Exploratory Data Analysis

Descriptive statistics and visualisation

Lecture outline

- Definitions
- 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

Definitions

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

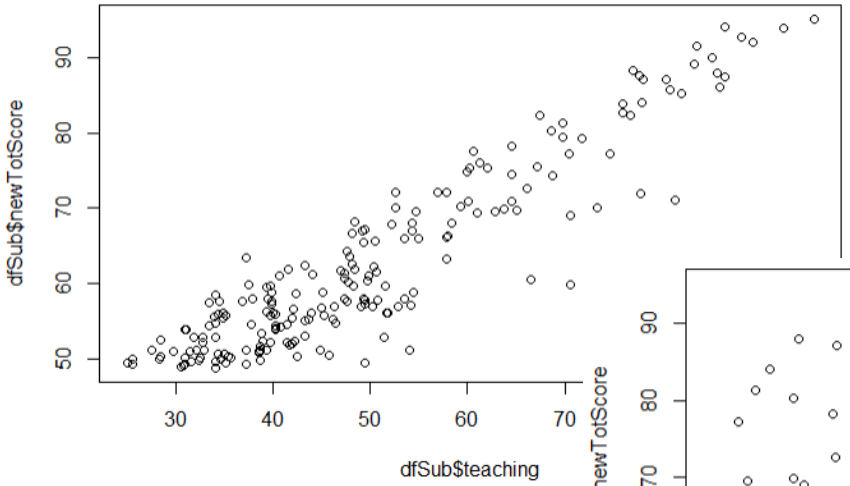
John Tukey, 1977, *Data Exploratory Analysis*, Addison-Wesley

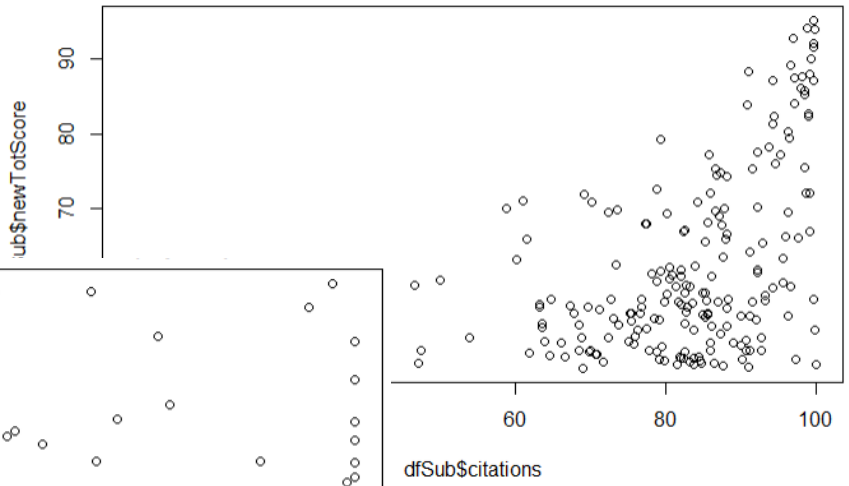
> Exploratory data analysis is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there.

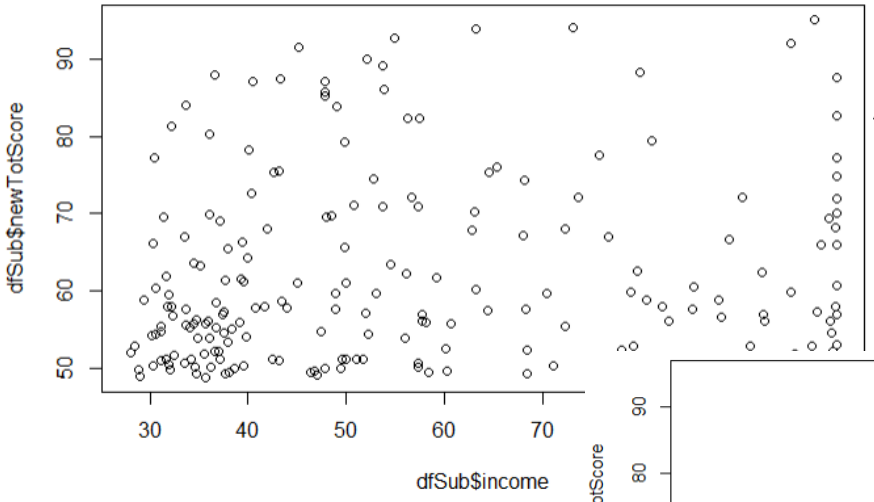
John Tukey, 1977, *Data Exploratory Analysis*, Addison-Wesley

