Data Mining and Machine Learning

Comp 3027J

Catherine Mooney

catherine.mooney@ucd.ie

Lectures and Text

- **Core Text:**
  Fundamentals of Machine Learning for Predictive Data Analytics
  By John D. Kelleher, Brian Mac Namee and Aoife D'Arcy
- Last week we covered Chapter 1, sections 1.1 – 1.4 in class
- This week we will cover Chapter 2, sections 2.3 and 2.4 and Chapter 3, sections 3.1 – 3.4
- Please read these sections of the book

Questions

- How many of you have taken the book out of the library?
- How many of you have enrolled for the course on Moodle?

Review of last week

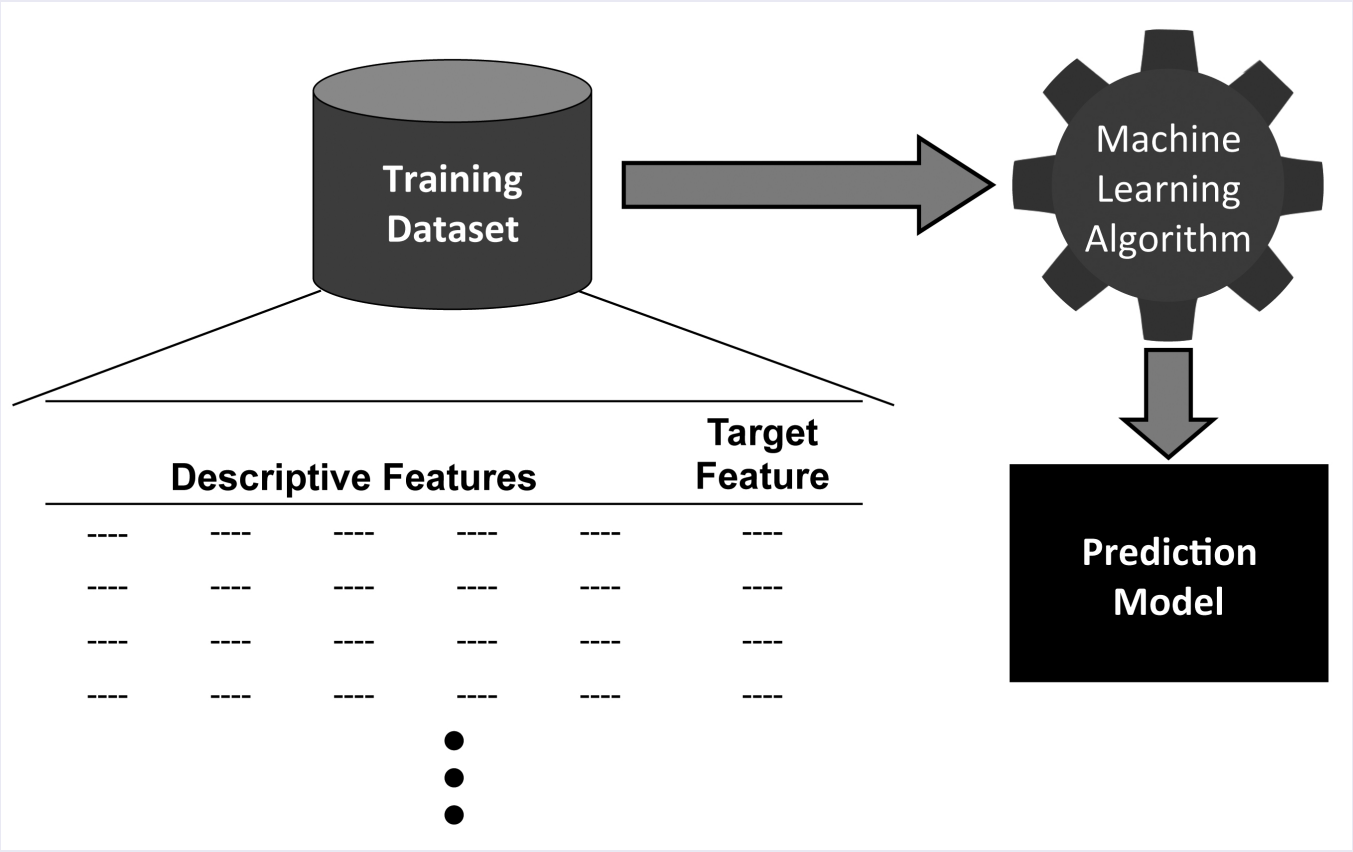
What is Data Mining/Machine Learning?

- The subfield of computer science that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959).
- **Unsupervised Learning:** An algorithm that finds patterns in data when **no manually labelled examples** are available as inputs. More focused on data exploration and knowledge discovery. e.g. Clustering, Graph partitioning algorithms
- **Supervised Learning:** An algorithm that learns a function from examples of its inputs and outputs. It requires **manually-labelled example data** to learn the correct answer for a given query input. e.g. Classification, Regression algorithms

What is the difference between Data Mining and Machine Learning?

- "The short answer is: None. They are ... concerned with the same question: how do we learn from data?"
  Larry Wasserman, Professor in Statistics and Machine Learning, Carnegie Mellon
- They cover almost exactly the same material and use almost exactly the same techniques
- They *emphasize* different things
  – The purpose of data mining is finding valuable **insights** in large databases
  – Machine learning is more focused on making accurate **predictions**

Supervised machine learning techniques automatically learn the relationship between a set of **descriptive features** and a **target feature** from a set of historical **instances** (referred to as a **training dataset**) to build a **prediction model**.



We can then use this **prediction model** to make predictions for new instances

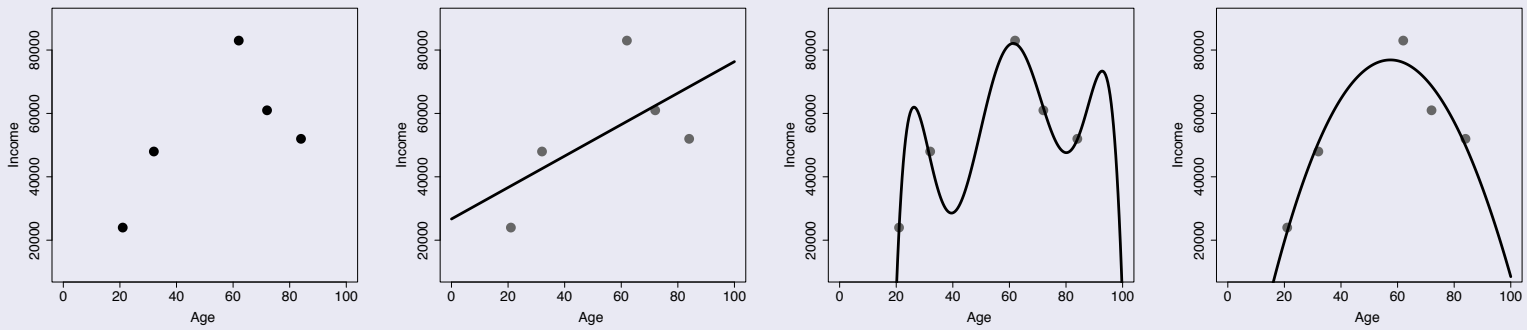
- A prediction model that makes the correct predictions for these queries is said to **generalise** well.
- The goal of machine learning is to find the predictive model that **generalises** best.
- To find the best prediction model, a machine learning algorithm must use some criteria for choosing among the candidate prediction models it considers during its search.

What criteria should we use?

- Lots of different machine learning algorithms.
- Each machine learning algorithm uses different model selection criteria to drive its search for the best **predictive model**.
- The set of assumptions that defines the model selection criteria of a machine learning algorithm is known as the **inductive bias** of the machine learning algorithm.
- It has been shown that there is no particular **inductive bias** that on average is the best one to use.
- **The ability to select the appropriate machine learning algorithm (and hence inductive bias) to use for a given predictive task is one of the core skills that a data analyst must develop!!**

What happens if we choose the wrong inductive bias:

- **Underfitting**
- **Overfitting**
- Striking the right balance between **model** simplicity and complexity (between underfitting and overfitting) is the hardest part of machine learning.



