Lazily Looking at pandas DataFrames

Eric_A_Moore@yahoo.com

8/21/18

“The Problem”

- See what’s in a DataFrame, toward identifying interesting relationships among variables
  - Concisely
  - Clearly
  - Don’t overwhelm
  - Don’t lie
- Code in order to learn: pandas, Altair, Vega
- Initial application: how much do athlete's bodies predict their position? (baseball, football, Olympics)

Key Ideas

Phases of analyses, series of views

Simple schema – permute over

Simple invocation

Embody visual judgment

Two modes:

- **Explore**: Design for screen first (over paper) – sequence, scroll, page, gesture
- **Explain**: Legible static image, highlight the “tell”

Phases of Analyses

1. ***Data***: Quality, Quantity, Qualify.
   - Formats; gaps, causes for gaps; need for transformation.
2. ***Variables***: what is potentially informative – similar, different, distinct?
   - See Variables – distribution of values
   - See Variable – Variable relationships
3. ***Model***: what is predictive?; what are useful abstractions?
   - See relationships in context of a Target Variable
   - Measures on Measures – toward statistical inference
   - “Fit” or abstract variable distributions and relationships
   - “Focus” a predictive model to most essential form

Simple Schema

Statistics

- Compare effects of a process across subpopulations
- Effects are measured, estimated given samples of a population
- *Category* – divides population
- *Measure* -- describes population

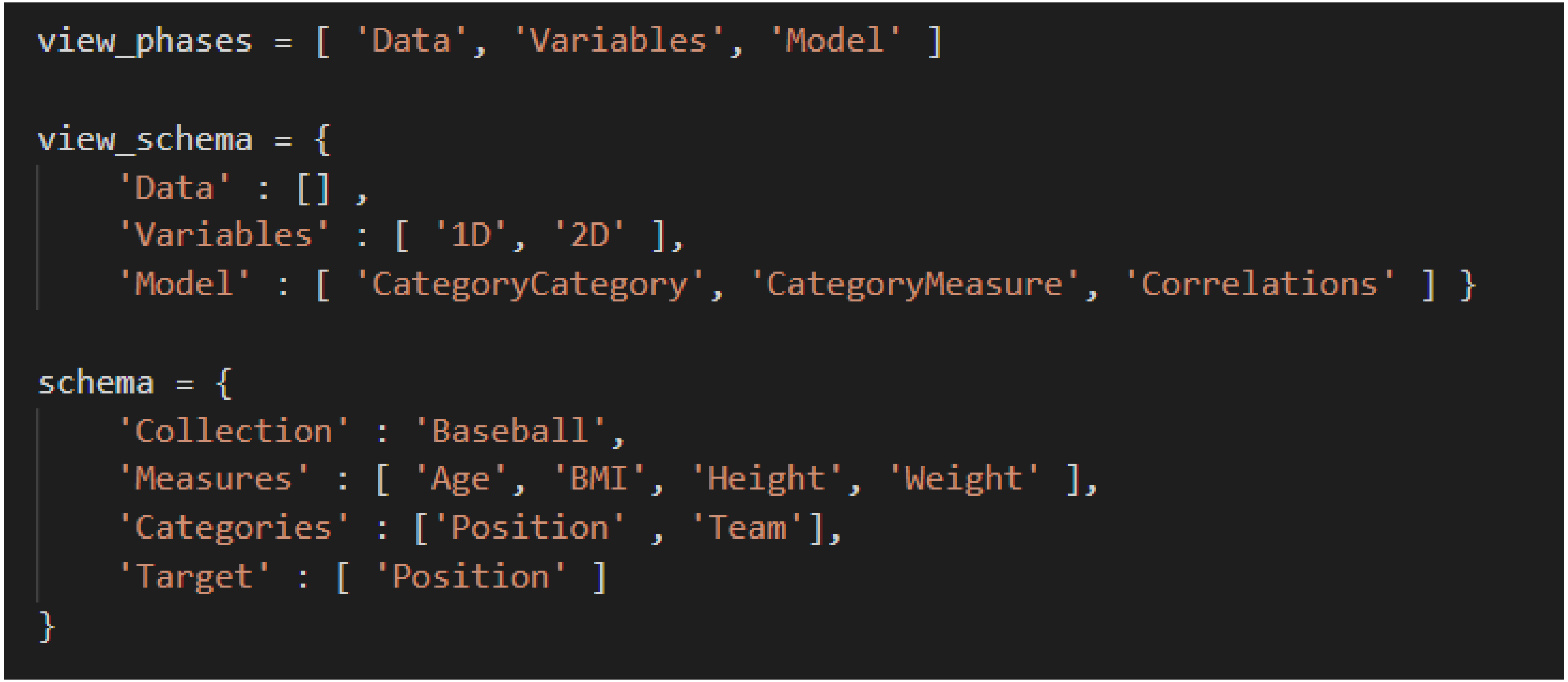
Data

- Symbolic -- *Categorical, Nominal, Ordinal, Key*
- Numeric -- *Quantitative, Temporal*
- *Set* - collection of records ~ Sample ~ Population
- Field ~ Variable ~ Feature

Embody visual judgment – describe, compare

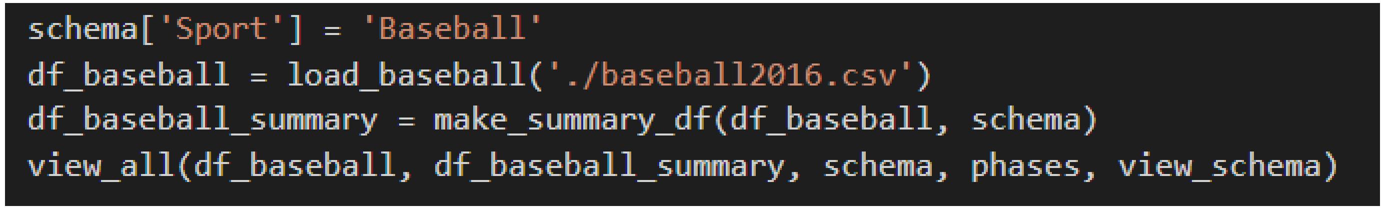
Subject	Chart	Comments
Categorical 1D	Horizontal Histogram	Group by category.
Scalar 1D	Vertical Histogram	Binned X. Fit later w/ parametric Distribution
Category vs Category	Crosstab Table	Counts
Measure vs Measure	Scatter Plot	Fit later w/ Loess or regression.
Measure over Category	Hozo Bar Chart	Measure, conditioned on each category.
Measures on Measure over Category	Hozo Box Chart	Using mean, std deviation vs median and quartiles. “Separation” in 1D, with one measure.
Category over Measures	Scatter Plot with Category	Separation in 2D, with two joint measures. Only compare 2 or 3 categories at once.
Category over Category differences	Crosstab Table with residual density	Highlight non uniformity in joint distribution
Set of Correlations	Hozo Bar Chart (Correlations)	Highlight strong or weak correlations

