THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Determining which variables contribute to shot success rate in soccer

Michael Womble, Daniel Diaz, Rose Mesina, and Ganesh Ghimire

March 8, 2022

G Data Set Overview:

2016:

1,974 Matches

3,251,294 events & 18 variables

2017 / 2018:

2018:

GW

Variables:

Event Name

- Shots
- Pass
- Free Kick

Subevent Name

- Simple Pass
- Cross
- Goal Kick

Event Time

Event Coordinates

- x, y
- start
- end

Player

Team

Match Date

Tags

- Goals
- Assist
- Key Passes

Variables:

Event Name

- Shots
- Pass
- Free Kick

Subevent Name

- Simple Pass
- Cross
- Goal Kick

Event Time

Event Coordinates

- x, y
- start
- end

Player

Team

Match Date

Tags

- Goals
- Assist
- Key Passes

M SMART Question:

Based on the matches played during the 2017/2018 season for Europe's top five leagues, the 2016 European Championship, and the 2018 World Cup, which match variables (ex: shot location, assist location, assist distance, etc…) result in the highest probability of a shot on goal being successful, result in a goal?

Approach:

Shots

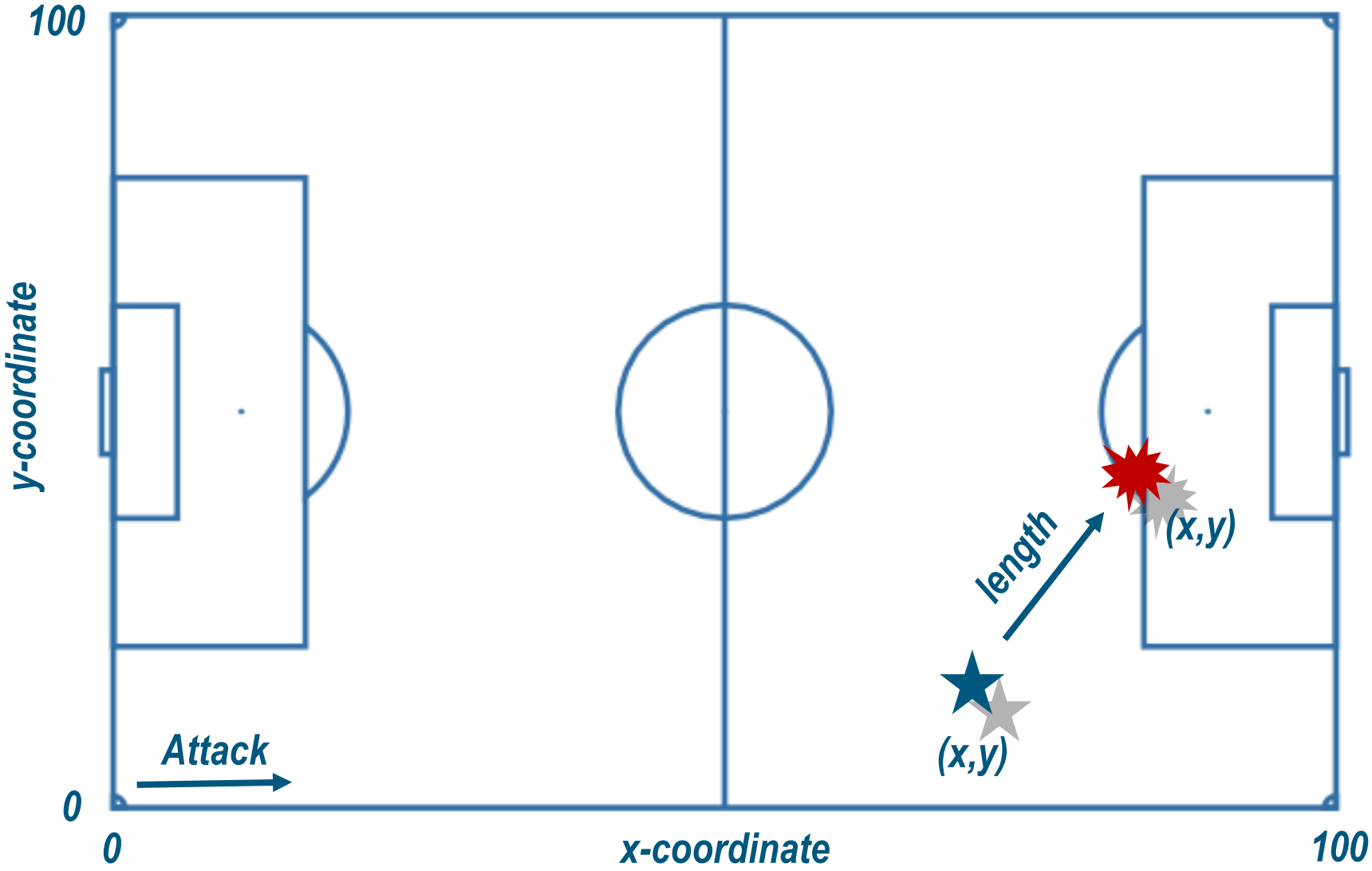
Assists

Key Passes

Location

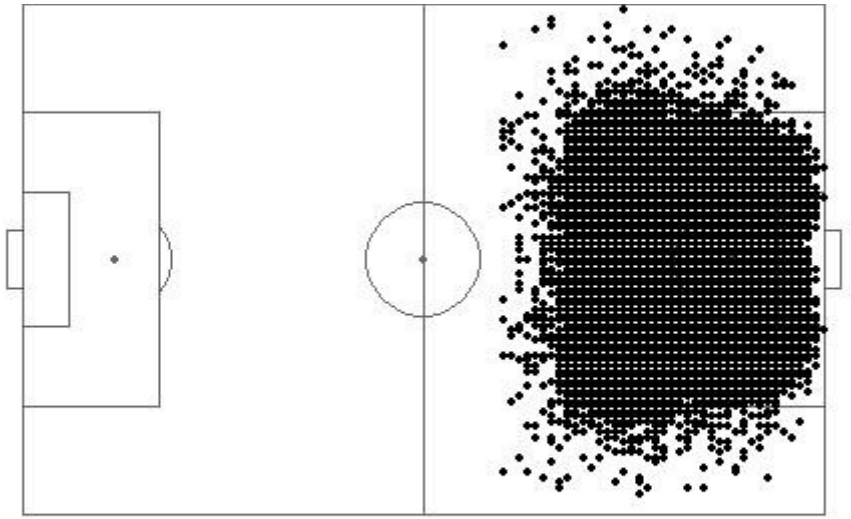
Distance

Understanding Field Coordinates:

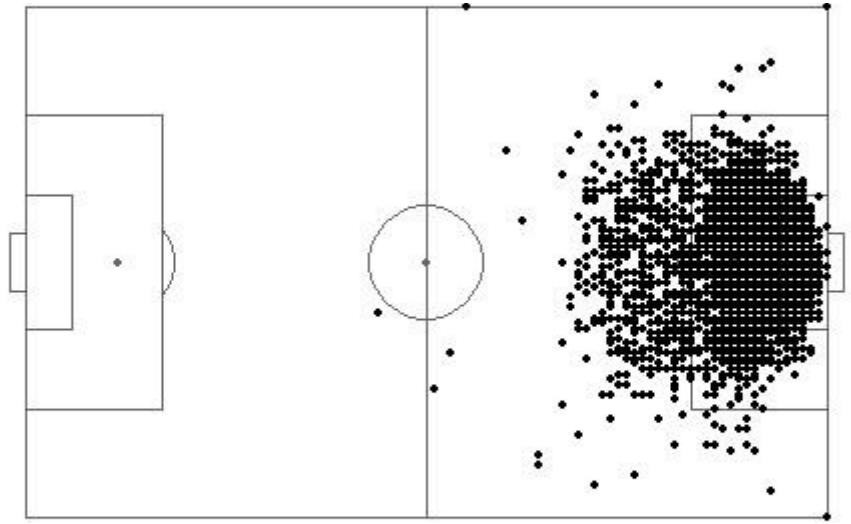


Starting Coordinates in the Data

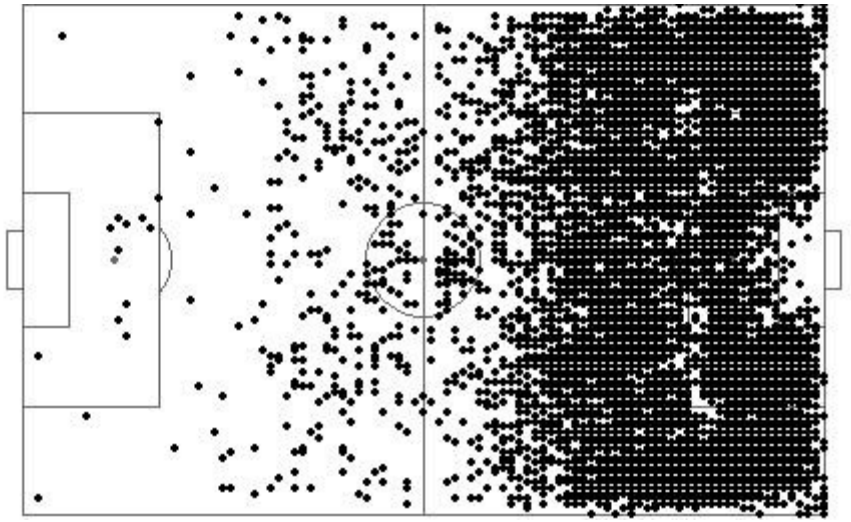
Shots



