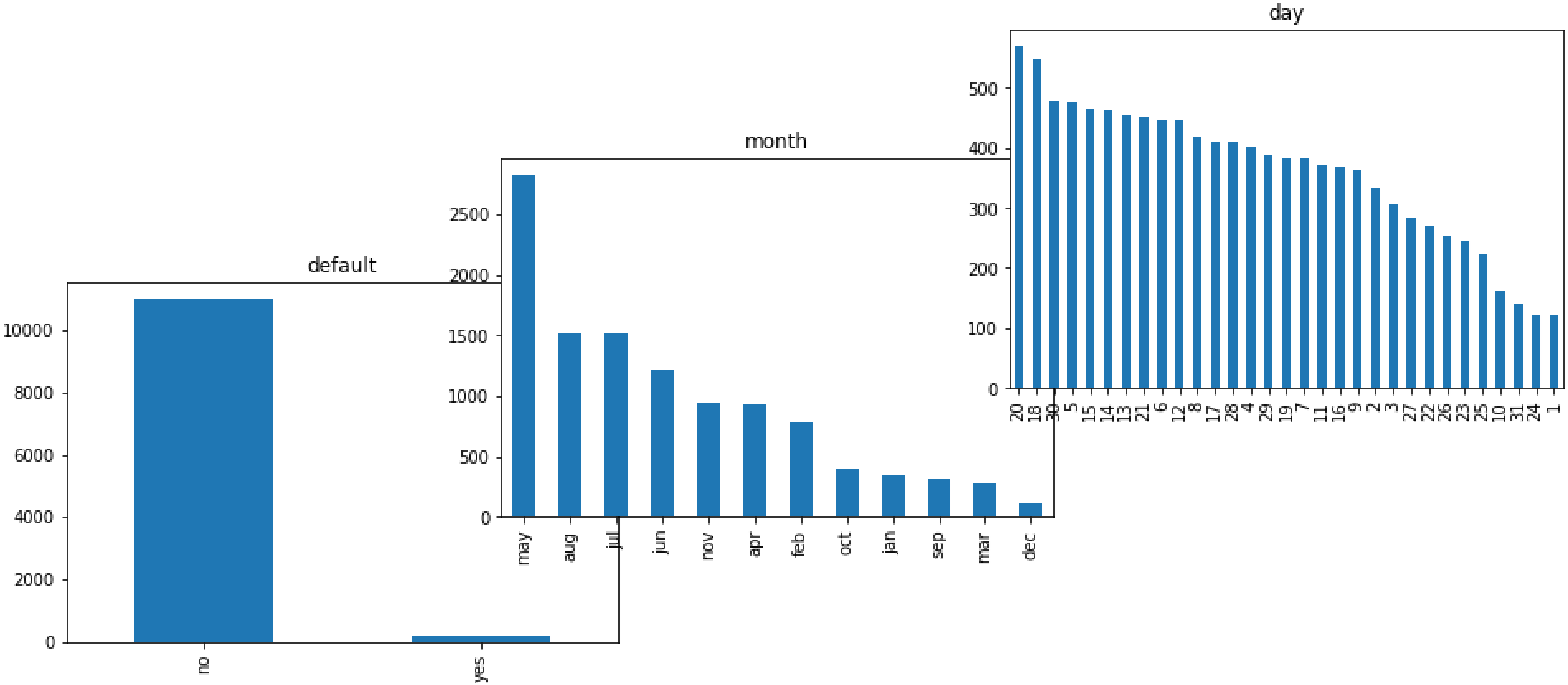
Categorical - Bar plots



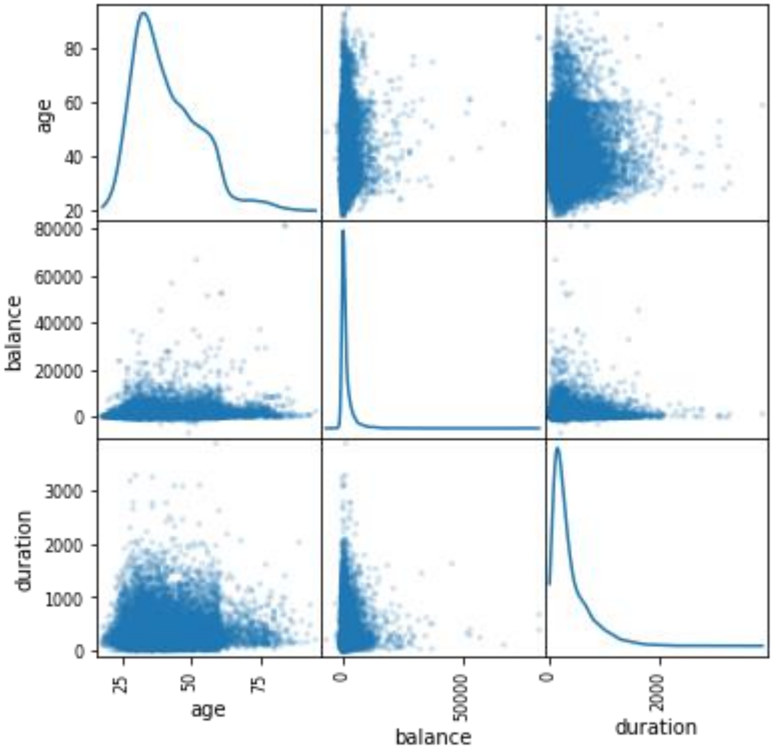


