Data exploration

Ludovic d'Estampes

Mineure - Science des données pour l'ingénieur

January 2024

Outline

Introduction

- Remember the quality of your inputs decide the quality of your output.
- 80% of a data scientist's time and effort is spent in collecting, cleaning and preparing the data for analysis because datasets come in various sizes and are different in nature.
- Data Preparation is the process of transforming raw data into useful information for data analysts
- Data Exploration is the process by which we can interactively explore the data presented to them. Data Exploration typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes.

Types of data sets

- Record: data matrix, document data, transaction data
- Graph: generic graph, HTML Links, chemical data
- Ordered: spatial data, temporal data, sequential data, genetic sequence data

Types of data sets

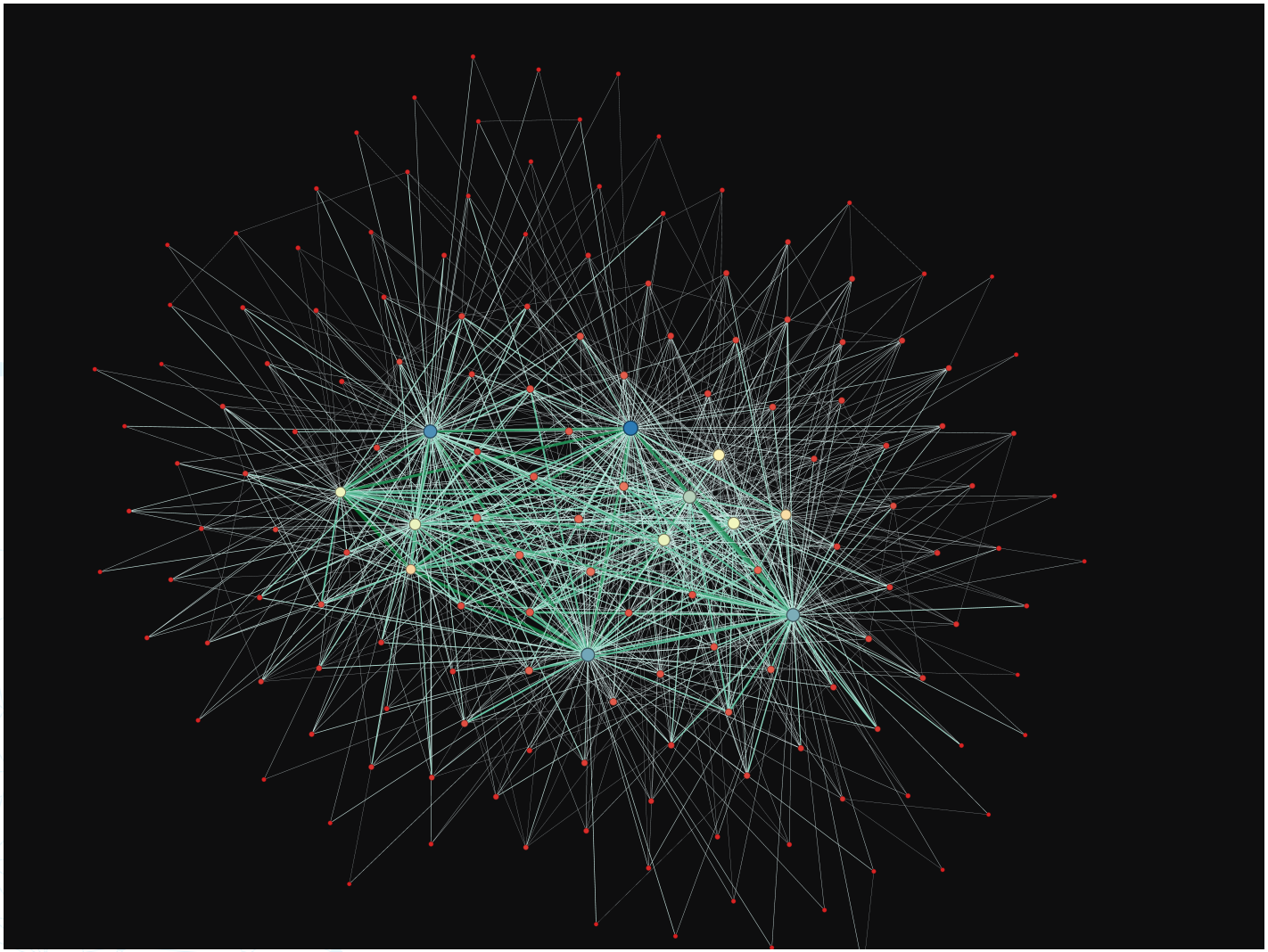
	Year	Airline	Code	Al.Code.ICAO	Country	Country.Code.OAG	Country.Code.IATA	IATA.I
1	2006	AER LINGUS	EI	EIN	IRISH REPUBLIC	441	IRLD	Europ
2	2007	AER LINGUS	EI	EIN	IRISH REPUBLIC	441	IRLD	Europ
3	2008	AER LINGUS	EI	EIN	IRISH REPUBLIC	441	IRLD	Europ
4	2009	AER LINGUS	EI	EIN	IRISH REPUBLIC	441	IRLD	Europ
5	2010	AER LINGUS	EI	EIN	IRISH REPUBLIC	441	IRLD	Europ
6	2011	AER LINGUS	EI	EIN	IRISH REPUBLIC	441	IRLD	Europ