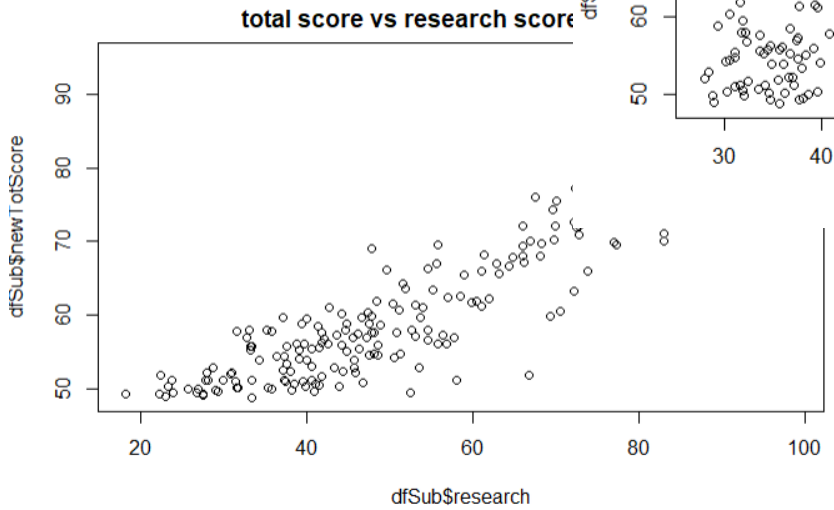
> The primary aim with exploratory data analysis is to examine the data for distribution, outliers and anomalies ... hypothesis generation by visualising and understanding the data.

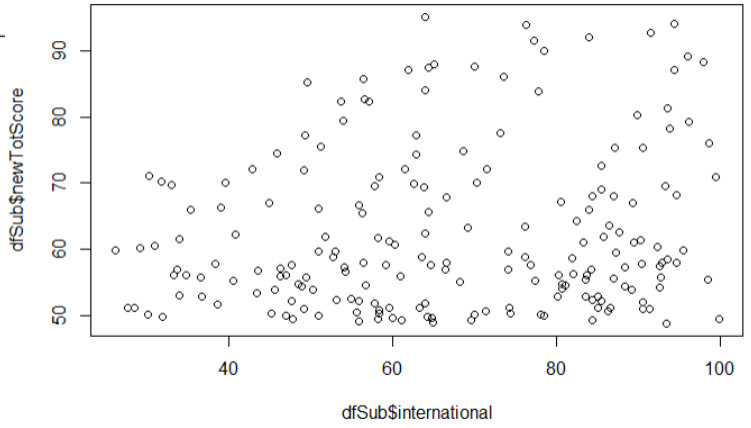
https://link.springer.com/chapter/10.1007/978-3-319-43742-2_15


dfSub$newTotScore
dfSub$teaching


dfSub$newTotScore
dfSub$citations


dfSub$newTotScore
dfSub$income


total score vs research score
dfSub$newTotScore
dfSub$research


dfSub$newTotScore
dfSub$international

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	

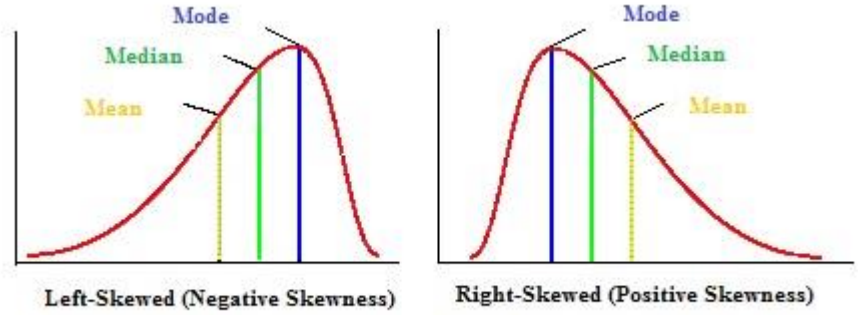
	day	month	duration	campaign	pdays	previous	poutcome	deposit
0	5	may	1042	1	-1	0	unknown	yes
1	5	may	1467	1	-1	0	unknown	yes
2	5	may	1389	1	-1	0	unknown	yes
3	5	may	579	1	-1	0	unknown	yes
4	5	may	673	2	-1	0	unknown	yes

```
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)
```