(a) Dataset (b) Underfitting (c) Overfitting (d) Just right

Any questions on what we covered last week?

Designing the Analytics Base Table

"The basic data requirements for predictive models are surprisingly simple. To build a predictive model, we need a large dataset of historical examples of the scenario for which we will make predictions. Each of these historical examples must contain sufficient data to describe the scenario and the outcome that we are interested in predicting." *Fundamentals of Machine Learning for Predictive Data Analytics,* pg 27

For example, if we are trying to predict whether or not insurance claims are fraudulent, we require:

- A large dataset of historical insurance claims
- For each claim we must know whether or not that claim was found to be fraudulent.

The basic structure in which we capture these historical datasets is the **analytics base table** (**ABT**).

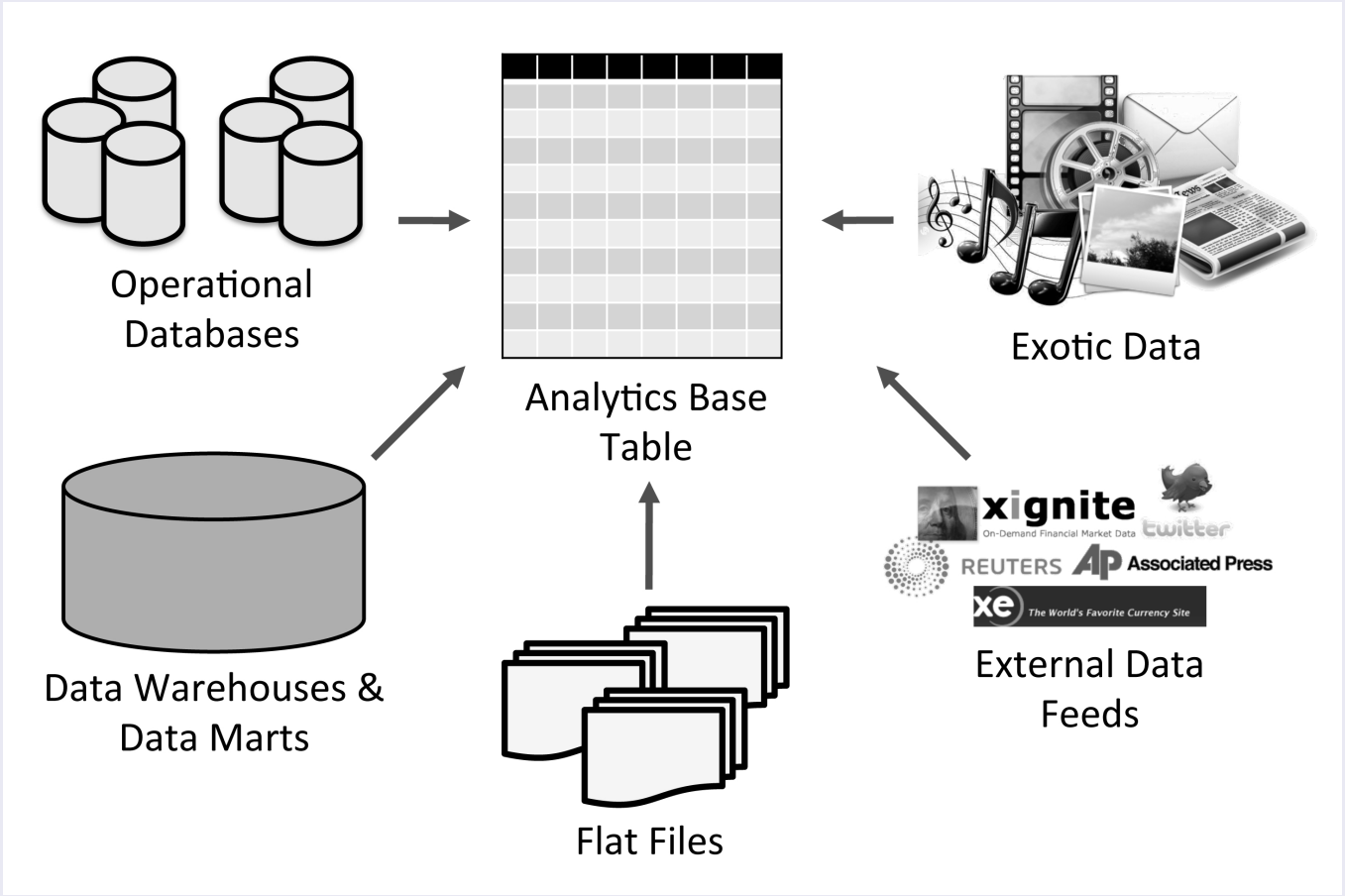
The Analytics Base Table

- A simple, flat, tabular data structure made up of rows and columns.
- The columns are divided into a set of descriptive features and a single target feature.
- Each row contains a value for each descriptive feature and the target feature.
- Each row represents an instance about which a prediction can be made.

Descriptive Features						Target Feature
----	----	----	----	----	----	----
----	----	----	----	----	----	----
----	----	----	----	----	----	----
----	----	----	----	----	----	----
----	----	----	----	----	----	----

Where does the data come from?

There are many different potential data sources and these typically combined to create an analytics base table.



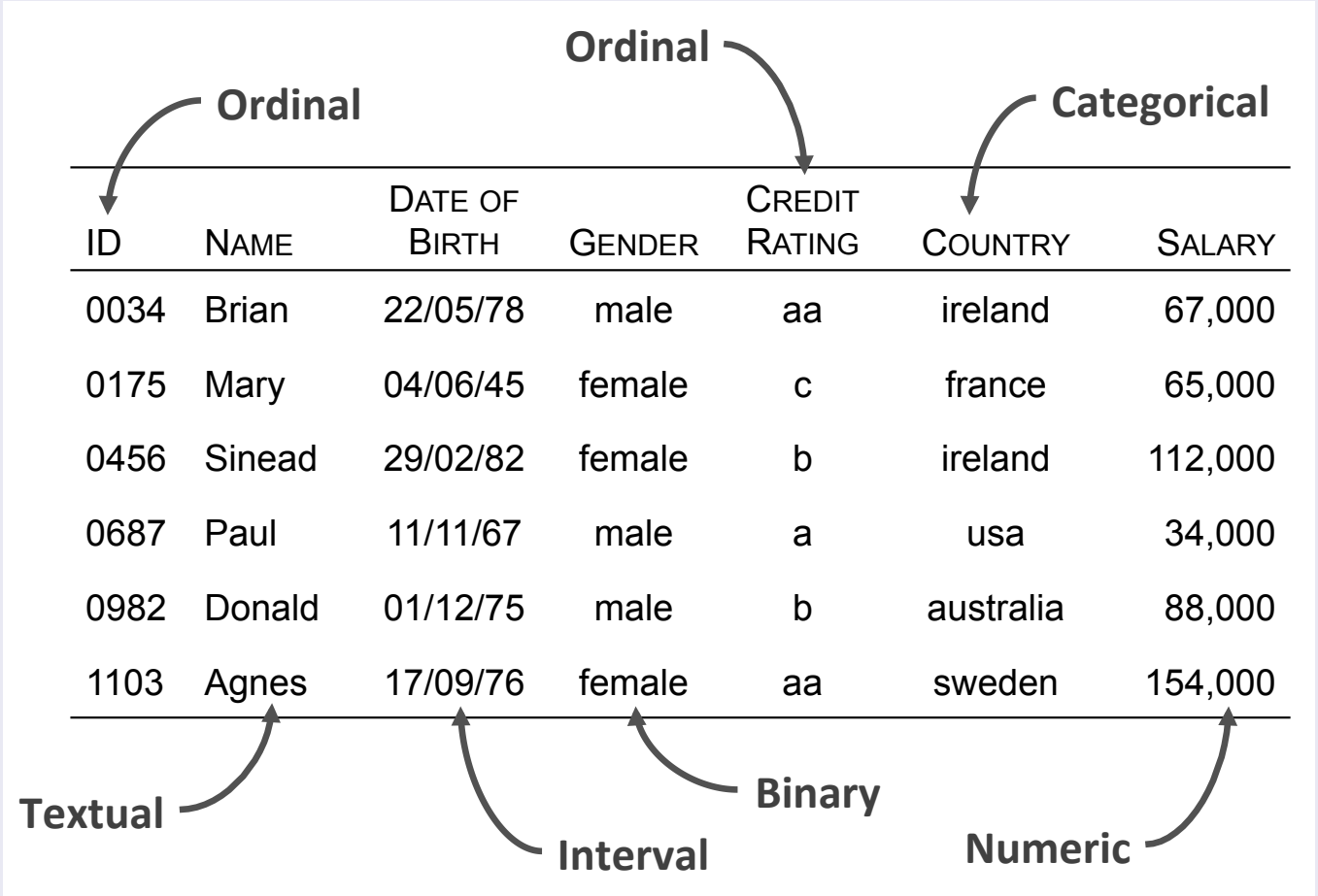
- The ultimate goal is to build a predictive model that predicts a target feature from a set of descriptive features
- We must try to identify the features of the prediction subject that are likely to be useful in making this prediction
- Defining features can be difficult!

