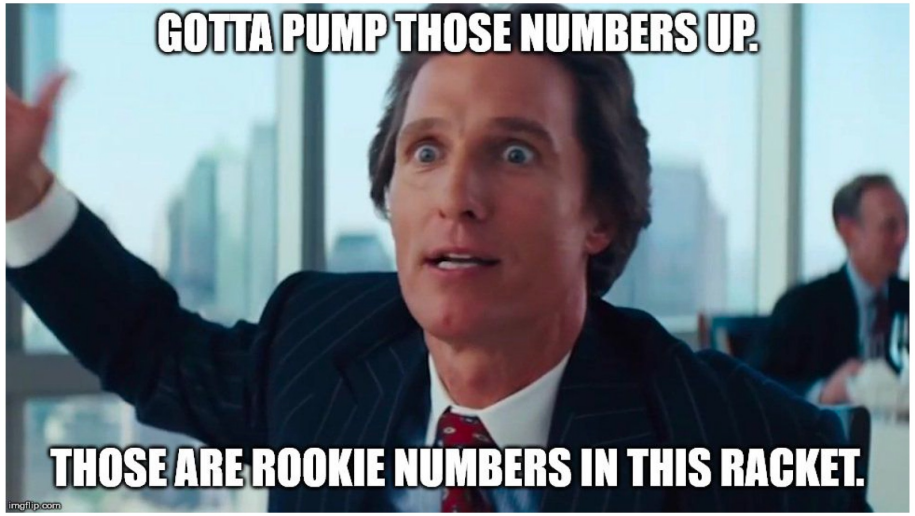
Performance management in practice: Using R to analyze performance data

Dr Patrick J. Ferguson
ACCT30002 – Enterprise Performance Management
University of Melbourne

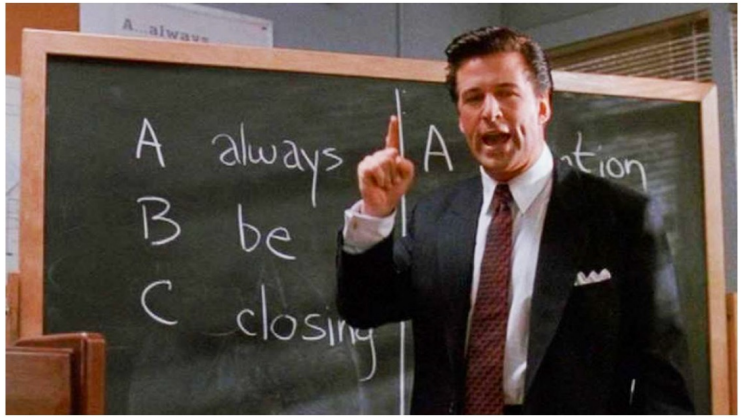


In the abstract, performance management seems straight-forward; in practice, it quickly gets complicated.

Performance measurement as an inference

The canonical agency problem: Firms hire and pay (self-interested) employees to do a job, but employees' performance is not 'perfectly observable'.

- How does the firm know that the employee did a 'good' job? That they exerted effort toward the 'right' tasks?
- For the firm and the individual under evaluation, the costs of getting this wrong are high…



There are good and bad ways to mitigate the agency problem. Shouting at subordinates only gets you so far…

Firms must *infer* employees' performance (effort), but from what?

- Outcomes (e.g., sales, profit) are often incomplete, noisy, and insensitive.
  - Noisy mappings complicate inference: low effort -> good outcome; high effort -> bad outcome
- Simple inputs (e.g., hours worked) can be gamed

The properties of performance measures

Managers use ('contract on') performance measures to infer employees' actions (and decisions). What properties should such measures have?