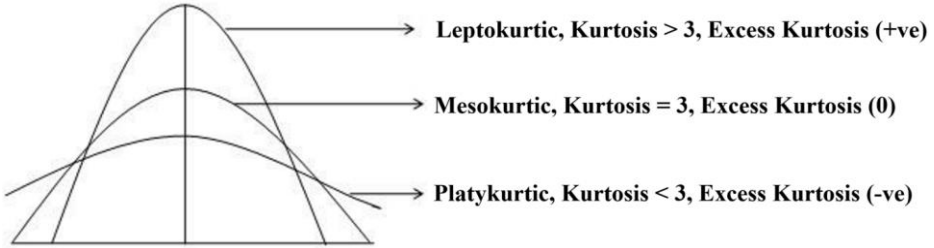
EDA. Descriptive statistics (univariate)

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 variables

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

Quantitative analysis

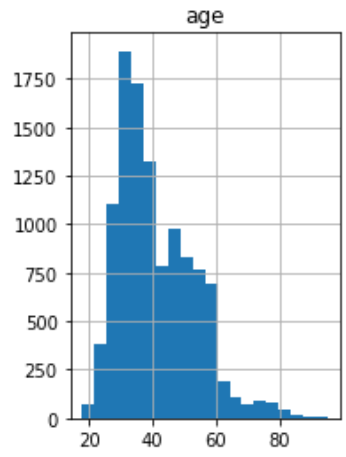
	Categorical	**Numerical**
Categorical	Chi-squared test	Student t-test, ANOVA, Logistic regression
Numerical	Student t-test, ANOVA, Logistic regression	Correlation, Linear regression

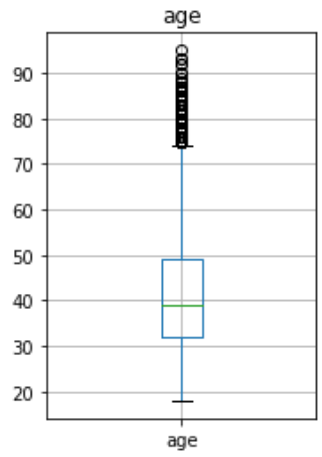
EDA. Exploratory visualisation (1D)

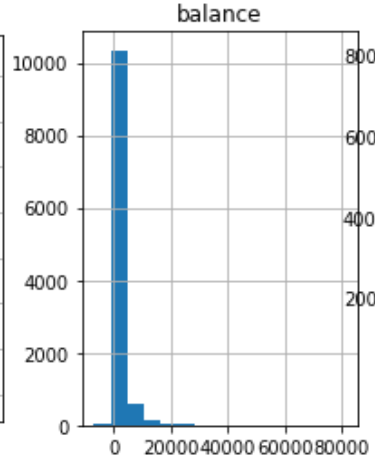
Numerical variables - histogram, boxplot

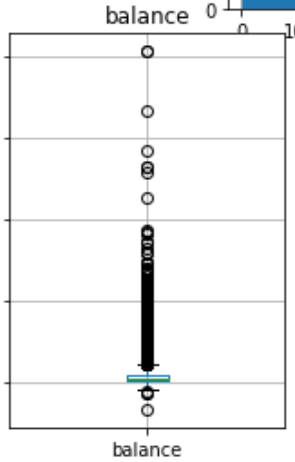
Freedman-Diaconis rule

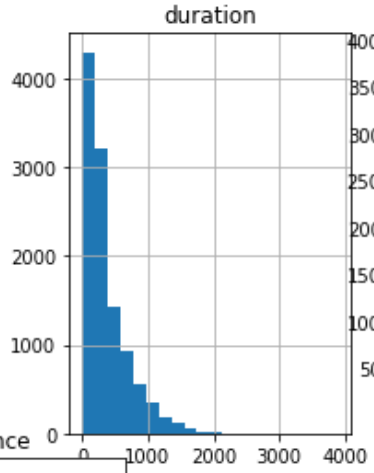
$$\text{Bin width} = 2\,\frac{\text{IQR}(x)}{\sqrt[3]{n}}$$