Designing & Implementing Features

Three key data considerations are particularly important when we are designing features.

- **Data availability** – we must have data available to implement any feature we would like to use.
- **Timing** – data that will be used to define a feature must be available before the event around which we are trying to make predictions occurs (e.g. attendance/soccer match outcome).
- **Longevity** – there is potential for features to go stale if something about the environment from which they are generated changes (e.g. salaries).
  - One way to extend the longevity of a feature is to use a derived ratio instead of a raw feature (e.g. ratio between salary and requested loan amount).

Different Types of Data



Different Types of Data

- **Ordinal**: Values that allow ordering but do not permit arithmetic (e.g. size measured as small, medium, or large)
- **Textual**: Free-form, usually short, text data (e.g. name, address)
- **Interval**: Values that allow ordering and subtraction, but do not allow other arithmetic operations (e.g. date, time)
- **Binary**: A set of just two values (e.g. yes/no)
- **Categorical**: A finite set of values that cannot be ordered and allow no arithmetic (e.g. country, product type)
- **Numeric**: True numeric values that allow arithmetic operations (e.g. price, age)

We often reduce this categorization to just two data types:

- **Continuous** (encompassing the numeric and interval types)
- **Categorical** (encompassing the categorical, ordinal, binary, and textual types)
  - We refer to the set of possible values that a categorical feature can take as the levels of the feature
  - For example, the levels of the CREDIT RATING feature are aa, a, b, c and the levels of the GENDER feature are male, female.

The presence of different types of descriptive and target features can have a big impact on how an algorithm works.

Different Types of Features

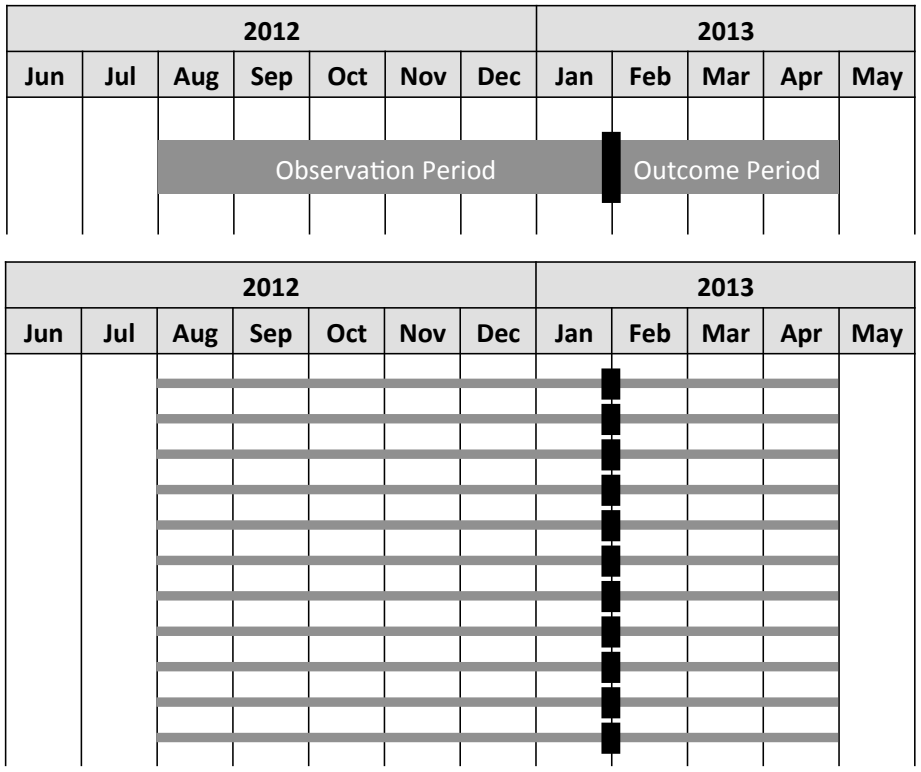
- The features in an ABT can be of two types:
  - **Raw features** – features that come directly from raw data sources (e.g. customer age, gender, loan amount)
  - **Derived features** – do not exist in any raw data source, so they must be constructed from data in one or more raw data sources (e.g. average customer purchases per month, loan-salary ratios)

Derived Feature Types

- There are a number of common derived feature types:
  - **Aggregates** – measures defined over a group or period and are usually defined as the count, sum, average, minimum, or maximum of the values within a group.
  - **Flags** – binary features that indicate presence or absence of some characteristic within a dataset (e.g. whether or not a bank account has ever been overdrawn).
  - **Ratios** – capture the relationship between two or more raw data values (e.g. loan-salary ratios).
  - **Mappings** – convert continuous features into categorical features and are often used to reduce the number of unique values that a model will have to deal with (e.g. salary –> low, medium, and high).
  - Other...

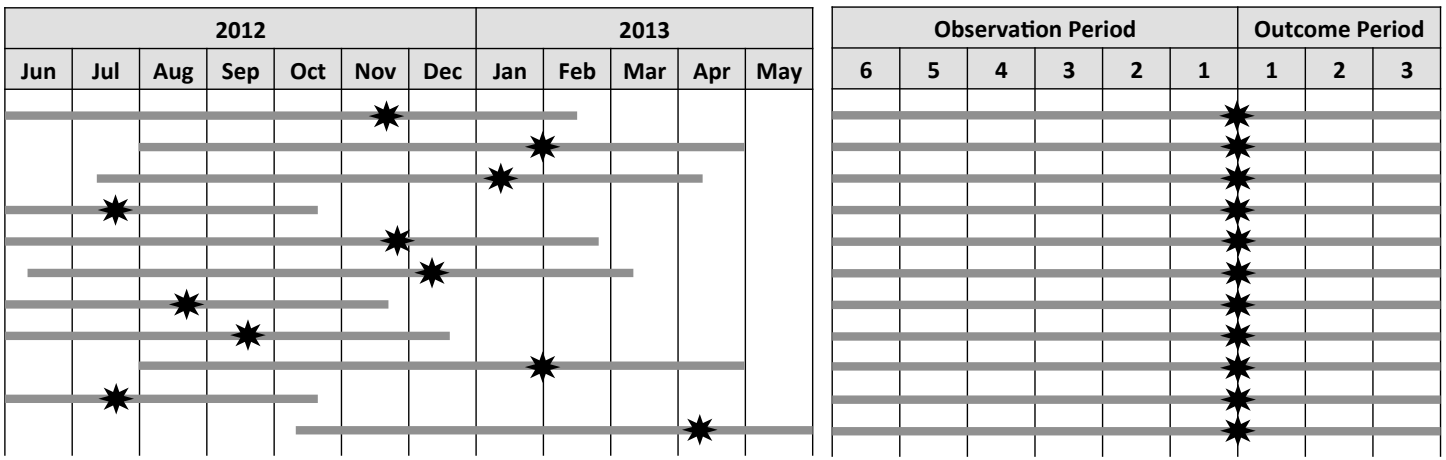
- Many of the predictive models that we build have a temporal element
- Two key periods:
  - the **observation period**
  - the **outcome period**

- In some cases the observation and outcome period are measured over the same time for all predictive subjects.