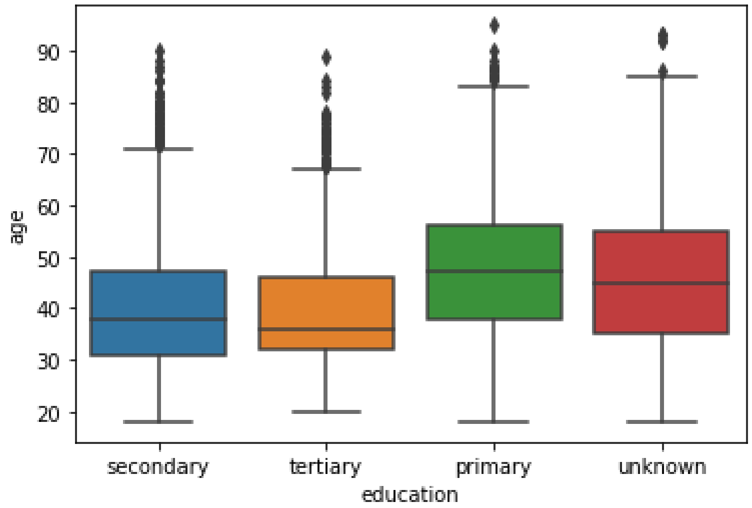


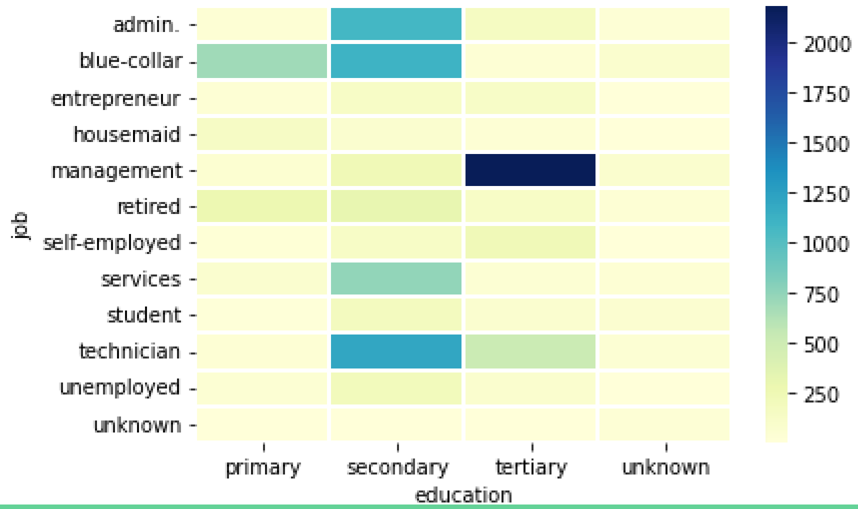
EDA - Exploratory visualisation (1 dimensional)