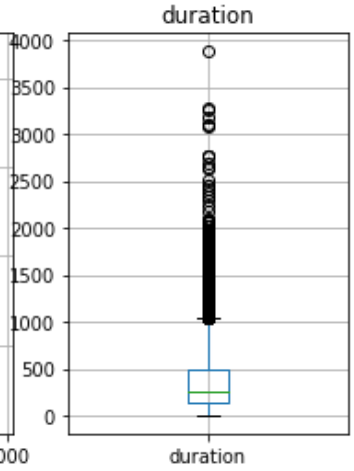






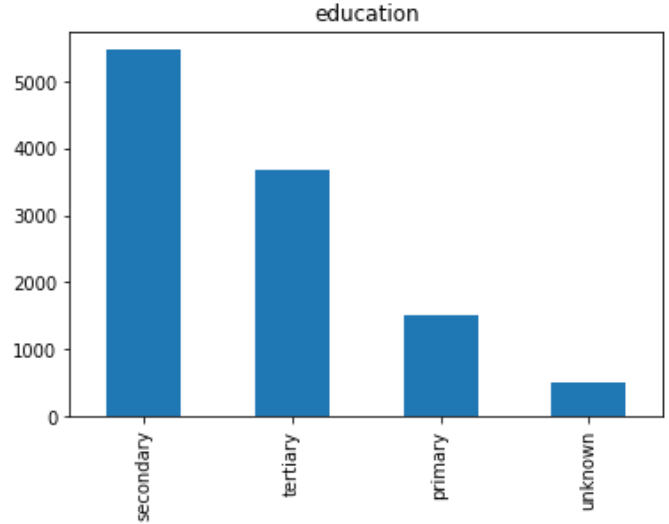


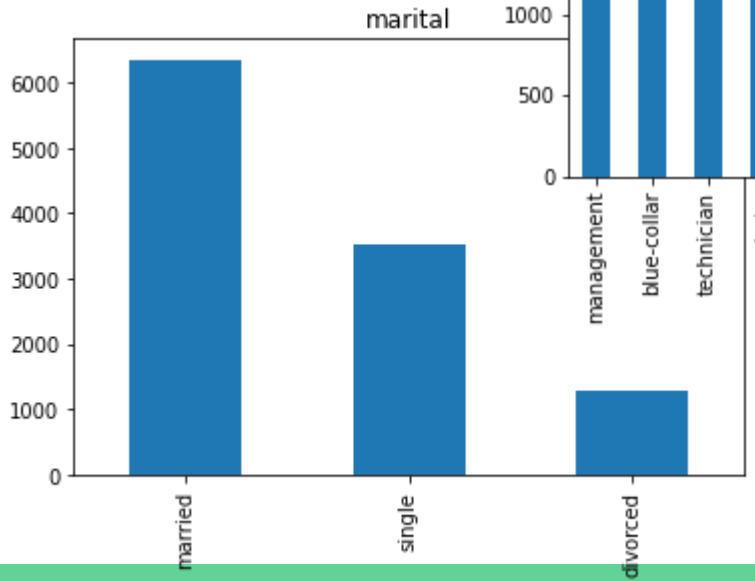


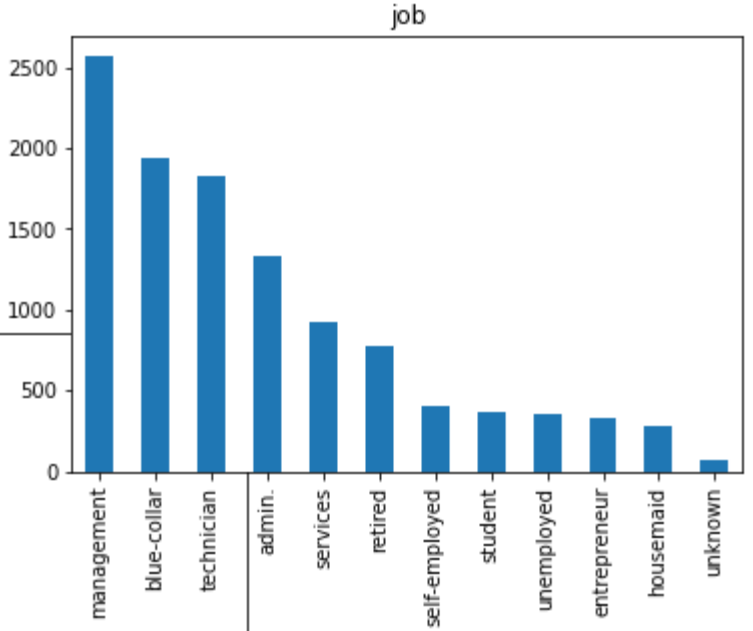


EDA. Exploratory visualisation (1D)