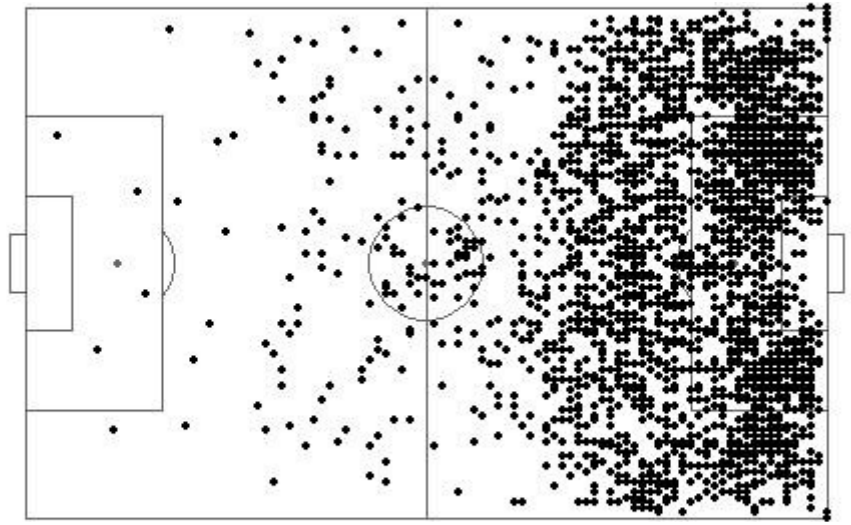
Goals



Key Passes

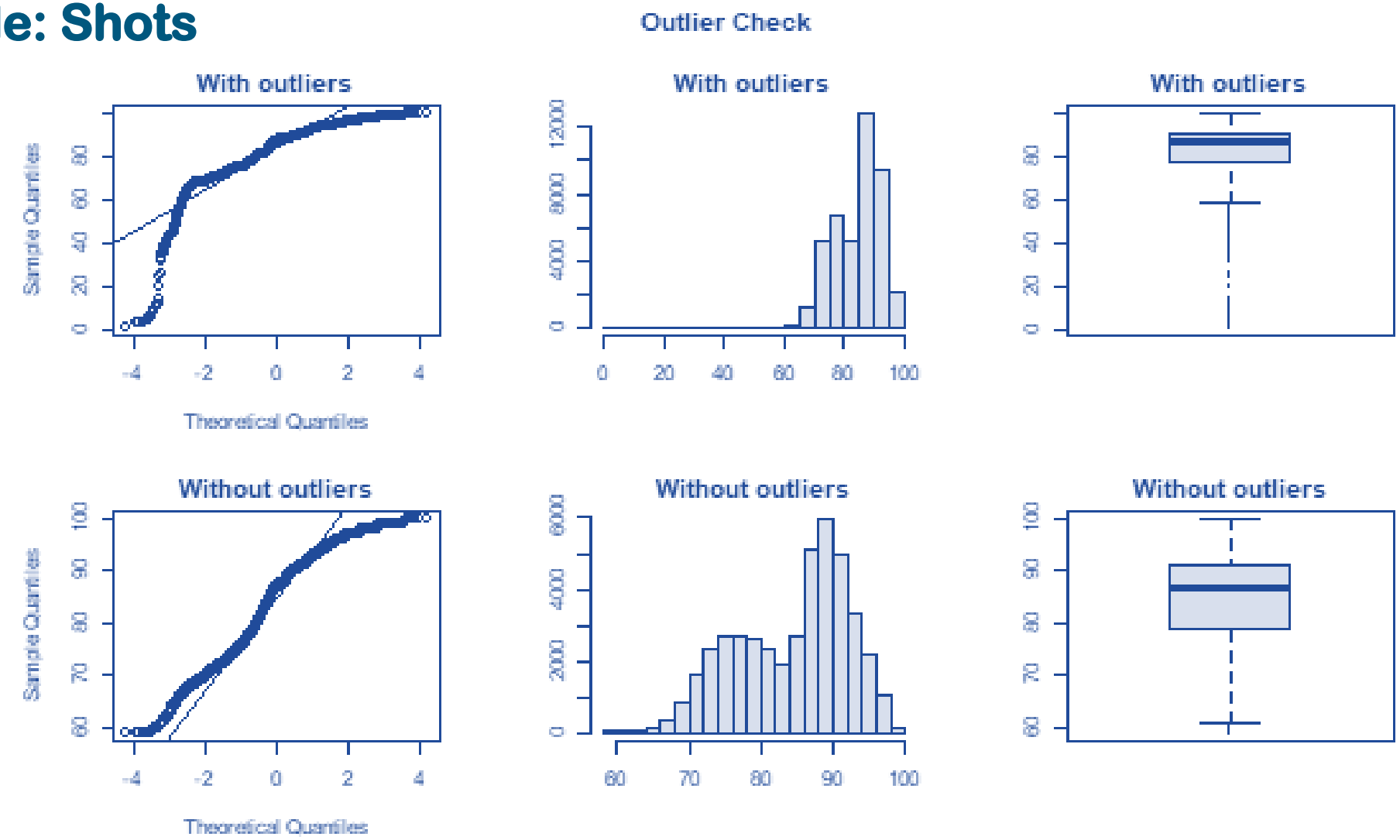


Assists



D Checking for Outliers:

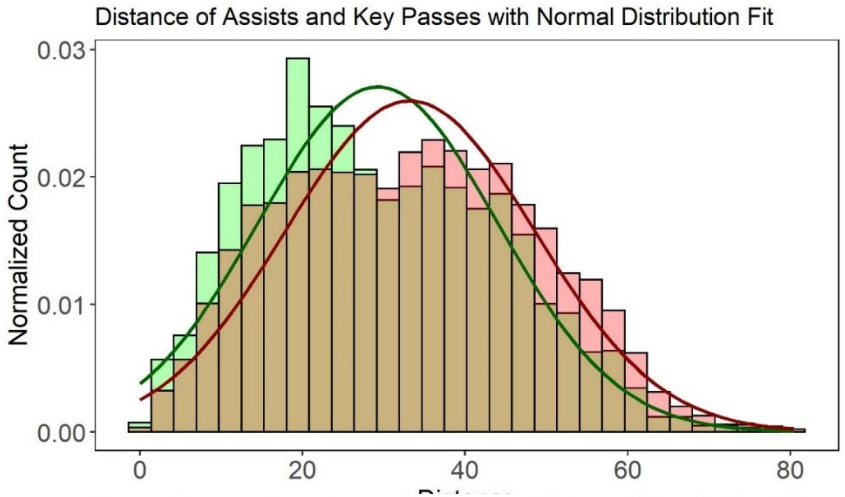
Example: Shots

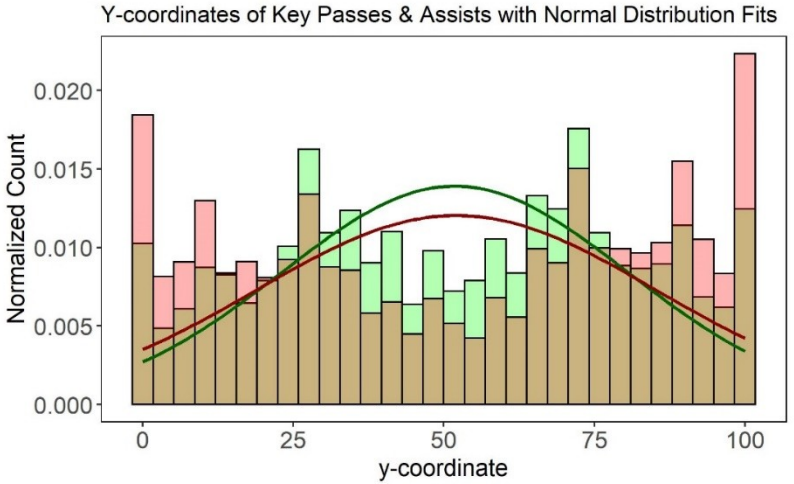


Hypothesis Testing: Assist vs Key Pass

Welch Two Sample t-test

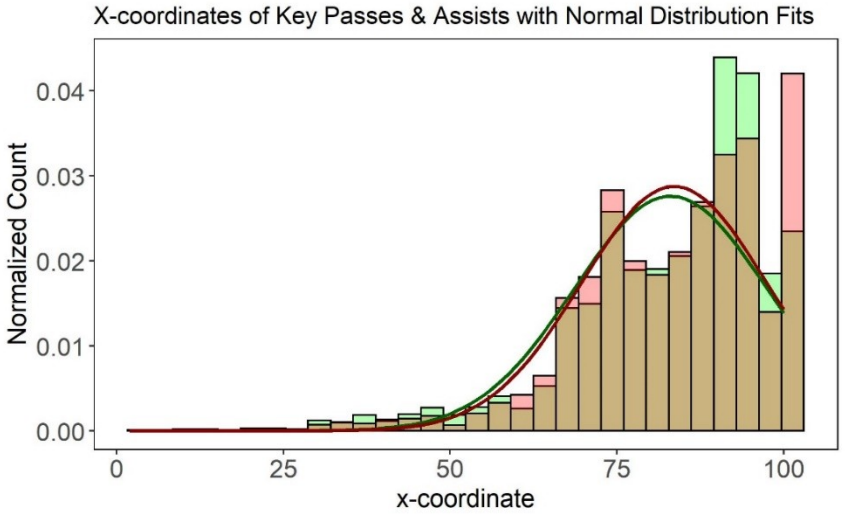
- ***p-value: <2e-16***
- ***Assist Mean: 29.2***
- ***Key Pass Mean: 33.2***





Welch Two Sample t-test

- ***p-value: 0.9***
- ***Assist Mean: 84.7***
- ***Key Pass Mean: 84.7***



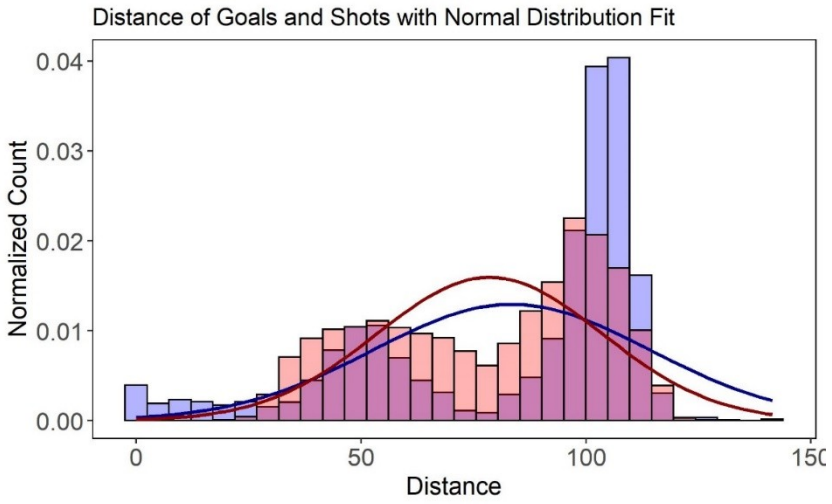
Welch Two Sample t-test

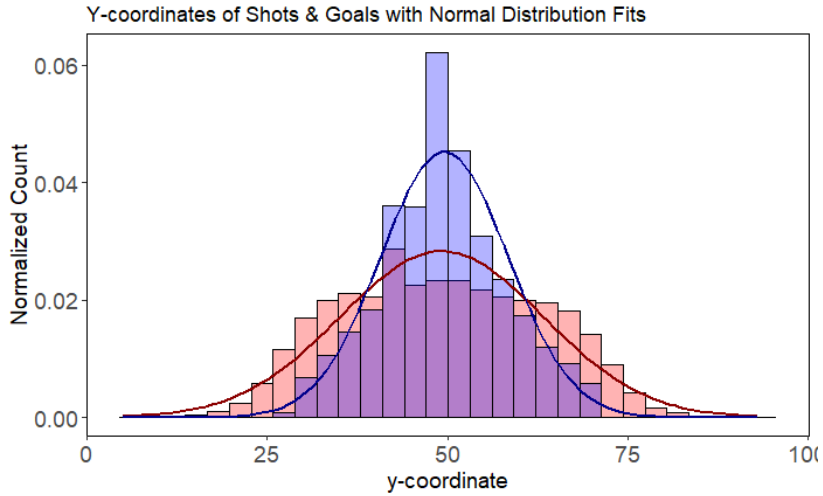
- ***p-value: 0.6***
- ***Assist Mean: 51.7***
- ***Key Pass Mean: 52.1***