EDA - Exploratory visualisation (2D)





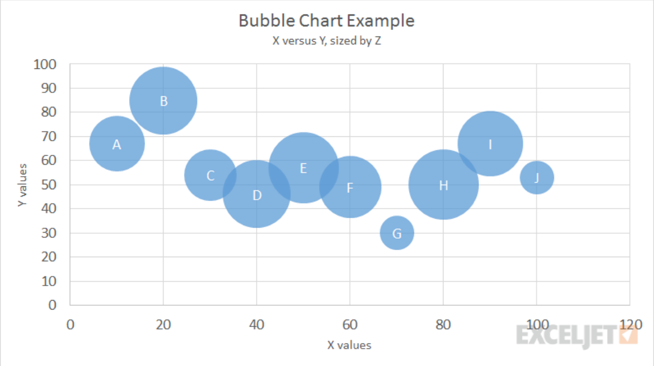


EDA - Statistics and visualisation summary

	Univariate		Bivariate		
	Numerical (N)	**Categorical (C)**	**N-N**	**N-C**	**C-C**
Statistics	- Mean, mode, median - Range, standard deviation - Quartiles, quintiles - Kurtosis, skewness	- Counts and frequencies	- Correlation coefficients - Linear regression	- Student T-test - ANOVA - Logistic regression	Chi-squared test
Visualisation	Histogram, box plot	Bar plot	Scatter plot	Box plot (for each category)	Heat map (of frequencies)

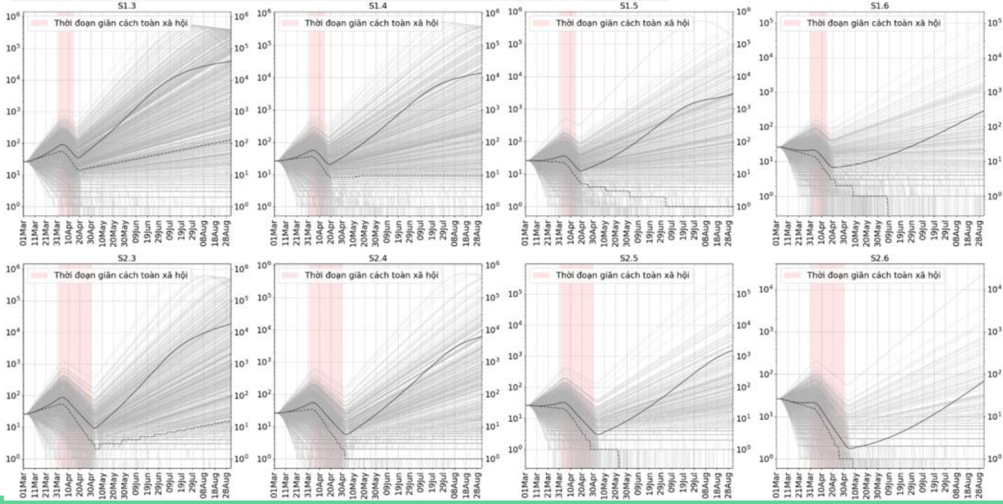
EDA - Exploratory visualisation (more than 2 variables)