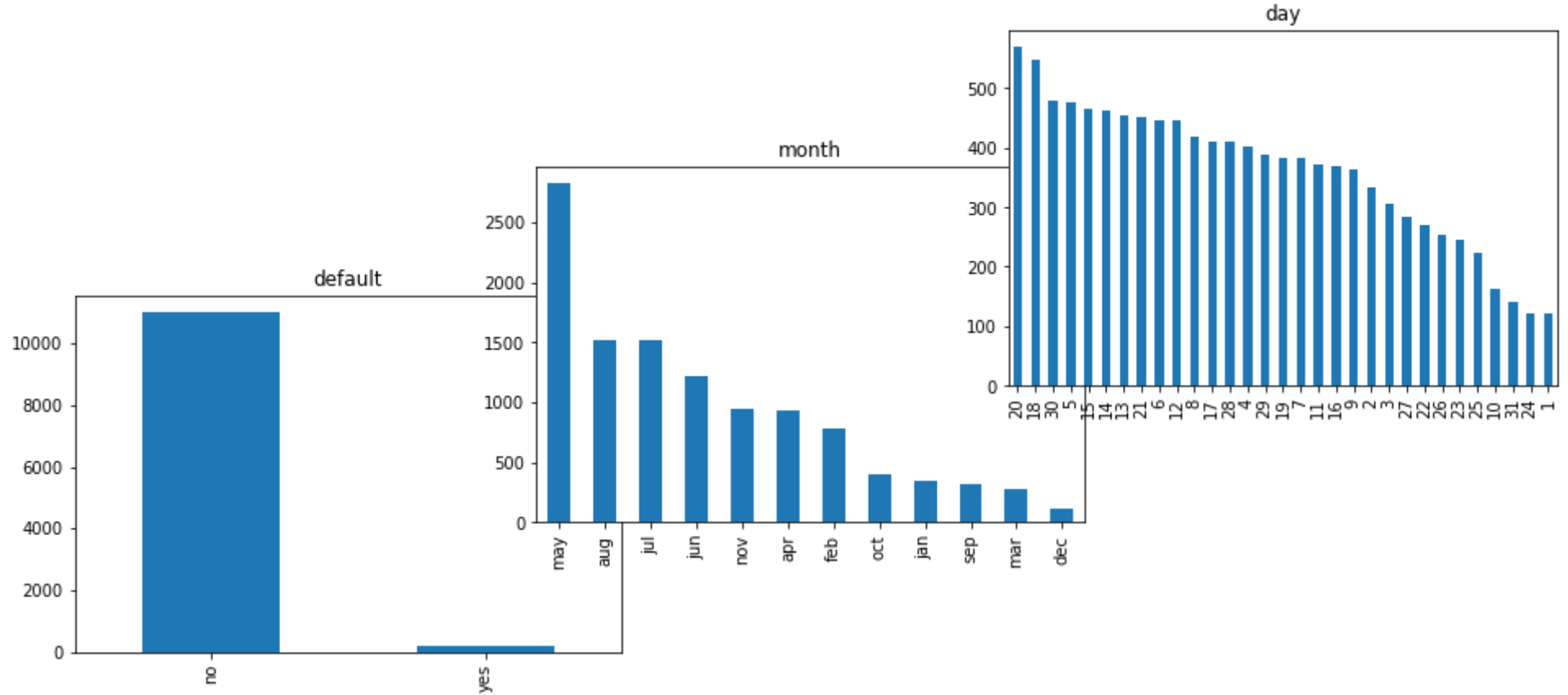
Categorical - Bar plots



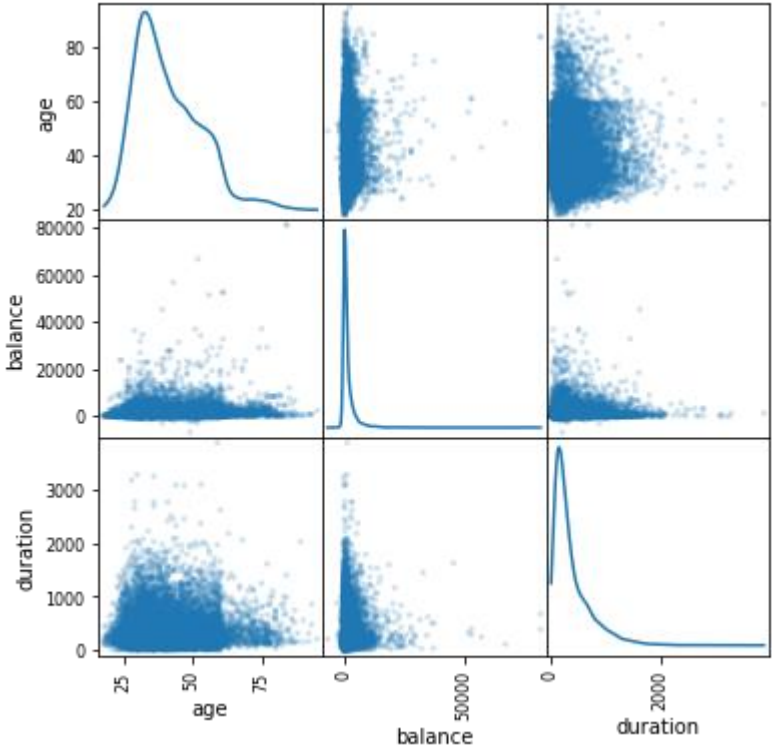


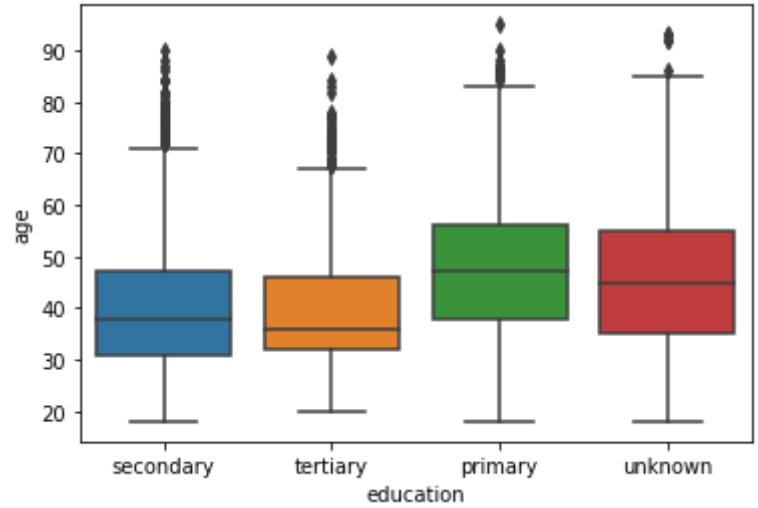