Hypothesis Testing: Shot vs Goal

Welch Two Sample t-test

- ***p-value: <2e-16***
- ***Goal Mean: 83.4***
- ***Shot Mean: 78.6***



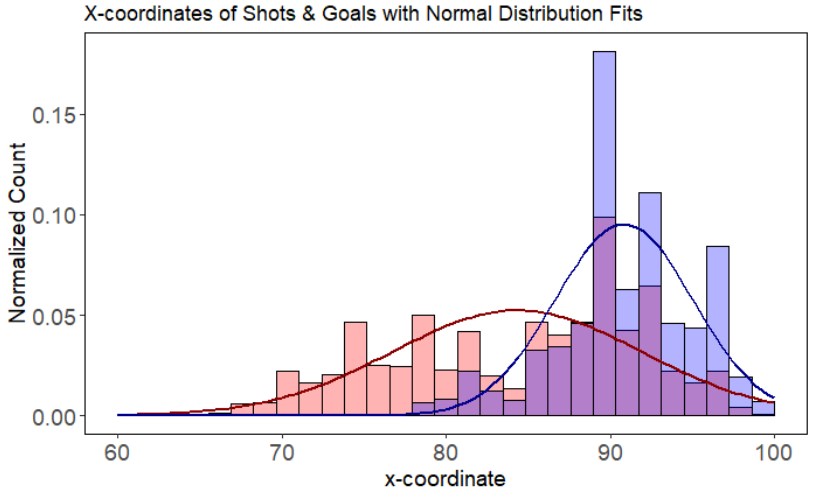


Welch Two Sample t-test

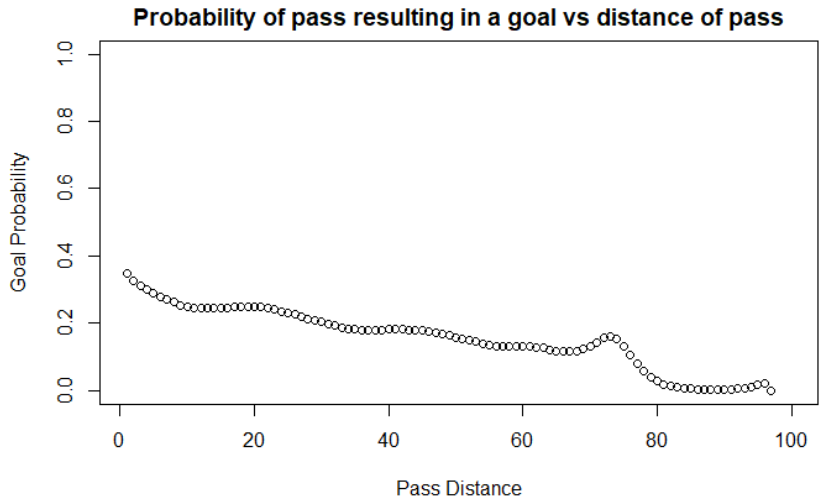
- ***p-value: <2e-16***
- ***Goal Mean: 68.1***
- ***Shot Mean: 49.2***

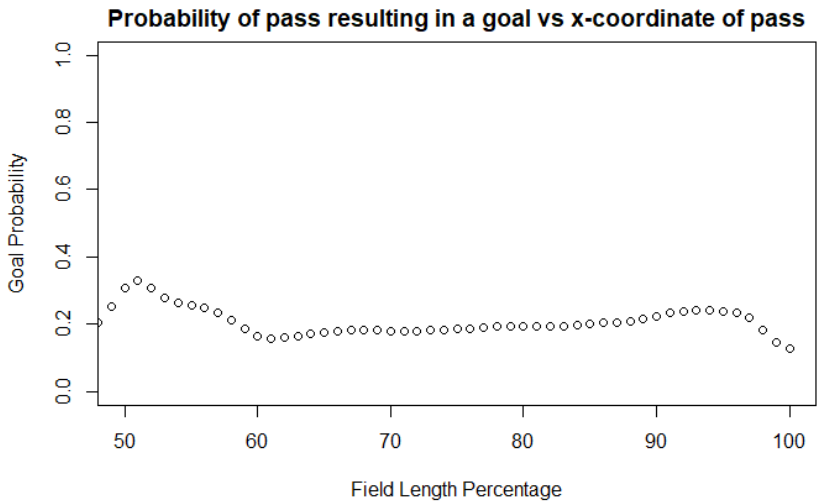
Welch Two Sample t-test

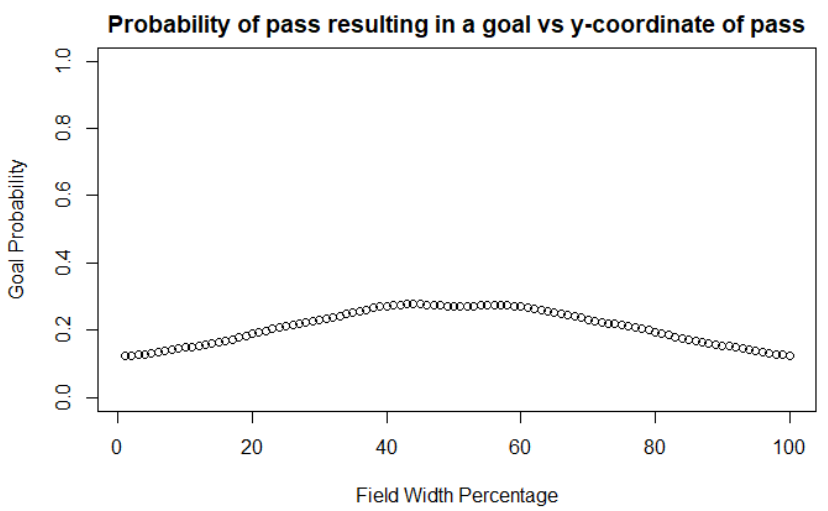
- ***p-value: <2e-16***
- ***Goal Mean: 93.5***
- ***Shot Mean: 84.3***