7	2012	AER LINGUS	EI	EIN	IRISH REPUBLIC	441	IRLD	Europ
8	2013	AER LINGUS	EI	EIN	IRISH REPUBLIC	441	IRLD	Europ
9	2014	AER LINGUS	EI	EIN	IRISH REPUBLIC	441	IRLD	Europ
10	2015	AER LINGUS	EI	EIN	IRISH REPUBLIC	441	IRLD	Europ
11	2006	AEROFLOT RUSSIAN AL	SU	AFL	RUSSIAN FEDERATION	475	RUSF	Europ
12	2007	AEROFLOT RUSSIAN AL	SU	AFL	RUSSIAN FEDERATION	475	RUSF	Europ
13	2008	AEROFLOT RUSSIAN AL	SU	AFL	RUSSIAN FEDERATION	475	RUSF	Europ
14	2009	AEROFLOT RUSSIAN AL	SU	AFL	RUSSIAN FEDERATION	475	RUSF	Europ
15	2010	AEROFLOT RUSSIAN AL	SU	AFL	RUSSIAN FEDERATION	475	RUSF	Europ
16	2011	AEROFLOT RUSSIAN AL	SU	AFL	RUSSIAN FEDERATION	475	RUSF	Europ
17	2012	AEROFLOT RUSSIAN AL	SU	AFL	RUSSIAN FEDERATION	475	RUSF	Europ
18	2013	AEROFLOT RUSSIAN AL	SU	AFL	RUSSIAN FEDERATION	475	RUSF	Europ
19	2014	AEROFLOT RUSSIAN AL	SU	AFL	RUSSIAN FEDERATION	475	RUSF	Europ
20	2015	AEROFLOT RUSSIAN AL	SU	AFL	RUSSIAN FEDERATION	475	RUSF	Europ
21	2006	AEROLINEAS ARGENTINAS	AR	ARG	ARGENTINA	303	ARGT	South
22	2007	AEROLINEAS ARGENTINAS	AR	ARG	ARGENTINA	303	ARGT	South
23	2008	AEROLINEAS ARGENTINAS	AR	ARG	ARGENTINA	303	ARGT	South
24	2009	AEROLINEAS ARGENTINAS	AR	ARG	ARGENTINA	303	ARGT	South
25	2010	AEROLINEAS ARGENTINAS	AR	ARG	ARGENTINA	303	ARGT	South
26	2011	AEROLINEAS ARGENTINAS	AR	ARG	ARGENTINA	303	ARGT	South
27	2012	AEROLINEAS ARGENTINAS	AR	ARG	ARGENTINA	303	ARGT	South