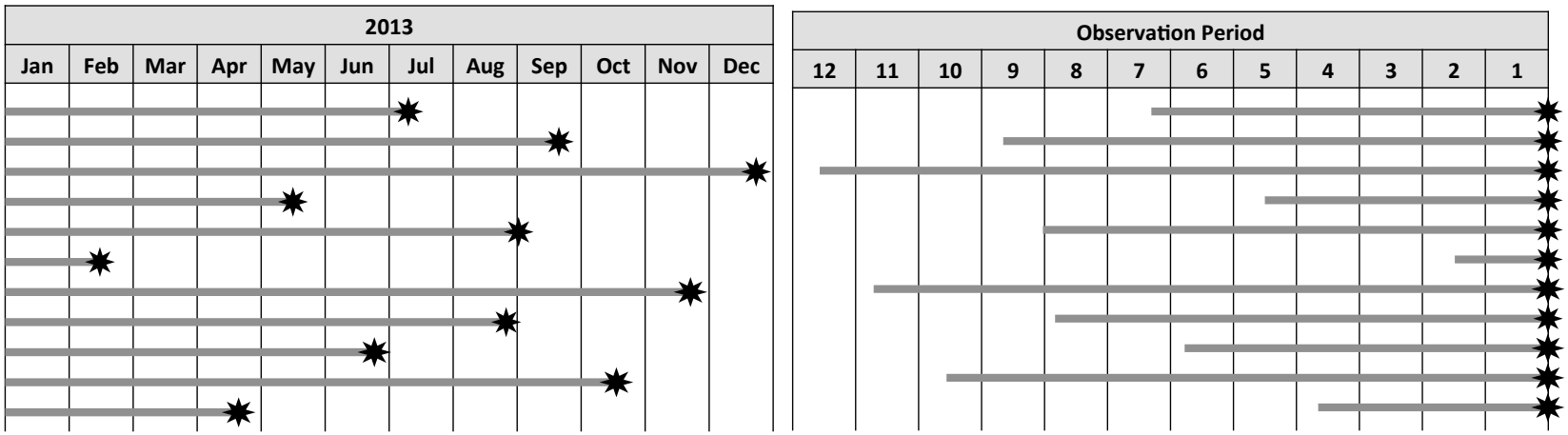


(each line represents a prediction subject)

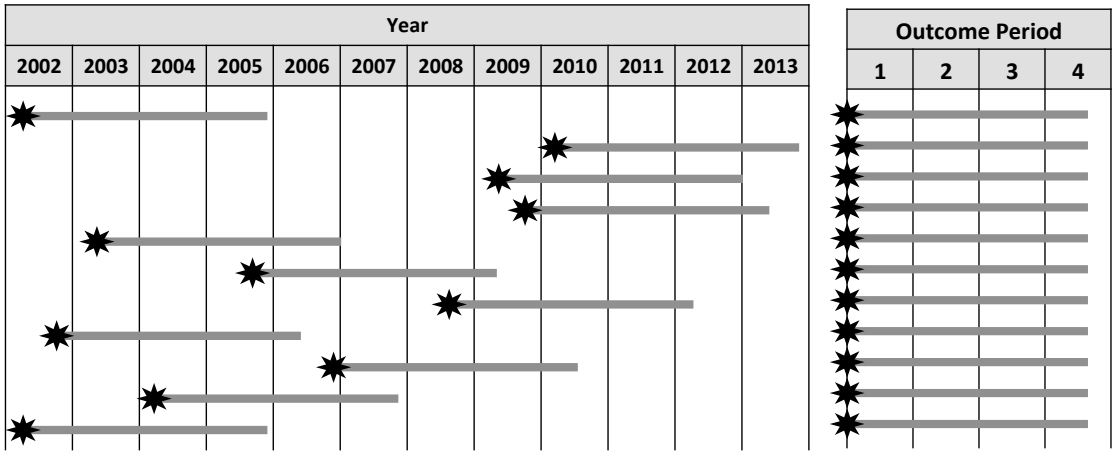
- Often the observation period and outcome period will be measured over different dates for each prediction subject.
- Observation and outcome periods defined by an event rather than by a fixed point in time (each line represents a prediction subject and stars signify events)



- In some cases only the descriptive features have a time component to them, and the target feature is time independent.
- Modeling points in time for a scenario with no real outcome period (each line represents a prediction subject, and stars signify events).



- Conversely, the target feature may have a time component and the descriptive features may not.
- Modeling points in time for a scenario with no real observation period (each line represents a prediction subject, and stars signify events).



Legal Issues

- Data analytics practitioners can often be frustrated by legislation that stops them from including features that appear to be particularly well suited to an analytics solution in an ABT.
- There are significant differences in legislation in different jurisdictions, but a couple of key relevant principles almost always apply.
  1. **Anti-discrimination legislation** in most jurisdictions prohibits discrimination on the basis of some set of the following grounds: sex, age, race, ethnicity, nationality, sexual orientation, religion, disability, and political opinions.
  2. **Data protection legislation** – rules surrounding the use of personal data.

Data protection legislation

- Although, data protection legislation changes significantly across different jurisdictions, there are some common tenets on which there is broad agreement which affect the design of ABTs
  - The **Collection limitation principle** – personal data should only be obtained by lawful means with the knowledge and **consent** of a data subject.
  - The **Purpose specification principle** – data subjects should be informed of the purpose for which data will be used at the time of its collection.
  - The **Use limitation principle** – collected data should not subsequently be used for purposes other than those stated at the time of collection.

Summary

- Predictive models built using machine learning techniques are tools that we can use to help make better decisions, not an end in themselves.
- It is important to fully understand the problem that a model is being constructed to address – this is the goal behind *converting problems into analytics solutions*

- Predictive data analytics models are reliant on the data that is used to build them – the **analytics base table** (**ABT**).
- The first step in designing an ABT is to decide on the **prediction subject**.
- An effective way in which to design ABTs is in collaboration with a domain expert to design **features**.

- Features (both descriptive and target) are concrete numeric or symbolic representations.
- It is useful to distinguish between **raw features** that come directly from existing data sources and **derived features** that are constructed by manipulating values from existing data sources.
- Common manipulations used in this process include aggregates, flags, ratios, and mappings, although any manipulation is valid.

Any questions so far?
5 minute break...

Data Exploration – The data quality report

Data Exploration

Before attempting to build predictive models based on an ABT it is important that we undertake some exploratory analysis, or data exploration, of the data contained in the ABT.

Data Exploration

There are two goals in data exploration:

1. To fully understand the characteristics of the data in the ABT.
   - It is important that for each feature in the ABT, we understand characteristics such as the types of values a feature can take, the ranges into which the values in a feature fall, and how the values in a dataset for a feature are distributed across the range that they can take. We refer to this as **getting to know the data**.
2. To determine whether or not the data in an ABT suffer from any **data quality issues** that could adversely affect the models that we build.
   - Examples of typical data quality issues include an instance that is *missing values* for one or more descriptive features, an instance that has an *extremely high value* for a feature, or an instance that has an *inappropriate level* for a feature.

Data Exploration

The most important tool used during data exploration is **the data quality report**.

The data quality report

- A data quality report includes tabular reports that describe the characteristics of each feature in an ABT using standard statistical measures of **central tendency** (mean, mode, and median) and **variation** (standard deviation and percentiles).
- The tabular reports are accompanied by data visualizations:
  - A **histogram** for each continuous feature in an ABT.
  - A **bar plot** for each categorical feature in an ABT.