Plotting 3 variables, e.g. bubble plots



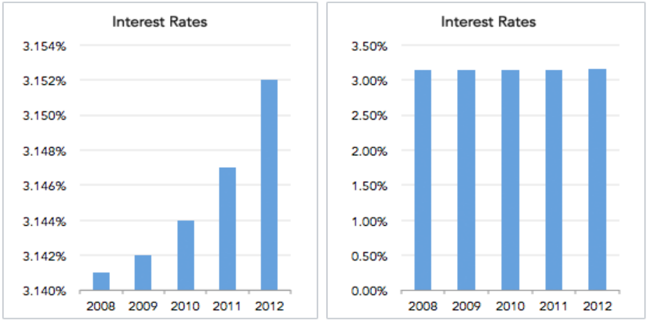
Plotting 4 variables, e.g. side-by-side plots

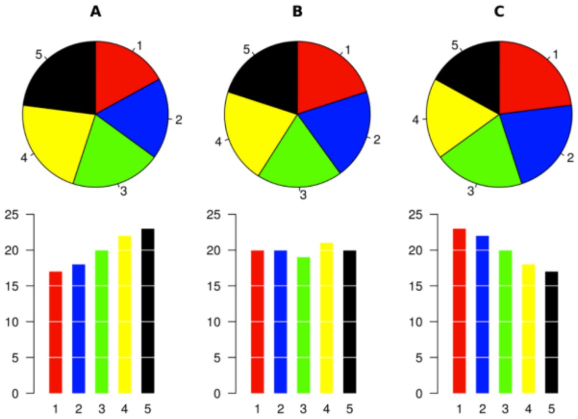
- Consistency - chart type, axis scale, colour scheme
- Arrangement - for easy comparison
- Sequence - following some natural orders

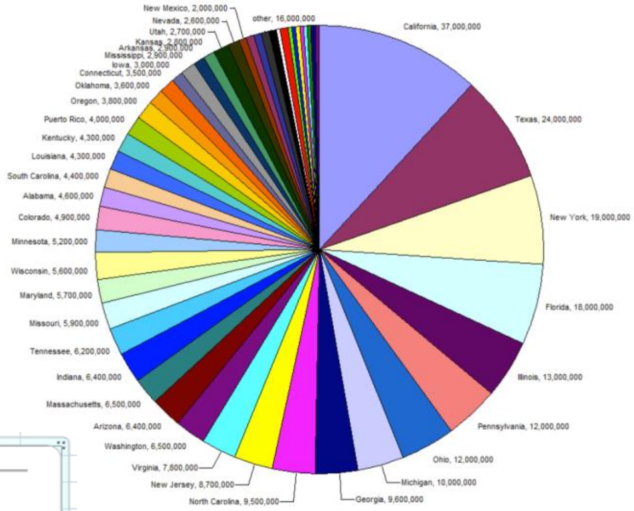


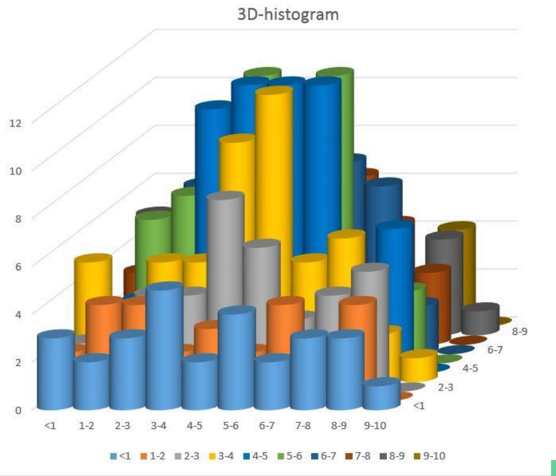
EDA - Exploratory visualisation - Plots to avoid

