COMP20008 Elements of Data Processing

Semester 2 2018

Lecture 4: Data Preprocessing and Cleaning: Missing Values and Outlier Detection

What we've covered so far!

Finished:

- Lecture 1: Introduction
- Lectures 2-3: Data formats: structured, unstructured and semi-structured

Next:

- Lectures 4-5: Data preprocessing and cleaning: missing values, outlier detection and recommender systems

Announcements

- Workshop Tuesday 9.00am moved to Thursday 4.15pm
  - Reason: venue is not good!
  - Starting from This Week!
- Workshop Thursday 6.15pm moved to Tuesday 4.15pm
  - Reason: very small registered students!
  - Starting from Next Week!

Why is pre-processing needed?

Name	Age	Date of Birth
"Henry"	20.2	20 years ago
Katherine	Forty-one	20/11/66
Michelle	37	5/20/79
Oscar@!!	"5"	13th Feb. 2019
-	42	-
Mike___Moore	669	-
巴拉克奥巴马	52	1961年8月4日

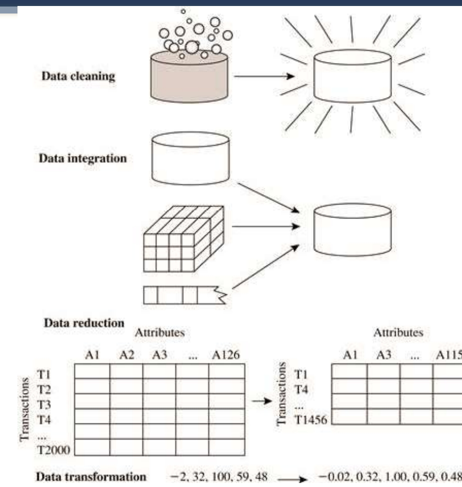
Why is pre-processing needed?

- Measuring data quality
  - Accuracy
    - Correct or wrong, accurate or not
  - Completeness
    - Not recorded, unavailable
  - Consistency
    - E.g. discrepancies in representation
  - Timeliness
    - Updated in a timely way
  - Believability
    - Do I trust the data is correct?
  - Interpretability
    - How easily can I understand the data?

	Date of Birth
1	20 years ago
2	20/11/66
3	5/20/79
4	13th Feb. 2019
5	-
6	-
7	1961年8月4日

THE UNIVERSITY OF MELBOURNE

Major data preprocessing activities



Data mining concepts and techniques, Han et al 2012

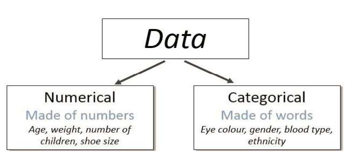
THE UNIVERSITY OF MELBOURNE

Terminology

Height	Weight	Age	Gender
1.8	80	22	Male
1.53	82	23	Male
1.6	62	18	Female

- The 4 columns (height, weight, age, gender) are *features or attributes*
- The data items (3 rows) are called *instances* or *objects*

- Height, Weight and Age are *continuous* features
- Gender is a *categorical* or *discrete* feature



THE UNIVERSITY OF MELBOURNE

Data cleaning – The Process

- Many tools exist (Google Refine, Kettle, Talend, …)
  - Data scrubbing
  - Data discrepancy detection
  - Data auditing
  - ETL (Extract Transform Load) tools: users specify transformations via a graphical interface
- Our emphasis will be to understand some of the methods employed by some of these tools

- Noisy data
- Inconsistent data
- Intentionally disguised data
- Incomplete (missing data)

Noisy data

- Truncated fields (exceeded 80 character limit)
- Text incorrectly split across cells (e.g. separator issues)
- Salary="-5"
- Some causes
  - Imprecise instruments
  - Data entry issues
  - Data transmission issues

Inconsistent data

- Different naming representations ("Melbourne University" versus "University of Melbourne") or ("three" versus "3")
- Different date formats ("3/4/2016" versus "3rd April 2016")
- Age=20, Birthdate="1/1/2002"
- Two students with the same student id
- Outliers
  - E.g. 62,72,75,75,78,80,82,84,86,87,87,89,89,90,999
    - No good if it is list of ages of hospital patients
    - Might be ok though for a listing of people number of contacts on Linkedin though
  - Can use automated techniques, but also need domain knowledge

