Home-Court Advantage In NCAA Basketball

Malcolm Gaynor and Parker Gibbons

TABLE OF CONTENTS

Project Motivation/Past Literature

- “The home-court advantage in NCAA Division-I men’s basketball” by Cabarkapa et al. (2023)
    - Randomly selected box scores from the 2018/2019 season
    - Concluded that playing at home gave a significant advantage
    - Assists, Field Goal percentage, and turnovers are especially significantly different in home vs. away games

- “Home sweet home: Quantifying home court advantages for NCAA basketball statistics” by van Bommela et al. (2021)
    - Looked at box scores from the 2011-2012 and the 2015-2016 seasons for DI, DII, and DIII men’s and women’s basketball
    - Found that home teams receive a boost as compare to the mean in almost every stat
    - Interestingly, found that higher attendance is significantly associated with referee bias towards the home team

- Both of these article only look at home vs. away games in general, they don’t dive into the differences in specific arenas.

Data Wrangling

- Dat
- Fut

	Venue	Season	AwayFGpct	Away3ptpct	AwayTotalRebounds	AwayAssists	AwaySteals	AwayBlocks	AwayTurnovers	AwayFouls
240	OSU	2020	0.414	0.400	23	12	3	3	9	21
241	OSU	2020	0.283	0.152	30	3	6	3	15	20
242	OSU	2020	0.397	0.261	31	9	5	1	4	21
243	OSU	2020	0.481	0.250	33	4	5	1	10	18
244	OSU	2020	0.441	0.474	24	9	6	0	11	19
245	OSU	2020	0.321	0.208	33	11	2	2	10	24
246	OSU	2020	0.420	0.333	25	13	3	0	14	20
247	OSU	2020	0.534	0.478	28	19	1	2	7	13
248	OSU	2020	0.468	0.417	31	22	5	1	5	13
249	OSU	2020	0.519	0.333	34	12	4	1	9	14
250	OSU	2021	0.338	0.231	25	7	7	5	8	19
251	OSU	2021	0.471	0.417	28	19	7	3	12	19
252	OSU	2021	0.396	0.273	27	8	5	4	7	25
253	OSU	2021	0.390	0.345	29	13	5	0	9	13
254	OSU	2021	0.308	0.263	27	6	8	2	12	12
255	OSU	2021	0.422	0.286	36	12	3	3	8	20
256	OSU	2021	0.344	0.316	33	12	6	4	11	21
257	OSU	2021	0.491	0.391	28	12	5	1	4	20
258	OSU	2021	0.453	0.500	26	11	3	3	13	15
259	OSU	2021	0.412	0.381	26	14	11	4	5	14
260	OSU	2022	0.519	0.294	25	14	7	4	13	19
261	OSU	2022	0.431	0.419	32	15	4	0	14	9
262	OSU	2022	0.500	0.318	32	12	3	6	9	14
263	OSU	2022	0.450	0.458	25	15	6	1	14	17
264	OSU	2022	0.418	0.300	24	9	8	0	7	15
265	OSU	2022	0.462	0.414	21	15	6	1	10	17
266	OSU	2022	0.458	0.318	37	15	4	2	10	12
267	OSU	2022	0.519	0.526	24	9	1	1	3	10
268	OSU	2022	0.361	0.207	25	8	4	4	11	13
269	OSU	2022	0.440	0.368	21	9	1	3	9	16

G	Date		Opp
21	2023-01-28	@	Indiana
22	2023-02-02		Wisconsin
23	2023-02-05	@	Michigan
24	2023-02-09		Northwestern
25	2023-02-12		Michigan State
26	2023-02-16	@	Iowa
27	2023-02-19	@	Purdue
28	2023-02-23		Penn State
29	2023-02-26		Illinois
30	2023-03-01		Maryland
31	2023-03-04	@	Michigan State
32	2023-03-08	N	Wisconsin
33	2023-03-09	N	Iowa
34	2023-03-10	N	Michigan State
35	2023-03-11	N	Purdue

t							
	ORB	TRB	AST	STL	BLK	TOV	PF
5	12	35	17	5	3	8	19
1	4	24	9	8	0	7	15
2	6	33	11	2	3	8	13
2	5	21	15	6	1	10	17
0	11	37	15	4	2	10	12
7	9	26	23	7	3	7	10
3	12	41	16	7	3	11	14
3	4	24	9	1	1	3	10
4	7	25	8	4	4	11	13
3	7	21	9	1	3	9	16
0	5	26	17	2	2	8	12
1	12	29	8	8	2	10	18
3	9	29	11	3	5	11	12
2	7	28	9	3	6	7	15
3	8	34	18	3	2	6	17