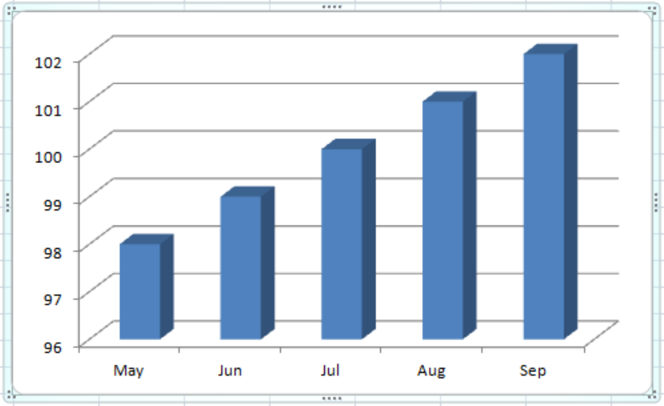
Same Data, Different Y-Axis











EDA - Anomalies - Outliers

> “… an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.

V. Barnett and T. Lewis. *Outliers in Statistical Data*. Wiley, 2nd edition, 1984

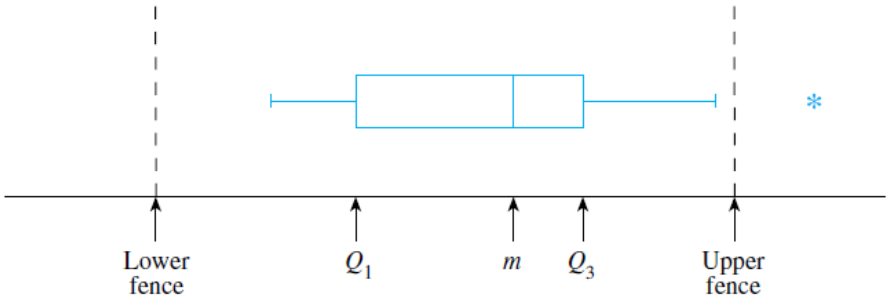
- Outliers significantly change the characteristics of a dataset
- They can be results of *gross data errors* or of *special cases*.

EDA - Anomalies - Outliers

Boxplot outlier identifier

A graphical tool "expressly designed" for isolating outliers from a sample.

$$x_k > Q_3 + 1.5IQR \text{ or } x_k < Q_1 - 1.5IQR$$



3-sigma outlier identifier

Based on the "Empirical rule"

$$|x_k - \bar{x}| > 3\sigma$$

σ is inflated by outliers

Larger outlier values -> larger σ -> larger the bound values -> less effective in identifying unusual values

