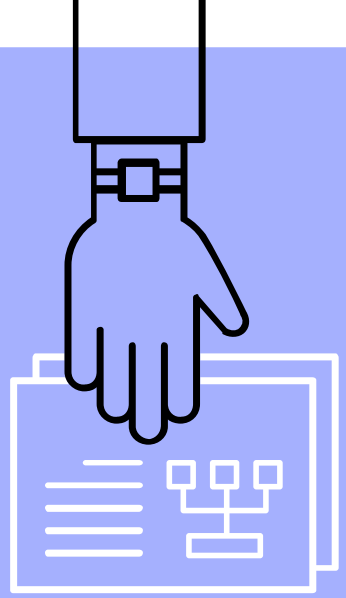
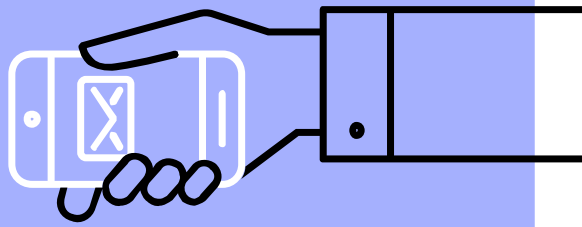
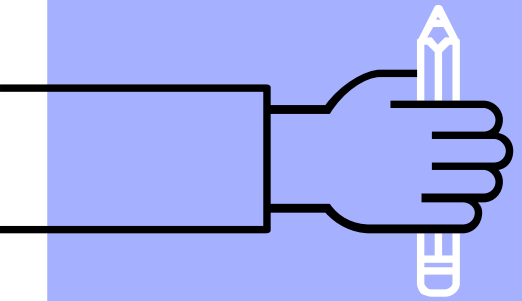
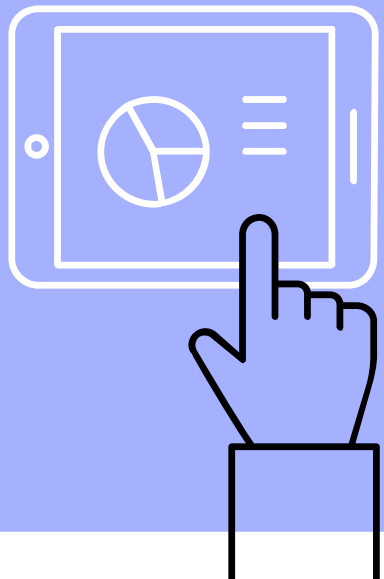




PREDICTING THE MADNESS!!