The data quality report – continuous features

Should include:

- the total number of instances in the ABT
- the percentage of instances in the ABT that are missing a value for each feature
- the cardinality of each feature (cardinality measures the number of distinct values present in the ABT for a feature)
- the minimum, 1st quartile, mean, median, 3rd quartile, maximum, and standard deviation statistics

Feature	Count	% Miss.	Card.	Min.	1^{st} Qrt.	Mean	Median	3^{rd} Qrt.	Max.	Std. Dev.
___	___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___	___

The data quality report – categorical features

Should include:

- the total number of instances in the ABT
- the percentage of instances in the ABT that are missing a value for each feature
- the cardinality of each feature
- for the two most frequent levels for the feature (the mode and 2nd mode) – the frequency with which these appear (both as raw frequencies and as a proportion of the total number of instances in the dataset)

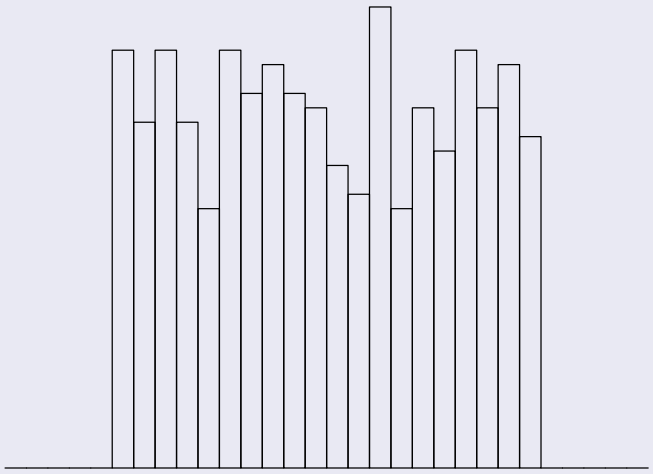
Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2^{nd} Mode	2^{nd} Mode Freq.	2^{nd} Mode %
___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___
___	___	___	___	___	___	___	___	___	___

Getting To Know The Data

- For categorical features, we should:
  - Examine the mode, 2^{nd} mode, mode %, and 2^{nd} mode % as these tell us the most common levels within these features and will identify if any levels dominate the dataset (these levels will have a very high mode %).
- For continuous 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.

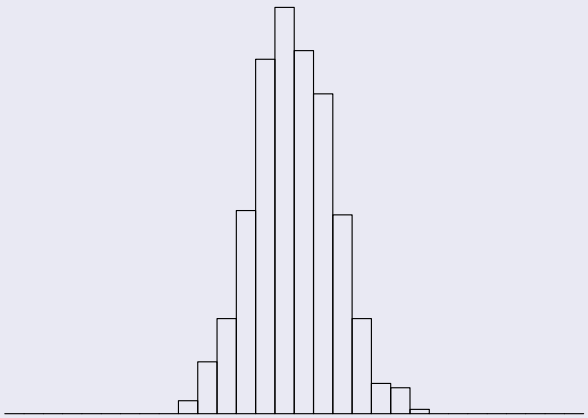
- The histograms for each continuous feature included in a data quality report are a very easy way for us to understand how the values for a feature are distributed across the range they can take.
- When we generate histograms of features there are a number of common, well understood shapes that we should look out for.

Uniform



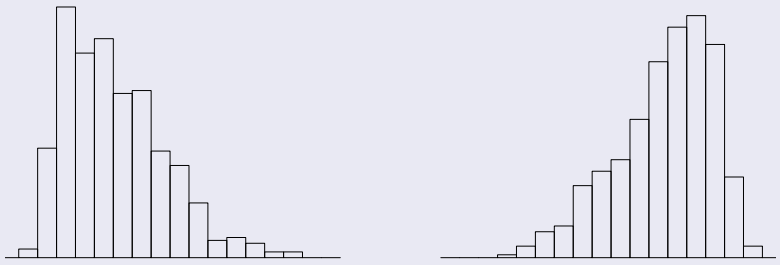
- A uniform distribution indicates that a feature is equally likely to take a value in any of the ranges present.

Normal (Unimodal)



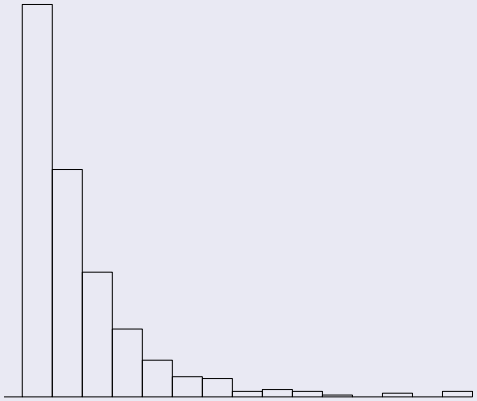
- Characterized by a strong tendency towards a central value and symmetrical variation to either side of this
- Naturally occurring phenomena tend to follow a normal distribution e.g. height or weight

Skew



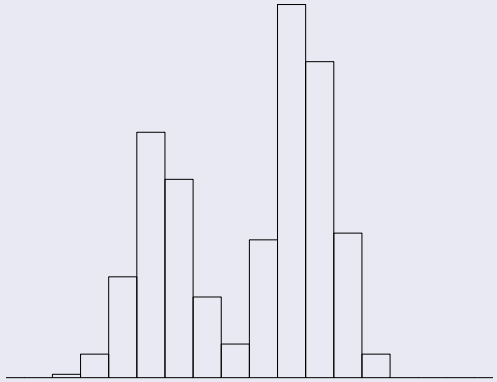
- Skew is simply a tendency towards very high (**right skew**) or very low (**left skew**) values.
- Features recording salaries often follow a right skewed distribution as most people are paid salaries near a well-defined central tendency, but there are usually a small number of people who are paid very large salaries.
- Skewed distributions are often said to have long tails.

Exponential



- In a feature following an **exponential distribution** the likelihood of occurrence of a small number of low values is very high, but sharply diminishes as values increase.
- The number of times a person has made an insurance claim or the number of times a person has been married tend to follow an exponential distribution.
- Warning sign that outliers are likely.

Multimodal



- A feature characterized by a **multimodal distribution** has two or more very commonly occurring ranges of values that are clearly separated.
- Caution – measures of central tendency and variation tend to break down for multimodal data.
- Optimism – the separate peaks may be associated with the different target levels we are trying to predict.

Data exploration

This stage of data exploration is mostly an information-gathering exercise, the output of which is just a better understanding of the contents of an ABT. It does, however, also present a good opportunity to discuss anything unusual that we notice about the central tendency and variation of features within the ABT.

Identifying and Handling Data Quality Issues

- A **data quality issue** is loosely defined as anything *unusual* about the data in an ABT.
- The most common data quality issues are:
  - **missing values**
  - **irregular cardinality**
  - **outliers**

- The data quality issues we identify from a data quality report will be of two types:
  - Data quality issues due to **invalid data** – take immediate action to correct them, regenerate the ABT, and recreate the data quality report.
  - Data quality issues due to **valid data** – record any data quality issues due to valid data in a data quality plan so that we remain aware of them and can handle them later if required.

The structure of a data quality plan

Feature	Data Quality Issue	Potential Handling Strategies
____	____	____
____	____	____
____	____	____
____	____	____

For each of the data quality issues found:

- include the feature it was found in
- the details of the data quality issue
- information on potential handling strategies

Handling Data Quality Issues

- Handling Missing Values
- Handling Irregular Cardinality
- Handling Outliers

Handling Missing Values

- The **% Miss.** columns in the data quality report highlight the percentage of missing values for each feature
- Why are the values missing?
- Approach 1: Drop any features that have missing value (if % Miss. $>$60%).
- Approach 2: **Imputation**

Handling Missing Values

- **Imputation** replaces missing feature values with a plausible estimated value based on the feature values that are present.
- The most common approach to imputation is to replace missing values for a feature with a measure of the central tendency of that feature.
- It is not recommended to use imputation on features missing in excess of 30% of their values

Handling Irregular Cardinality

- The **Card.** column in the data quality report shows the number of distinct values for a feature within an ABT.
- A data quality issue arises when the cardinality for a feature does not match what we expect, a mismatch called an irregular cardinality.
- Features with a cardinality of 1 – Vaild?