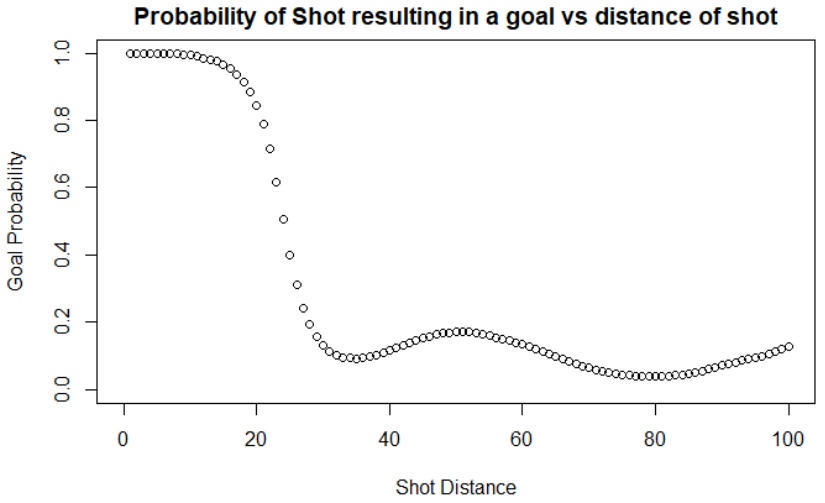


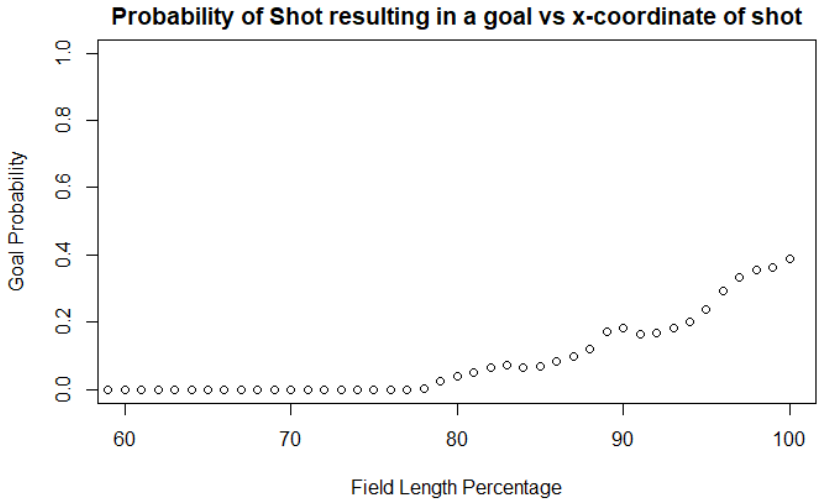
M Success Probability Plots

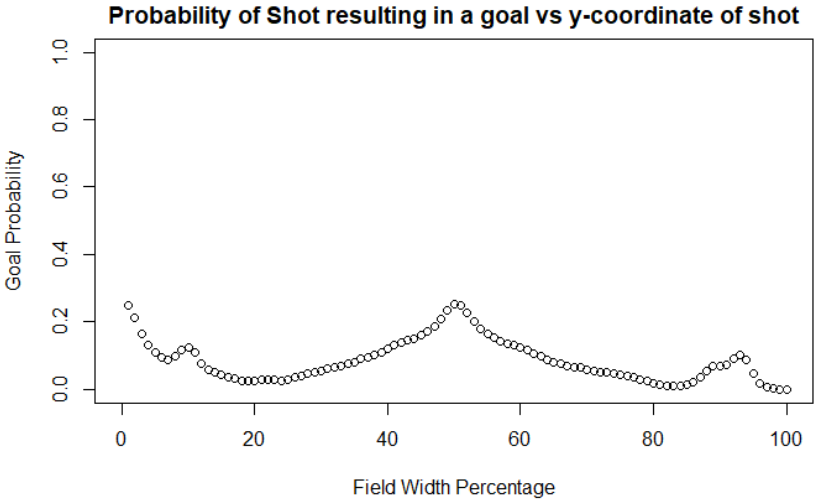












R Hypothesis Testing Recap

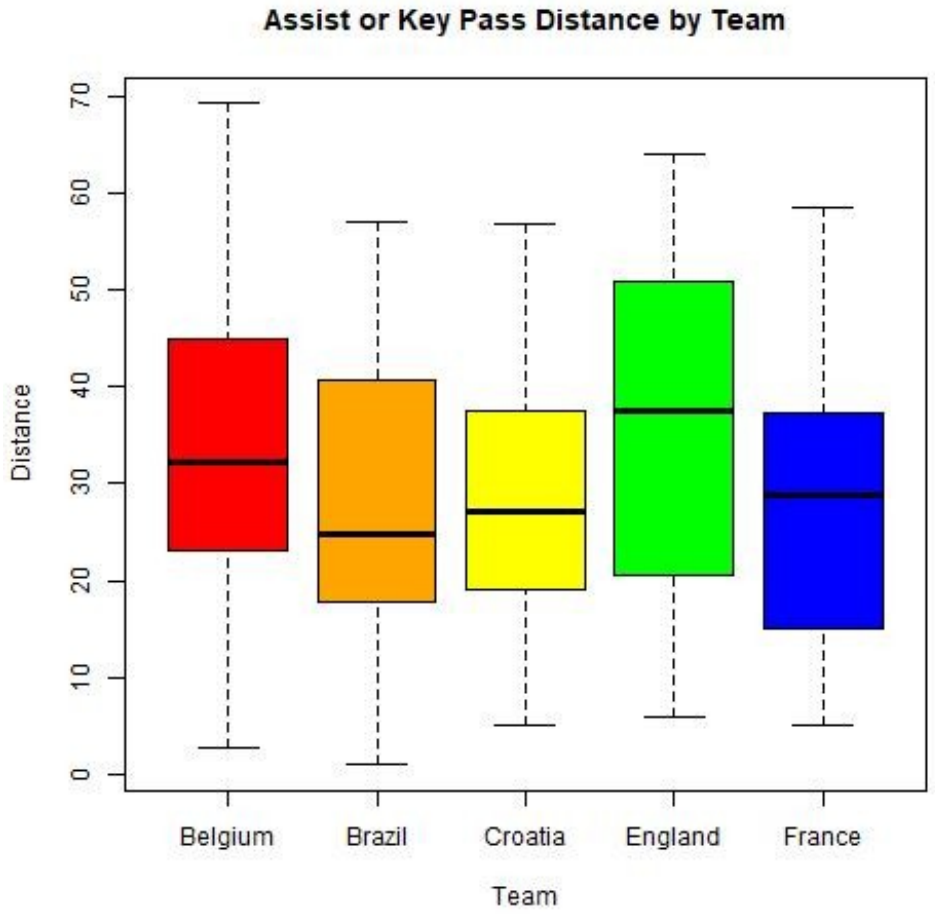
Significant Difference in Means

1. *Assist v. Key Pass Distance (higher mean for key pass)*