Disguised data

- Everyone's birthday is January 1st?
- Email address is xx@xx.com
- Adriaans and Zantige
  - "*Recently, a colleague rented a car in the USA. Since he was Dutch, his post-code did not fit the fields of the computer program. The car hire representative suggested that she use the zip code of the rental office instead.*"
- How to handle
  - Look for "unusual" or suspicious values in the dataset, using knowledge about the domain

Missing or incomplete data

- Lacking feature values
    - Name=""
    - Age=null
- Some types of missing data (Rubin 1976)
  - Missing completely at random: Data are missing independently of observed and unobserved data.
    - E.g/ Coin flipping to decide whether or not to answer an exam question.
  - Missing not completely at random
    - I create a dataset by surveying the class about how healthy they feel. What is the meaning of missing values for those who don't respond?

THE UNIVERSITY OF MELBOURNE

Missing Completely at Random: Example

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	121
29	91
30	
30	110
31	
44	118
46	93
48	
51	
51	116
54	

Missing data are MCAR when the probability of missing data on a variable is unrelated to any other measured variable and is unrelated to the variable with missing values itself.

THE UNIVERSITY OF MELBOURNE

Missing Not at Random: Example

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	133
26	121
29	
30	
30	110
31	
44	118
46	
48	141
51	
51	116
54	

Data are MNAR when the missing values on a variable are related to the values of that variable itself, even after controlling for other variables.

For example, when data are missing on IQ and only the people with low IQ values have missing observations for this variable.

THE UNIVERSITY OF MELBOURNE

Example: USA Salary survey data

Name	Salary
Person C	$59k
Person D	$63k
Person H	$99k
Person E	$102k
Person G	$140k
Person F	$150k
Person A	$180k
Person B	-

- Is Person B’s salary missing at random?
- Very difficult to determine reasons for missingness.
  - In practice report assumptions about missingness.

THE UNIVERSITY OF MELBOURNE

Causes of missing data

- Why does it occur?
  - Malfunction of equipment (e.g. sensors)
  - Not recorded due to misunderstanding
  - May not be considered important at time of entry
  - Deliberate

Dealing with missing data

- What are the consequences of missing data?
  - May break application programs not expecting it
  - Less power for later analysis analysis
  - May bias later analysis

- *So, how to handle it?*

Strategy 1: Delete all instances with a missing value

- Sometimes called case deletion
- Effects
  - Easy to analyse the new (complete data)
  - May produce bias on analysis if new sample size small or structure exists in the missing data.

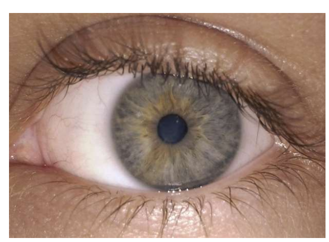
Case deletion

Person	Star Wars	Batman	Jurassic World	The Martian	The Revenant	Lego Movie	Selma
Mandy	1	2	1	3	3	2	3
James	3	2	-	-	-	1	-
John	-	-	1	2	-	-	-
Jill	1	-	-	3	2	1	-

Person	Star Wars	Batman	Jurassic World	The Martian	The Revenant	Lego Movie	Selma
Mandy	1	2	1	3	3	2	3

Strategy 2: Manually correct

- A human eyeballs the missing value and fills it in using their expert knowledge



https://en.wikipedia.org/wiki/Eye

Strategy 3: Imputation

- Impute a value (replace the missing value with a substitute one)
- After imputing all missing values, can use standard analysis techniques for complete datasets

Person	Star Wars	Batman	Jurassic World	The Martian	The Revenant	Lego Movie	Selma	....
James	3	2	-	-	-	1	-	
John	-	-	1	2	-	-	-	
Jill	1	-	-	3	2	1	-	

⇩

Person	Star Wars	Batman	Jurassic World	The Martian	The Revenant	Lego Movie	Selma	....
James	3	2	2	2	1	1	1	
John	3	2	1	2	2	1	1	
Jill	1	1	1	3	2	1	1	

Imputation: Fill in with zeros (or similar)

