Manipulating data tables II

Overview

Lab 2 discussion

Discussion of chapter 2 of Astroball

Review of data manipulation methods and steps

Working together on:

- Warm up exercises
- Calculating the run expectancy matrix

Lab 3

- Calculate the probability a team will win a game given the inning/outs and run differential

Lab 2: questions?

How did it go?

Astroball discussion

Let’s discuss the chapter for 7 minutes in breakout rooms and then have a larger conversation as a group

- Discuss your quote and reaction to chapter 2

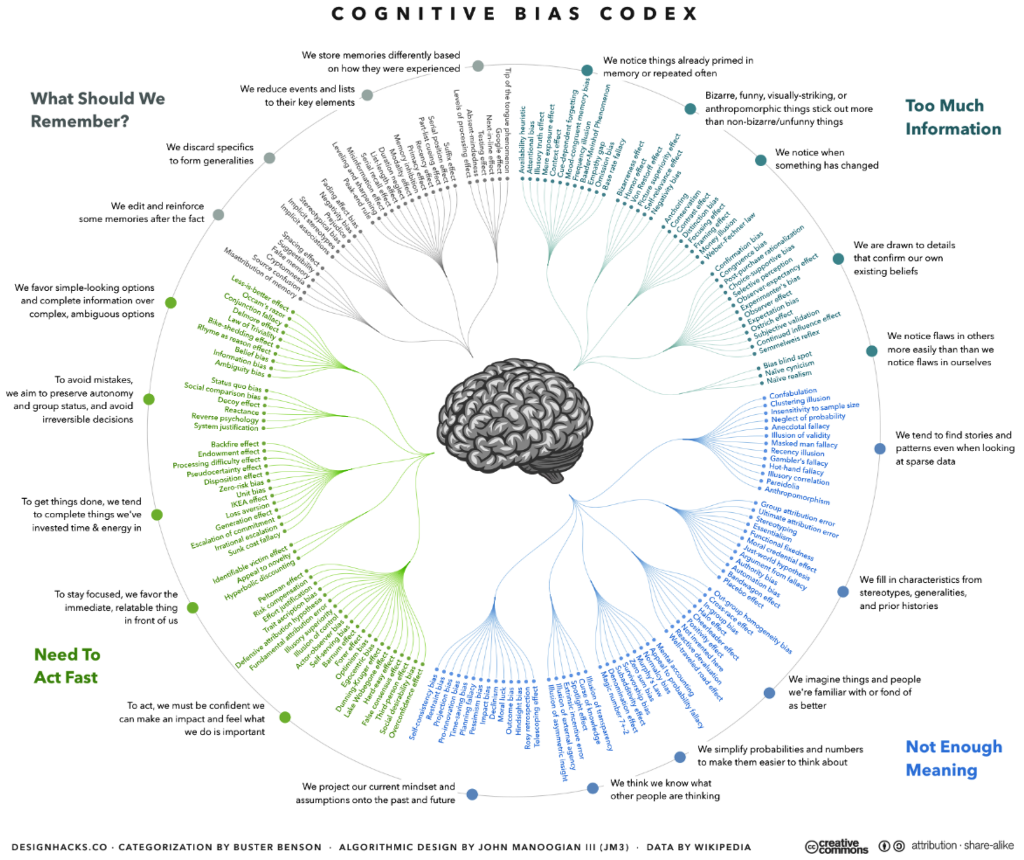
Thoughts on chapter 2 of Astroball?

Astroball: making the best decisions

Using linear regression to combining performance statistics and scouting reports.