1. Valid – captures the underlying construct of interest
2. Precise/low noise
3. Sensitive – responds to behavioral changes
4. Low-cost to measure and verify

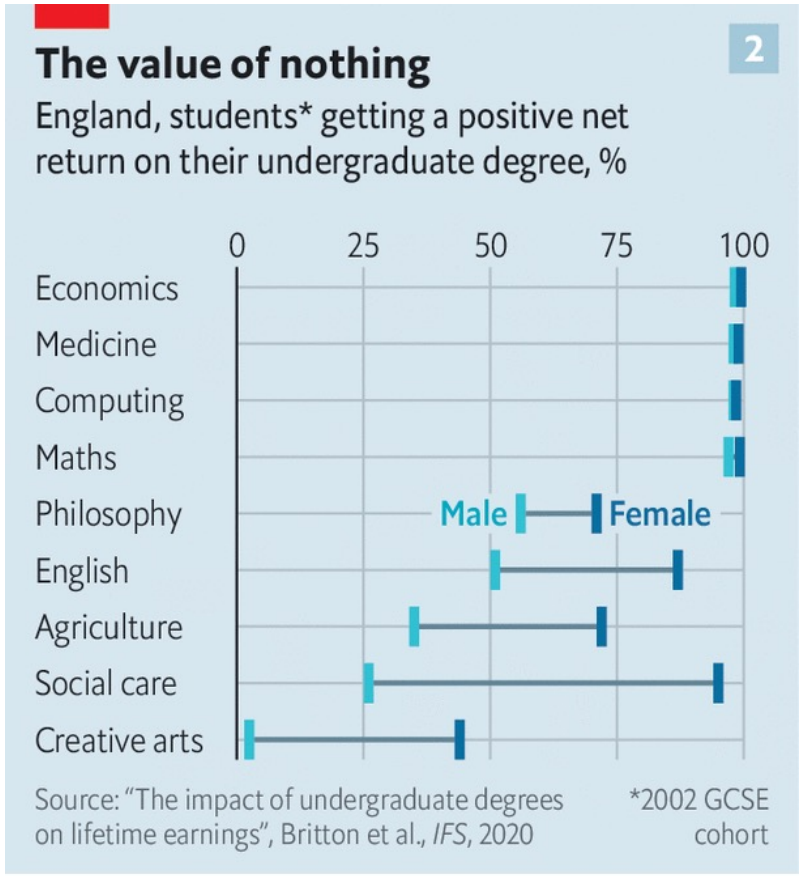
These are attributes of data. Performance management is about using 'good' data to learn about an uncertain and unverifiable world.

Performance evaluation requires managers to 'work' with data.

- Data is costly to collect and store, tricky to clean and arrange, and challenging to meaningfully analyze
- Firms now need graduates who can 'do' data science to help design and execute performance management systems.
    - Excel doesn't always (often?) cut it…

According to this framework, are VCE results 'good' performance measures?

What does this look like in practice?



The ability to produce and interpret visualized data is a lucrative skill set

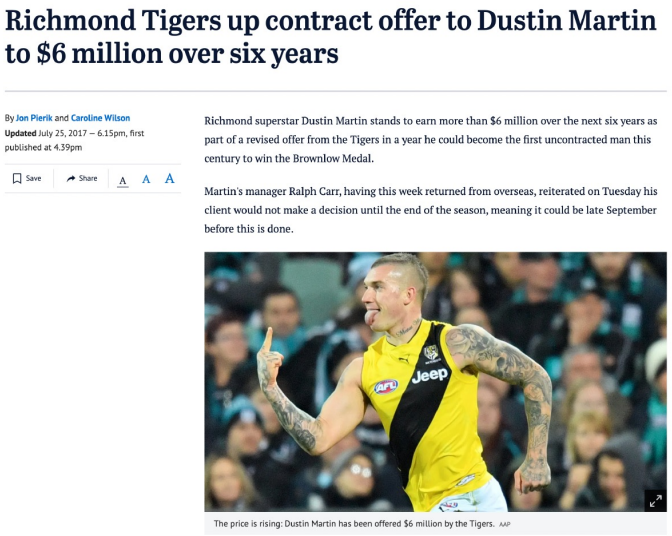
We are going to get our hands dirty working with raw performance data. We will write code to explore and analyze this data set.

- Performance measurement in the 'real world' is messy (even in very structured, transparent settings).
- Informed and careful analysis using basic statistical techniques can generate a wealth of insights

We will be using R, the open-source programming language.