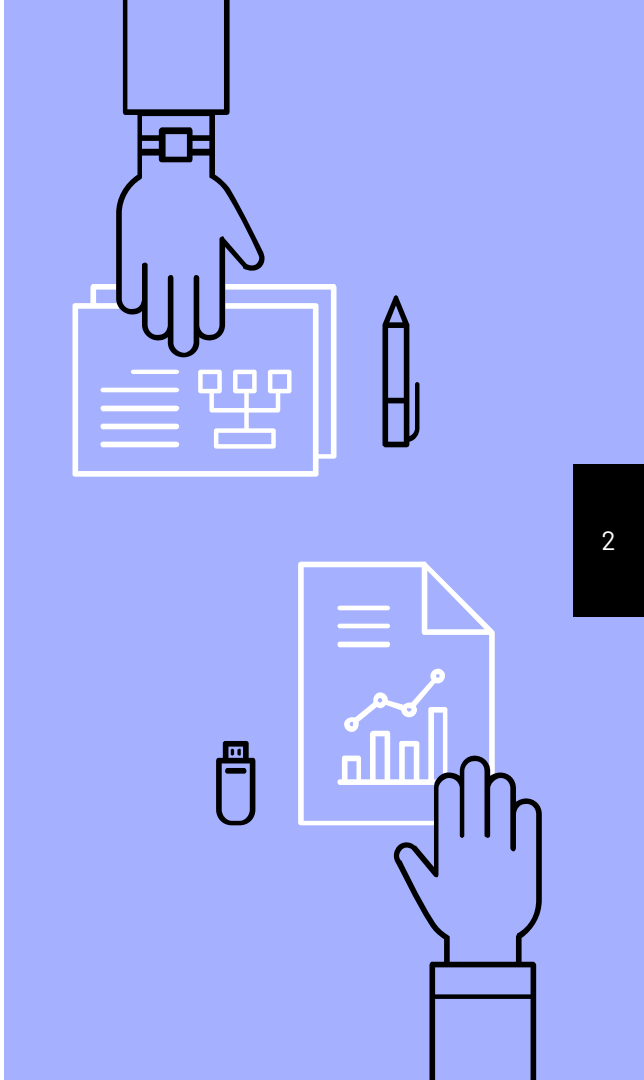




By: Harvir Singh Virk, Yu Zhong, Yi-Cheng Lu

Agenda

1. Introduction & Research Questions
2. Exploratory data analysis
3. Methodology
4. Conclusions & Discussions
5. Future Work



March Madness is a college basketball tournament. It starts with 68 teams and the regular season determines the seeding and they enter a single elimination tournament.

Our data set comes with 6 csv files from 2003 to 2016.

- Regular Season
- Seasons
- Teams
- Tournament
- Other relational table files

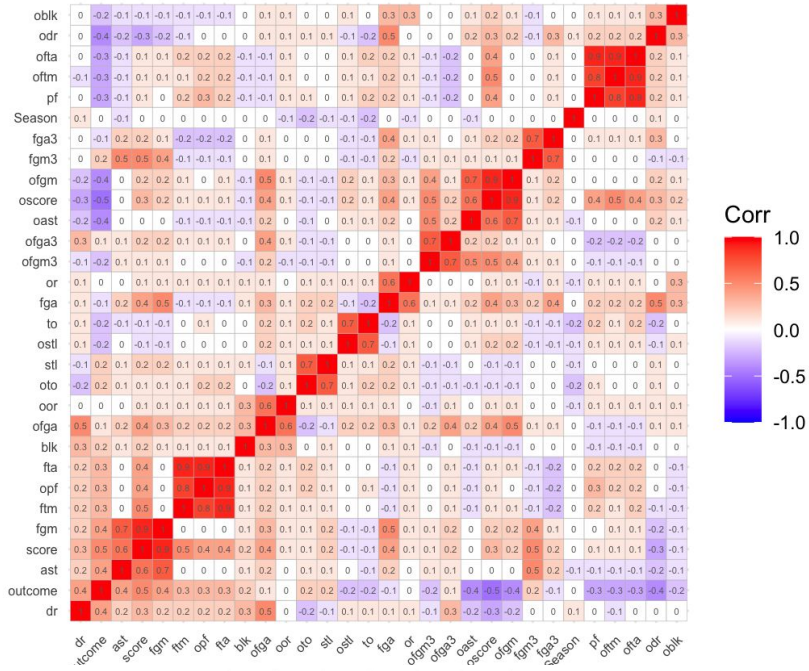
First look... Hard to predict because 1) no player data, 2) single elimination