EDA - Anomalies - Outliers

Hampel outlier identifier

$$|x_k - median| > 3MADM$$

$$MADM = 1.4826 * median(|x_k - median(x)|)$$

x
15
20
25
25
26.32
...
388.06
660
848.48
969.70

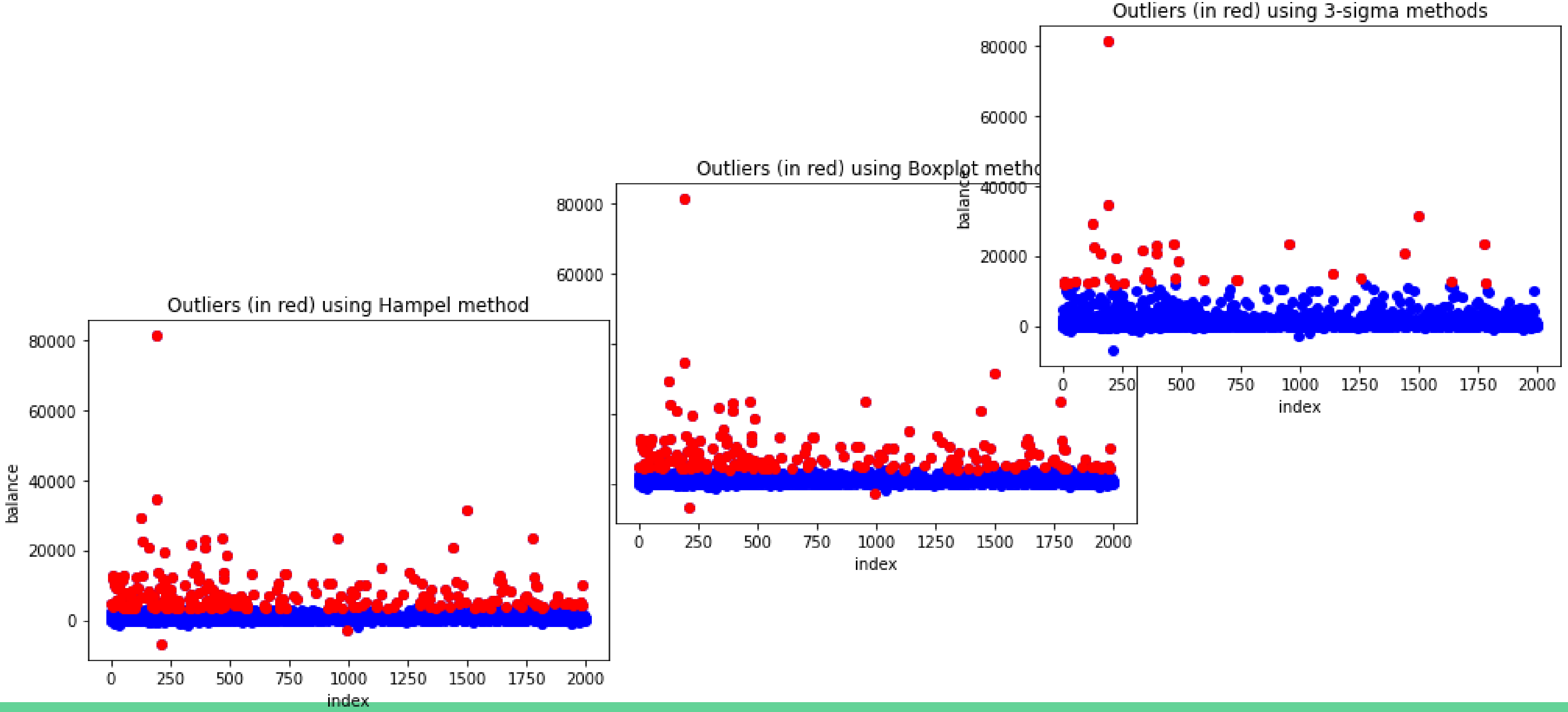
$\longrightarrow$ $median(x) = 96.59$ $\longrightarrow$

y = x-median(x)
81.59
76.59
71.59
71.59
70.28
...
291.47
563.41
751.89
873.11

$\longrightarrow$

$$MADM = 1.4826 * median(y) \\ = 98.73$$

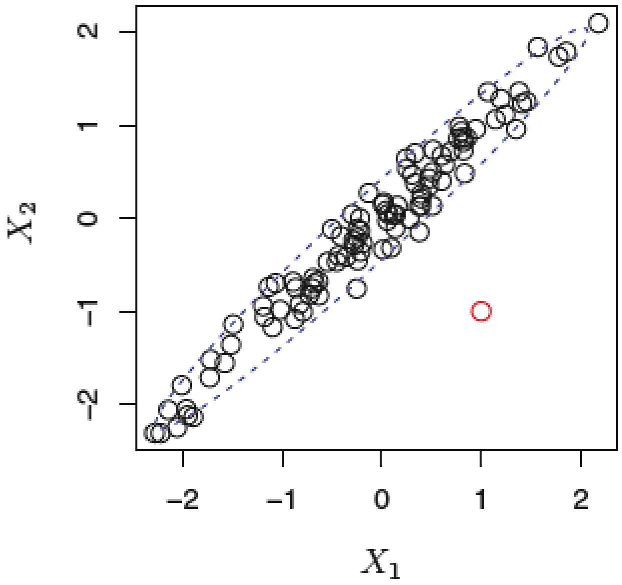
EDA - Anomalies - Outliers



EDA - Anomalies - Outliers

- The three procedures described above may identify different sets of outliers.
- A suggested strategy:
    - Apply all three procedures and compare (i) the number and the value of outliers identified by each procedure, and (ii) the range of the data values not declared as outliers.
    - Apply application-specific assessments, i.e. does the nominal range (excluded outliers) make sense? Do outliers seem extreme enough to be excluded?
    - Visualise the data either with different colours for nominal values and for outliers, or with indication of outlier detection thresholds.
- *Identifying* outliers can be a mathematical procedure – *interpreting* the outliers is NOT.
- Outliers are not necessarily bad data that should be removed/rejected – they simply need further investigation.