- (Re)applying basic data wrangling and analysis you should have covered in QM courses
- The objective is to identify analysis that needs to be performed, and then understand and interpret the output from this analysis.
  - I will provide resources to introduce and explain the code. <u>You will not be assessed on writing code.</u>
  - Those who want to engage more on the technical details, I will give you the resources to do so.

Workshop data task: Setting



Richmond Tigers up contract offer to Dustin Martin to $6 million over six years

By Jon Pierik and Caroline Wilson
Updated July 25, 2017 – 6.15pm, first published at 4.39pm

Save Share A A A

Richmond superstar Dustin Martin stands to earn more than $6 million over the next six years as part of a revised offer from the Tigers in a year he could become the first uncontracted man this century to win the Brownlow Medal.

Martin's manager Ralph Carr, having this week returned from overseas, reiterated on Tuesday his client would not make a decision until the end of the season, meaning it could be late September before this is done.

The price is rising: Dustin Martin has been offered $6 million by the Tigers. AAP

Players can be paid $1M+ per season. These are high-stakes performance reviews.

The Australian Football League (AFL) is the national Australian-rules football competition.

- In 2019, earnings exceeded $1 billion.
- If you are unfamiliar, a quick primer on the sport: http://www.youtube.com/watch?v=XMZYZcoAcU0

In the AFL, coaches (managers) evaluate players (subordinates) after each game.

- The evaluations determine which players get selected for upcoming games, and which players receive contract renewals at the end of the season.
- In conducting these reviews, coaches consult many 'stats' (quantitative metrics) that capture different dimensions of each player's performance during a game.

Workshop data task: Data and objectives

You will be working with a sample of the actual data that AFL coaches use to evaluate players.

- This data set is at the player-game level
- ~1,500 observations; nine different metrics
- 60+ unique players; data from games across three seasons

Game_ID	Player_ID	Position	GameTotalMins	GameTotalDistance_km	Disposals	Disposal_efficiency	Goals
1	2	DEFENCE	108.5	13.2552	17	58.8	0
1	5	DEFENCE	115.4	14.2282	21	66.7	0
1	6	MIDFIELD	90.1	12.2865	12	66.7	0
1	7	DEFENCE	109.8	13.1969	19	78.9	1
1	8	FORWARD	109.3	15.2498	14	78.6	3
1	13	DEFENCE	116.5	13.2598	15	86.7	0
1	14	MIDFIELD	107.8	14.2644	17	76.5	0
1	16	MIDFIELD	95.8	11.8750	21	71.4	3
1	21	DEFENCE	108.2	13.7973	11	81.8	0
1	22	MIDFIELD	91.8	12.0372	14	71.4	1

Bonus points if you can identify the team…

Across the preparation material and workshop, we will analyze this data set.

- Examine performance management concepts from prior weeks. No new theory per se.
- How do we describe the properties of performance measures?
- How can we determine if a measure is sensitive? Noisy?
- How do we measure performance in team settings?

Workshop data task: Prep and workshop

I have broken this data task into two components: preparation task and a workshop.

- Preparation task is a self-directed exercise. I give you instructions, code, data, output, etc to work through.
- You will run and install basic commands on our data set.
- You can also preview questions we will explore in workshop.

The workshop is a live coding session. I will analyze the data set and we will work through a set of performance management questions.

- You only broadly need to 'get' how the code works; but you should absolutely understand what we are doing with the analysis and why we are doing it.

Summarizing our data

When working with data, one of the very first things you will want to do is summarize the variables of interest in your data set. This is because it is often not feasible to look at the value a variable takes for every observation in the data set (and even if doing so was feasible, it is not clear what you would learn by just 'eye-balling' the data). Statistics tells us that a good way to summarize a variable is describe its distribution. I will show you two common, easy-to-interpret ways to do this.