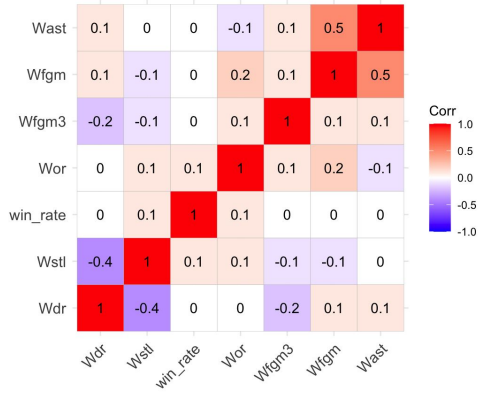
INTRODUCTION

Research Questions

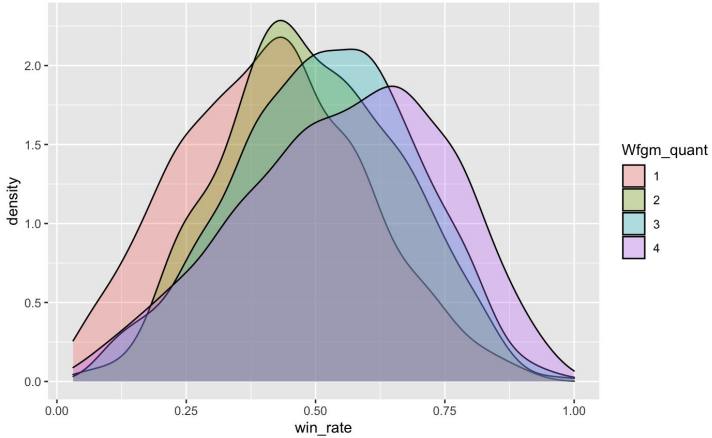
Question 1: Important Features for Winning

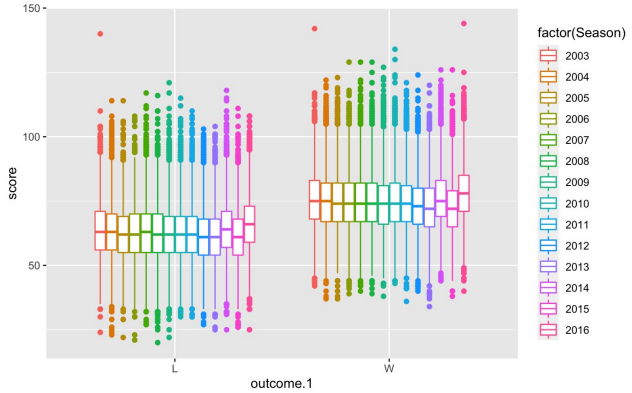
Question 2: How can teams make it to the playoffs?

