Managing Data

March 2019

Dr. Susan McKeever

CeADAR

Dublin Institute of Technology

School of Computing

CeADAR Centre for Applied Data Analytics

OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
TU DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

Managing Data -

- Types of Data
- Exploring your data
- Data quality
- Storing data
- Risk free management of Data

Data - “Life blood” of your project

Data set Sources

- Private datasets from company?
- Public datasets (Kaggle or equivalent etc), to allow benchmarking and comparison?
- Other researchers’ datasets on request?
- Synthesise data (more on this..)
- Your own data collection efforts?

To do machine learning

- Need to process your data into instances (examples)
- Often need to reduce it to an **Analytics Base Table -**
  - And with Labels for supervised learning

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Dataset - Types

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotitaw	28

Structured data

- Often tackled first

- Tabular data (usually)

- Features are transparent/ predefined
  - although feature engineering might still be needed!

Dataset - Types

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Structured data

- Often tackled first

*Un*Structured data

(e.g. images, text, audio…)

- Often a source of rich insights – and explored after the easier structure data

*Un*Structured data

(images, text, audio…) - Harder to turn into tabular data

- Feature engineering required – especially if used with traditional machine learning

  e.g. Texts (term document matrix), image (matrix of pixels)

- More recently, used with deep learning

- Recent improvements in how this data is represented
  - E.g. Text docs: Term document matrix → Word vectors, sentence vector (word2Vec, glove etc)
  - State of the Art is developing rapidly!

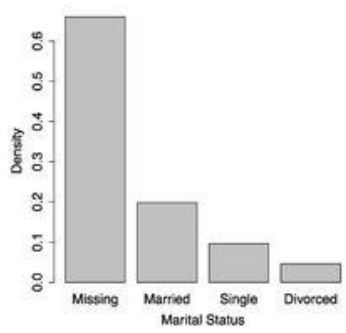
Data Exploration of features

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

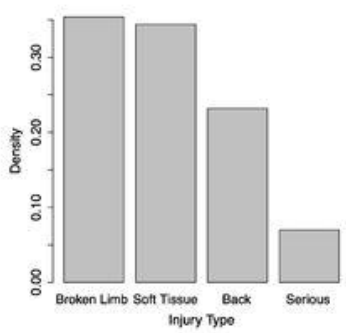
For categorical **features**, we should:

- Examine the mode, 2nd mode, mode %, and 2nd mode % as these tell us the most common levels within these features and will identify if any levels dominate the dataset.
  - E.g. City:
- For continuous (numerical) features we should:
  - Examine the mean and standard deviation of each feature to get a sense of the central tendency and variation of the values within the dataset for the feature.
  - Examine the minimum and maximum values to understand
  - the range that is possible for each feature.
    - E.g. AGE