Types of data sets

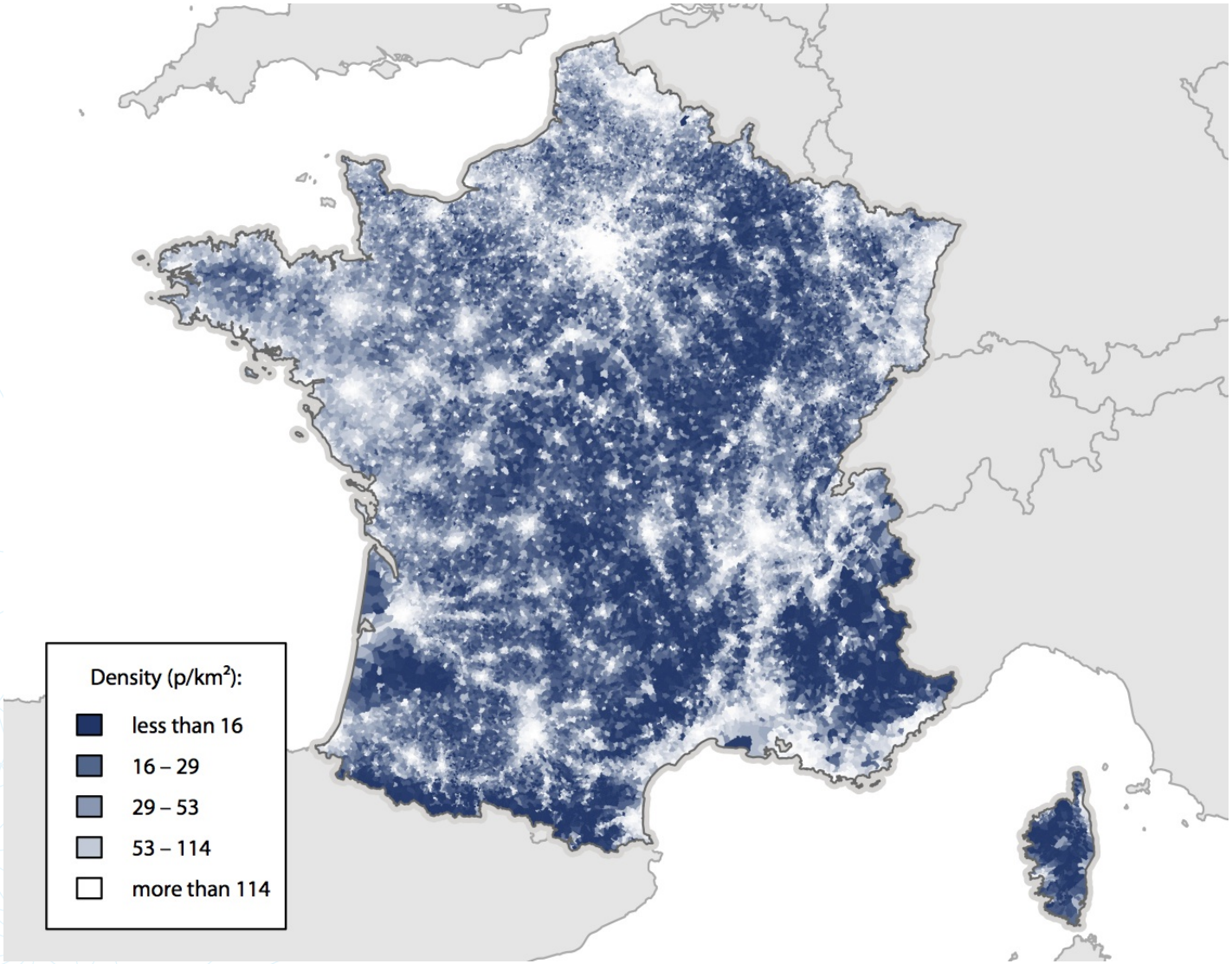
Document Term Matrix

	intelligent	applications	creates	business	processes	bots	are	i	do	intelligence
Doc 1	2	1	1	1	1	0	0	0	0	0
Doc 2	1	1	0	0	0	1	1	0	0	0
Doc 3	0	0	0	1	0	0	0	1	1	1