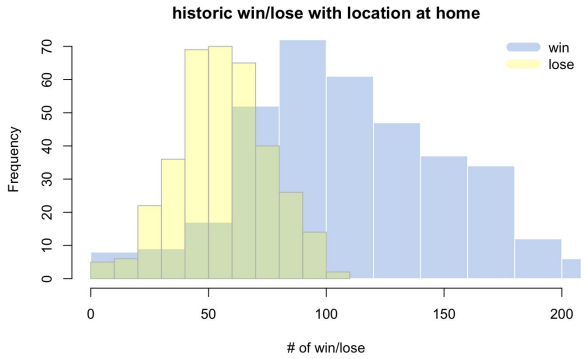




EXPLORATORY DATA ANALYSIS



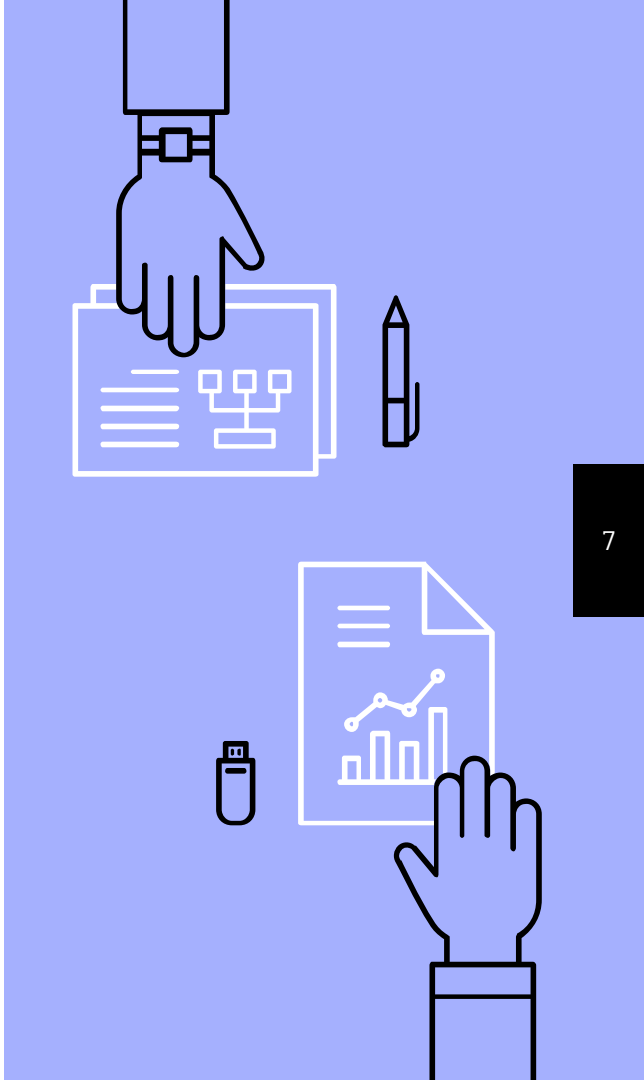




EXPLORATORY DATA ANALYSIS

Question 1: Important Features for Winning

- Predicting who can win games is important!
- Which features are the best predictors of winning.
- Variable Selection - Backwards, Forwards, LASSO Classification
- Model: Logistic Regression, Decision Tree, Random Forest + Boost



QUESTION 1.1

DATA CLEANING

- **Data**: Each row represents a game from one teams perspective
- **Independent Variables:**
  - 29 independent variables, such as field goals made, 3 pointers made, free throws made, etc.
  - Reduced based on variable selection
- **Target Variables:**
  - Win or lose the game.

QUESTION 1.1

Methodology

- **Data Models**
  - Logistic Regression w/o variable selection:
    - Precision = 1.00
    - Recall = 1.00
  - Logistic Regression:
    - Precision = 0.83
    - Recall = 0.83
  - Decision Tree:
    - Precision = 0.74
    - Recall = 0.71

QUESTION 1.1

Conclusions and Discussion

- **Variable Selection**
  - LASSO Classification
  - Forward Variable Selection
  - Backward Variable Selection
  - Decision Tree

QUESTION 1.1

Conclusions and Discussions

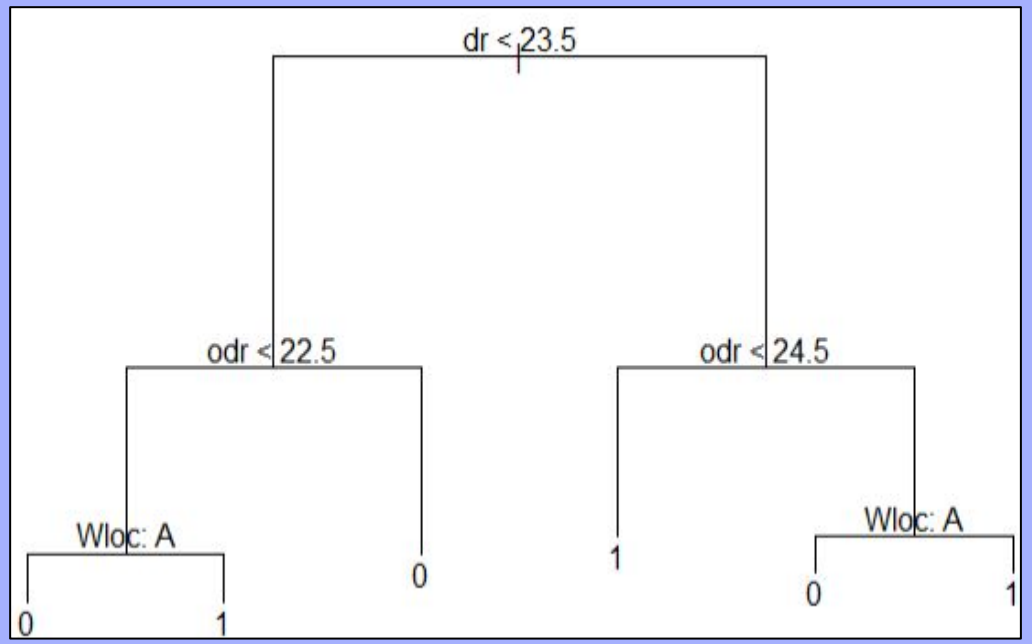
- LASSO Classification
- Forward Variable Selection
- Backward Variable Selection