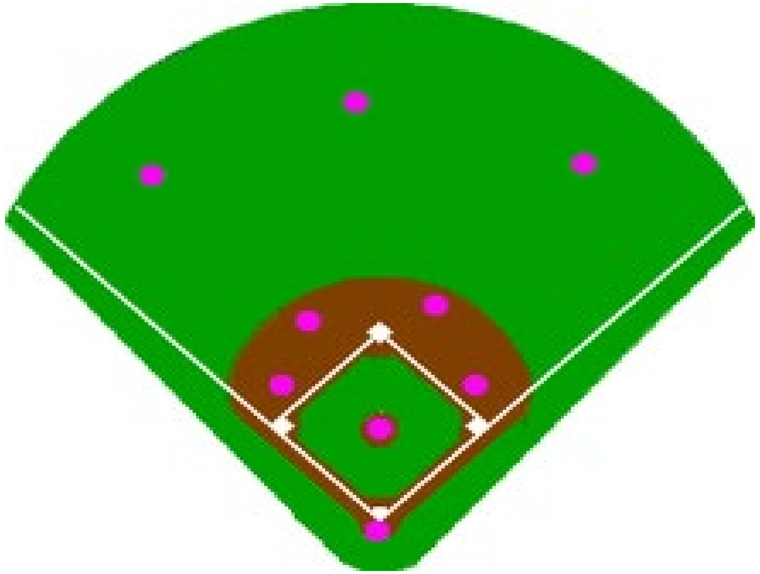
Kahneman and Tversky

- Behavioral economics vs. assuming rational agent.
- Cognitive biases

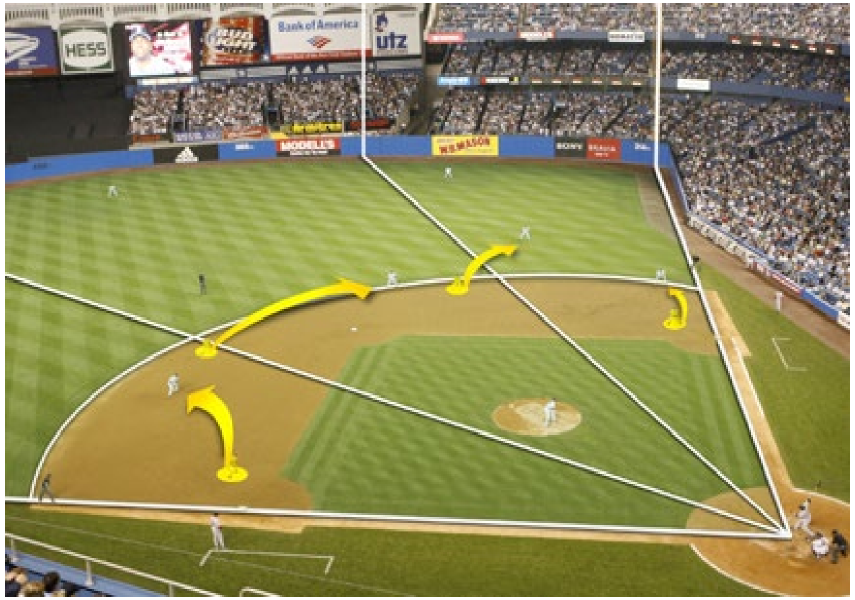

COGNITIVE BIAS CODEX
What Should We Remember?
Too Much Information
Need To Act Fast
Not Enough Meaning
We store memories differently based on how they were experienced
We reduce events and lists to their key elements
We discard specifics to form generalities
We edit and reinforce some memories after the fact
We notice things already primed in memory or repeated often
Bizarre, funny, visually-striking, or anthropomorphic things stick out more than non-bizarre/unfunny things
We notice when something has changed
We are drawn to details that confirm our own existing beliefs
We notice flaws in others more easily than than we notice flaws in ourselves
We favor simple-looking options and complete information over complex, ambiguous options
To avoid mistakes, we aim to preserve autonomy and group status, and avoid irreversible decisions
To get things done, we tend to complete things we've invested time & energy in
To stay focused, we favor the immediate, relatable thing in front of us
To act, we must be confident we can make an impact and feel what we do is important
We tend to find stories and patterns even when looking at sparse data
We fill in characteristics from stereotypes, generalities, and prior histories
We imagine things and people we're familiar with or fond of as better
We simplify probabilities and numbers to make them easier to think about
We think we know what other people are thinking
We project our current mindset and assumptions onto the past and future
DESIGNHACKS.CO · CATEGORIZATION BY BUSTER BENSON · ALGORITHMIC DESIGN BY JOHN MANOOGIAN III (JM3) · DATA BY WIKIPEDIA
creative commons attribution · share-alike