Types of data sets



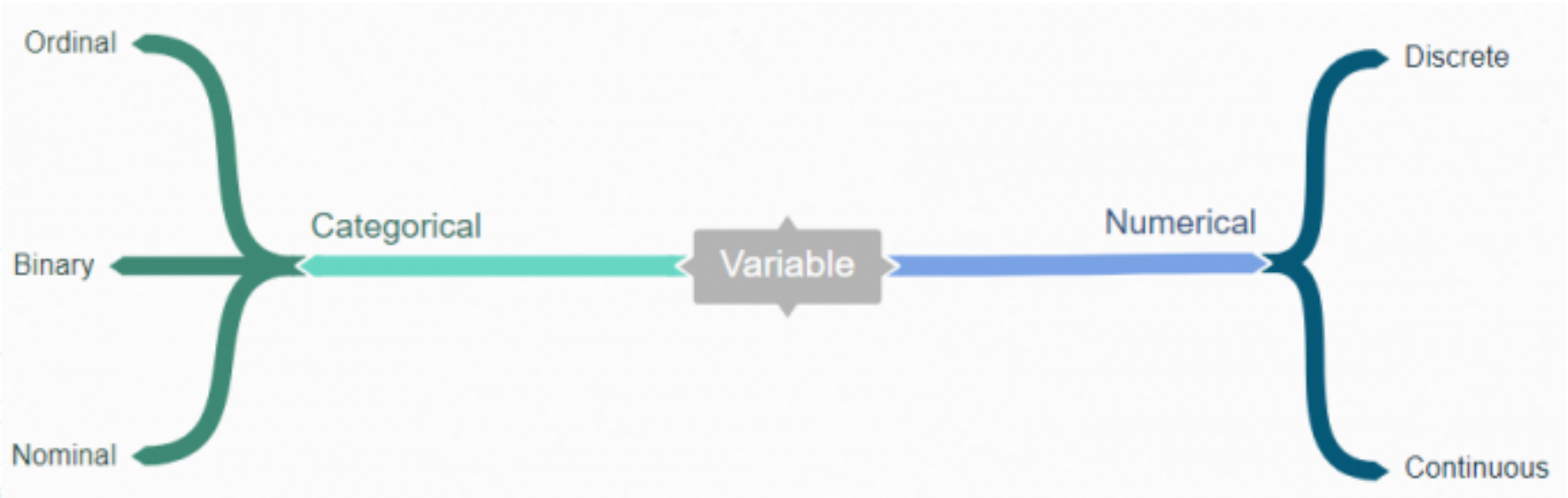
Types of data sets



Types of data sets

```
>PB22887-RA class=Sequence position=scf7180000350335:1110961..1112267 (+ strand)
ATTCTTTCTT TCAGCGTTCA TCCTACGAGA AGGTATGTCA TATCAGATTA TATTTCGAAA TATGTATATC GTGTTTTATA
TATAGTTTTC TAAGTATGTA GAGGATTAAT AAATTATAAA TGTTATTACA GTCGAGGAAG CATAAAACCA CAAAAATGGT
AGCTCAAAAG AAACAGGTAA TTACACTGTT GCACGTCCAC GTGTTCGTCC GTAAGTTTTC TCGTAGCAAA ATTTCATGCT
TTTCAATATA TGCGTAGATA TGATCTTGTT ATATCATTAT TATAATTGTT TTAACTGTAA ATCGTTGCTT TTACATTGAT
TGAAATAATA TAAAACTCGA GTGTACCATA GTAATCTGAA AGGAACCGCT TGATTTTGCT ATTCTCTTTC TAAATTTTGC
TTCTATCTAC TATGAGAAAT ATAGATTCTA TTTTCAATTT CTCAAATCTA CAAATCGTAT CGCTATTCTG ATCGACAATT
TATAATTTCT TGTTACTTCA GAACGTGAGT TATTAAGTTA ATCTAAATAT ATCAATGTAT GTAATGAGTT TTATTACCTG
ATCATATAGA AAAAGGCTCA GGAGAGCATC AACACCAGAT TAGCACTCGT CATGAAATCT GGTAAATACG TCCTTGGTTA
TAAACAAACT CTGAAGTCAC TCCGCCAAGG CAAAGCTAAA TTGGTCATCA TTGCTAGCAA TACGCCACCG CTAAGGTGAG
ACTGAGAAGT AGCACTCCTT TTGTATTGGC AAATTGACAT ATTTAAAATA AATAATTTCT TTTATAAGCG AATTAATTCA
AAATTTAATG AATATAATCT TTGATAAATG TATACATATA TATATACATA TATTTGTAAA AATTATATTC TATTGTAATT
TAATTTTATT AGAATCTAGT GAAATATTAA ATAAATAATT ACATTAGAGA TGCAGTAATA GAAACTAGTA ATATTGATAT
AATCATCAAT TTGCATATTG GAAAACAAAT AGGAGTACTA TGGAGTAGTC TCCAGCTCCT GCTTGTAAGT GCATGGGGCA
TTAAGATTTA ATTTTTTAAT GTTTTAAACA TAATTTATCT CGATTTAACA GGAAGTCGGA GATTGAATAC TATGCAATGT
TAGCGAAGAC TGGTGTGCAT CATTACACGG GGAATAACAT CGAACTGGGT ACAGCTTGTG GTAAATATTT CCGTGTCTGT
ACACTCTCGA TCACAGATCC TGGTAATTCT GACATTATAA AATCTATGCC AACTGGTGAT CAAGCGTAAT GTACAGTTTT
TAATCCAATA AATAATTCAA AACGTTT
```

Types of data (I)



Types of data (II)

- Categorical data represents characteristics; e.g gender, questionnaires
- Nominal values are used to labels variables, that have no quantitative values. Nominal data has no order.
- Ordinal values represent discrete and ordered units; e.g ranking of product, level of educational background
- Numerical data is information that is measurable, and it is, of course, data represented as numbers and not words or text.
- Continuous numbers are numbers that does not have a logical end to them; e.g money or height.
- Discrete numbers are the opposite; they have a logical end to them; e.g days in the month, or number of bugs logged.

Outline

What is Data exploration?

- Visualization and calculation to better understand characteristics of data.
- Key motivations of data exploration
  - Helping to select the right tool for preprocessing or analysis
  - Making use of human's abilities to recognize patterns (people can recognize patterns not captured by data analysis tools)
- Below are the steps involved to understand, clean and prepare your data for building your predictive model:
  - Variable Identification
  - Univariate Analysis
  - Bi-variate Analysis
  - Missing values treatment
  - Outlier treatment
  - Variable transformation
  - Variable creation

Exploratory Data Analysis Methods?

EDA Methods

Descriptive statistics

Central tendency of a distribution

Mean	$\frac{\text{Sum of all values}}{\text{Total number of values}}$
Median	Middle value(when data are arranged in order
Mode	Most common value

Measure of Variation

Variance	**how far a set of numbers are spread out from mean**
Interquartile range	divides a data set into quartiles.
Standard deviation	dispersion of a set of data from mean

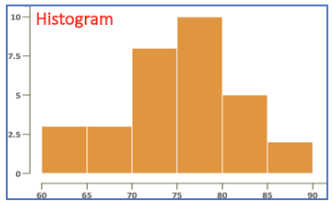
Skewness & Kurtosis

Skewness	**Measure of symmetry**
Kurtosis	**Kurtosis is a measure of "peakedness" relative to a Gaussian shape**