Simple Schema

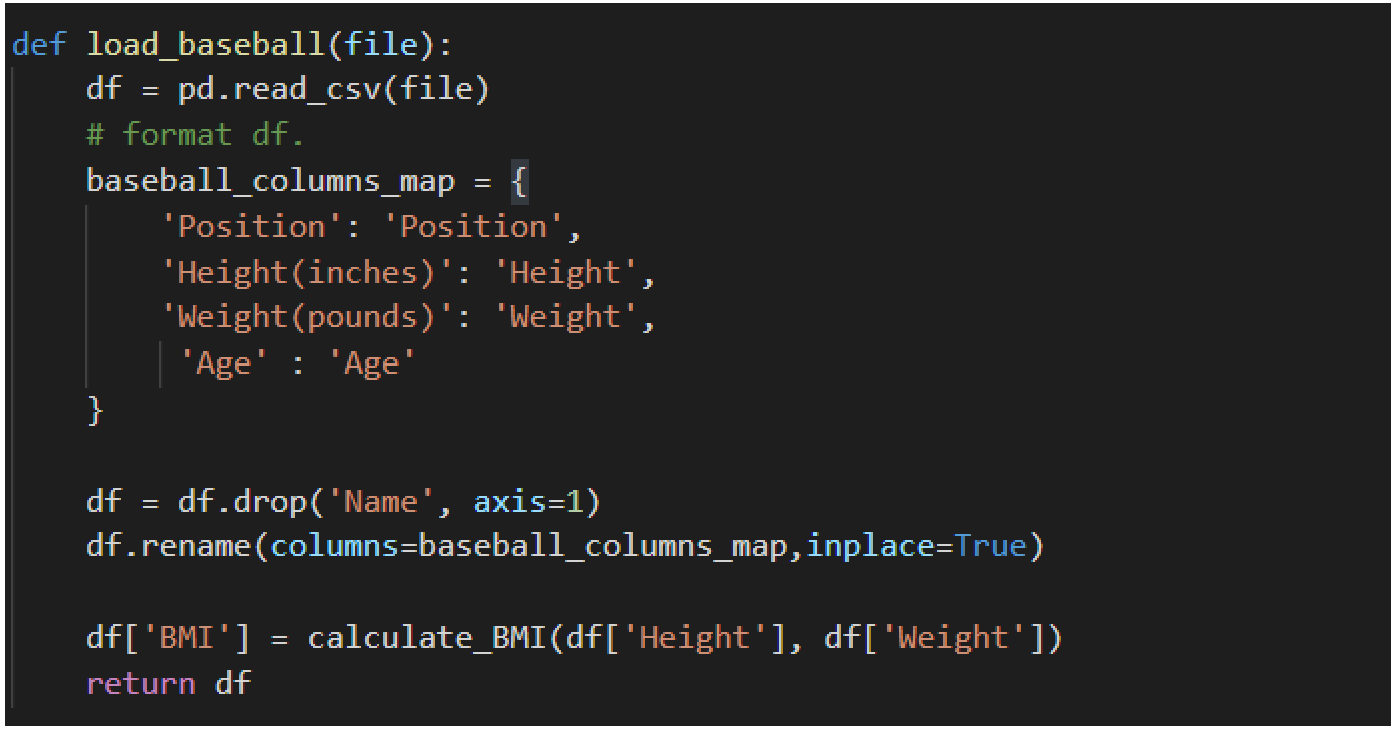


```
view_phases = [ 'Data', 'Variables', 'Model' ]

view_schema = {
    'Data' : [] ,
    'Variables' : [ '1D', '2D' ],
    'Model' : [ 'CategoryCategory', 'CategoryMeasure', 'Correlations' ] }

schema = {
    'Collection' : 'Baseball',
    'Measures' : [ 'Age', 'BMI', 'Height', 'Weight' ],
    'Categories' : ['Position' , 'Team'],
    'Target' : [ 'Position' ]
}
```