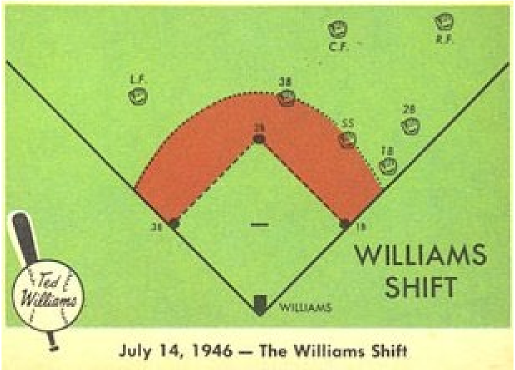
Astroball: infield shifts

Normal positioning

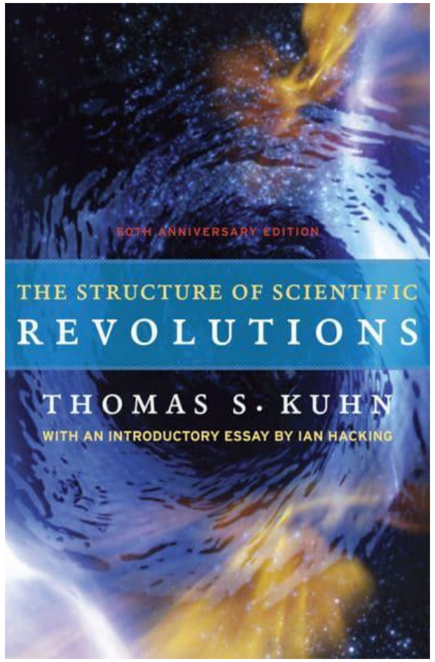


Defensive shift





July 14, 1946 — The Williams Shift


50TH ANNIVERSARY EDITION
THE STRUCTURE OF SCIENTIFIC
REVOLUTIONS
THOMAS S. KUHN
WITH AN INTRODUCTORY ESSAY BY IAN HACKING

Astroball: the butt slide

Deadspin from every angle

MLB analysis

Data manipulation part II

Let’s continue getting practice with data manipulation/wrangling by working in groups to do more advanced exercises.

- Quick review of key datascience method and data manipulation steps
- Warm up exercises: examining batting order and lefty/righty matchups
- Run expectancy matrix
- Lab 3: probability of winning a game given the current situation

Very quick review of datascience methods

Selecting and filtering

Description	datascience package	Pandas
Select a single column as a Table/Series	tb.select("col1")	df.col1 also df["col1"]
Select a single column as a ndarray	tb["col1"]	df["col1"].to_numpy()
Select multiple columns as a Table/DataFrame	tb.select("col1", "col2")	df[["col1", "col2"]]
Select 20th row	tb.take(20)	df.iloc[[20]]
Filtering rows equal to cond	tb.where("col", cond)	df[df.col == cond]
Sort descending by column	tb.sort("col", descending = True)	df.sort_values("col", ascending = False)

Data aggregation

Data aggregation/summarization consists of reducing a data table to a summary, often conditioned on a groping variable.

Q: Have we seen a data aggregation function yet?

A: tb.group('grouping_var', aggregation_function)

- tb.group('age') # counts how many players are a particular age
- tb.group('yearID', max) # get the max statistic value for each year

Pandas

- df.groupby('grouping_var’).aggfunc().reset_index()