Outliers - multidimensional



EDA - Anomalies - Inliers

> “… a data value that lies in the interior of a statistical distribution and is an error.

D. DesJardins. Paper 169: Outliers, inliers and just plain liars – new eda+ techniques for understanding data. In *Proceedings SAS User’s Group International Conference*, SUG126. Cary, NC, USA, 2001

Inliers often represent in the form of values which *repeat unusually frequently*.

	Chile$statusquo	Frequency
1	-1.80301	1
2	-1.74401	1
...	...	...
19	-1.29617	201
20	-1.29293	2
21	-1.28924	1
22	-1.28897	1
23	-1.27876	3
24	-1.27556	1
25	-1.2727	5
...	...	...
2092	2.04859	1
2093	NA	17

EDA - Anomalies - Inliers

Because the majority of numerical values in Chile$statusquo appears only *once*,

- the majority of values in Frequency is 1, median of Frequency is 1, MADM of Frequency is 0 => we cannot use Hampel identifier to detect inliers.
- Quartiles of Frequency are as below

0%	25%	50%	75%	100%
1	1	1	1	201

- Both Hampel and boxplot procedures would declare that all data points in Frequency are outliers!

	Chile$statusquo	Frequency
1	-1.80301	1
2	-1.74401	1
...	...	...
19	-1.29617	201
20	-1.29293	2
21	-1.28924	1
22	-1.28897	1
23	-1.27876	3
24	-1.27556	1
25	-1.2727	5
...	...	...
2092	2.04859	1
2093	NA	17

Frequency

1	2	3	4	5	6	8	9	13	17	18	21	61	201
1955	72	22	19	8	5	4	1	1	2	1	1	1	1

EDA - Anomalies - Inliers

Applying the three-sigma procedure to identify outliers in ‘Frequency’.

- Mean $\bar{x} = 1.29$
- Standard deviation $\sigma = 4.67$
- A value x_k in Frequency is considered outlier if $|x_k - \bar{x}| > 3\sigma$ or $x_k > 15.3$

	Chile$statusquo	Frequency
19	-1.29617	201
39	-1.25795	21
61	-1.21834	18
137	-1.14049	17
2074	1.5877	61
2093	NA	17

Similar to outliers, inliers are not necessarily bad data and need to be rejected/removed – they simply need further investigation.