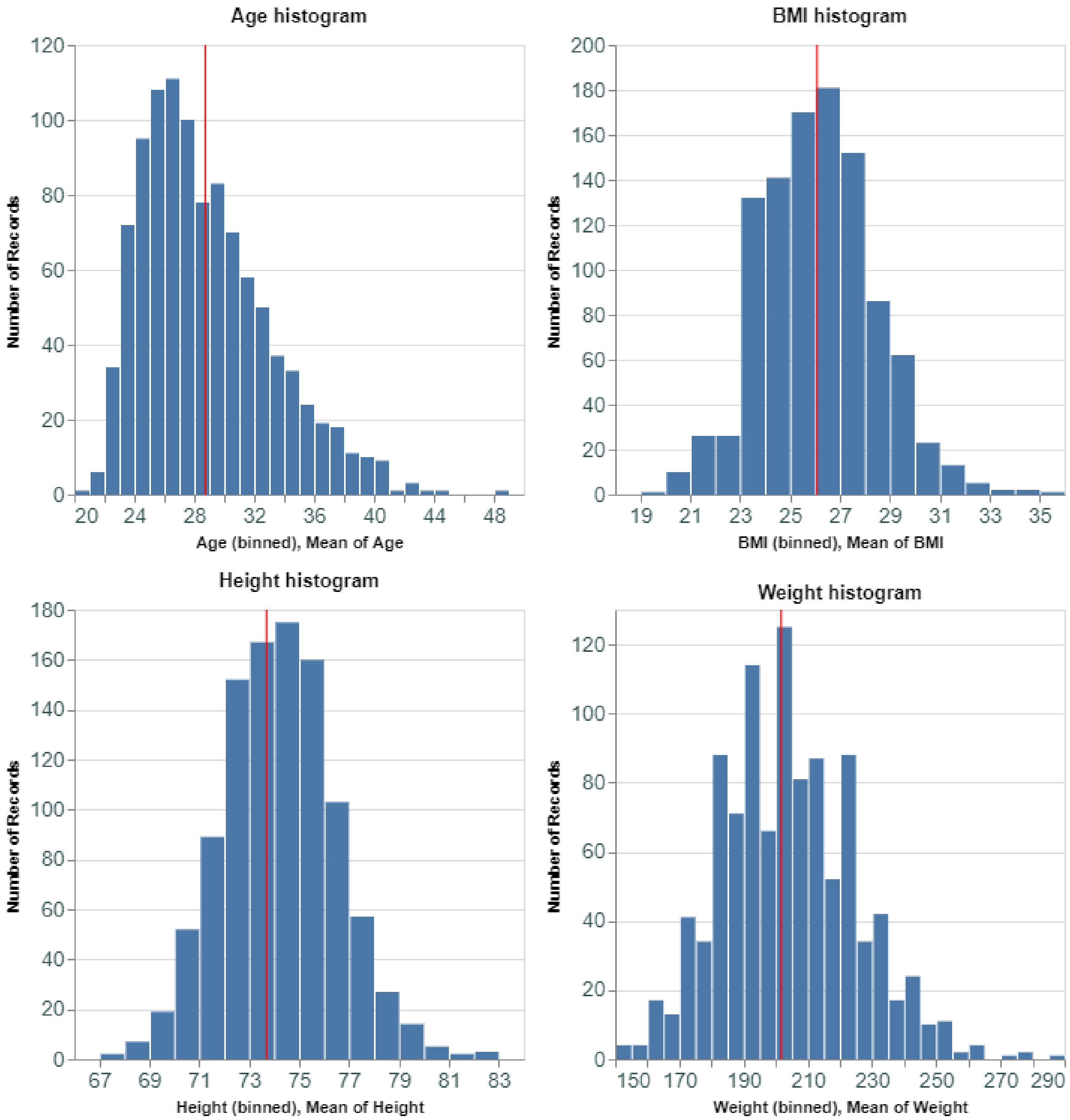
Simple invocation

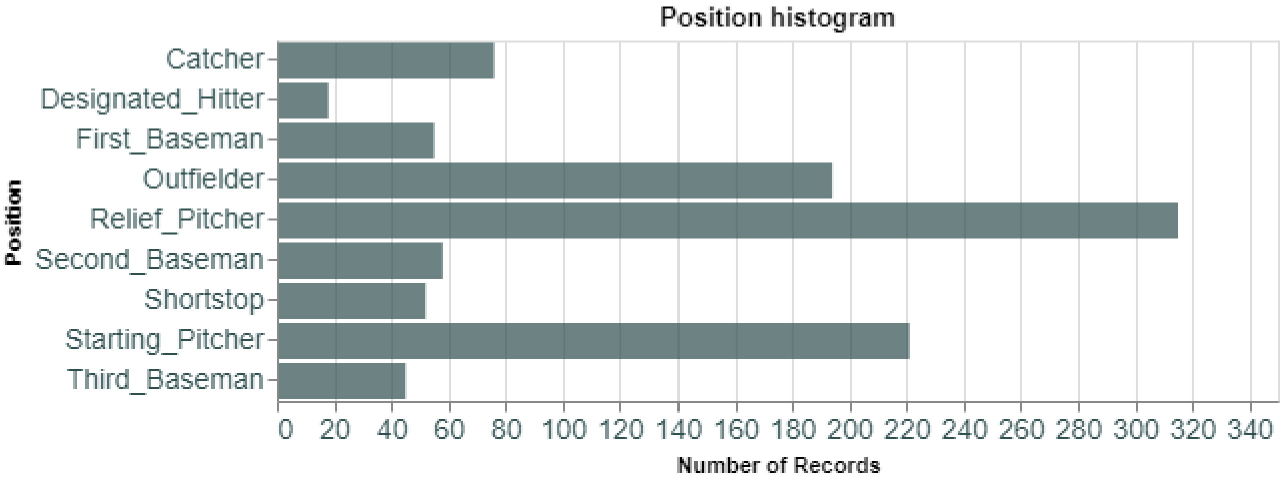


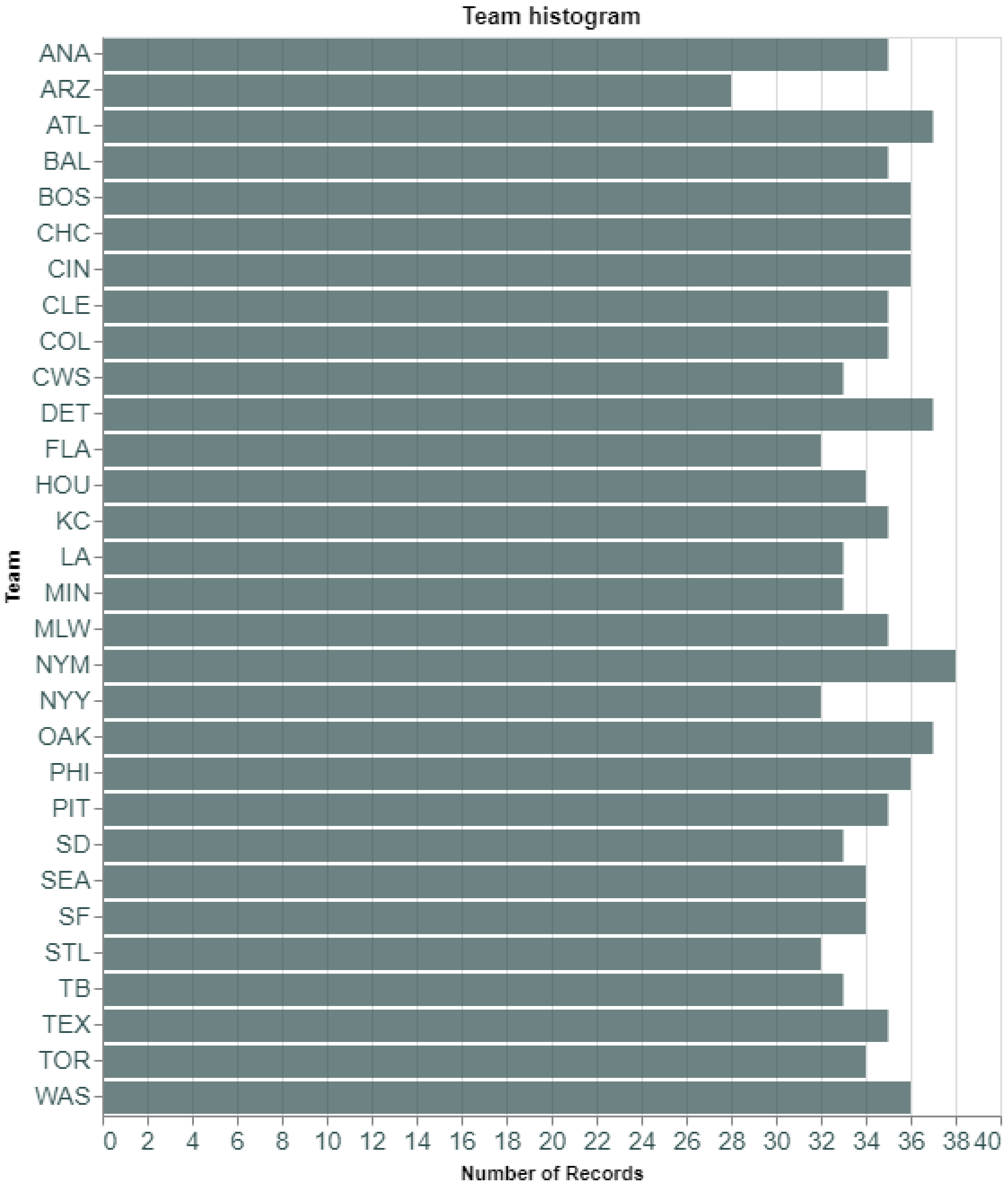
```python
schema['Sport'] = 'Baseball'
df_baseball = load_baseball('./baseball2016.csv')
df_baseball_summary = make_summary_df(df_baseball, schema)
view_all(df_baseball, df_baseball_summary, schema, phases, view_schema)
```



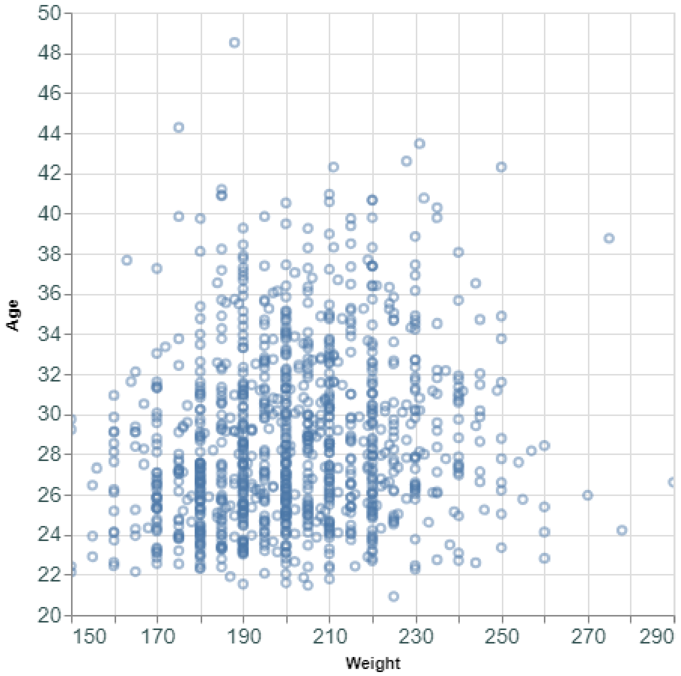
```python
def load_baseball(file):
    df = pd.read_csv(file)
    # format df.
    baseball_columns_map = {
        'Position': 'Position',
        'Height(inches)': 'Height',
        'Weight(pounds)': 'Weight',
         'Age' : 'Age'
    }

    df = df.drop('Name', axis=1)
    df.rename(columns=baseball_columns_map,inplace=True)

    df['BMI'] = calculate_BMI(df['Height'], df['Weight'])
    return df
```