You can calculate the summary statistics of a variable. I will get you to use the summary () function to do this for the variable GameTotalDistance_km:

```
summary(AFL_data_set$GameTotalDistance_km)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1829 12.2714 13.1155 12.9939 14.0226 17.0754
```

A downside of the summary() function is that it lacks many of the statistics we commonly use in economics and data science (e.g, standard deviation, etc). An alternative approach is to use the stargazer() function from the stargazer package (to use this papproach you will need to install and load the stargazed package). Stargazer produces the following output, which provides a compact overivew of the summary statistics for all the variables in our data set (NB Stargazer does not take tibbles as an input, hence why you need to convert the data to a data frame within the function):

```
stargazer(data.frame(AFL_data_set), type = "html")
```

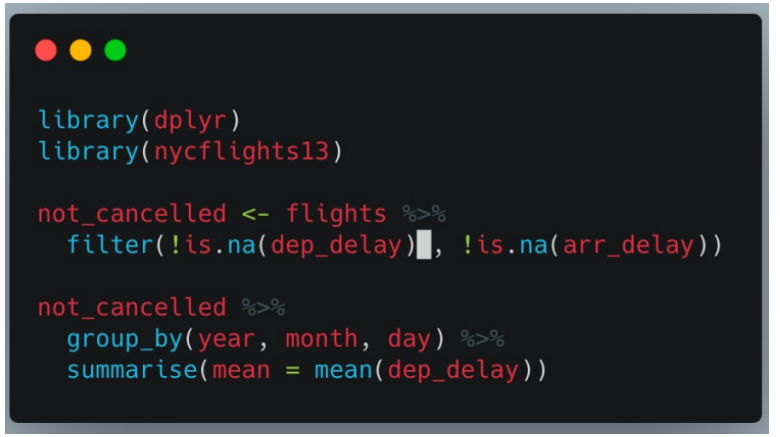
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Game_ID	1,474	35.307	20.521	1	17	54	70
Player_ID	1,474	27.988	15.152	1	15	42	53
GameTotalMins	1,474	99.827	13.892	1.320	93.422	108.295	129.520
GameTotalDistance_km	1,474	12.994	1.703	0.183	12.271	14.023	17.075
Disposals	1,474	17.389	7.364	0	12	22	48
Disposal_efficiency	1,474	73.396	13.012	0.000	65.875	82.325	100.000
Goals	1,474	0.624	1.023	0	0	1	7
Tackles	1,474	3.212	2.424	0	1	4	18
Marks	1,474	4.102	2.464	0	2	6	14
Clearances	1,474	1.716	2.293	0	0	2	13
Margin	1,474	20.258	38.755	-51	-11	42	133
Rainfall_mm	1,474	2.821	5.169	0.000	0.000	2.800	26.200
Wind_mph	1,474	7.430	5.267	0	2	11	20
Meters_per_min	1,474	130.702	9.913	93.902	124.020	137.629	173.652
best_on_ground	1,474	0.033	0.179	0	0	0	1

The preparation task looks intimidating, but it's actually fairly gentle.

Workshop data task: R and RStudio

We will code in R, the ‘defacto language’ of data science.

- Open-source programming language, widely-used in academia and industry
- Well-supported and documented – large, online community



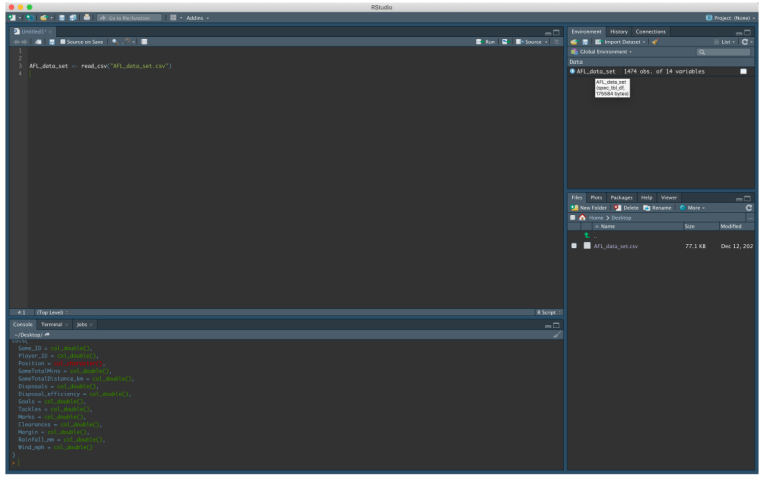
```r
library(dplyr)
library(nycflights13)

not_cancelled <- flights %>%
  filter(!is.na(dep_delay), !is.na(arr_delay))

not_cancelled %>%
  group_by(year, month, day) %>%
  summarise(mean = mean(dep_delay))
```

We can write and run R in any editor (even directly in the terminal), but we will do so in Rstudio.