EDA - Missing values

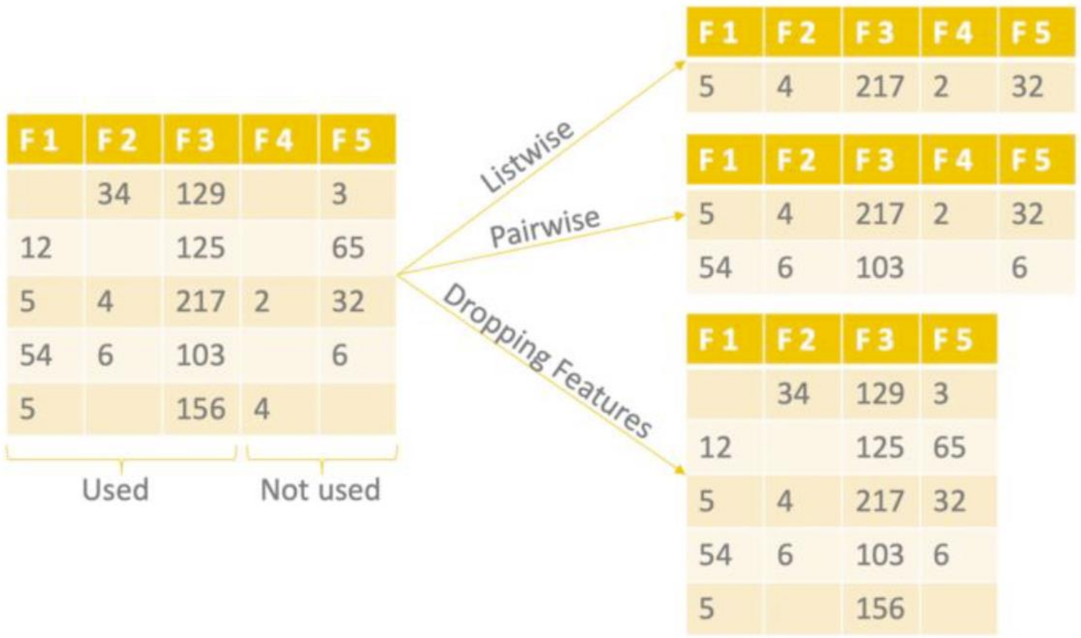
- Sampling
- Data processing errors, e.g. data entry, software engineering, version incompatibility (in apps)
- Data sources
  - 3rd party data, e.g. Tax vs Telco for demographics data
  - 1st party data, e.g. missing required data fields

EDA - Missing values

- **Missing completely at random (MCAR)** - the probability of an instance being missing does not depend on known values nor the missing value itself.
- **Missing at random (MAR)** - the probability of an instance being missing may depend on known values (of other variables), but not on the variable having missing values.
- **Missing not at random (MNAR)**
  - the probability of an instance being missing depends on other variables which also have missing values, or
  - The probability of missingness depends on the very variable itself

EDA - Missing values

Deletion



F1	F2	F3	F4	F5
	34	129		3
12		125		65
5	4	217	2	32
54	6	103		6
5		156	4	

F1	F2	F3	F4	F5
5	4	217	2	32

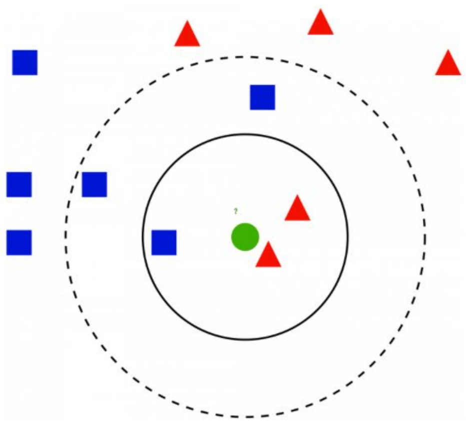
F1	F2	F3	F4	F5
5	4	217	2	32
54	6	103		6

F1	F2	F3	F5
	34	129	3
12		125	65
5	4	217	32
54	6	103	6
5		156	

EDA - Missing values

Single imputation

- A fixed value
- Minimum, maximum, mean (or moving average), median, most frequent value
- Previous or next value (ordered data or time series data)
- K-nearest neighbours
- Regression



EDA - Missing values

Multiple imputation by Chained Equations

- Creates multiple replacements for each missing value, multiple versions of the complete dataset
- Step 1. Make a simple imputation (e.g. mean) for all missing values in the dataset
- Step 2. Set missing values in a variable ‘A’ back to missing
- Step 3. Train a model to predict missing values in ‘A’ using available values of ‘A’ as dependent and other variables in the dataset as independent
- Step 4. Predict missing values in ‘A’ using the trained model in Step 3
- Step 5. Repeat step 2-4 for all other variables with missing values
- Step 6. Repeat step 2-5 for a number of cycles until convergence (or a preset maximum cycles)
- Step 7. Repeat steps 1-6 multiple times with different random number settings to create different versions of the complete/imputed dataset.