Feature	Rank of Importance
Opponent Field Goals Made	1
Field Goals Made	2
Free Throws Made	3
Opponent Free Throws Made	4
Opponent 3 Point Field Goals Made	5
3 Point Field Goals Made	6

QUESTION 1.1

Conclusions and Discussions

- Decision Tree
  - Defensive Rebounds
  - Opponent Defensive Rebounds
  - Winning Location

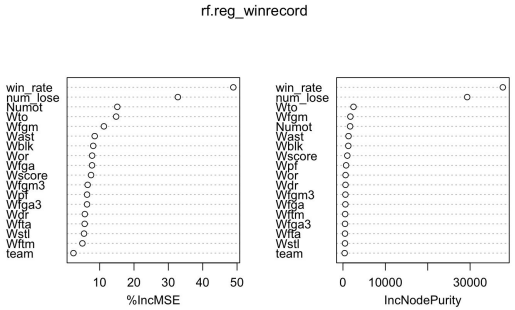


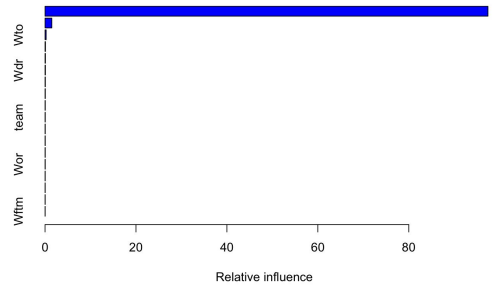
QUESTION 1.2

DATA CLEANING

- **Data**: performance(averages) of each team each year
- **Independent Variables:**
  - field goals made, 3 pointers made, free throws made, offensive rebounds, defensive rebounds, assists, turnovers, steals, blocks, personal fouls
- **Target Variables:**
  - Number of wins

Question 1.2: Important Features for Winning





Random Forest to see which feature is more important to predict the number of wins for teams