Handling Irregular Cardinality

- Categorical features incorrectly labeled as continuous (if the cardinality of a continuous feature is significantly less than the number of instances in the dataset)
- If a categorical feature has a much higher cardinality than we would expect given the definition of the feature (gender with a cardinality of 6 – male, female, m, f, M, and F)
- If a categorical feature has a very high number of levels – anything over 50 (machine learning algorithms will struggle)

Handling Outliers

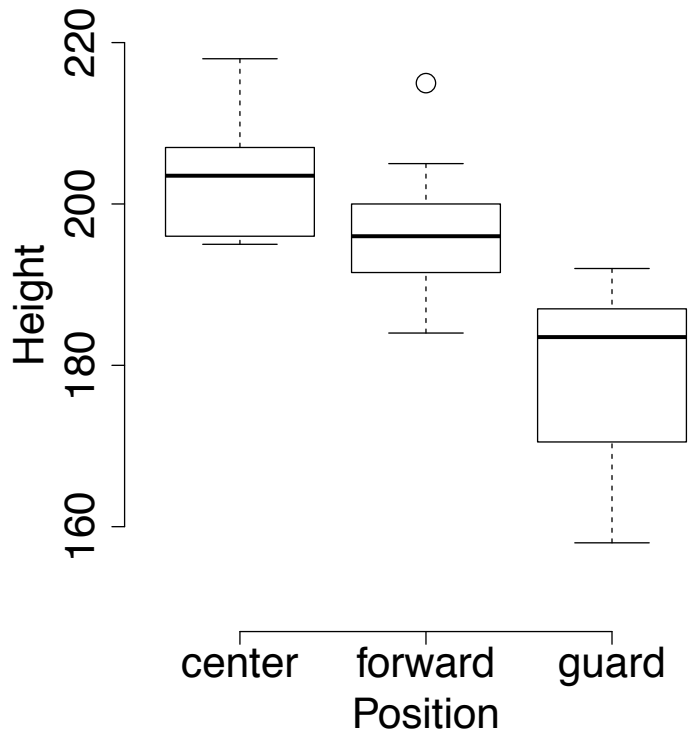
- Outliers are values that lie far away from the central tendency of a feature.
- There are two kinds of outliers that might occur in an ABT: invalid outliers and valid outliers.
- Invalid outliers are values that have been included in a sample through error and are often referred to as noise in the data.
- Valid outliers are correct values that are simply very different from the rest of the values for a feature, for example, a billionaire who has a massive salary compared to everyone else in a sample.

Identifying Outliers

- Examine the minimum and maximum values for each feature and use domain knowledge to determine whether these are plausible values.
- Invalid outliers and should immediately be either corrected, if data sources allow this, or removed and marked as missing values if correction is not possible.
- In some cases we might even remove a complete instance from a dataset based on the presence of an outlier.

Identifying Outliers

- Compare the gaps between the median, minimum, maximum, 1st quartile, and 3rd quartile values.
- Box plots can help to make this comparison.
- Exponential or skewed distributions in histograms are also good indicators of the presence of outliers.



- The median marks the mid-point of the data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value and half are less.
- The middle “box” represents the middle 50% of scores for the group. The range of scores from lower to upper quartile is referred to as the inter-quartile range.
- Seventy-five percent of the scores fall below the upper quartile.
- Twenty-five percent of scores fall below the lower quartile.
- The upper and lower whiskers represent scores outside the middle 50%.

Handling Outliers

- The easiest way to handle outliers is to use a **clamp transformation** that clamps all values above an upper threshold and below a lower threshold to these threshold values, thus removing the offending outliers

$$a_i = \begin{cases} lower & \text{if } a_i < lower \\ upper & \text{if } a_i > upper \\ a_i & otherwise \end{cases} \quad (1)$$

where a_i is a specific value of feature a, and *lower* and *upper* are the lower and upper thresholds.

Case Study: Motor Insurance Fraud

Case Study: Motor Insurance Fraud

The following slides show a portion of the ABT that has been developed for a motor insurance claims fraud detection and a data quality report

Portions of the ABT for the motor insurance claims fraud detection problem

ID	Type	Inc.	Marital Status	Num Clmnts.	Injury Type	Hospital Stay	Claim Amnt.	Total Claimed	Num Claims	Num Soft Tiss.	% Soft Tiss.	Claim Amt Rcvd.	Fraud Flag
1	CI	0		2	Soft Tissue	No	1,625	3250	2	2	1.0	0	1
2	CI	0		2	Back	Yes	15,028	60,112	1		0	15,028	0
3	CI	54,613	Married	1	Broken Limb	No	-99,999	0	0	0	0	572	0
4	CI	0		4	Broken Limb	Yes	5,097	11,661	1	1	1.0	7,864	0
5	CI	0		4	Soft Tissue	No	8869	0	0	0	0	0	1
6	CI	0		1	Broken Limb	Yes	17,480	0	0	0	0	17,480	0
7	CI	52,567	Single	3	Broken Limb	No	3,017	18,102	2	1	0.5	0	1
8	CI	0		2	Back	Yes	7463	0	0	0	0	7,463	0
9	CI	0		1	Soft Tissue	No	2,067	0	0	0	0	2,067	0
10	CI	42,300	Married	4	Back	No	2,260	0	0	0	0	2,260	0
		⋮				⋮					⋮		
300	CI	0		2	Broken Limb	No	2,244	0	0	0	0	2,244	0
301	CI	0		1	Broken Limb	No	1,627	92,283	3	0	0	1,627	0
302	CI	0		3	Serious	Yes	270,200	0	0	0	0	270,200	0
303	CI	0		1	Soft Tissue	No	7,668	92,806	3	0	0	7,668	0
304	CI	46,365	Married	1	Back	No	3,217	0	0		0	1,653	0
		⋮				⋮					⋮		
458	CI	48,176	Married	3	Soft Tissue	Yes	4,653	8,203	1	0	0	4,653	0
459	CI	0		1	Soft Tissue	Yes	881	51,245	3	0	0	0	1
460	CI	0		3	Back	No	8,688	729,792	56	5	0.08	8,688	0
461	CI	47,371	Divorced	1	Broken Limb	Yes	3,194	11,668	1	0	0	3,194	0
462	CI	0		1	Soft Tissue	No	6,821	0	0	0	0	0	1
		⋮				⋮					⋮		
491	CI	40,204	Single	1	Back	No	75,748	11,116	1	0	0	0	1
492	CI	0		1	Broken Limb	No	6,172	6,041	1		0	6,172	0
493	CI	0		1	Soft Tissue	Yes	2,569	20,055	1	0	0	2,569	0
494	CI	31,951	Married	1	Broken Limb	No	5,227	22,095	1	0	0	5,227	0
495	CI	0		2	Back	No	3,813	9,882	3	0	0	0	1
496	CI	0		1	Soft Tissue	No	2,118	0	0	0	0	0	1
497	CI	29,280	Married	4	Broken Limb	Yes	3,199	0	0	0	0	0	1
498	CI	0		1	Broken Limb	Yes	32,469	0	0	0	0	16,763	0
499	CI	46,683	Married	1	Broken Limb	No	179,448	0	0		0	179,448	0
500	CI	0		1	Broken Limb	No	8,259	0	0	0	0	0	1

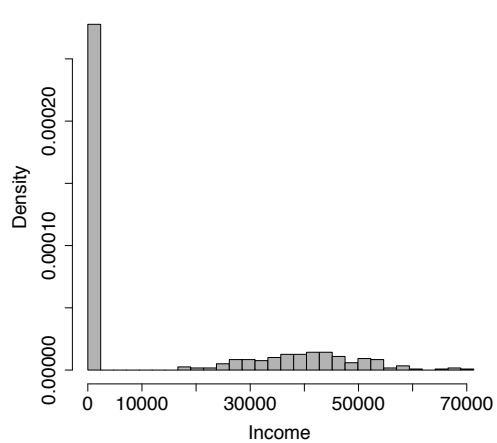
A data quality report for the motor insurance claims fraud detection ABT