Person	Star Wars	Batman	Jurassic World	The Martian	The Revenant	Lego Movie	Selma	....
James	3	2	0	0	0	1	0	
John	0	0	1	2	0	0	0	
Jill	1	0	0	3	2	1	0	

- Simple
- Won't break application programs
- Limited utility for analysis

Imputation: Fill in with mean value

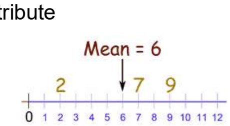
- Popular method
  - Can be good for supervised classification
  - Apply separately to each attribute

Name	Age
Daisy	10
Maisy	15
Harry	2
Jackie	-

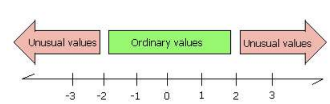
Jackie's age is imputed to be (10+15+2)/3=9

Imputation: Fill in with mean value cont

- Drawbacks
  - Reduces the variance of the feature
  - Incorrect view of the distribution of that attribute
  - Relationships to other features changes



- Can also use median instead of mean (if distribution is skewed)



1, 3, 3, **6**, 7, 8, 9

Median = 6

1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

= 4.5

- Use mode (most frequent value) imputation for categorical features

Fill in with category mean

- Take categories/clusters and compute the mean ….

Name	Age	Gender
Daisy	10	Female
Maisy	15	Female
Harry	2	Male
Jackie	-	Female

Jackie's age is imputed to be (10+15)/2=12.5 (considering the category "Female")

THE UNIVERSITY OF MELBOURNE

Example: The effect of data cleaning

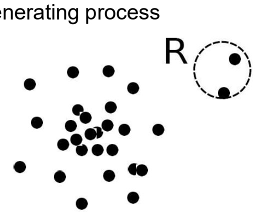
- Math grades of sample group of students
- Download csv file from this link
- Imagine 50 out of 350 marks are missing!

Let's see an ipynb example

THE UNIVERSITY OF MELBOURNE

Outlier analysis

- **Outlier**: A data object that **deviates significantly** from the normal objects as if it were **generated by a different mechanism** (Hawkins, 1980)
  - Ex.: Unusual credit card purchase, sports: Michael Jordon, Lance Franklin, …
- From a statistics perspective
  - Normal (non-outlier) objects are generated using some statistical process
  - The outlier objects deviate from this generating process



THE UNIVERSITY OF MELBOURNE

Example: Hadlum vs Hadlum paternity case

- Paternity case: "The study of outliers", V. Barnett, Journal of the Royal Statistical Society, 27(3), 1978

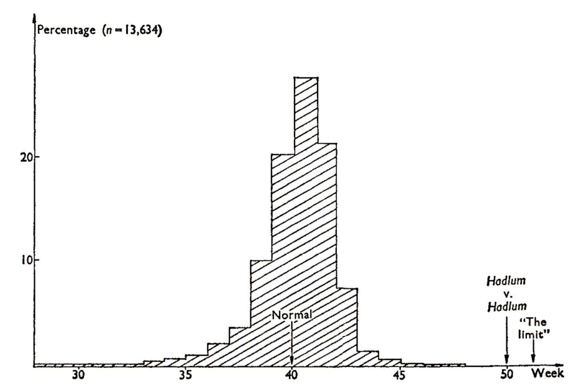


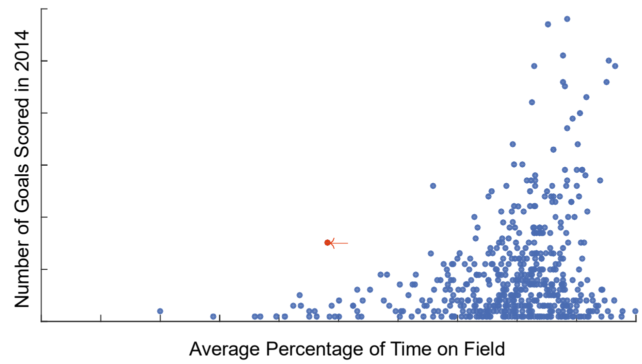
FIG. 1. Distribution of human gestation periods.