Boosting and CV

cv.error = `0.1381571`

Wto is significant

Methodology

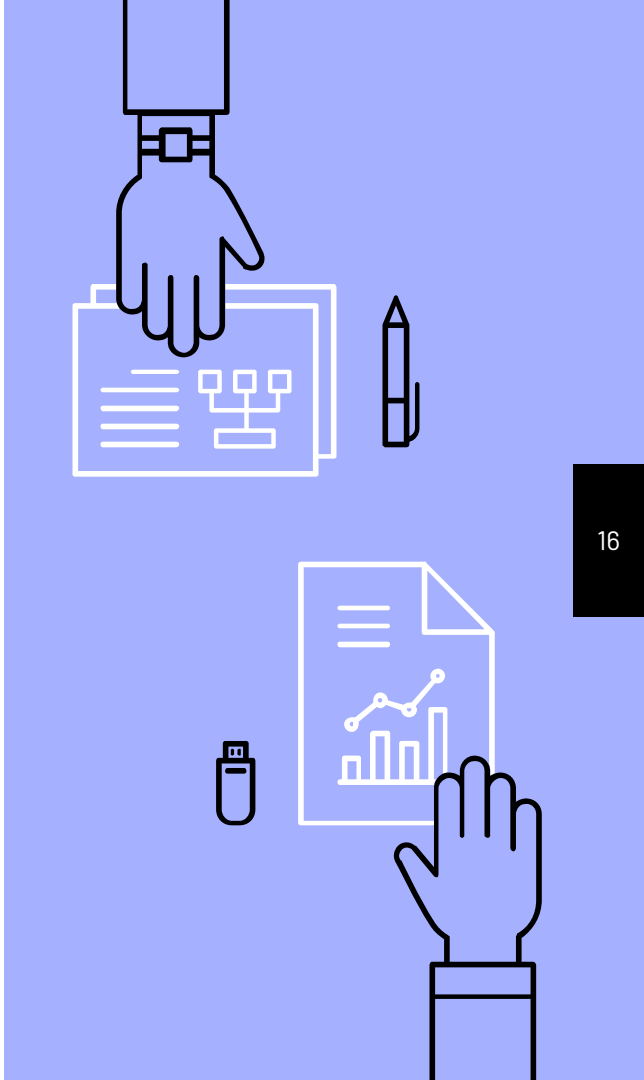
QUESTION 1.1 - 1.2

Conclusions and Discussions

- Logistic Regression performed well
  - Outcome was based on scoring
- Decision Tree
  - Did not perform as well
  - Provided great insight to important variables not related to scoring
- Random Forest
  - One variable comes up important with overall low error rate

Question 2: How can teams make it to the playoffs?

- College basketball teams care much about playoffs!
- Data: performance of each team each year
- Model: KNN, lasso, random forests
- Unsupervised: hierarchical clustering



QUESTION 2

DATA CLEANING

- **Data**: performance(averages) of each team each year
- **Independent Variables:**
  - field goals made, 3 pointers made, free throws made, offensive rebounds, defensive rebounds, assists, turnovers, steals, blocks, personal fouls
- **Target Variables:**
  - Enter playoff season or not

QUESTION 2

METHODOLOGY

- **Data Models**
  - KNN: 0.866
  - Lasso Regression: 0.878
  - Random Forests: 0.867
- **Unsupervised learning:**
  - Hierarchical Clustering
    - Use wss method to cut the tree at 4

QUESTION 2

plots

```
12 x 1 sparse Matrix of class "dgCMatrix"
                             1
(Intercept) -11.89750716
(Intercept)   .
fgm            0.02529763
fgm3           .
ftm            0.14511682
fta            .
or             .
dr             0.29001376
ast            0.31392607
to            -0.38466384
stl            0.24333355
blk            0.09715161
[1] 0.8776065
```

QUESTION 2

plots

fgm	23.219	23.130	25.694	23.664
fgm3	5.989	6.807	6.560	5.829
ftm	13.891	12.160	15.190	17.751
fta	20.286	17.398	21.813	25.597
or	11.299	9.670	12.011	12.026
dr	22.849	21.938	24.306	23.787
ast	12.270	13.026	14.243	12.209
to	14.542	12.896	13.497	15.042
stl	6.602	6.211	6.945	6.867
blk	3.168	2.783	3.897	3.370
pf	19.121	17.659	18.231	20.733
tour	0.082	0.124	0.409	0.202