If this isn’t used then the grouping column becomes the Index

Data aggregation

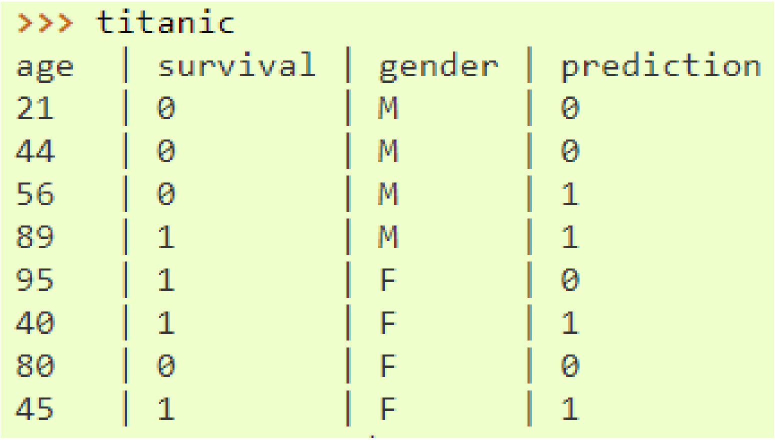
Other datascience aggregation functions:

tb.groups([‘col1’, ‘col2’], aggregation function)

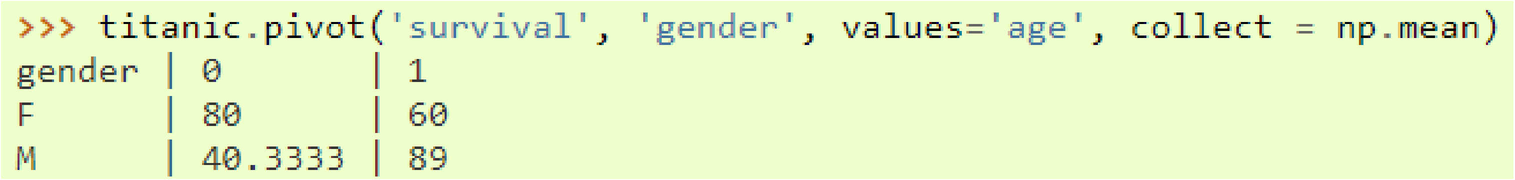
“long format”

tb.pivot('col_var', ‘row_var', values = 'fill_var' , collect = agg_function)

“wide format”



```
>>> titanic
age   | survival | gender | prediction
21    | 0        | M      | 0
44    | 0        | M      | 0
56    | 0        | M      | 1
89    | 1        | M      | 1
95    | 1        | F      | 0
40    | 1        | F      | 1
80    | 0        | F      | 0
45    | 1        | F      | 1
```

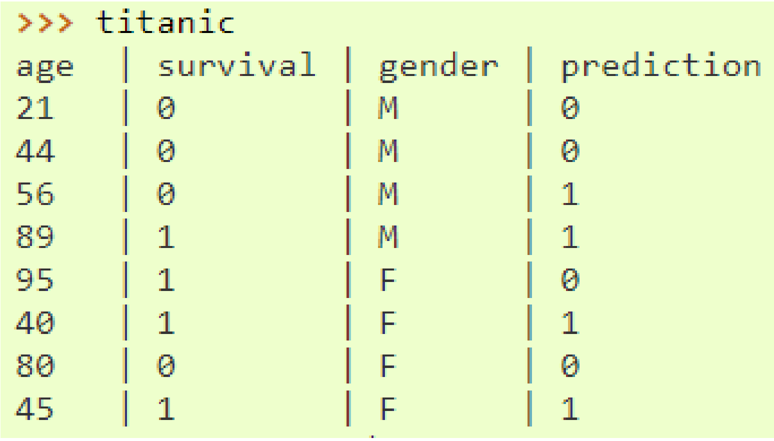


```
>>> titanic.pivot('survival', 'gender', values='age', collect = np.mean)
gender | 0       | 1
F      | 80      | 60
M      | 40.3333 | 89
```