Age histogram
Number of Records
120
100
80
60
40
20
0
20
24
28
32
36
40
44
48
Age (binned), Mean of Age
BMI histogram
Number of Records
200
180
160
140
120
100
80
60
40
20
0
19
21
23
25
27
29
31
33
35
BMI (binned), Mean of BMI
Height histogram
Number of Records
180
160
140
120
100
80
60
40
20
0
67
69
71
73
75
77
79
81
83
Height (binned), Mean of Height
Weight histogram
Number of Records
120
100
80
60
40
20
0
150
170
190
210
230
250
270
290
Weight (binned), Mean of Weight

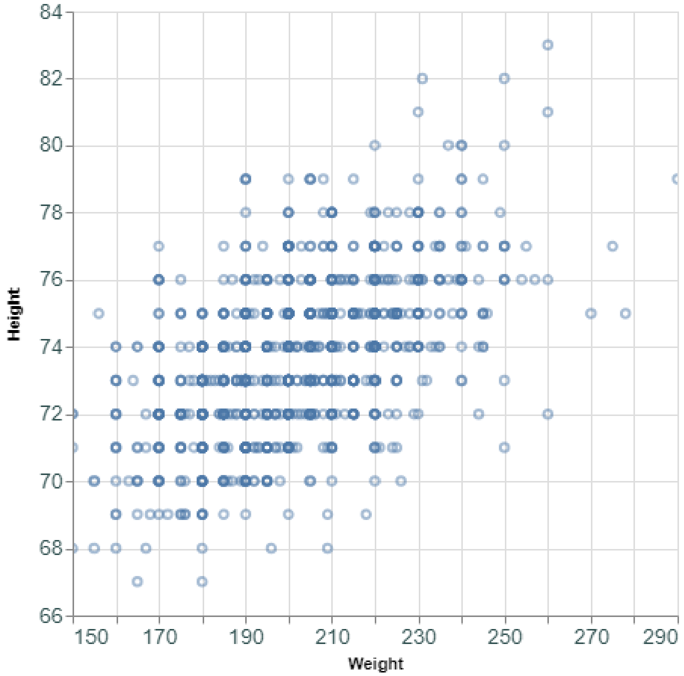

Position histogram
Position
Catcher
Designated_Hitter
First_Baseman
Outfielder
Relief_Pitcher
Second_Baseman
Shortstop
Starting_Pitcher
Third_Baseman
0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340
Number of Records


Team histogram
Team
ANA
ARZ
ATL
BAL
BOS
CHC
CIN
CLE
COL
CWS
DET
FLA
HOU
KC
LA
MIN
MLW
NYM
NYY
OAK
PHI
PIT
SD
SEA
SF
STL
TB
TEX
TOR
WAS
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
Number of Records