Outlier analysis

- Outliers can be different from the noise data
  - Noise is random error or variance in a measured variable
  - Noise should be removed before outlier detection
- Outliers are interesting: Violation of the mechanism that generates the normal data
- Applications:
  - Credit card fraud detection (change in behaviour)
  - Telecom fraud detection
  - Medical analysis (unusual test results)
  - Sports (identifying exceptional talent)

Australian Rules Football

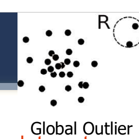
- Daniel Giansiracusa



Why do we care?

- Compute the average age of people in this room
  - Skewed results
- Compute the average salary of people in this room
  - What if Donald Trump is in the audience?

Types of Outliers



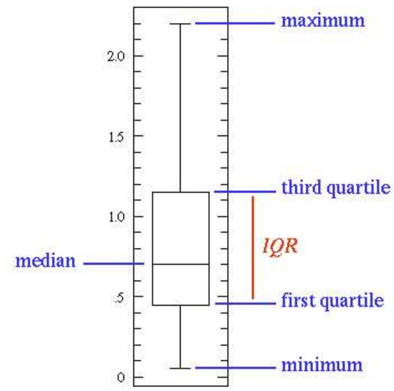
- **Global outlier** (or point anomaly)
  - Object is O_g if it significantly deviates from the rest of the data set
  - Ex. Intrusion detection in computer networks
  - Issue: Find an appropriate measurement of deviation
- **Contextual outlier** (or *conditional outlier*)
  - Object is O_c if it deviates significantly based on a selected context
  - Is 5^o in Melbourne an outlier? (depending on summer or winter?)
  - Attributes of data should be divided into two groups
    - Contextual attributes: defines the context, e.g., time & location
    - Behavioral attributes: characteristics of the object, used in outlier evaluation, e.g., temperature
  - Issue: How to define or formulate meaningful context?

How to detect outliers

- 1-D data
  - Boxplot
  - Histogram
  - Statistical tests
- 2-D Data: Scatter plot and eyeball
- 3-D data: Can also use scatter plot and eyeball
- >3-D data: Statistical or algorithmic methods

Box and whisker plot (diagram from http://www.physics.csbsju.edu/stats/box2.html)

From sample compute
- Minimum and maximum (the whiskers)
- Median
- First quartile(Q1): middle number between median and minimum
- Third quartile(Q3): middle number between median and maximum
- IQR=interquartile range =Q3-Q1

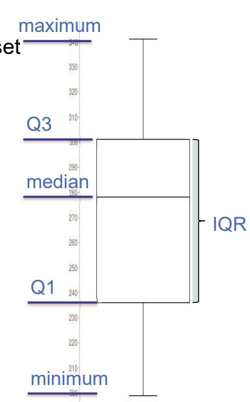


Identifying Outliers: Example

Let n be the number of data values in the dataset.
Example: 199, 201, 236, 269, 271, 278, 283, 291, 301, 303, 341

Steps to draw the boxplot:

1. The median (Q2) is the middle value of the data set
   - $Q2 = \frac{1}{2}(n+1)\ th\ term$ → 6th term → 278
2. The lower quartile (Q1) is the median of the lower half of the data set
   - $Q1 = \frac{1}{4}(n+1)\ th\ term$ → 3rd term → 236
3. The upper quartile (Q3) is the median of the upper half of the data set
   - $Q3 = \frac{3}{4}(n+1)\ th\ term$ → 9th term → 301
4. The interquartile range (IQR) is the spread of the middle 50% of the data values
   - $IQR = Q3 - Q1$ → 9th term → 301-236 → 65

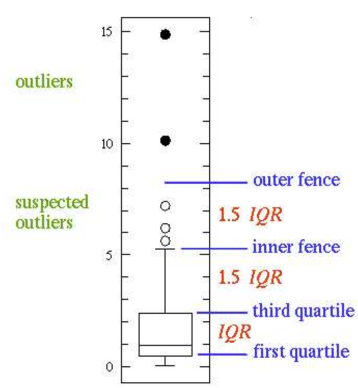


Outliers and Tukey Boxplots(diagram from http://www.physics.csbsju.edu/stats/box2.html)

Whiskers
- Lowest point still within 1.5IQR of lower quartile
- Highest point still within 1.5 IQR of upper quartile