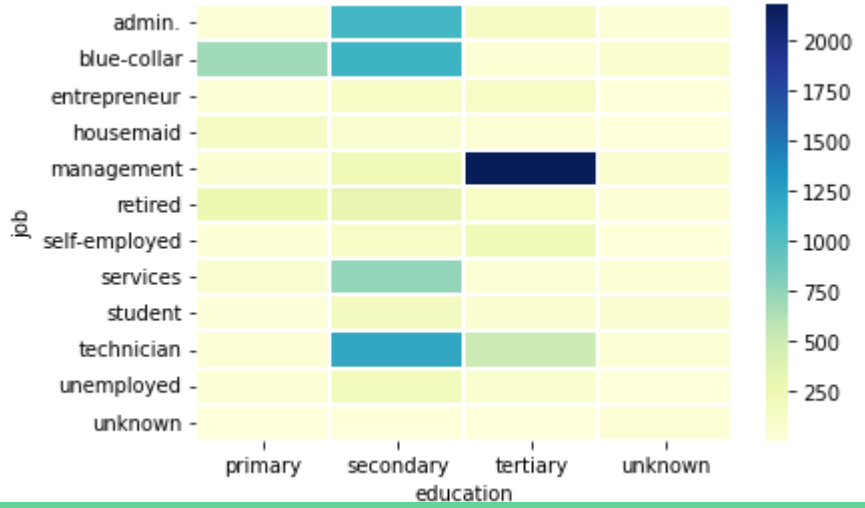
EDA. Exploratory visualisation (1D)



EDA. Exploratory visualisation (2D)





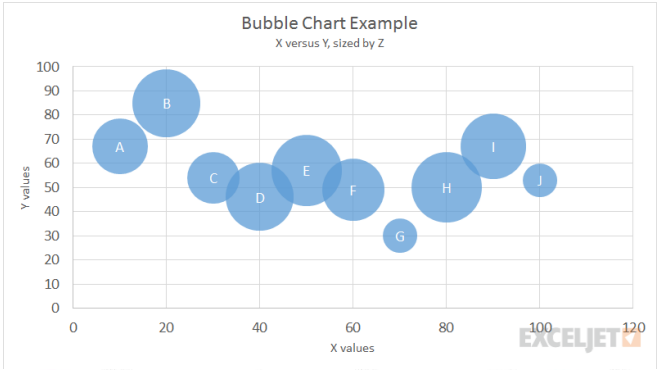


EDA statistics and visualisation summary