2. *Goal v. Shot Distance (higher mean for goals)*
3. *Goal v. Shot x.Start (higher mean for goals)*
4. *Goal v. Shot y.Start (higher mean for goals)*

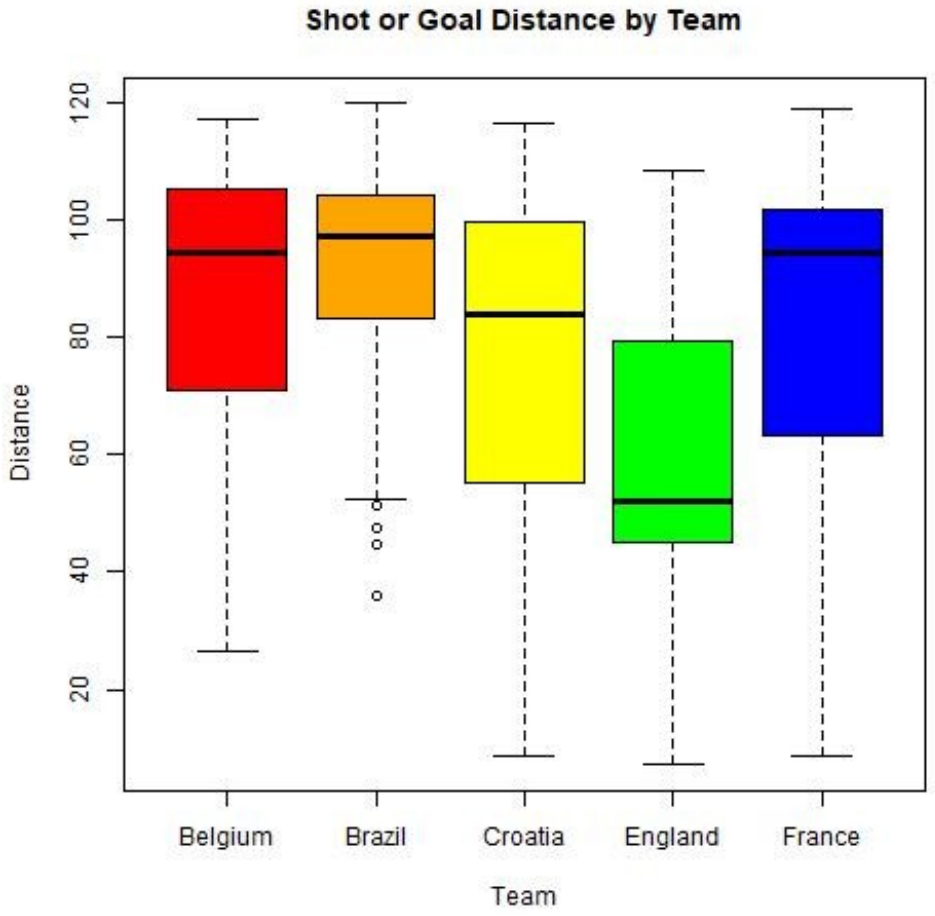
- *Do teams differ in their strategies when it comes to the above variables? Let's take a look at some Teams in the World Cup.*