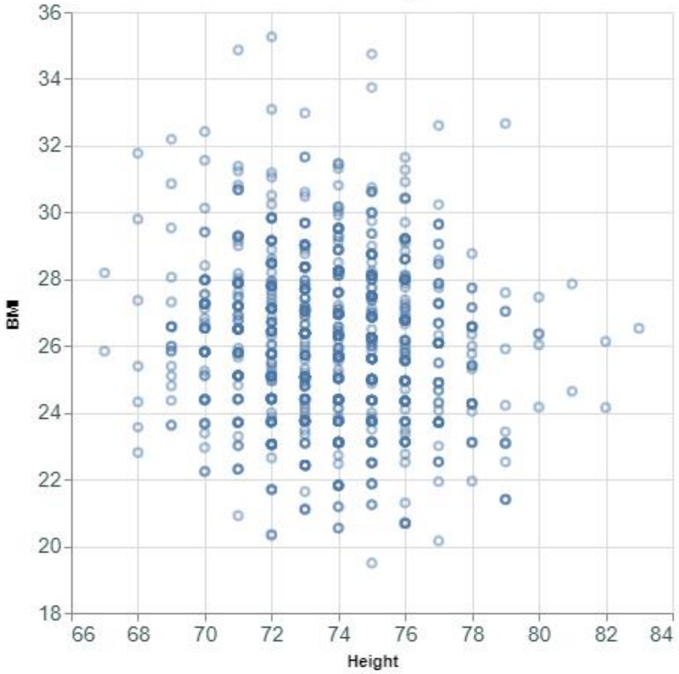
Age vs Weight



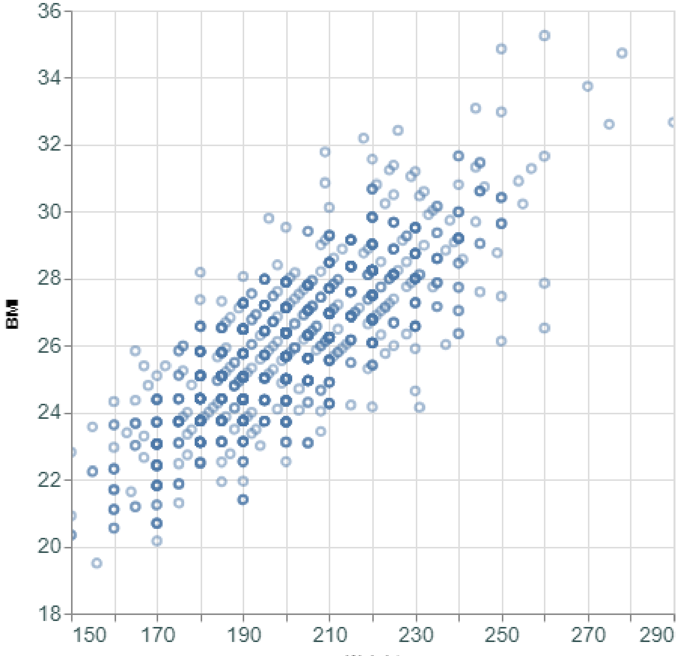
Height vs Weight



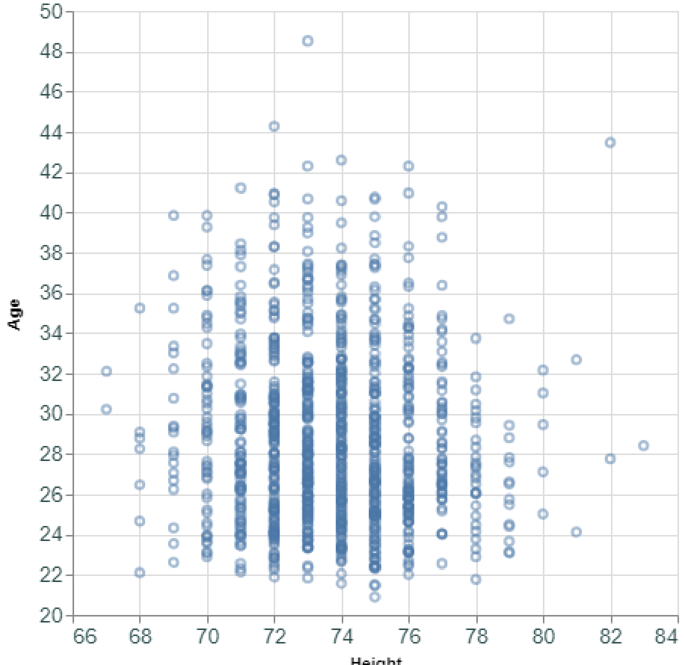
BMI vs Height



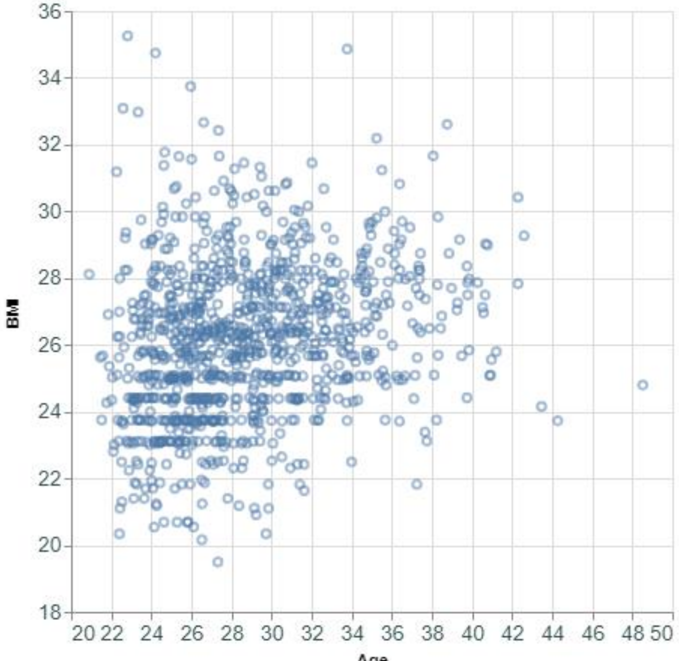
BMI vs Weight

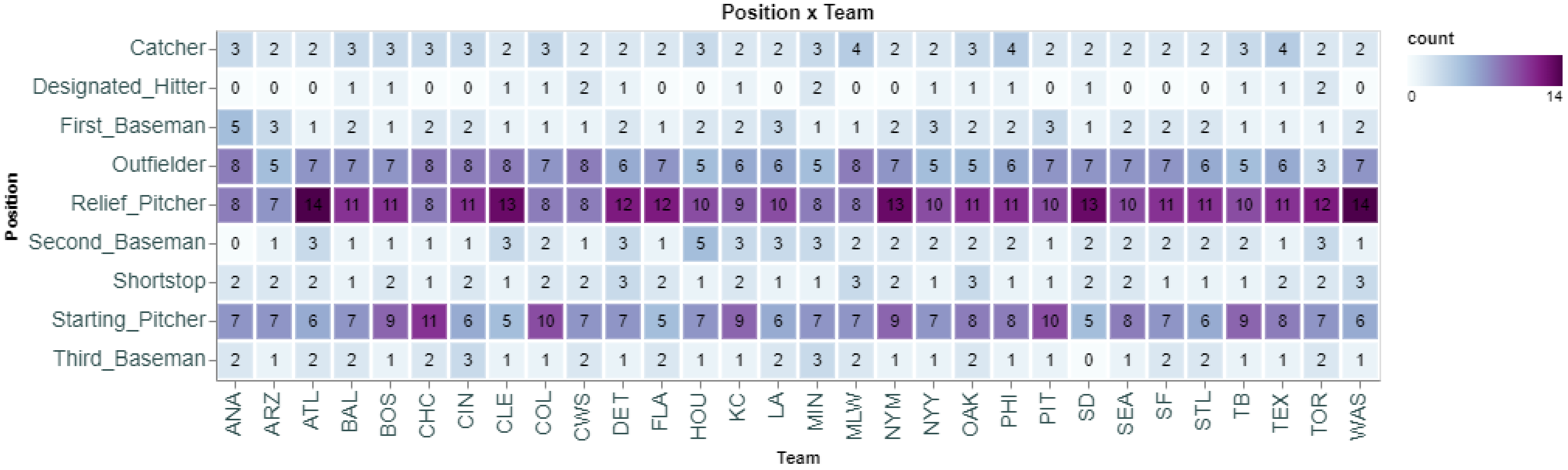


Age vs Height



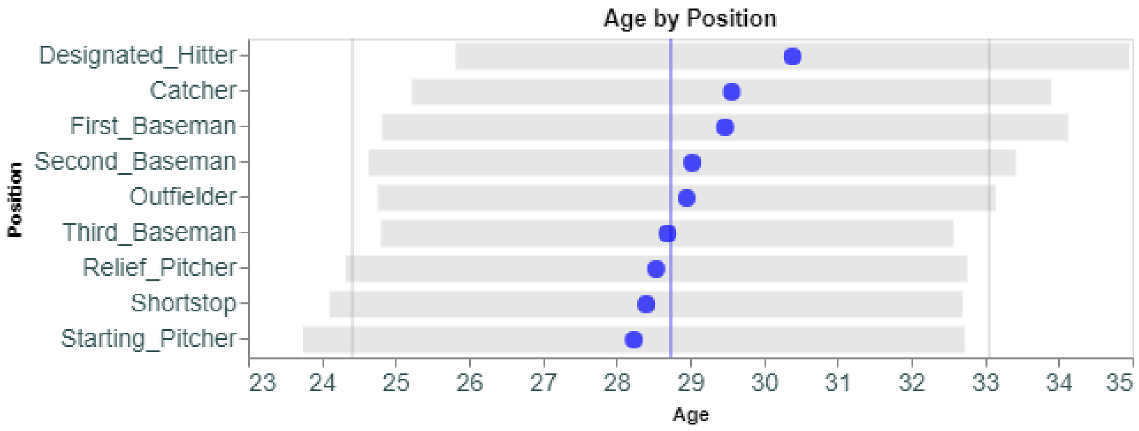
BMI vs Age

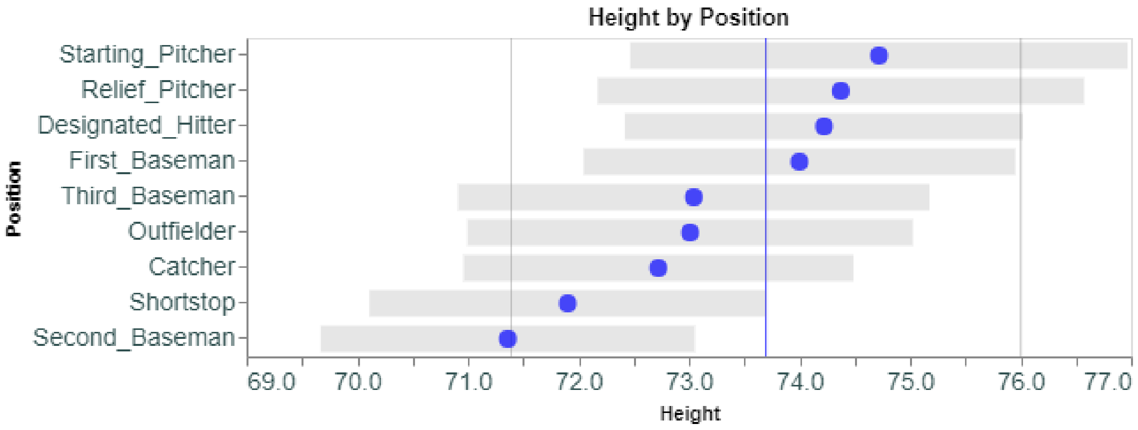


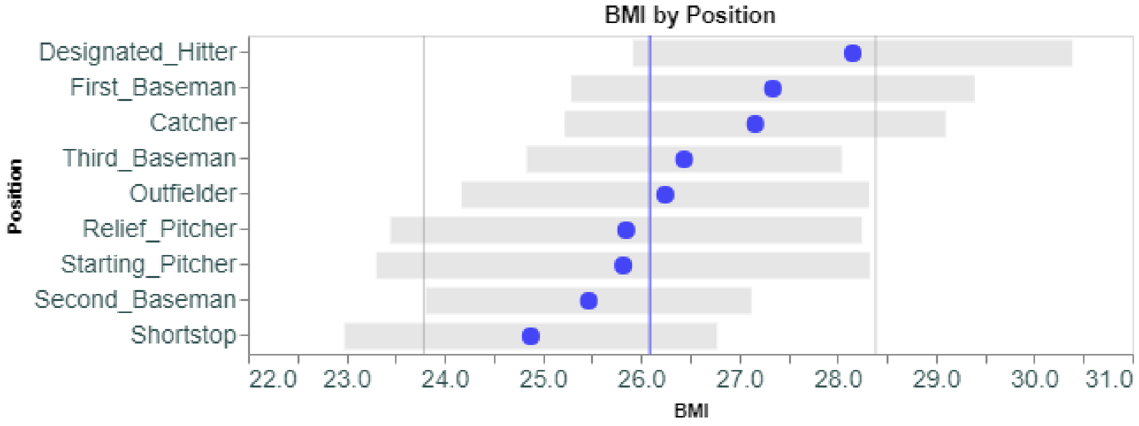


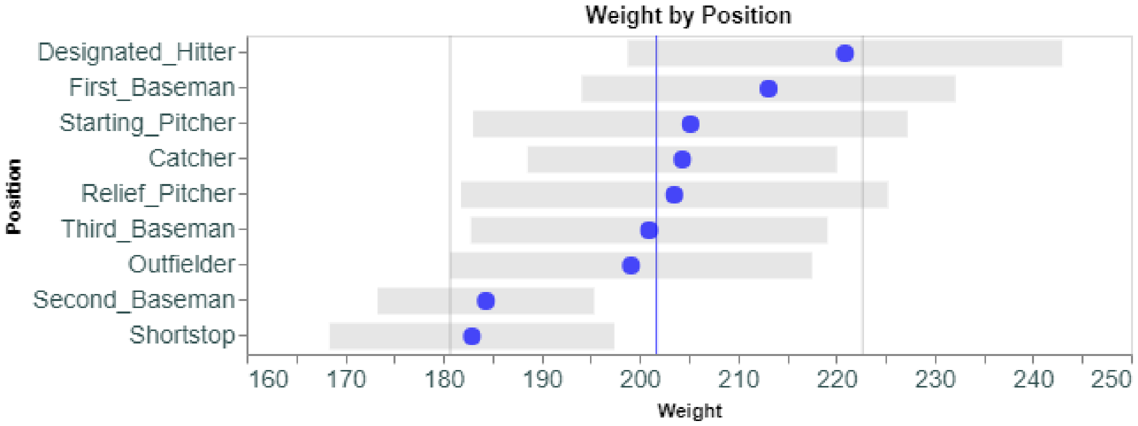
	ANA	ARZ	ATL	BAL	BOS	CHC	CIN	CLE	COL	CWS	DET	FLA	HOU	KC	LA	MIN	MLW	NYM	NYY	OAK	PHI	PIT	SD	SEA	SF	STL	TB	TEX	TOR	WAS
Catcher	3	2	2	3	3	3	3	2	3	2	2	2	3	2	2	3	4	2	2	3	4	2	2	2	2	2	3	4	2	2