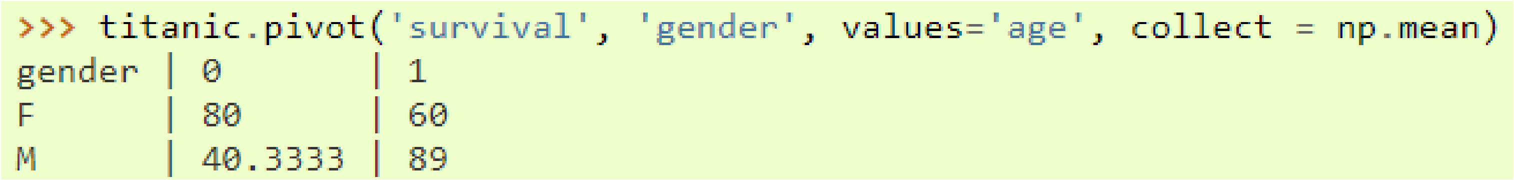
Data aggregation

Pandas

pd.pivot_table(input_table, values = 'fill_var', index = 'row_var',
columns = 'col_var', aggfunc = np.function)



```
>>> titanic
age   | survival | gender | prediction
21    | 0        | M      | 0
44    | 0        | M      | 0
56    | 0        | M      | 1
89    | 1        | M      | 1
95    | 1        | F      | 0
40    | 1        | F      | 1
80    | 0        | F      | 0
45    | 1        | F      | 1
```



```
>>> titanic.pivot('survival', 'gender', values='age', collect = np.mean)
gender | 0       | 1
F      | 80      | 60
M      | 40.3333 | 89
```

Numpy operations

Combining ndarrays

- combo_ndarray = ndarray1.append(ndarray2)

Casting: converts from one type to another type

- ndarray.astype('int')   # could be useful to convert Booleans to numbers
- ndarray.astype('str')   # could be useful to convert numbers to characters

Combining strings in ndarrays

- np.char.add(strings_ndarray, '---',)
- np.char.add(strings_ndarray1, strings_ndarray2)

Data processing steps

Data processing steps

1. Data cleaning: getting data into a table-like format (tidy)

- 80% of Data Scientists time is spent cleaning data
- 20% of Data Scientists time is spent complaining about cleaning data

2. Data manipulation: get columns you need for subsequent analyses

- Creating new columns from existing columns
- Joining data tables
- Etc.

3. Data aggregation: extract key summary statistics from data

4. Visualize: create clear figures to reveal insights

Thinking through data manipulation steps

2. Data manipulation: get columns you need for subsequent analyses

3. Data aggregation: extract key summary statistics from data

To think through data manipulation steps it is useful to:

1. Think about final aggregated result one wants

2. Think about the data table (i.e., columns needed) that could be aggregated to get the final result
   - Second to last step

3. Think about how to get data table in step 2 from starting data table

Data manipulation exercises

Calculating home run rates a function of batting order

It is widely believed that the best home run hitters usually bat third or fourth in the line up.

Let's check to see if this is the case!

Let’s use the play-by-play retro data from 2019

Calculating home run rates a function of batting order

Let’s start with the retrosheet play-by-play data and go through these steps:

1. Think about final aggregated result one wants, what would it look like?

2. Think about the data table (i.e., columns needed) that could be aggregated to get the final result
    - Second to last step

3. Think about how to get data table in step 2 from starting data table

Calculating home run rates a function of batting order

Let’s start with the retrosheet play-by-play data

1. Think about final aggregated result one wants, what would it look like?
   - i.e., what are the variables in the final aggregated table?
   - Sketch it

Batting order	HR frequency
1	
2	
3	
4	
5	
6	
7	
8	
9	

These values filled in

Calculating home run rates a function of batting order

Let’s start with the retrosheet play-by-play data

2. Think about the data table (i.e., columns needed) that could be aggregated to get the final result
    - Second to last step

What variables are needed?

Batting order	HR frequency
1	
2	
3	
4	
5	
6	
7	
8	
9	

Calculating home run rates a function of batting order

Let's start with the retrosheet play-by-play data

3. Think about how to get data table in step 2 from starting data table

Batting order	HR frequency
1	
2	
3	
4	
5	
6	
7	
8	
9	

Calculating home run rates a function of batting order

Let's go into breakout rooms and solve these together…

Get in class_04 Jupyter notebook

Pandas people together:

- Gabby, Harry, Teddy,   anyone else?