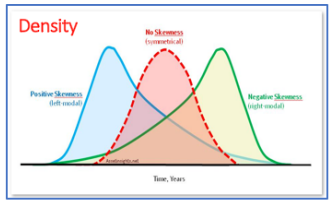
Visualizations

1-dimension

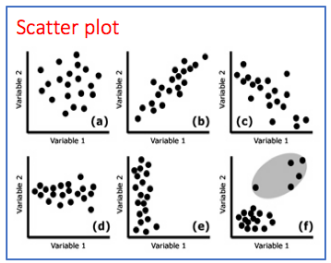
Few data points



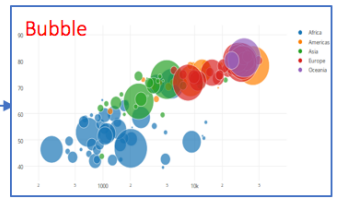
Many data points



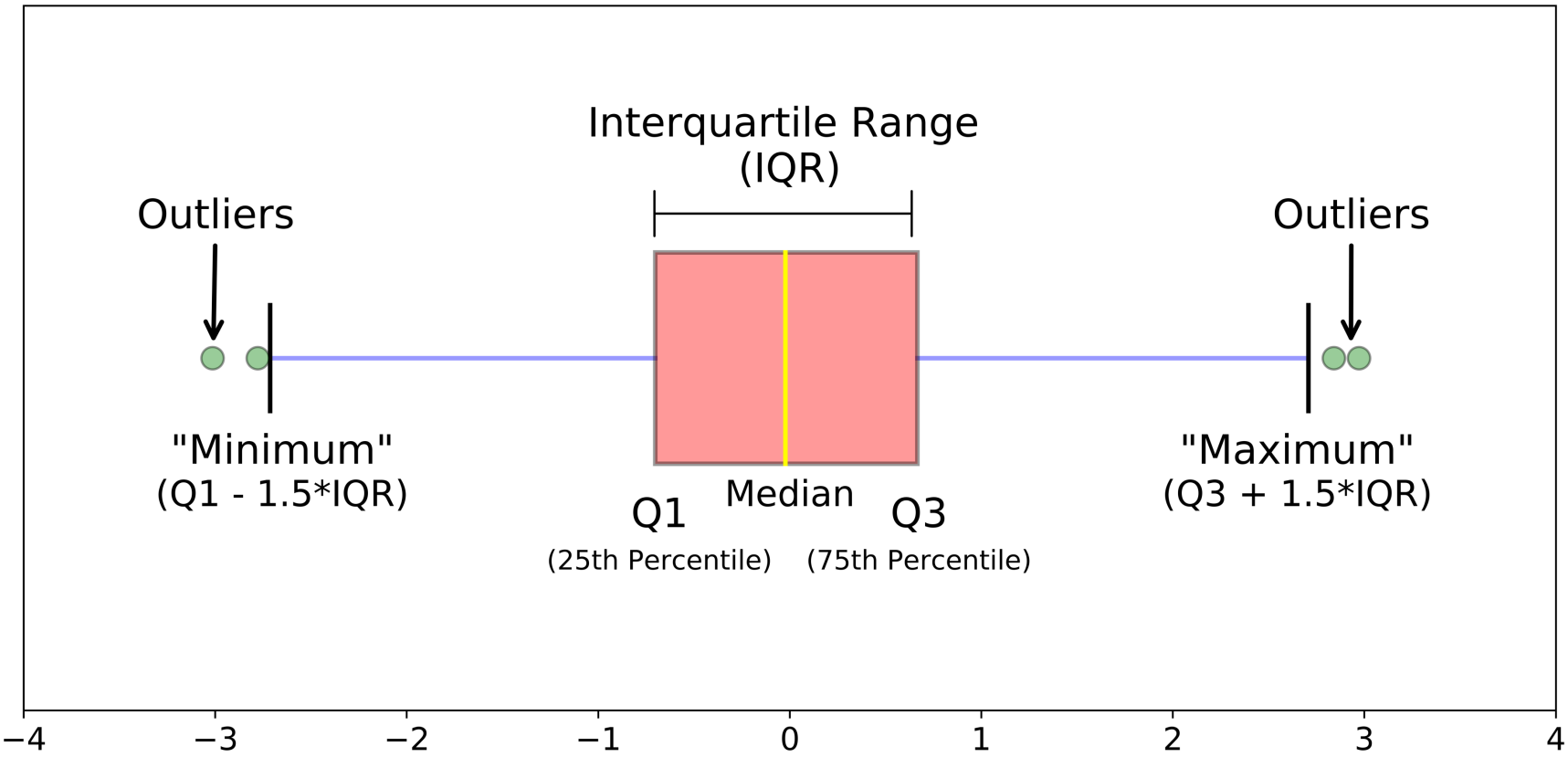
2-dimension



3-dimension



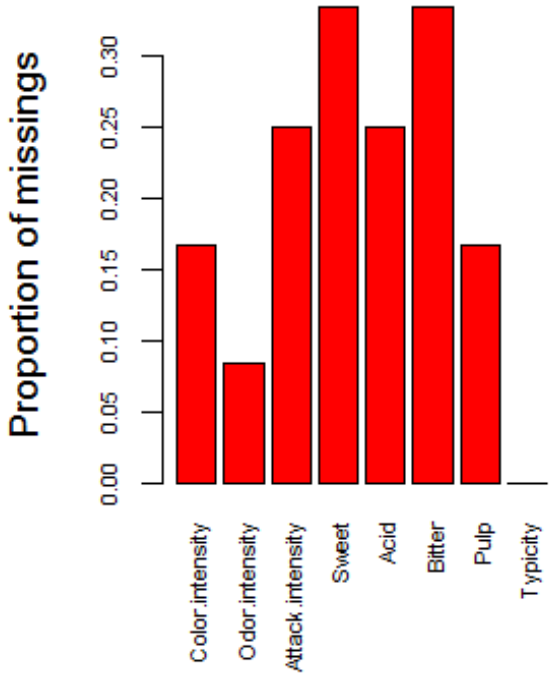
Boxplot

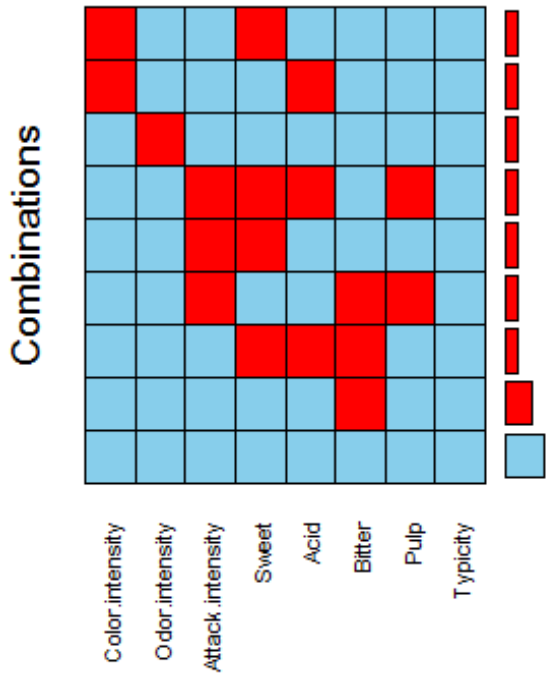

Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
Q1
(25th Percentile)
Median
Q3
(75th Percentile)
"Maximum"
(Q3 + 1.5*IQR)
−4
−3
−2
−1
0
1
2
3
4

Outline

Why missing values treatment is required?