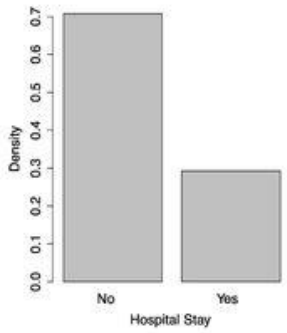
Explore the value distributions



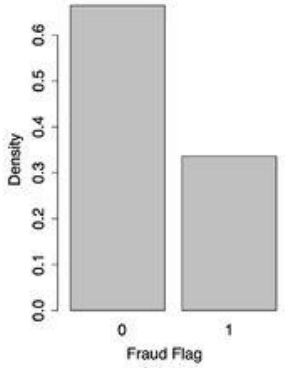
(a) MARITAL STATUS



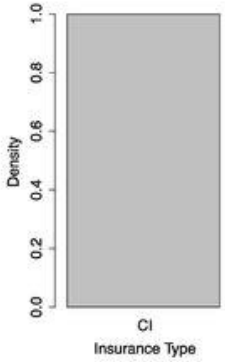
(b) INJURY TYPE



(c) HOSPITAL STAY

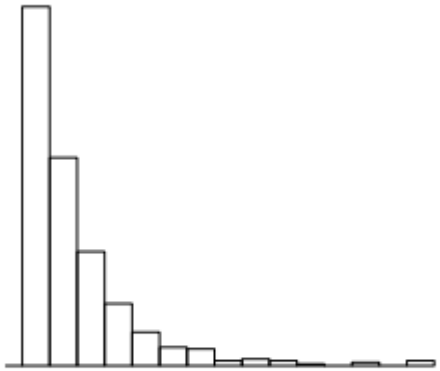


(a) FRAUD FLAG

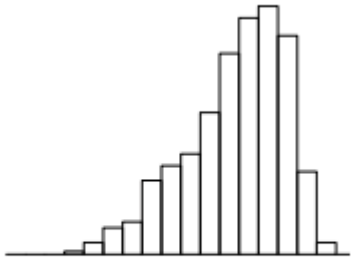


(b) INSURANCE TYPE

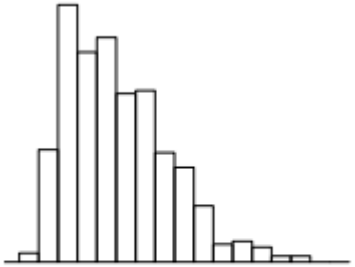
Data distributions

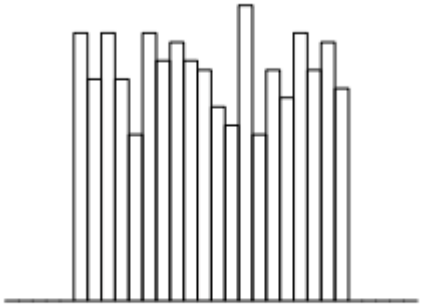


Exponential

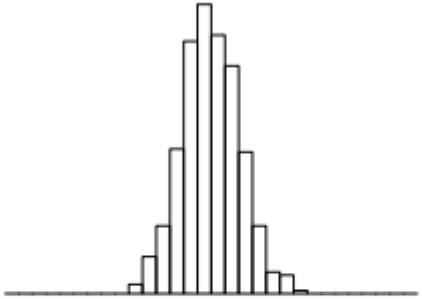


Unimodal (skewed left)

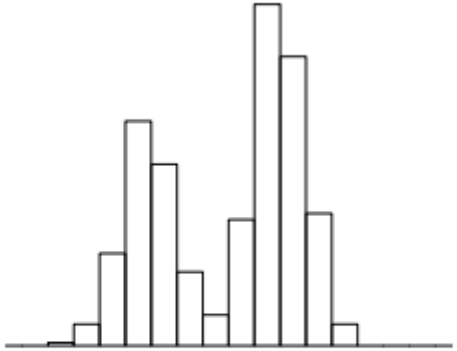




Uniform



Normal (Unimodal)



Multimodal

Data quality issues –

First Name	Last Name	Address	City	Age
Mickey	Mouse	123 Fantasy Way	Anaheim	73
Bat	Man	321 Cavern Ave	Gotham	54
Wonder	Woman	987 Truth Way	Paradise	39
Donald	Duck	555 Quack Street	Mallard	65
Bugs	Bunny	567 Carrot Street	Rascal	58
Wiley	Coyote	999 Acme Way	Canyon	61
Cat	Woman	234 Purrfect Street	Hairball	32
Tweety	Bird	543	Itotltaw	28

Most common ones

- **Outliers** (valid or invalid?)
  - Very high or low : e.g. age = 150 , Salary = 400K
  - Treatment: Clamp transformation + domain knowledge..
- Missing values
  - Heuristic: Remove anything with 50% or more missing
- Irregular Cardinality
  - Cardinality of 1 – everything instance has the same value – no predictive value!
  - Cardinality high - every instance has a different values
  - Investigate

Plenty of tools to help

Python and Data Pandas

- # Select the rows that have at least one missing value
  peopledata[peopledata.isnull().any(axis=1)].head()

```
peopledata[peopledata.isnull().any(axis=1)].head()
```

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	rac
0	39	State-gov	77516	Bachelors	13.0	Never-married	Adm-clerical	Not-in-family	Whi
1	50	Self-emp-not-inc	83311	Bachelors	13.0	Married-civ-spouse	Exec-managerial	Husband	Whi
2	38	Private	215646	HS-grad	NaN	Divorced	Handlers-cleaners	Not-in-family	Whi
3	53	Private	234721	11th	NaN	Married-civ-spouse	Handlers-cleaners	Husband	Bla
4	28	Private	338409	Bachelors	13.0	Married-civ-spouse	Prof-specialty	Wife	Bla

R packages, Tabula, Data Wrangler…

Labelling your data

If you are going to train prediction models:

- Does your client have *labelled* data ?

e.g. insurance claim data for fraud prediction – do you have labelled of which ones are fraudulent so you can "train" a model?