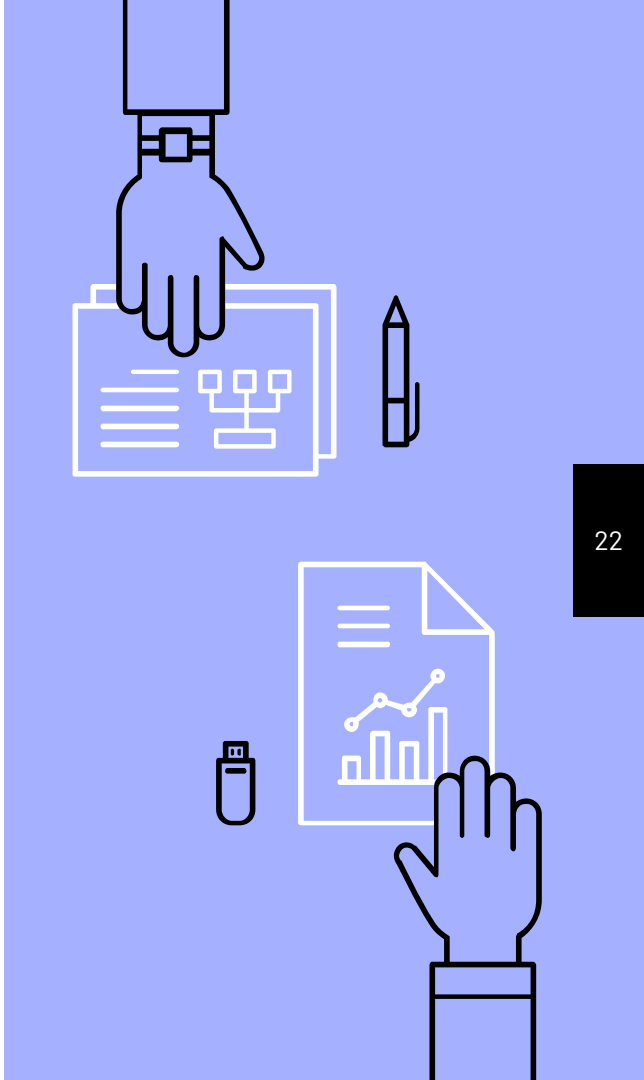
QUESTION 2

CONCLUSION

- **Supervised learning:**
  - Lasso gives the best prediction results!
  - Turnover is the most important variable
- **Unsupervised learning:**
  - Hierarchical clustering gives guidance to teams for earning spot in playoff

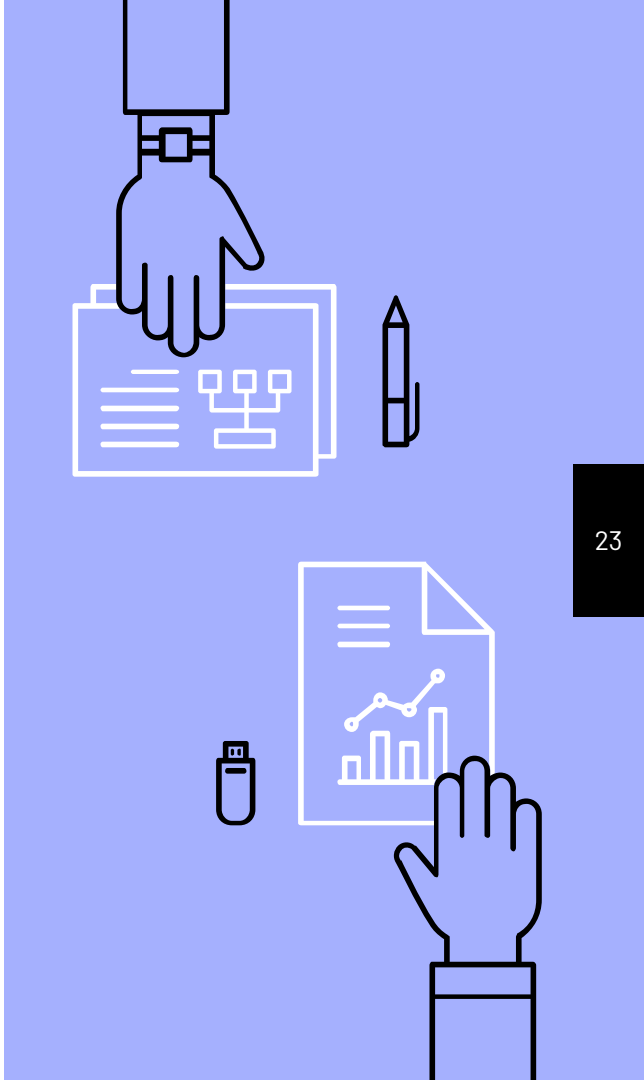
Future work

1. Time Series Analysis
2. Which variables not related to scoring produce high results
3. Player specific data
4. Viewing tournament and regular season statistics separately



References

1. https://www.kaggle.com/c/march-machine-learning-mania-2016/rules
2. An Introduction to Statistical Learning with Applications in R
3. Special thanks to Daniel Sun, a crazy basketball fan that provides precious suggestions!



Thank You!