- Missing data in the training data set can reduce the power/fit of a model or can lead to a biased model because we have not analysed the behaviour and relationship with other variables correctly.
- It can lead to wrong prediction or classification.





Why my data has missing values?

- Problems with extraction process, these errors are relatively easy to find and can be corrected as well.
- Errors occurring at time of data collection and are harder to correct.
  - **Missing completely at random (MCAR)**: the probability of missing variable is same for all observations.
  - **Missing at random (MAR)**: the probability of being missing is the same only within groups defined by the observed data. For example: We are collecting data for age and female has higher missing value compare to male.
  - **Missing not at random (MNAR)**: the probability of being missing varies for reasons that are unknown to us. An example of MNAR in public opinion research occurs if those with weaker opinions respond less often. Strategies to handle MNAR are to find more data about the causes for the missingness, or to perform what-if analyses to see how sensitive the results are under various scenarios.

See : Little, R. J. A., and D. B. Rubin. 2002. Statistical Analysis with Missing Data. 2nd ed. New York: John Wiley & Sons.

List wise or pair wise deletion

- The most common approach to the missing data is to simply omit those cases with the missing data and analyse the remaining data. This approach is known as listwise deletion.
- Some researchers insist that it may introduce bias in the estimation of the parameters. However, if the assumption of MCAR is satisfied, a listwise deletion is known to produce unbiased estimates and conservative results. When the data do not fulfil the assumption of MCAR, listwise deletion may cause bias in the estimates of the parameters.

User	Device	OS	Transactions
~~A~~	~~Mobile~~	~~NA~~	~~5~~
B	Mobile	Android	3
~~C~~	~~NA~~	~~iOS~~	~~2~~
D	Tablet	Android	1
E	Mobile	iOS	4

Pairwise deletion

- Pairwise deletion eliminates information only when the particular data-point needed to test a particular assumption is missing. Since a pairwise deletion uses all information observed, it preserves more information than the listwise deletion, which may delete the case with any missing data.
- Pairwise deletion is known to be less biased for the MCAR or MAR data, and the appropriate mechanisms are included as covariates. However, if there are many missing observations, the analysis will be deficient.

User	Device	OS	Transactions
A	Mobile	~~NA~~	5
B	Mobile	Android	3
C	~~NA~~	iOS	2
D	Tablet	Android	1
E	Mobile	iOS	4

Missing Data Imputation

For categorical variables, there are 2 methods to impute the data.

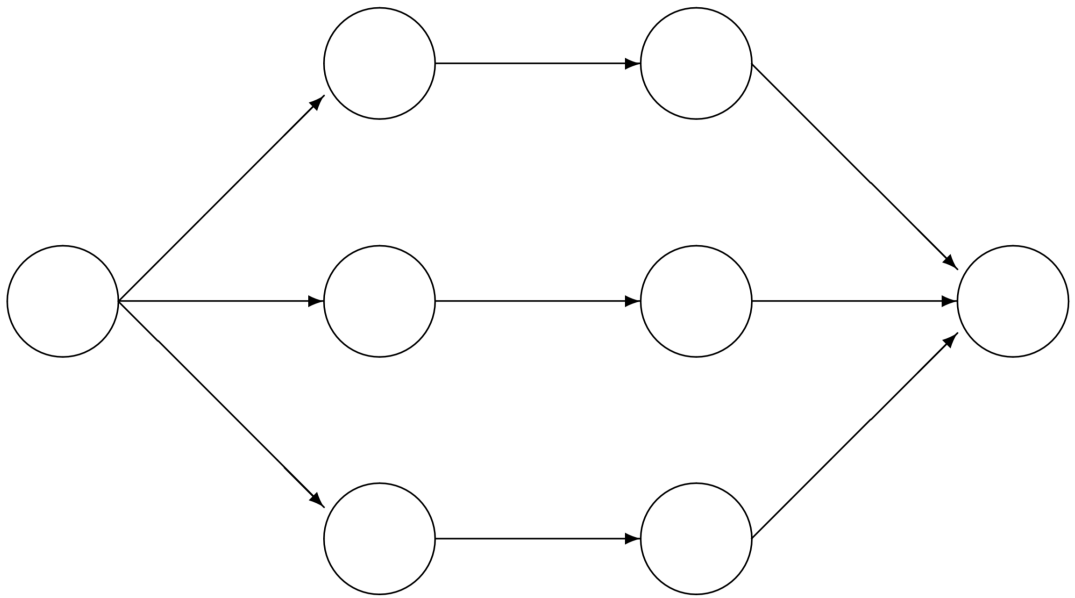
- Create a new level out of the missing values
- Use predictive models such as logistic regression, K Nearest Neighbors (KNN) to estimate the data

For continuous variables, there are 2 methods to impute the data.

- Use Mean, Median, Mode. Mean imputation is a fast and simple but it will underestimate the variance, disturb the relations between variables, bias almost any estimate other than the mean and bias the estimate of the mean when data are not MCAR. Mean imputation should be avoided in general.
- Use predictive models such as linear regression. Regression imputation is a dangerous method since it artificially strengthens the relations in the data. Correlations are biased upwards. Variability is underestimated. Imputations are too good to be true.