Designated_Hitter	0	0	0	1	1	0	0	1	1	2	1	0	0	1	0	2	0	0	1	1	1	0	1	0	0	0	1	1	2	0
First_Baseman	5	3	1	2	1	2	2	1	1	1	2	1	2	2	3	1	1	2	3	2	2	3	1	2	2	2	1	1	1	2
Outfielder	8	5	7	7	7	8	8	8	7	8	6	7	5	6	6	5	8	7	5	5	6	7	7	7	7	6	5	6	3	7
Relief_Pitcher	8	7	14	11	11	8	11	13	8	8	12	12	10	9	10	8	8	13	10	11	11	10	13	10	11	11	10	11	12	14
Second_Baseman	0	1	3	1	1	1	1	3	2	1	3	1	5	3	3	3	2	2	2	2	2	1	2	2	2	2	2	1	3	1
Shortstop	2	2	2	1	2	1	2	1	2	2	3	2	1	2	1	1	3	2	1	3	1	1	2	2	1	1	1	2	2	3
Starting_Pitcher	7	7	6	7	9	11	6	5	10	7	7	5	7	9	6	7	7	9	7	8	8	10	5	8	7	6	9	8	7	6
Third_Baseman	2	1	2	2	1	2	3	1	1	2	1	2	1	1	2	3	2	1	1	2	1	1	0	1	2	2	1	1	2	1

