

An aerial photograph of a university campus. In the foreground, there is a large, curved green lawn with several small trees and spherical stone ornaments. A paved road curves around the lawn. In the background, there is a large, multi-story white building with a central tower, surrounded by more greenery and hills under a cloudy sky. A large yellow rectangular box is overlaid on the right side of the image, containing the text "DATA - COLLECTION, SAMPLING & PREPROCESSING" in white, bold, sans-serif capital letters.

DATA - COLLECTION, SAMPLING & PREPROCESSING



1. INTRODUCTION

Data is everywhere

Introduction



List all data sources that are of potential interest before starting the analysis

Real life data is mostly dirty

Messy data will yield messy analytical models (GIGO)

Every data preprocessing step must be carefully justified, carried out, validated, and documented

An aerial photograph of a university campus. In the background, a large, multi-story white building with a central tower and a green roof stands against a backdrop of misty, green mountains. In the middle ground, there is a paved parking lot with several cars and a small green-roofed pavilion. The foreground features a large, well-maintained green lawn with a curved concrete path and several young trees planted in rows.

2. DATA SOURCES

2.1. Transactional Data



Structured, low-level, detailed information capturing the key characteristics of a customer transaction (e.g., purchase, claim)

Stored in massive online transaction processing (OLTP) relational databases.

It can be summarized over longer time horizons by aggregating it into averages, absolute/relative trends, maximum/minimum values, and so on.

2.2. Unstructured Data



Emails, web pages, claim forms, etc.

Needs extensive pre-processing. **Why? [Find Out]**

2.3. Qualitative Data



Expert based data

Steers the modeling in the right direction and allows to interpret the analytical results from the right perspective

How confidently can you interpret a cancer dataset?



Dun & Bradstreet

Bureau Van Dijck

Thomson Reuters

**Search what they do and
post in the group...**

2.4. Publicly Available Data



Social Media

Some Health Data

Economic Data

Sources such as data world, Kaggle, medium, etc.

Make sure that all data gathering respect both local and international privacy regulations



3. SAMPLING

Sampling



Subset of past customer data, used to build an analytical model

A good sample should be representative of the future customers

The optimal time window for the sample involves a trade-off between lots of data and recent data

At the same time, average business period must be considered



4. DATA ELEMENTS

Types



Continuous: Defined on limited or unlimited intervals (income, sales)

Categorical

Nominal: Limited set of values with no meaningful ordering in between (marital status)

Ordinal: Limited set of values with meaningful ordering in between (age coding)

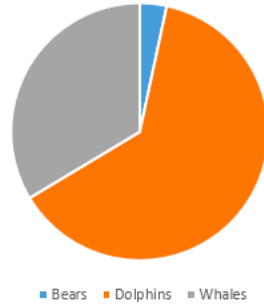
Binary: (Gender, employment status)



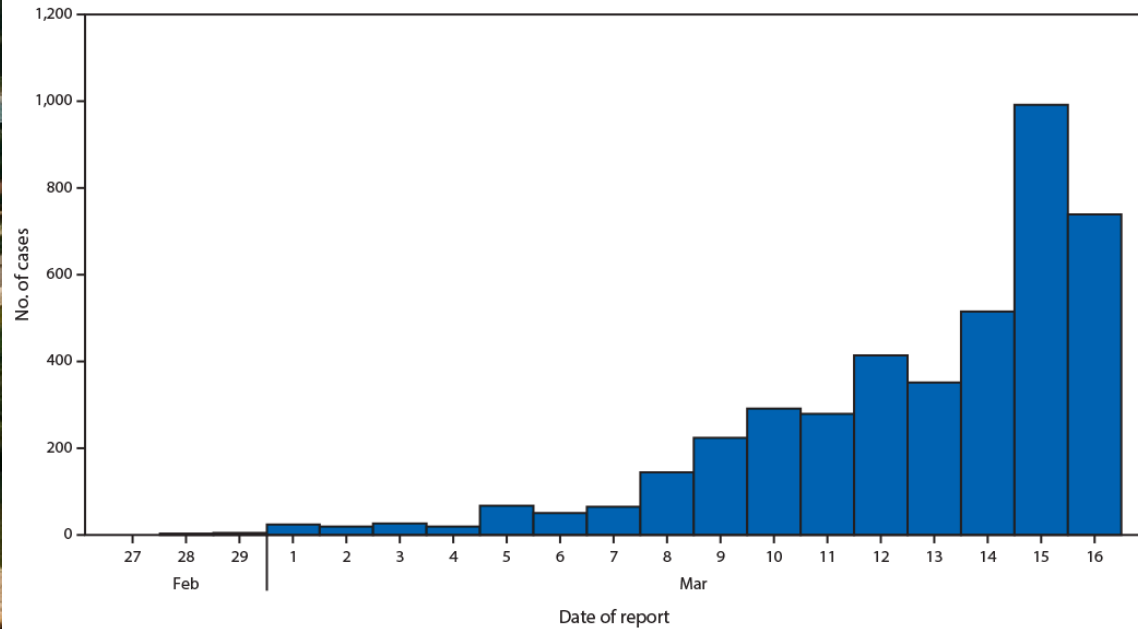
5. VDA & ESA

Visual Data Exploration & Exploratory Statistical Analysis

Wildlife Population in 2017



Visual Analysis





6. MISSING VALUES

Sources



Non applicable information

Undisclosed information

Errors during data collection or merging

Techniques



Impute: Replacing missing values with a known one.