Missing Data Imputation

A more convenient solution is multiple imputation. Multiple imputation creates $m > 1$ complete datasets. Each of these datasets is analysed by standard analysis software. The m results are pooled into a final point estimate plus standard error by pooling rules



Missing Data imputation using PCA

	Color.intensity	Odor.intensity	Attack.intensity	Sweet	Acid	Bitter	Pulp	Typicity
1	4.791667	5.291667	NA	NA	NA	2.833333	NA	5.208333
2	4.583333	6.041667	4.416667	5.458333	4.125000	3.541667	4.625000	4.458333
3	4.708333	5.333333	NA	NA	4.291667	3.166667	6.250000	5.166667
4	6.583333	6.000000	7.416667	4.166667	6.750000	NA	1.416667	3.416667
5	NA	6.166667	5.333333	4.083333	NA	4.375000	3.416667	4.416667
6	6.333333	5.000000	5.375000	5.000000	5.500000	3.625000	4.208333	4.875000
7	4.291667	4.916667	5.291667	5.541667	5.250000	NA	1.291667	4.333333
8	NA	4.541667	4.833333	NA	4.958333	2.916667	1.541667	3.958333
9	4.416667	NA	5.166667	4.625000	5.041667	3.666667	1.541667	3.958333
10	4.541667	4.291667	NA	5.791667	4.375000	NA	NA	5.000000
11	4.083333	5.125000	3.916667	NA	NA	NA	7.333333	5.250000
12	6.500000	5.875000	6.125000	4.875000	5.291667	4.166667	1.500000	3.500000

	Color.intensity	Odor.intensity	Attack.intensity	Sweet	Acid	Bitter	Pulp	Typicity
1	4.791667	5.291667	4.077034	5.527352	4.177564	2.833333	5.711715	5.208333
2	4.583333	6.041667	4.416667	5.458333	4.125000	3.541667	4.625000	4.458333
3	4.708333	5.333333	4.158054	5.442936	4.291667	3.166667	6.250000	5.166667
4	6.583333	6.000000	7.416667	4.166667	6.750000	4.702509	1.416667	3.416667
5	6.271605	6.166667	5.333333	4.083333	5.455805	4.375000	3.416667	4.416667
6	6.333333	5.000000	5.375000	5.000000	5.500000	3.625000	4.208333	4.875000
7	4.291667	4.916667	5.291667	5.541667	5.250000	3.214232	1.291667	4.333333
8	4.460613	4.541667	4.833333	5.479128	4.958333	2.916667	1.541667	3.958333
9	4.416667	5.136550	5.166667	4.625000	5.041667	3.666667	1.541667	3.958333
10	4.541667	4.291667	4.176991	5.791667	4.375000	2.735255	4.026062	5.000000
11	4.083333	5.125000	3.916667	5.703297	3.900164	2.815857	7.333333	5.250000
12	6.500000	5.875000	6.125000	4.875000	5.291667	4.166667	1.500000	3.500000

Outline

Outliers

Outliers

"Observation which deviates so much from other observations as to arouse suspicion it was generated by a different mechanism" — Hawkins (1980)

Most common causes

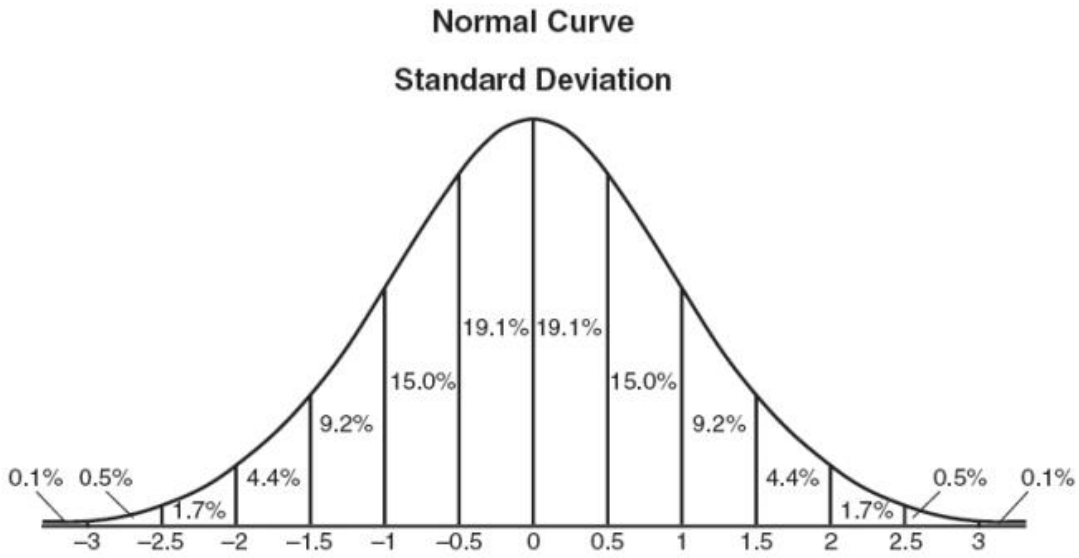
- Data entry errors (human errors)
- Measurement errors (instrument errors)
- Experimental errors (data extraction or experiment planning/executing errors)
- Intentional (dummy outliers made to test detection methods)
- Data processing errors (data manipulation or unintended mutations)
- Sampling errors (extracting or mixing data from wrong or various sources)

Detecting outliers

- Detecting outliers is of major importance for almost any quantitative discipline.
- When trying to detect outliers in a dataset it is very important to keep in mind the context and try to answer the question: "Why do I want to detect outliers?" The meaning of your findings will be dictated by the context.
- A continuous outliers is not necessarily an aberrant value, it may correspond to a particular profile, of which it is necessary to judge whether or not to keep.
- It can also correspond to a rare profile, interesting to detect, whose suppression impoverishes the sample.