- Rstudio is a development environment/GUI that provides an intuitive integration with R
- Shows us the code we’ve written (via the editor), the output of that code (via the console), and the ‘objects’ we interact with (via the environmental pane)



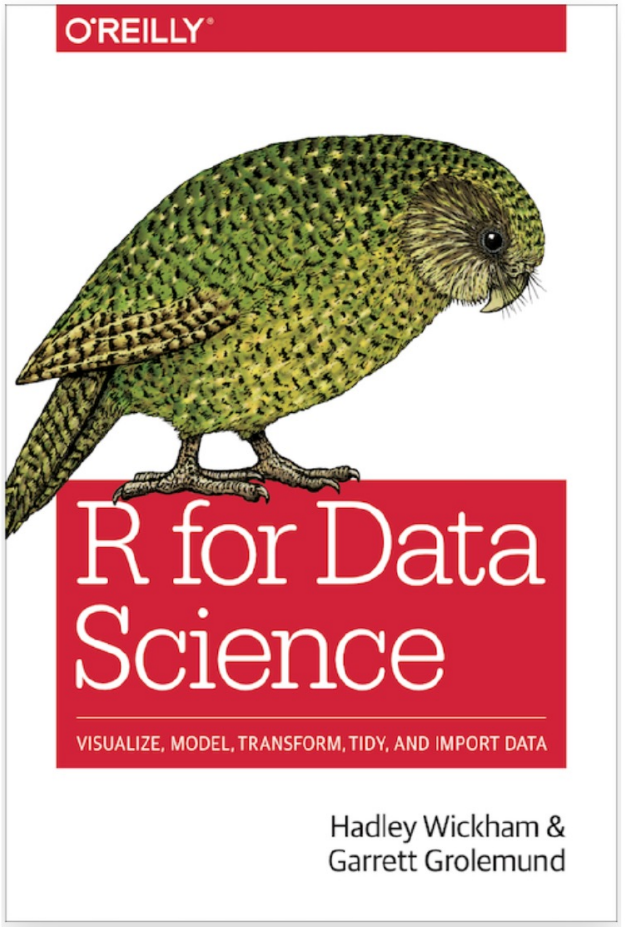
Workshop data task: An R primer

In R, everything (vectors, lists, data frames, functions, etc) is an 'object', and you give every object a name.

- Multiple objects (e.g., data frames) can exist in the same environment (unlike Stata).
- Makes it much easier to merge data sets, work with transformed versions of a data set, etc.

We use 'functions', many of which come pre-written in 'packages', to perform operations on these objects (e.g., cleaning, visualizing, and analyzing data).

- 'Base' R has many useful functions, but the best data science tools come from external packages written by users (just like yourself).
- We will use the Tidyverse suit of packages (dplyr, tidyr, ggplot2, etc).
    - 'Tidy' data – each variable is a column, each observation forms a row, each type of observation unit forms a table.



'The Bible' - R for Data Science: https://r4ds.had.co.nz/

Recap of lecture

What have we covered?

- Agency problem and performance management as inference using data
- Empirical properties of performance measures
- Data science as a performance management tools
- R as a tool for cleaning, arranging, visualizing, and analyzing performance data
- Objectives of our workshop data task

Where are we going?

- Strategic capital investments
- Market revenue analysis
- Profit planning

A few key takeaways

- Firms must use data to infer employees' actions and decisions.
- Data is messy and unstructured – so making inferences is tough.
- Firms place a premium on graduates with data skills…