Outliers (filled black)
- >3*IQR above third quartile, or
- >3*IQR below 1st quartile

Suspected outliers (open black)
- >1.5*IQR above third quartile, or
- >1.5*IQR below 1st quartile



Example of box plot outlier detection

- **Continuing the previous example:** 199, 201, 236, 269, 271, 278, 283, 291, 301, 303, 341
  - $Q1 = 236, Q3 = 301, IQR = Q3 - Q1 = 65$

Suspected outliers

- 1.5*IQR = 1.5 * 65 = 97.5
- >1.5*IQR above third quartile or >1.5*IQR below 1st quartile

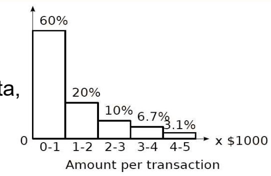
Outliers (filled black)

- 3*IQR = 3 * 65 = 195
- >3*IQR above third quartile, or >3*IQR below 1st quartile

- **Another example from**
  - http://www.alcula.com/calculators/statistics/box-plot
  - 10,20,30,40,50,60,70,80,90,100,120,130,140,150,160,180,999

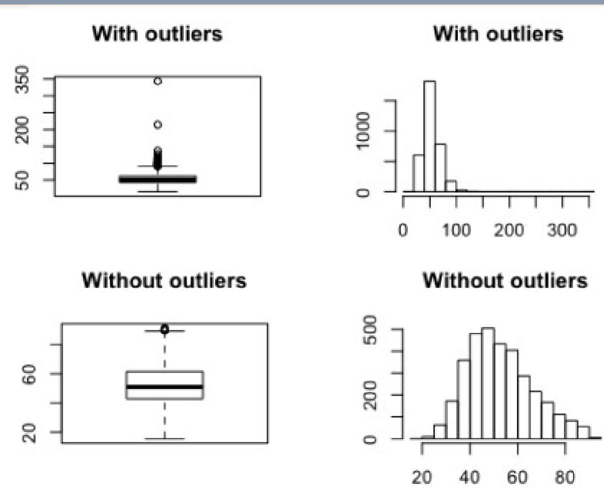
Detection Using Histogram

- The model of normal data is learned from the input data without any *a priori* structure.
- Often makes fewer assumptions about the data, and thus can be applicable in more scenarios
- Outlier detection using histogram:



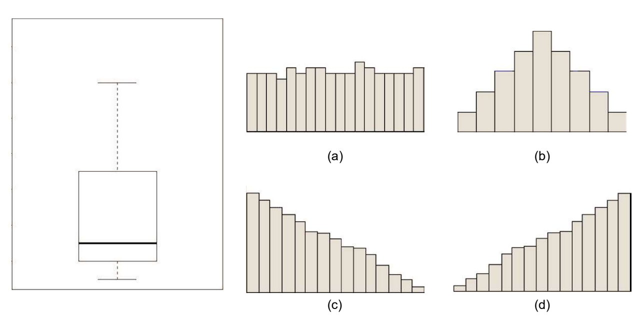
  - Figure shows the histogram of purchase amounts in transactions
  - A transaction in the amount of $7,500 is an outlier, since only 0.2% transactions have an amount higher than $5,000
- Problem: Hard to choose an appropriate bin size for histogram
  - Too small bin size → normal objects in empty/rare bins, false positive
  - Too big bin size → outliers in some frequent bins, false negative

Outlier Check

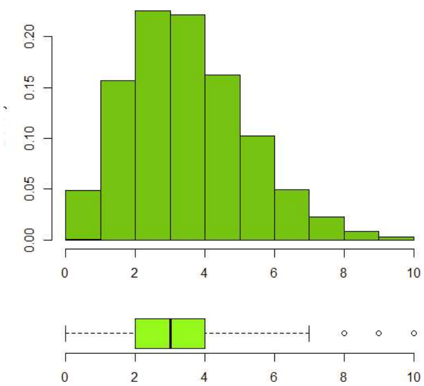


Exercise

- Which histogram is the best representation of the boxplot?