Delete: Remove observations with lots of missing values. [**Missing at random**]

Keep: Some missing values can be meaningful



7. OUTLIERS

Detection & Treatment

Definition

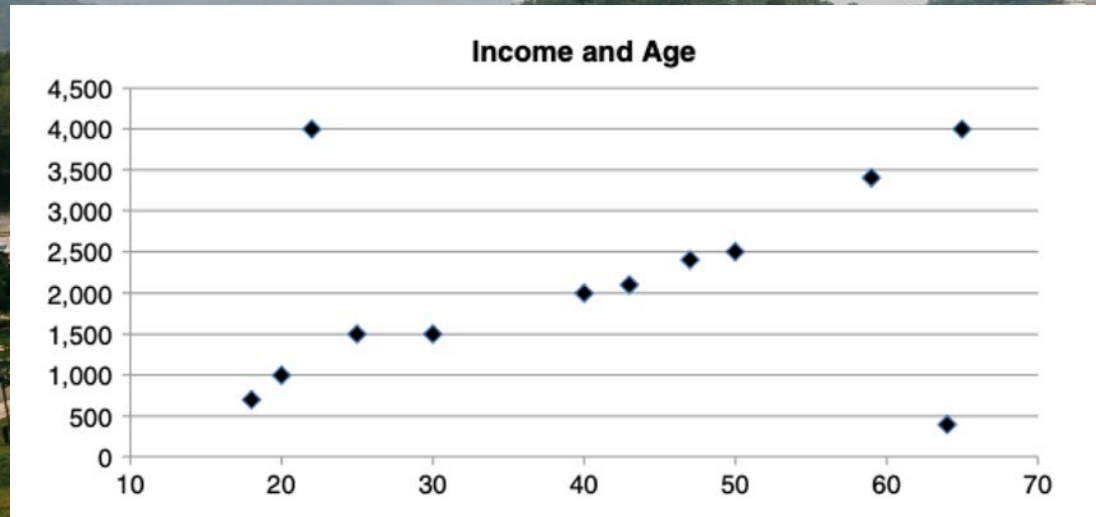


Extreme observations

Valid (CEO salary is \$1 Million)

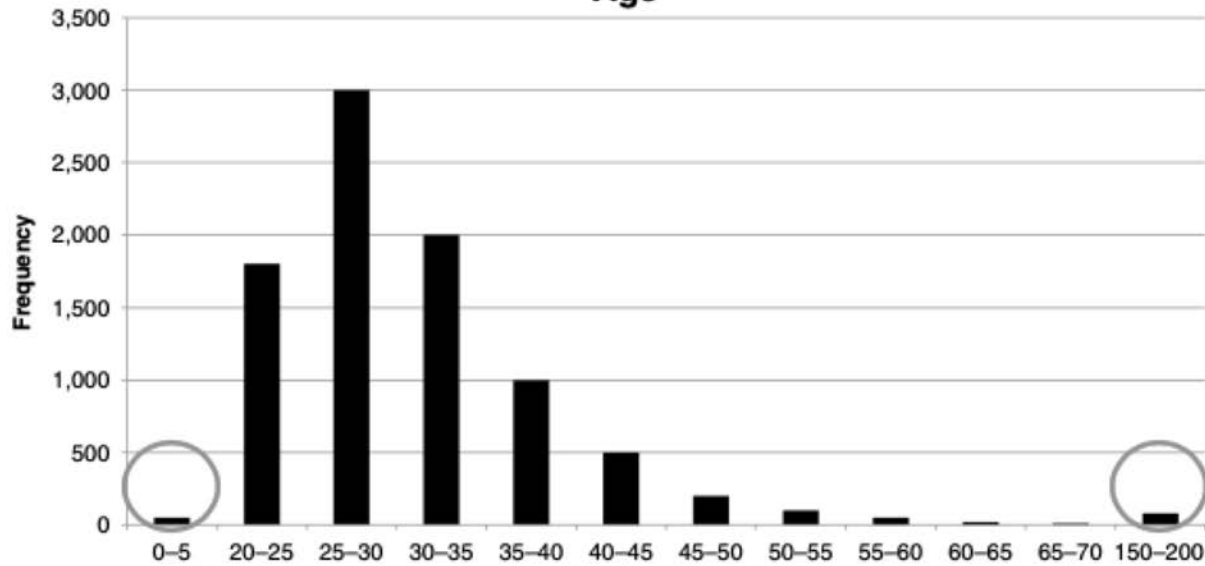
Invalid (Age is 300 years)

Univariate & Multivariate outliers



Multivariate Outliers

Age



**Detection -
Histograms**

Z scores



z-scores ($z = (x - \mu) / \sigma$), where **x** is the **test score**, μ is the **mean** & σ refers to **standard deviation**

Measure of how many standard deviations below or above the population mean a raw score is.