Baseball…

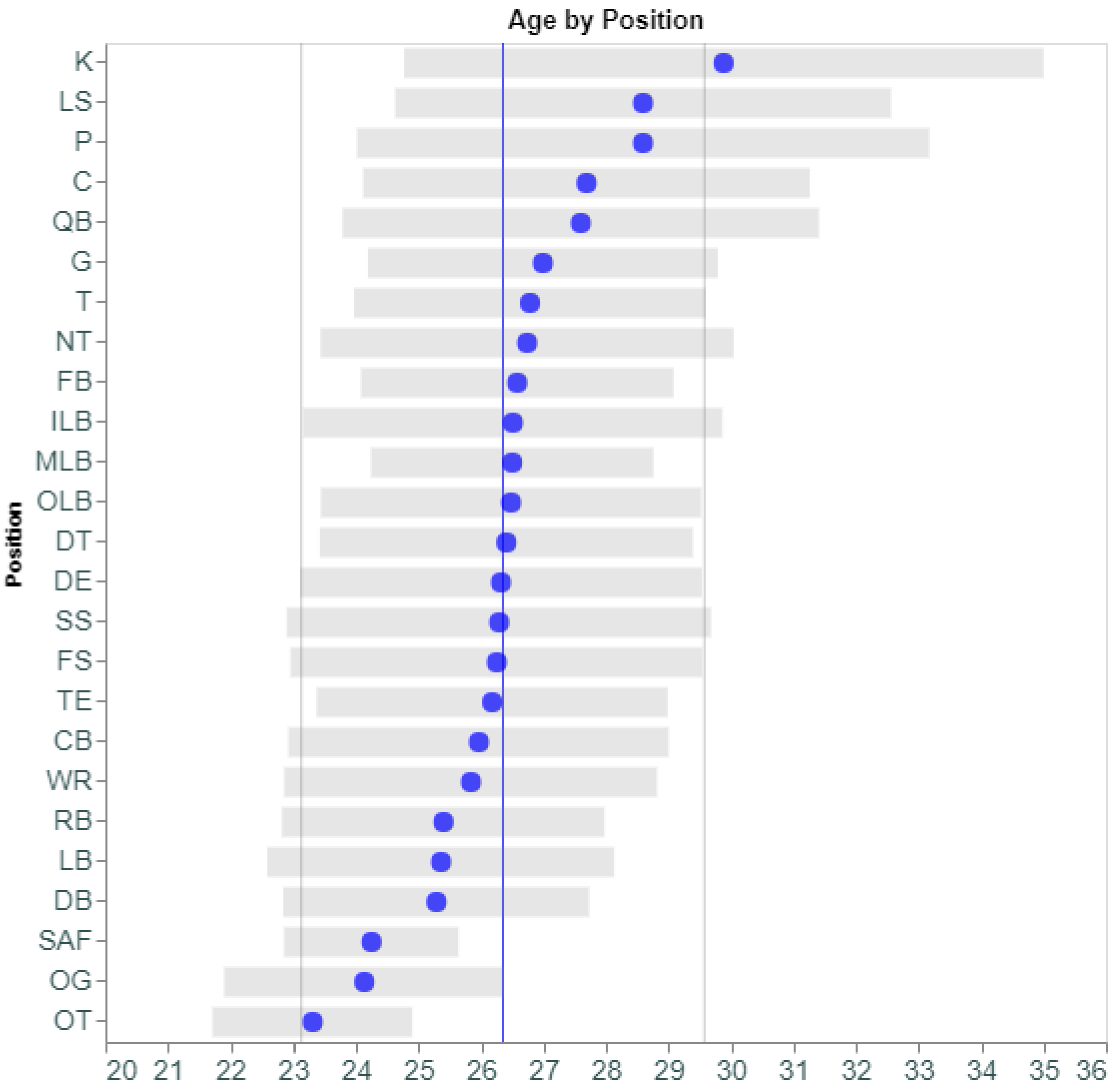

Age by Position

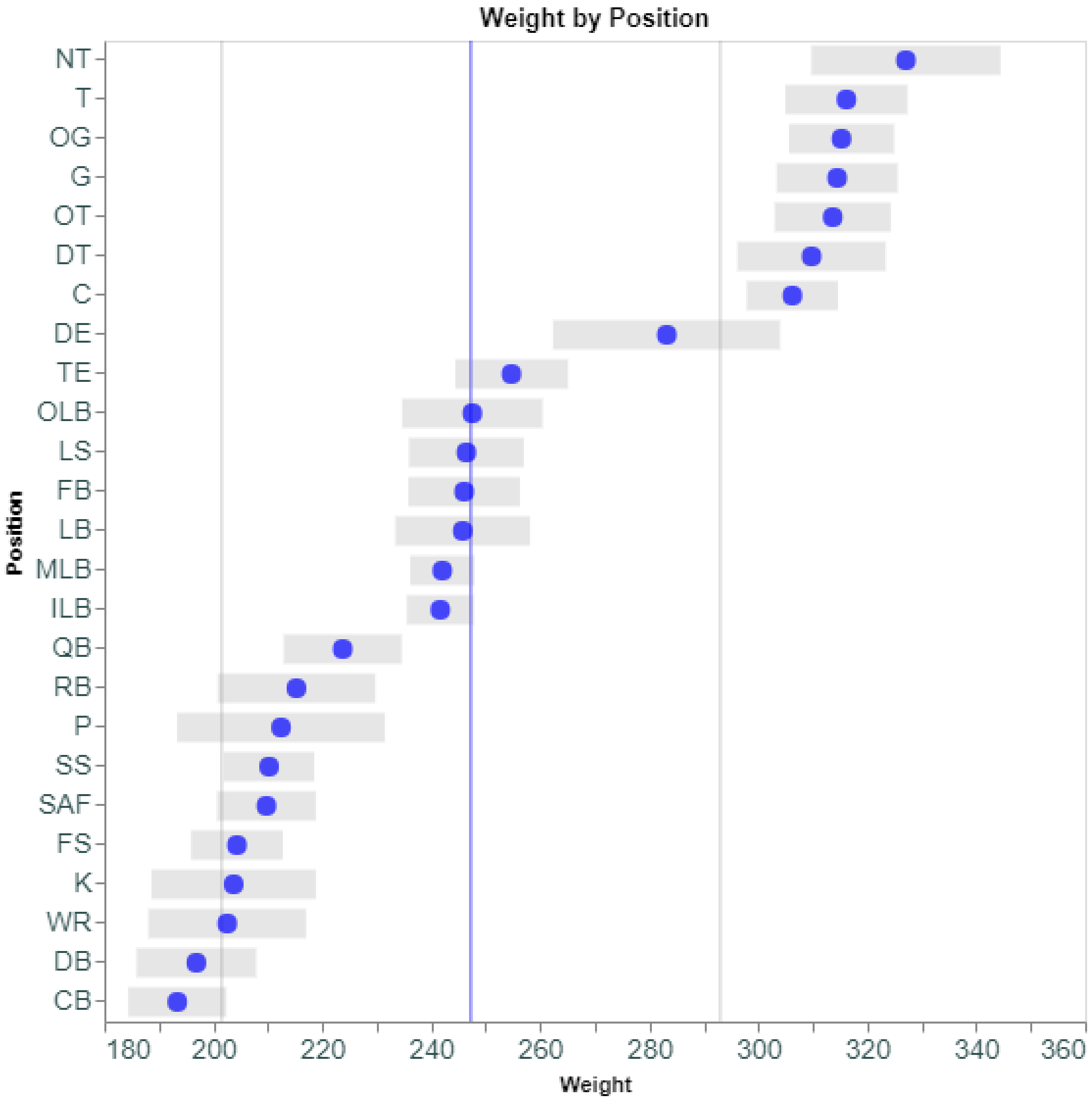

Height by Position


BMI by Position

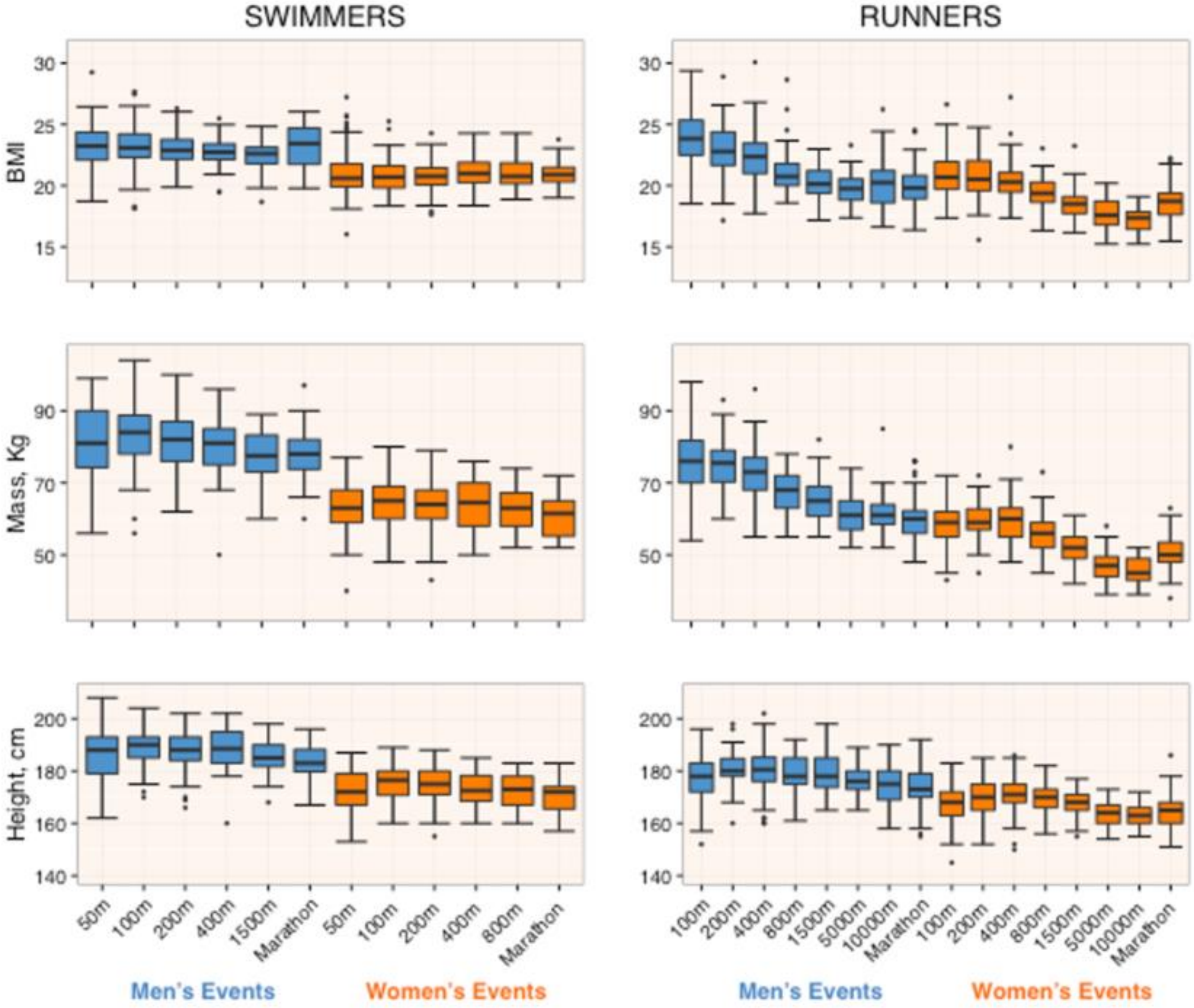

Weight by Position

Football …

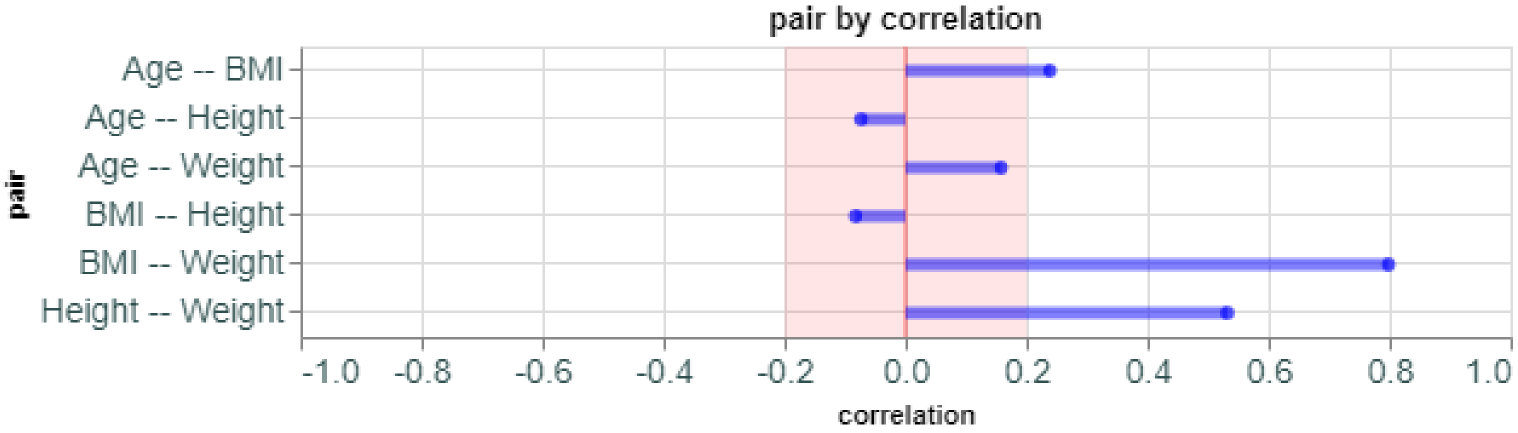

Age by Position
K
LS
P
C
QB
G
T
NT
FB
ILB
MLB
OLB
DT
DE
SS
FS
TE
CB
WR
RB
LB
DB
SAF
OG
OT
Position
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36

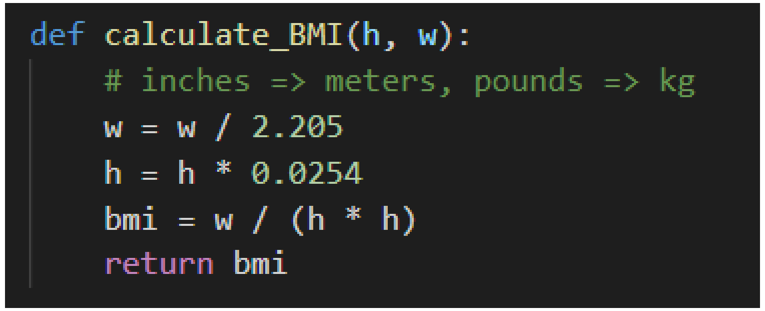

Weight by Position
NT
T
OG
G
OT
DT
C
DE
TE
OLB
LS
FB
LB
MLB
ILB
QB
RB
P
SS
SAF
FS
K
WR
DB
CB
Position
180
200
220
240
260
280
300
320
340
360
Weight

Swimmers vs Runners, Gagnon, et al (2018)



Correlations … not enough by themselves





```
def calculate_BMI(h, w):
    # inches => meters, pounds => kg
    w = w / 2.205
    h = h * 0.0254
    bmi = w / (h * h)
    return bmi
```

Charts in Altair

(DataFrame + field) + (axis + encoding + mark) = chart

- Tidy data preferred
- Composition of chart by layers
- Peered charts – cross select
- Data transformation – binning, sort, grouping
- Stack of Altair, Vega Lite, Vega. Backed by visionaries … UW IDL.

Challenges (for me)

- Composition of scenes (working outside Notebooks)
- Navigation events handled outside the chart (how to "click through") **
- Resolution of conflicting attributes in scope hierarchy – axes, domains
- Should know how Vega Lite, Vega work

References

altair-viz.github.io

github.com/jakevdp/altair-examples