	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)

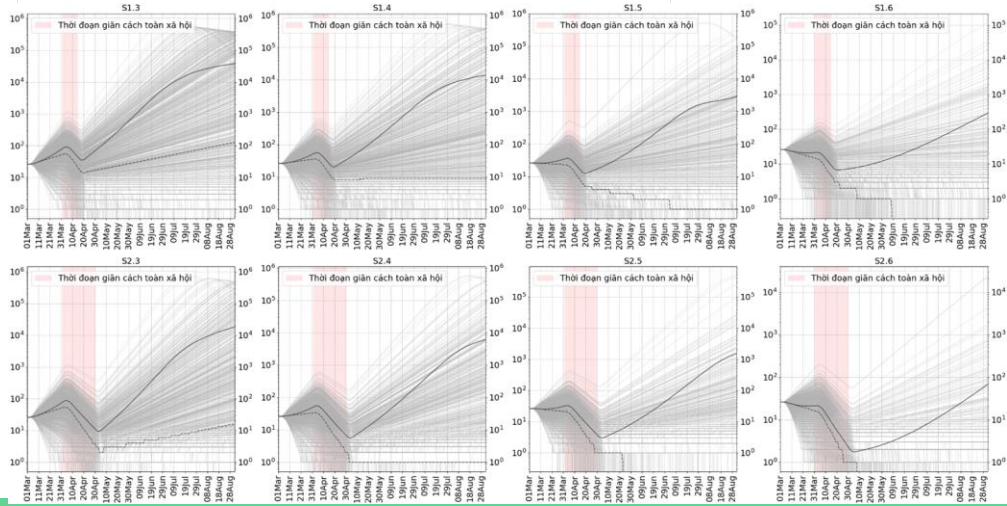
EDA. Exploratory visualisation of more than 2 variables