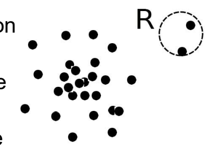


Align Histogram with BoxPlot



Outlier Detection: Statistical Methods

- Statistical methods assume that the normal data follow some statistical model
  - The data not following the model are outliers.
- Example (right figure): First use Gaussian distribution to model the normal data
  - For each object y in region R, estimate $g_D(y)$, the probability of y fits the Gaussian distribution
  - If $g_D(y)$ is very low, y is unlikely generated by the Gaussian model, thus an outlier



- Effectiveness of statistical methods: highly depends on whether the assumption of statistical model holds in the real data
- There are rich alternatives to use various statistical models

Univariate case: Grubb's Test

- Univariate outlier detection: Detect one outlier at a time and repeat.
  - Compute the following statistic where x_i is a data instance

$$\frac{\max_{i=1,\ldots,N} |x_i - \mu|}{\sigma}$$

where μ is the sample mean and

σ is the sample standard deviation

Then assume population is normally distributed and do a statistical hypothesis test (Python package outlier_utlis). Farthest point is an outlier if unlikely to have occurred under normal distribution assumption. Throw away outlier if test indicates that instance is "too far" from the mean.

Grubb's Test: Example

Here are 8 spectrometer measurements on a uranium isotope:

199.31	199.53	200.19	200.82	201.92	201.95	202.18	245.57

$$\frac{\max_{i=1,\ldots,N} |x_i - \mu|}{\sigma}$$

- Step1: Calculate μ and σ

Average (μ)	206.434
Std. dev. (σ)	15.853

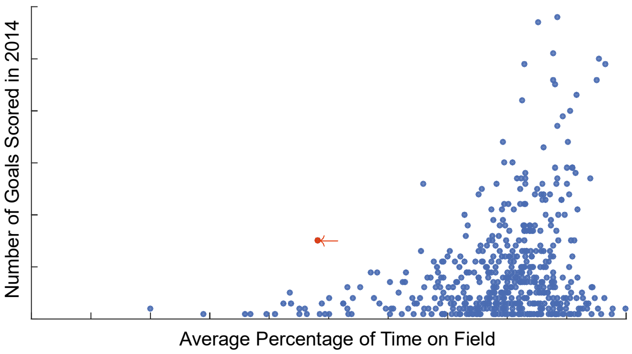
- Step2: Grubb's test values

Data point	199.31	199.53	200.19	200.82	201.92	201.95	202.18	245.57
Index i	1	2	3	4	5	6	7	8
Grubb's test	0.449	0.436	0.394	0.354	0.285	0.283	0.268	2.469

THE UNIVERSITY OF MELBOURNE

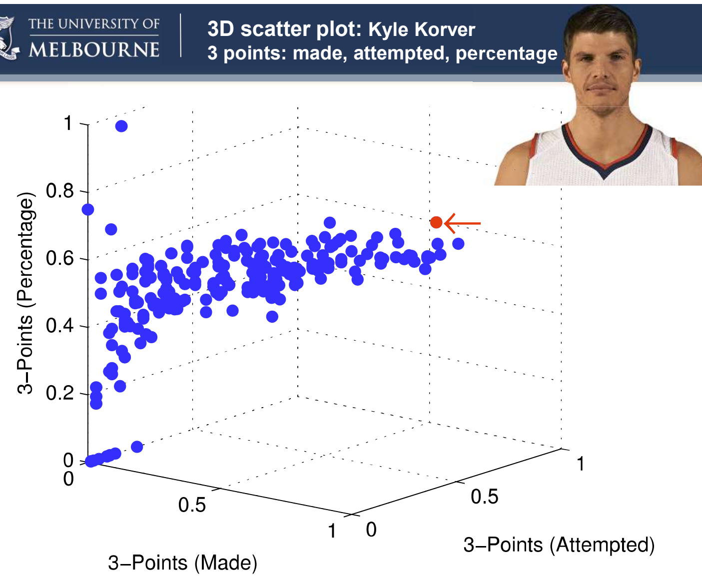
2D scatter plot

- Daniel Giansiracusa



THE UNIVERSITY OF MELBOURNE

3D scatter plot: Kyle Korver
3 points: made, attempted, percentage



[Stats Main][AFL Main]

[2013 Stats][2015 Stats]

2014 Player Stats

[2014 Stats Summary]