GW

Hypothesis Testing: Means Between Teams Using ANOVA

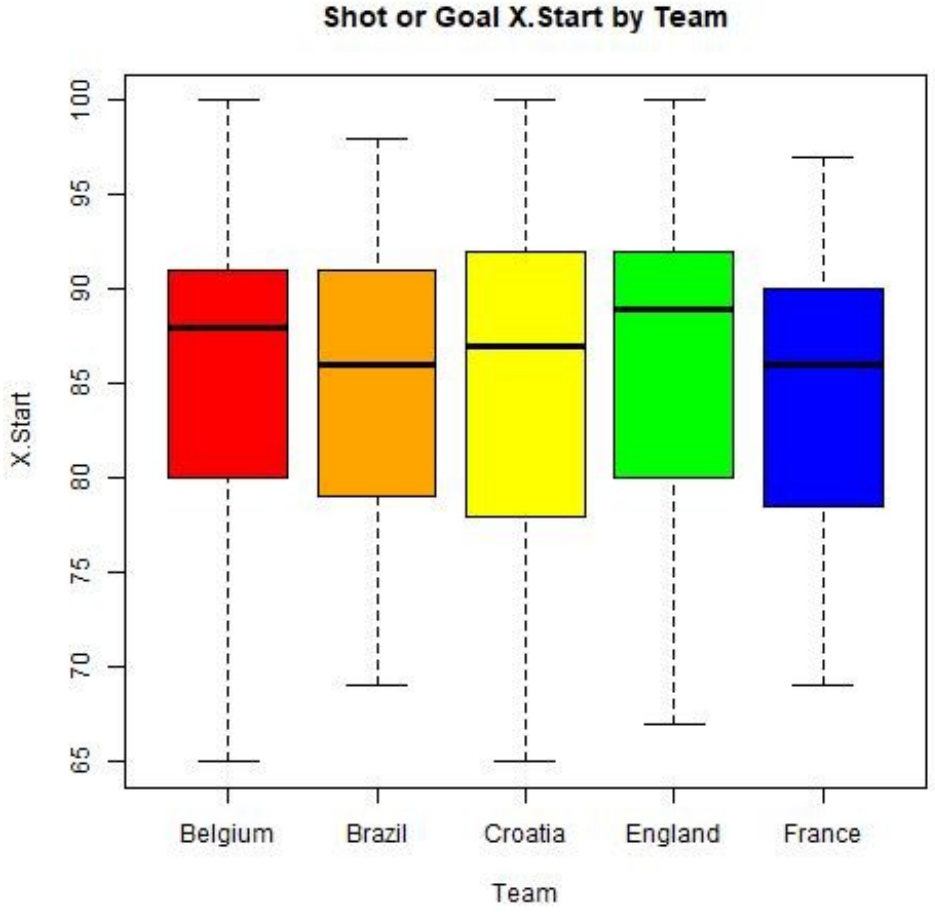


ANOVA
- p-value: 0.07

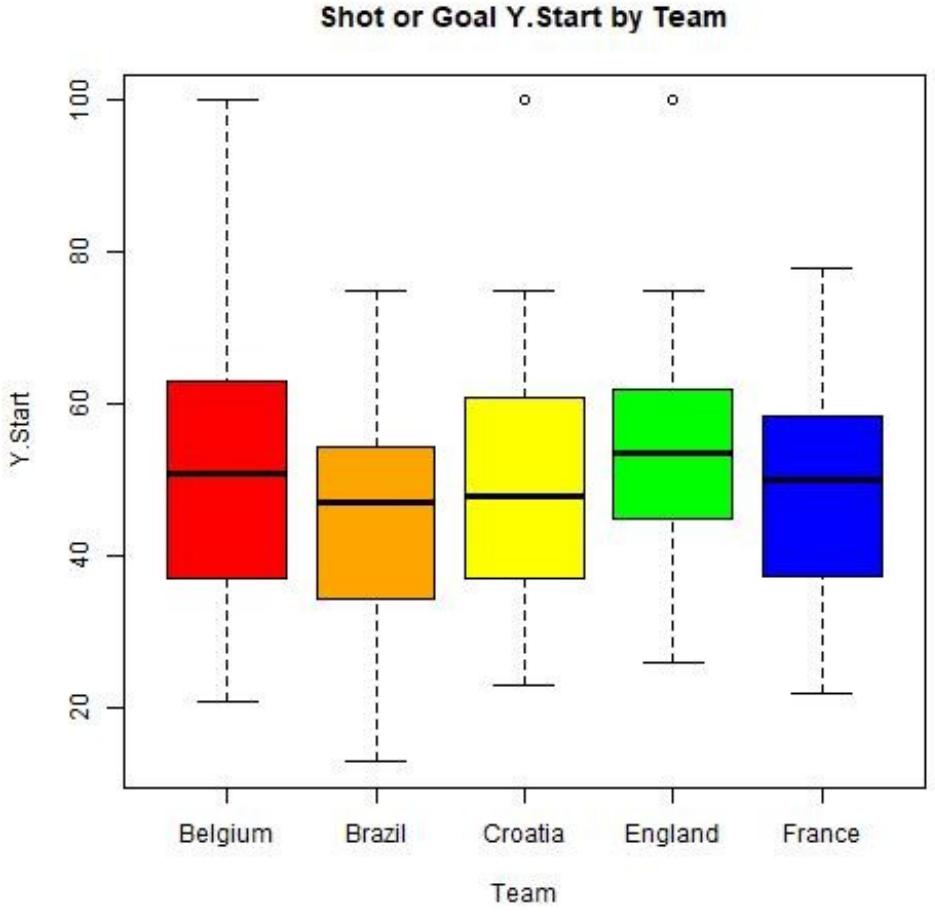


ANOVA
- p-value: 7.74e-18

Hypothesis Testing: Means Between Teams Using ANOVA



ANOVA
- p-value: 0.153



ANOVA
- p-value: 0.002

Conclusions:

- *Based on t-tests done between Shots vs Goals, and Assists vs Key Passes, there is evidence of distance and start coordinate differences between successful and unsuccessful plays.*

- *Success probability plots show that all variables selected play a significant role in the success of shot or assist/key pass resulting in a goal.*
  - *Most influential appears to be shot distance*
  - *Least influential appears to by the x-coordinate of final pass (assist/key pass)*

- *Applying this finding to the top 5 teams in the World Cup, we see further that there are teams that are significantly different in their tendencies for distance and Y.Start when it comes to attempting a goal.*

GW

Next steps

- *Adjust fitting to better account for non-normal distributions*
- *Further exploration of data by groups (by league, etc?)*
- *Post-hoc comparisons based on ANOVA significant results*
- *Improving visualizations (3D plots)*
- *Linear model*

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Q&A?

Thank You!