Plotting 3 variables, e.g. bubble plots

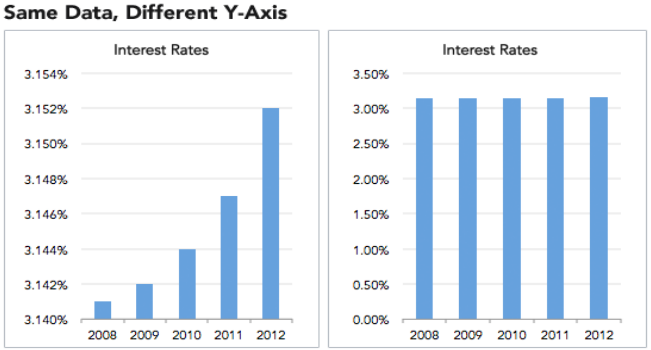


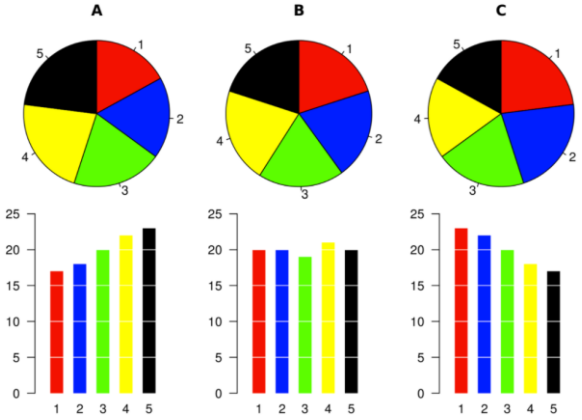
Plotting 4 variables, e.g. side-by-side plots

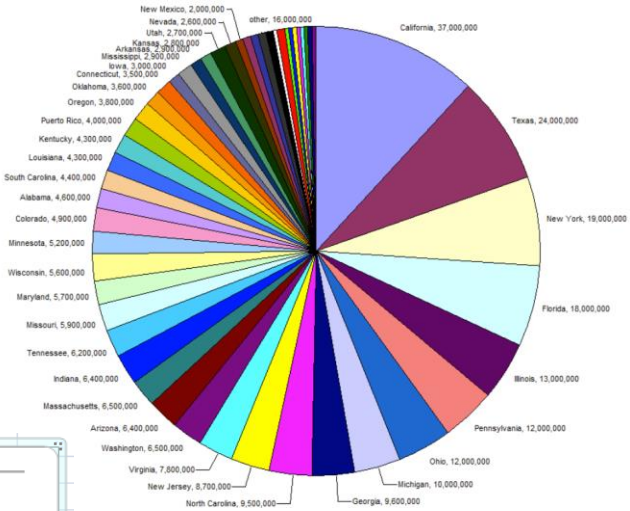
- Consistency - chart type, axis scale, colour scheme
- Arrangement - for easy comparison
- Sequence - following some natural orders

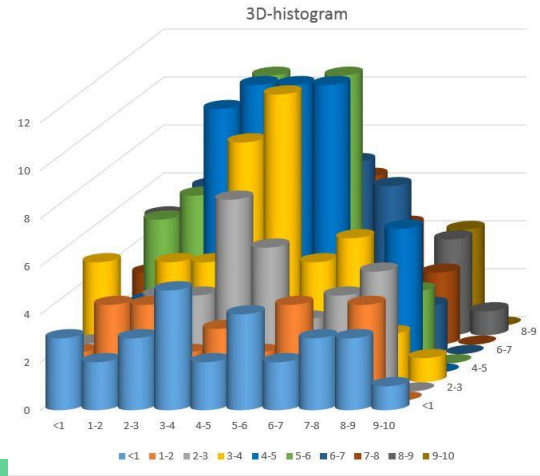


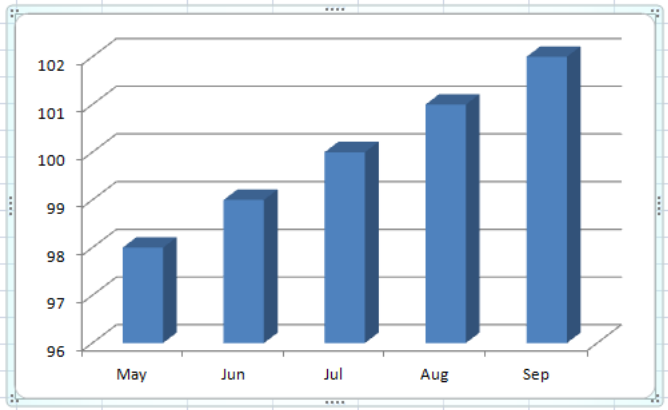
EDA. Exploratory visualisation - Plots to avoid











EDA. Other preprocessing considerations

- Data transformation, e.g. centering and scaling
- Adding variables, e.g. one hot encoding
- Remove variables, those with zero or near zero variance