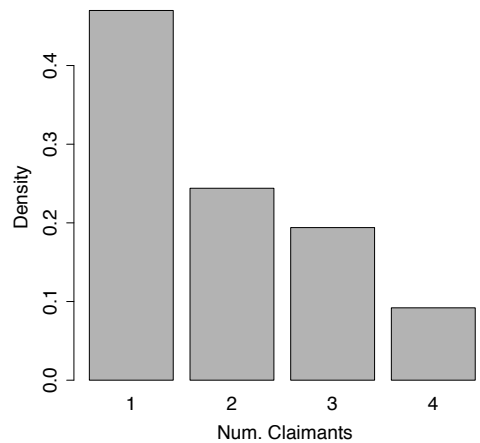
(a) Continuous Features

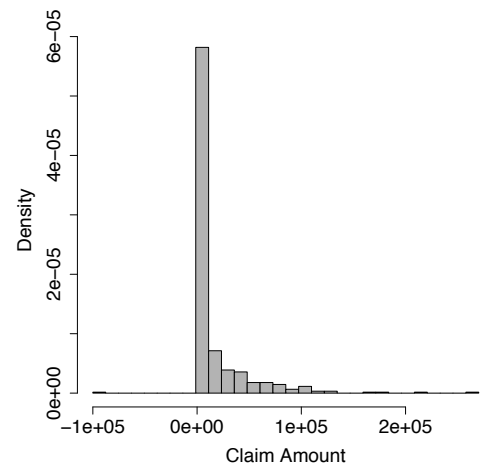
Feature	Count	% Miss.	Card.	Min	1^{st} Qrt.	Mean	Median	3^{rd} Qrt.	Max	Std. Dev.
INCOME	500	0.0	171	0.0	0.0	13,740.0	0.0	33,918.5	71,284.0	20,081.5
NUM CLAIMANTS	500	0.0	4	1.0	1.0	1.9	2	3.0	4.0	1.0
CLAIM AMOUNT	500	0.0	493	-99,999	3,322.3	16,373.2	5,663.0	12,245.5	270,200.0	29,426.3
TOTAL CLAIMED	500	0.0	235	0.0	0.0	9,597.2	0.0	11,282.8	729,792.0	35,655.7
NUM CLAIMS	500	0.0	7	0.0	0.0	0.8	0.0	1.0	56.0	2.7
NUM SOFT TISSUE	500	2.0	6	0.0	0.0	0.2	0.0	0.0	5.0	0.6
% SOFT TISSUE	500	0.0	9	0.0	0.0	0.2	0.0	0.0	2.0	0.4
AMOUNT RECEIVED	500	0.0	329	0.0	0.0	13,051.9	3,253.5	8,191.8	295,303.0	30,547.2
FRAUD FLAG	500	0.0	2	0.0	0.0	0.3	0.0	1.0	1.0	0.5

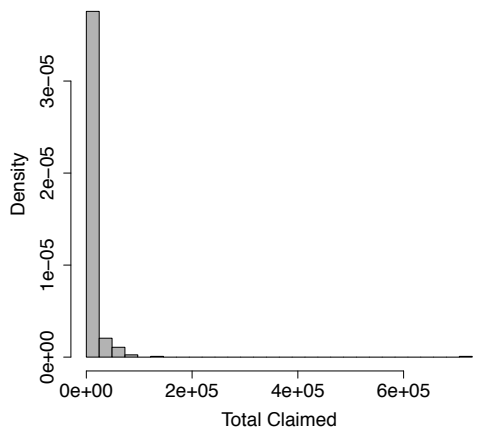
(b) Categorical Features

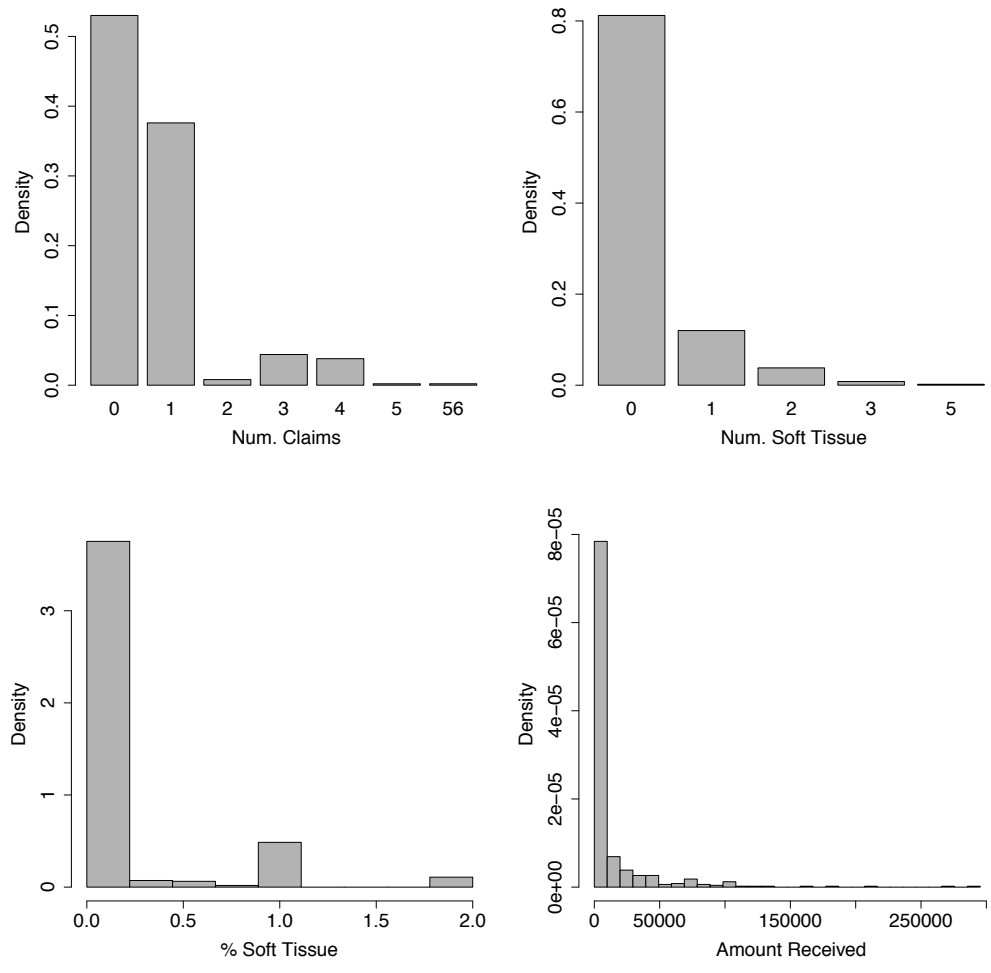
Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2^{nd} Mode	2^{nd} Mode Freq.	2^{nd} Mode %
INSURANCE TYPE	500	0.0	1	CI	500	1.0	–	–	–
MARITAL STATUS	500	61.2	4	Married	99	51.0	Single	48	24.7
INJURY TYPE	500	0.0	4	Broken Limb	177	35.4	Soft Tissue	172	34.4
HOSPITAL STAY	500	0.0	2	No	354	70.8	Yes	146	29.2