Outliers: Absolute value of **z score** is more than **3**

Outliers [5]



Previous methods focus on univariate outliers

Multivariate outliers can be detected by fitting regression lines and inspecting the observations with large errors (using a residual plot)

Other methods include clustering or calculating Mahalanobis distance, etc.

It is typically not considered in many modeling exercises due to the typical marginal impact on model performance

Outliers [6]



Invalid observations can be treated using impute, delete or keep schemes

Valid observations can be treated with techniques such as truncation/capping/winsorizing

This implies imposing a lower and upper limit on a variable
Anything out of the limits are brought back to these limits
Z-scores or IQR can be used

Upper/lower limit = $M \pm 3s$, with M = median and $s = IQR / (2 \times 0.6745)$.

Expert-based limits based on business knowledge and/or experience can be imposed.



8. Standardizing Data

Definition



Process of putting different variables on the same scale

Compare scores between different types of variables

Consider two columns in a dataset: age and salary

Age ranges from 30 – 70

Salary ranges from 55000 – 100000

The range of salary is much larger than range of age

Techniques



Min/max standardization

$$X_{new} = \frac{X_{old} - \min(X_{old})}{\max(X_{old}) - \min(X_{old})} (\text{newmax} - \text{newmin}) + \text{newmin}$$

,where **newmax** and **newmin** are the newly imposed maximum and minimum (e.g., 1 and 0)

Z-score standardization

Decimal scaling $X_{new} = \frac{X_{old}}{10^n}$,where **n** is the number of digits of the maximum absolute value.



9. Categorization

Classification



Coarse classification, classing, grouping, binning, etc.

Used to minimize the effects of small observation errors

Applicable to both categorical and continuous variables

Methods include interval binning, equal frequency binning, chi-squared analysis, etc.

Binning



Equal Frequency Binning : bins have equal frequency

Data : [5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215]

EFB : [5, 10, 11, 13][15, 35, 50, 55][72, 92, 204, 215]

Equal Width Binning : bins have equal width with a range of each bin are defined as $[\text{min} + w]$, $[\text{min} + 2w]$ $[\text{min} + nw]$ where **$w = (\text{max} - \text{min}) / (\text{no of bins})$**

Data : [5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215]

EWB : [10, 11, 13, 15, 35, 50, 55, 72][92][204]

Chi-squared analysis



Attribute	Owner	Rent Unfurnished	Rent Furnished	With Parents	Other	No Answer	Total
Goods	6,000	1,600	350	950	90	10	9,000
Bads	300	400	140	100	50	10	1,000
Good: bad odds	20:1	4:1	2.5:1	9.5:1	1.8:1	1:1	9:1

Suppose we want to categorize as follows:

- Option 1:** owner, renters, others
- Option 2:** owner, with parents, others

Chi-squared analysis [2]



Attribute	Owner	Renters	Others	Total
Goods	6,000	1,950	1,050	9,000
Bads	300	540	160	1,000
Total	6,300	2,490	1,210	10,000

Attribute	Owner	Renters	Others	Total
Goods	5,670	2,241	1,089	9,000
Bads	630	249	121	1,000
Total	6,300	2,490	1,210	10,000



10. Variable Selection

Variable Selection [1]



Analytical modeling starts with tons of variables

Typically only a few actually contribute to the prediction of the target variable

Filtering techniques such as Pearson correlation ρ or Fisher score, Cramer's V etc.

Pearson correlation



Measures linear dependency between two variables

Varies between -1 and 1

First step is to draw a scatter plot and check for existence of linearity

Don't calculate if the relationship is non-linear

$$\rho_P = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Cramer's V



Based on chi-squared analysis

$$\text{Cramer's } V = \sqrt{\frac{\chi^2}{n}}$$

where **n** is the number of observations in the data set

Bound between 0 & 1, higher values indicate better predictive power

Typically a cut-off value of **0.1** is adopted

An aerial photograph of a university campus. In the background, a large, multi-story white building with a central tower and a green roof stands against a backdrop of misty, green mountains. In the middle ground, there is a paved parking lot with several cars and a small pavilion with a green roof. The foreground features a large, well-maintained green lawn with a curved, light-colored stone or concrete path. Several young trees and large, rounded stone planters are scattered across the lawn.

11. Segmentation

Reasons



Strategic reasons [banks might want to adopt special strategies to specific segments of customers]

Operational reasons [new customers must have separate models because the characteristics in the standard model do not make sense operationally for them]

Significant variable interaction [if one variable strongly interacts with a number of others, it might be sensible to segment according to this variable]

Techniques



Business expert based

Statistical analysis

Decision Trees

K- means

Self-organizing maps

Caution: Segmentation might lead to increased number of analytical models



BARAKALLAH FEEKUM!

Any questions?

Feel free to contact me using designated channels