[Adelaide][Brisbane Lions][Carlton][Collingwood][Essendon][Fremantle][Geelong][Gold Coast][Greater Western Sydney][Hawthorn]
[Melbourne][North Melbourne][Port Adelaide][Richmond][St Kilda][Sydney][West Coast][Western Bulldogs]
[All Teams]

Abbreviations key

Adelaide [Game by Game]

#	Player	GM	KI	MK	HB	DI	DA	GL	BH	HO	TK	RB	IF	CL	CG	FF	FA	BR	CP	UP	CM	MI	1%	BO	GA	%P	SU
32	Dangerfield, Patrick	22	276	74	272	548	24.91	17	22	28	78	33	104	136	66	34	19	21	341	210	25	16	35	18	10	83.7	
9	Sloane, Rory	22	269	105	252	521	23.68	13	9	19	147	45	99	92	50	26	15	10	275	256	9	7	64	5	21	87.2	
5	Thompson, Scott	19	257	69	262	519	27.32	3	7	2	86	28	77	118	61	19	22	14	224	280	3	5	21	1	7	81.7	0/2
33	Smith, Brodie	22	287	108	209	496	22.55	11	8		35	109	76	18	45	9	6	4	142	319	7	2	56	46	7	87.0	
10	Jaensch, Matthew	22	297	126	166	463	21.05	7	5		54	89	54	7	34	19	10		106	325	16	1	57	34	3	81.3	
26	Douglas, Richard	19	266	52	147	413	21.74	11	8	4	91	21	96	91	38	22	17		182	228	2	6	36	13	11	86.4	
11	Wright, Matthew	20	224	89	150	374	18.70	14	8		68	22	47	39	27	30	6		141	227	4	12	26	6	17	80.0	1/2
24	Jacobs, Sam	22	193	90	165	358	16.27	7	3	763	46	20	40	69	33	11	15	6	150	189	19	4	63	1	10	87.9	0/1
14	Mackay, David	19	168	58	174	342	18.00	11	7		77	30	62	32	31	22	13		127	224	5	3	34	37	8	81.1	0/2
18	Betts, Eddie	22	167	53	123	290	13.18	51	22		74	8	37	30	39	19	16	4	149	136	3	29	21	8	29	87.7	
1	Podsiadly, James	21	189	119	101	290	13.81	26	14	2	37	17	52	2	63	14	25	4	132	165	41	35	60	1	16	90.1	
16	Brown, Luke	22	138	55	148	286	13.00	1	1		54	37	16	8	18	13	5		81	205	1	1	42	1	4	84.5	
2	Crouch, Brad	11	125	26	147	272	24.73	5	6	1	61	22	40	56	30	8	6		114	156	1	2	17	9	6	83.7	0/1
36	Martin, Brodie	17	155	65	109	264	15.53	8	15		45	30	38	23	40	13	11		97	174	7	12	34	11	4	69.2	2/1
12	Talia, Daniel	22	167	105	93	260	11.82		1		24	45	25		29	11	12		79	183	13		149	1	2	90.0	0/1
29	Laird, Rory	16	126	65	129	255	15.94	2	2		37	21	34	15	31	8	8		81	177	1	1	25	2	2	75.4	2/0
4	Jenkins, Josh	20	170	86	64	234	11.70	40	26	55	27	13	46	11	36	12	8	3	97	140	21	32	48	10	7	90.6	
13	Walker, Taylor	15	138	84	82	220	14.67	34	22		24		50		47	10	21	5	102	120	23	31	20		17	90.3	
3	Reilly, Brent	10	130	65	63	193	19.30				19	32	17	8	30	3	13		46	139	7		19	24	1	81.0	
17	Kerridge, Sam	14	72	33	84	156	11.14	10	1		52	10	23	26	25	3	14		54	97	2	9	9	4	5	83.7	0/1

Multidimensional case: Who are the outliers? [From http://afltables.com/afl/stats/2014.html]

THE UNIVERSITY OF MELBOURNE

Acknowledgements

– Data Mining Concepts and Techniques. Han, Kamber and Pei. 3rd edition (chapter 3 and 12). Available through library as ebook.
– Data analysis using regression and multilevel hierarchical models. Gelman and Hill (chapter 25), 2006.