vallandingham.me/altair_intro.html

U Washington Interactive Data Lab https://idl.cs.washington.edu/

Making Data Visual A Practical Guide to Using Visualization for Insight

By Miriah Meyer, Danyel Fisher

How Body Type May Determine Runners’ and Swimmers’ Destinies – NYTimes 08-14-18

Development Observations

- Libraries … as collective works, and as works in progress
  - Many Channels: docs, git issues (open & closed), project google group/slack, project contributor blogs, user blogs, contributor presentations
  - Different ways to do the same thing … bushy interfaces
- Library as a lever, vs utility
  - Great to copy from Examples gallery
  - “Why did that break?” Value in knowing what it really does, and how.
- VS Code
  - Code folding, Vim plugin support, a lot of “push”.

DataViz design issues

What am I looking at?

→ clearly display clear labels for fields

Is this “a thing”? (is there an ‘effect’ here?)

→ visualize statistical measures too

Scale # fields

→ groups, hierarchies in schema

→ rank and trim display via measures over measures

Scale # records

→ measure, aggregate, bin *before* handing off to display substrate

Analyses – humans in the loop

To Judge –

to test a hypothesis,
to ask and answer a question in some medium

Operations toward judgment:

Describe, Compare, Abstract, Infer.

Visual "medium":

Person, through DataViz.
Look → Judge

Computational "medium":

System, through Algorithms.
Compute → Look → Judge.