Righties vs. lefties

1. Do righthanded batters hit more home runs than lefthanded batters?

2. Do right/lefthanded batters hit more home runs off right/lefthanded pitchers?

- A platoon is a method of sharing playing time, where two players are selected to play a single defensive position. Usually, one platoon player is right-handed and the other is left-handed.

Los Angeles Daily News

SIGN UP FOR NEWSLETTERS
E-EDITION
SUBSCRIBE + SUBSCRIBER SERVICES

TRENDING: LAUSD Daily Pass COVID app | Fans at Dodger Stadium in 2021? | Kim Kardashian files for divorce | How to grow tomatoes

SPORTS › MLB › LOS ANGELES DODGERS

LA Dodgers' use of platooning backed by data, players

Run expectancy matrix

The **run expectancy matrix** contains the expected number of runs given the number of outs, and base runners

- E[P(R|base runners, outs)]
- History:
  - First described in 1963 by George Lindsey in an article in Operations Research
  - Discussed extensively in The Book by Tango, Lichtman and Dolphin

Q: How many out-runner game states are there in half an inning?

- A: 24

Empty run expectancy matrix

Bases	**0 Outs**	**1 Out**	**2 Outs**
_ _ _			
_ _ 3			
_ 2 _			
_ 2 3			
1 _ _			
1 _ 3			
1 2 _			
1 2 3			

Run expectancy matrix

The **run expectancy matrix** contains the expected number of runs given the number of outs, and base runners

Let's try to fill in the run expectancy matrix with your own best guesses

- Go to: bit.ly/runexpect
- Copy over the template to your own sheet
- Fill in with your best guess

Empty run expectancy matrix

Bases	**0 Outs**	**1 Out**	**2 Outs**
_ _ _			
_ _ 3			
_ 2 _			
_ 2 3			
1 _ _			
1 _ 3			
1 2 _			
1 2 3			

How can we calculate the run expectancy matrix?

2. Data manipulation: get columns you need for subsequent analyses

3. Data aggregation: extract key summary statistics from data

To think through data manipulation steps it is useful to:

1. Think about final aggregated result one wants

2. Think about the data table (i.e., columns needed) that could be aggregated to get the final result
    - Second to last step

3. Think about how to get data table in step 2 from starting data table

Do part 2.1.1, 2.1.2, 2.1.3 in a breakout room and then let's discuss these steps as a class

Bases	**0 Outs**	**1 Out**	**2 Outs**
_ _ _			
_ _ 3			
_ 2 _			
_ 2 3			
1 _ _			
1 _ 3			
1 2 _			
1 2 3			

How can we calculate the run expectancy matrix?

Question 2.1.1 - Final goal: What are the...?

1. **rows: runners on base (8 states)**
2. **columns: outs (3 states)**
3. **fill values:** expected number of runs scored by the end of the inning given the current game state.

Bases	0 Outs	1 Out	2 Outs
_ _ _			
_ _ 3			
_ 2 _			
_ 2 3			
1 _ _			
1 _ 3			
1 2 _			
1 2 3			

How can we calculate the run expectancy matrix?

Question 2.1.2 - The aggregating methods are:

1. **pivot**: Generate a table with a column for each unique value in columns, with rows for each unique value in rows.
2. **group**: Group rows by unique values in a column; count or aggregate others.
3. **groups**: Group rows by multiple columns, count or aggregate others.

How can we calculate the run expectancy matrix?

Question 2.1.3 - The columns that are needed in the 2nd to last step are:

1. The state of the base runners
   - e.g., bases empty, runner on first, etc.
2. The number of outs
3. The number of runs scored that remain to be scored in the inning (i.e., the final score in the inning minus the current score).

How would you aggregate these results to get the final run expectancy matrix?

BASE_RUNNER_STATE	**OUTS_CT**	**RUN_ROI**
000	0	0
001	0	0
001	1	0
011	1	0
000	1	0
…	…	…

Let's work in groups to try to calculate the run expectancy matrix

Lab 3

Calculate the probability of winning a game given the inning and score difference

Might find it challenging so start early!!!

- Use Ed discussions for questions and answer other's questions