- Expensive, time consuming
- Various options
- Also unsupervised learning – no labels – clustering

Labelling

Most common

PROS AND CONS OF LABELING APPROACHES

Approach	Description	Pros	Cons
Internal labeling	Assignment of tasks to an in-house data science team	✓ Predictable results ✓ High accuracy of labeled data ✓ The ability to track progress	× It takes much time
Outsourcing	Recruitment of temporary employees on freelance platforms, posting vacancies on social media and job search sites	✓ The ability to evaluate applicants' skills	× The need to organize workflow
Crowdsourcing	Cooperation with freelancers from crowdsourcing platforms	✓ Cost savings ✓ Fast results	× Quality of work can suffer
Specialized outsourcing companies	Hiring an external team for a specific project	✓ Assured quality	× Higher price compared to crowdsourcing
Synthetic labeling	Generating data with the same attributes of real data	✓ Fewer constraints for using sensitive and regulated data ✓ Training data without mismatches and gaps ✓ Cost- and time-effectiveness	× High computational power required
Data programming	Using scripts that programmatically label data to avoid manual work	✓ Automation ✓ Fast results	× Lower quality dataset

altexsoft
software r&d engineering

What storage do you need for your Data ?

- Flat files – e.g. CSV, JSON, spreadsheet
  - Simple to use/ No querying / low volume
- Database - especially if merging data
  - Relational dB -?
  - NoSQL – e.g. couchdB, Mongo – for fast applications X
- Big data – extremely large datasets Hadoop / Map Reduce etc

Database storage

- When to use Relational SQL
  - Structured unchanging database
  - Do you need to query /
    - Structured Query Language (SQL)
    - E.g. “Which customers, living in Dublin have an income great than 10,000

```
SELECT name, ID from customers where
city = Dublin and income >1000
```

  - Candidates…
    - MySQL (free), Oracle… many more
    - Access Microsoft – lightweight!

GDPR / Data

- Important to know about GDPR
  - EU General Data Protection regulation
  - No individuals identifiable in your data
  - Anonymise if possible

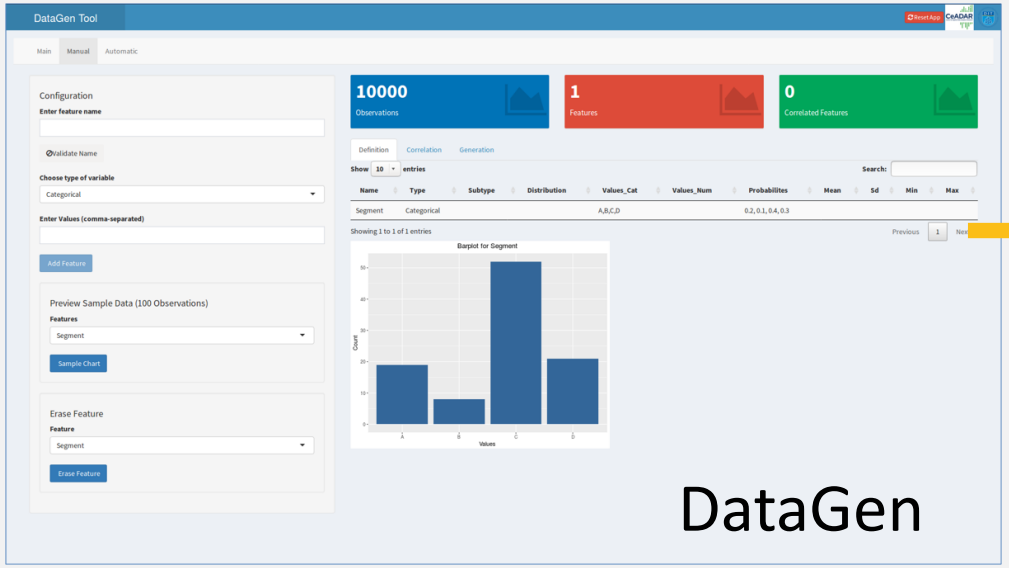
- Data management procedure – is this already sorted in your project **ethical approval?**
- Sample procedures / Data Transfer Agreement
- Make sure to agree with your company!

Earlier – mentioned Data Synthesis

- CeADAR has built a demonstrator to generate data

Company Dataset

Need a similar copy



DataGen

Synthetic Copy Dataset

Same features; Relationships etc

Data - Summary

- Source it
- Explore it
- Clean it
- Keep it safe, in way that meets your need
- Be careful of GDPR!

Questions