Popular Methods for Outlier Detection: Z-Score

- The z-score, $z = (x - \mu)/\sigma$, of an observation is a metric that indicates how many standard deviations a data point is from the sample's mean, assuming a Gaussian distribution. For data points not described by a Gaussian distribution, we can applied transformations to data (i.e. scaling). By "tagging" the data points that lay beyond a given threshold, we are classifying data into outliers and not outliers.

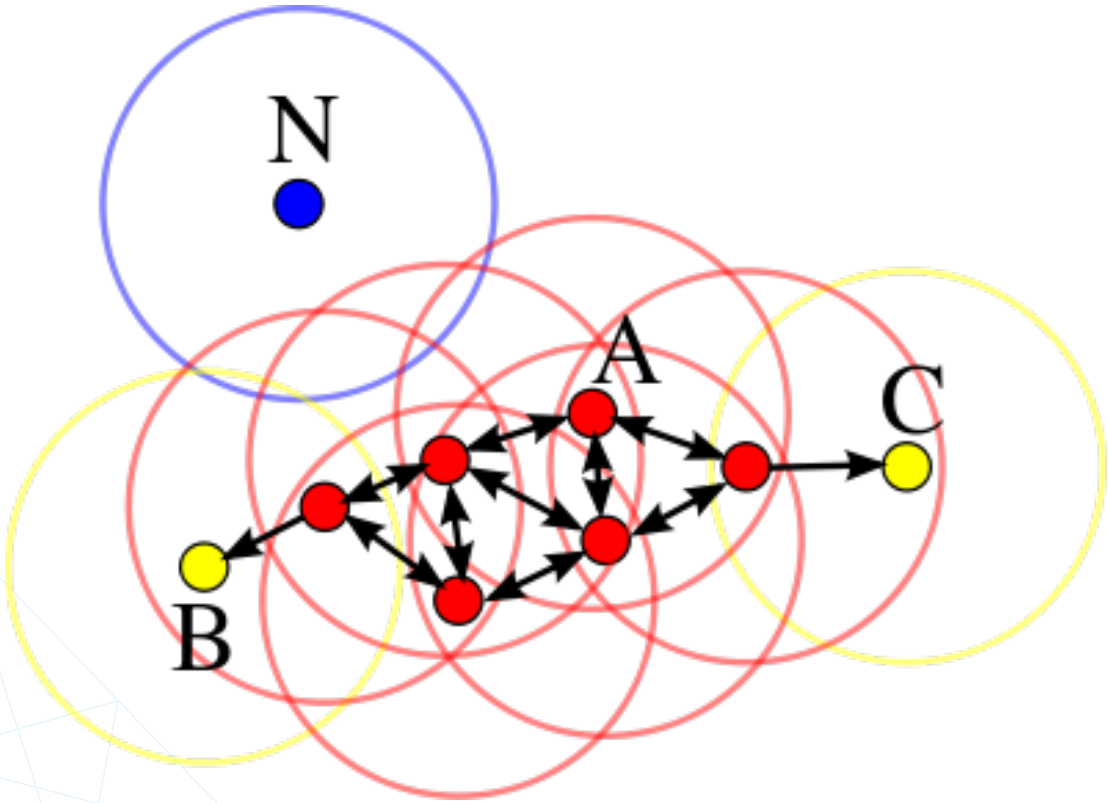


Popular Methods for Outlier Detection: Dbscan

Dbscan is a density based clustering algorithm, it is focused on finding neighbours by density (MinPts) on an "n-dimensional sphere" with radius ϵ. A cluster can be defined as the maximal set of "density connected points" in the feature space. Dbscan defines 3 classes of points:

- **Core point**: A is a core point if its neighbourhood (defined by ϵ) contains at least MinPts points.
- **Border point**: C is a border point that lies in a cluster and its neighbourhood contain less points than MinPts, but it is still "density reachable" by other points in the cluster.
- **Outlier**: N is an outlier point that lies in no cluster and it is not "density reachable" nor "density connected" to any other point. Thus this point will have "his own cluster".

Popular Methods for Outlier Detection: Dbscan



Popular Methods for Outlier Detection: Isolation Forest

- Isolation forest is a method based on binary decision trees. Isolation forest's basic principle is that outliers are few and far from the rest of the observations.
- To build a tree (training), the algorithm randomly picks a feature from the feature space and a random split value ranging between the maximums and minimums. This is made for all the observations in the training set. To build the forest, a tree ensemble is made averaging all the trees in the forest.
- An outlier score can computed for each observation: $s(x, n) = 2^{-\frac{E(h(x))}{c(n)}}$, where $h(x)$ is the path length of the sample x, and $c(n)$ is the maximum path length of a binary tree from root to external node and n is the number of external nodes.
- After giving each observation a score ranging from 0 to 1; 1 meaning more outlyingness and 0 meaning more normality. A threshold can be specified (ie. 0.55 or 0.60).

Popular Methods for Outlier Detection: Isolation Forest