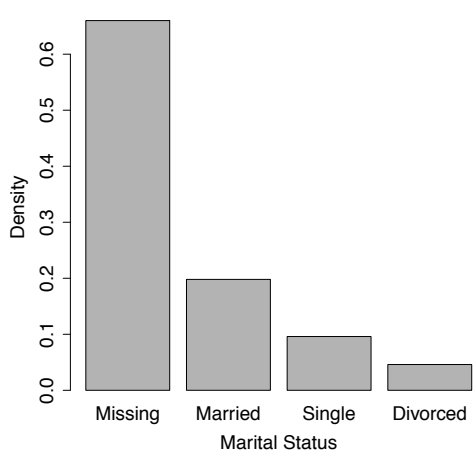

Density
0.00000
0.00010
0.00020
0
10000
30000
50000
70000
Income

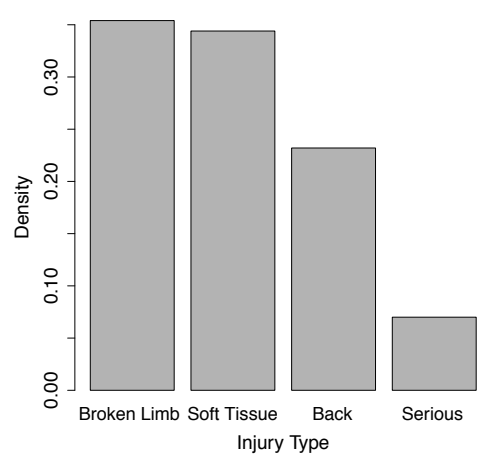

Density
0.0
0.1
0.2
0.3
0.4
1
2
3
4
Num. Claimants

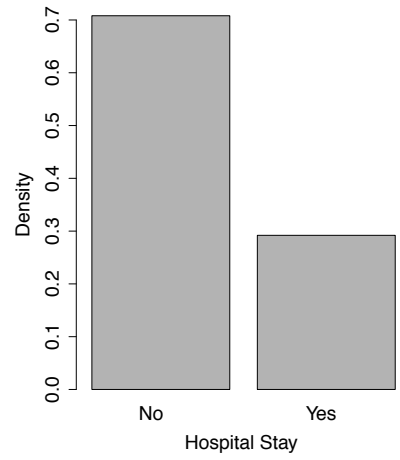

Density
0e+00
2e−05
4e−05
6e−05
−1e+05
0e+00
1e+05
2e+05
Claim Amount

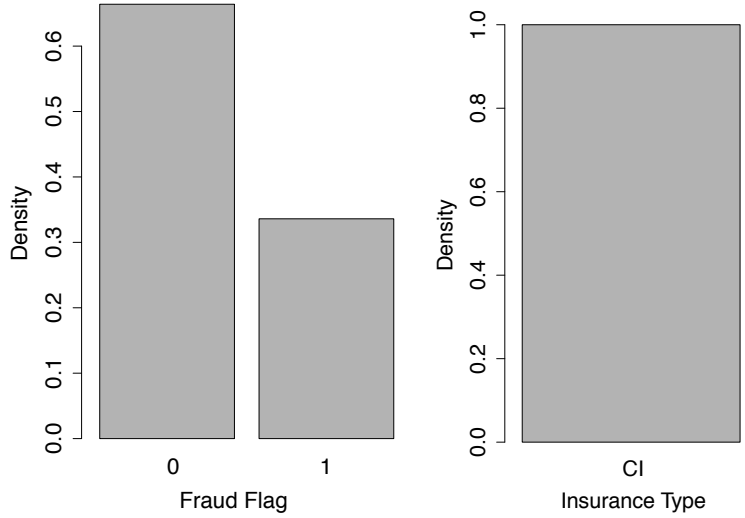

Density
0e+00
1e−05
2e−05
3e−05
0e+00
2e+05
4e+05
6e+05
Total Claimed


Density
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
56
Num. Claims
Density
0.0
0.2
0.4
0.6
0.8
0
1
2
3
5
Num. Soft Tissue
Density
0
1
2
3
0.0
0.5
1.0
1.5
2.0
% Soft Tissue
Density
0e+00
2e−05
4e−05
6e−05
8e−05
0
50000
150000
250000
Amount Received


Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Missing
Married
Single
Divorced
Marital Status


Density
0.00
0.10
0.20
0.30
Broken Limb
Soft Tissue
Back
Serious
Injury Type


Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
No
Yes
Hospital Stay


Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
Fraud Flag
Density
0.0
0.2
0.4
0.6
0.8
1.0
CI
Insurance Type

Case Study: Motor Insurance Fraud

What handling strategies would you recommend for the data quality issues found in the motor Insurance fraud ABT?

A data quality report for the motor insurance claims fraud detection ABT

(c) Continuous Features

Feature	Count	% Miss.	Card.	Min	1st Qrt.	Mean	Median	3rd Qrt.	Max	Std. Dev.
INCOME	500	0.0	171	0.0	0.0	13,740.0	0.0	33,918.5	71,284.0	20,081.5
NUM CLAIMANTS	500	0.0	4	1.0	1.0	1.9	2	3.0	4.0	1.0
CLAIM AMOUNT	500	0.0	493	-99,999	3,322.3	16,373.2	5,663.0	12,245.5	270,200.0	29,426.3
TOTAL CLAIMED	500	0.0	235	0.0	0.0	9,597.2	0.0	11,282.8	729,792.0	35,655.7
NUM CLAIMS	500	0.0	7	0.0	0.0	0.8	0.0	1.0	56.0	2.7
NUM SOFT TISSUE	500	2.0	6	0.0	0.0	0.2	0.0	0.0	5.0	0.6
% SOFT TISSUE	500	0.0	9	0.0	0.0	0.2	0.0	0.0	2.0	0.4
AMOUNT RECEIVED	500	0.0	329	0.0	0.0	13,051.9	3,253.5	8,191.8	295,303.0	30,547.2
FRAUD FLAG	500	0.0	2	0.0	0.0	0.3	0.0	1.0	1.0	0.5

(d) Categorical Features

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
INSURANCE TYPE	500	0.0	1	CI	500	1.0	–	–	–
MARITAL STATUS	500	61.2	4	Married	99	51.0	Single	48	24.7
INJURY TYPE	500	0.0	4	Broken Limb	177	35.4	Soft Tissue	172	34.4
HOSPITAL STAY	500	0.0	2	No	354	70.8	Yes	146	29.2

Missing Values

- MARITAL STATUS and NUM. SOFT TISSUE are the only features with an obvious problem with missing values.
- 60% of the values for MARITAL STATUS are missing, so this feature should almost certainly be removed from the ABT
- Only 2% of the values for the NUM. SOFT TISSUE feature are missing, so removal would be extreme in this case.
- This issue should be noted in the data quality plan.
- Histogram for the INCOME feature – zero values

Irregular Cardinality

- Cardinality of the INSURANCE TYPE feature is 1 (*ci*)
- *ci* refers to car insurance – removed from the ABT
- Many of the continuous features in the dataset have very low cardinality values
- NUM. CLAIMANTS, NUM. CLAIMS, NUM. SOFT TISSUE, % SOFT TISSUE, and FRAUD FLAG all have cardinality less than 10
- The cardinality of 2 for the FRAUD FLAG feature highlights the fact that this is really a categorical feature

Outliers

- CLAIM AMOUNT jumps out as having an unusual minimum value of -99,999
- Treat as a invalid outlier and replace with a missing value.
- CLAIM AMOUNT, TOTAL CLAIMED, NUM. CLAIMS and AMOUNT RECEIVED all seem to have unusually high maximum values
- When investigated TOTAL CLAIMED and NUM. CLAIMS both came from same policy – company policy rather than an individual policy – removed from the ABT
- When investigated large maximums for CLAIM AMOUNT and AMOUNT RECEIVED – a valid outlier and represents an unusually large claim for a very serious injury

Case Study: Motor Insurance Fraud

The data quality plan for the motor insurance fraud prediction

Feature	Data Quality Issue	Potential Handling Strategies
NUM SOFT TISSUE	Missing values (2%)	Imputation (median: 0.0)
CLAIM AMOUNT	Outliers (high)	Clamp transformation (manual: 0, 80 000)
AMOUNT RECEIVED	Outliers (high)	Clamp transformation (manual: 0, 80 000)

Summary

- The key outcomes of the **data exploration** process are that the practitioner should
  1. Have *gotten to know* the features within the ABT, especially their central tendencies, variations, and **distributions**.
  2. Have identified any **data quality issues** within the ABT, in particular **missing values**, **irregular cardinality**, and **outliers**.
  3. Have corrected any data quality issues due to **invalid data**.
  4. Have recorded any data quality issues due to **valid data** in a **data quality plan** along with potential handling strategies.
  5. Be confident that enough good quality data exists to continue with a project.

Any questions?

Recommended Reading

- **Core Text:**
  Fundamentals of Machine Learning for Predictive Data Analytics
  By John D. Kelleher, Brian Mac Namee and Aoife D'Arcy
- This week we covered Chapter 2, sections 2.3 and 2.4 and Chapter 3, sections 3.1 – 3.4
- I would suggest that you would read over these sections again
- Email me if you have any questions and I will cover them at the beginning of class next week
- Next week we will cover the rest of Chapter 3

Labs

- There will be a lab on Thursday at 1.30 pm this week.
- We will be using R and RStudio during the labs.
- R `https://www.r-project.org/about.html`
- RStudio `https://www.rstudio.com/`
- Try to install these on your laptops
- If you have any problem we will try to help you on Thursday
- You will need these for your assignments