Exploratory Data Analysis

- Correlation Matrix - What statistics in our data best correlate to winning games?
  - FG made (36.78)
  - FG% (0.4664)
  - 3pt% (0.3661)
  - FT made (0.3087)
  - Total rebounds (0.3526)
  - Assists (0.3017)
- Multiple Linear Regression
  - Full Model (Adjusted R-squared = 0.5435)
    - Significant variables (at 99%): Total rebounds, steals, blocks, turnovers, fouls
  - Final model obtained through backwards elimination and best subsets (Adjusted R-squared = 0.5482)
    - Variables: FG made, FG%, FT attempted, Offensive rebounds, total rebounds, steals, blocks, turnovers, fouls)
    - All significant except for offensive rebounds
    - All significant at 99% except for FT attempted and blocks

EDA Takeaways

- Findings reveal variables that are potentially predictive of winning games
  - FG made
  - FG%
  - 3pt%
  - Total Rebounds
  - Assists
  - Steals
  - Blocks
  - Turnovers committed
  - Fouls
- Ideally most impactful on whether a team wins and thus are worth looking at differences across different venues

Why MANOVA?

Our question: does venue (one independent variable) impact away team's statistics (multiple dependent variables)?

BACKGROUND:

- MANOVA handles multiple dependent variables
- Null hypothesis: group VECTORS are equal
- Alternative hypothesis: at least one group MEAN is different

$$\mu_i = \begin{bmatrix} \text{Mean FG\%} \\ \text{Mean Rebounds} \\ \vdots \\ \text{Mean Turnovers} \end{bmatrix}$$

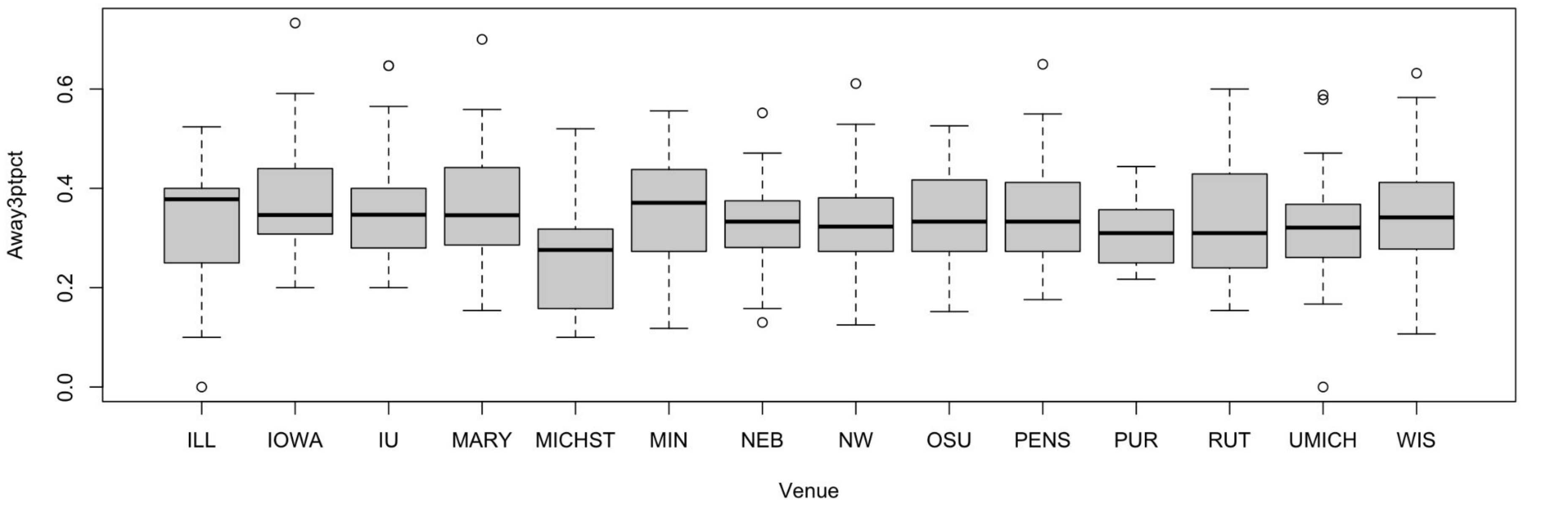
TEST STATISTIC:

- Pillai Trace: *trace*$(H(H+E)^{-1})$
    - *trace* is the sum of the diagonals of a matrix
    - H is a matrix where element (a,b) is:
    - E is a matrix where element (a,b) is:
        - Considering j observations per i groups
- Reject the null when the value is large

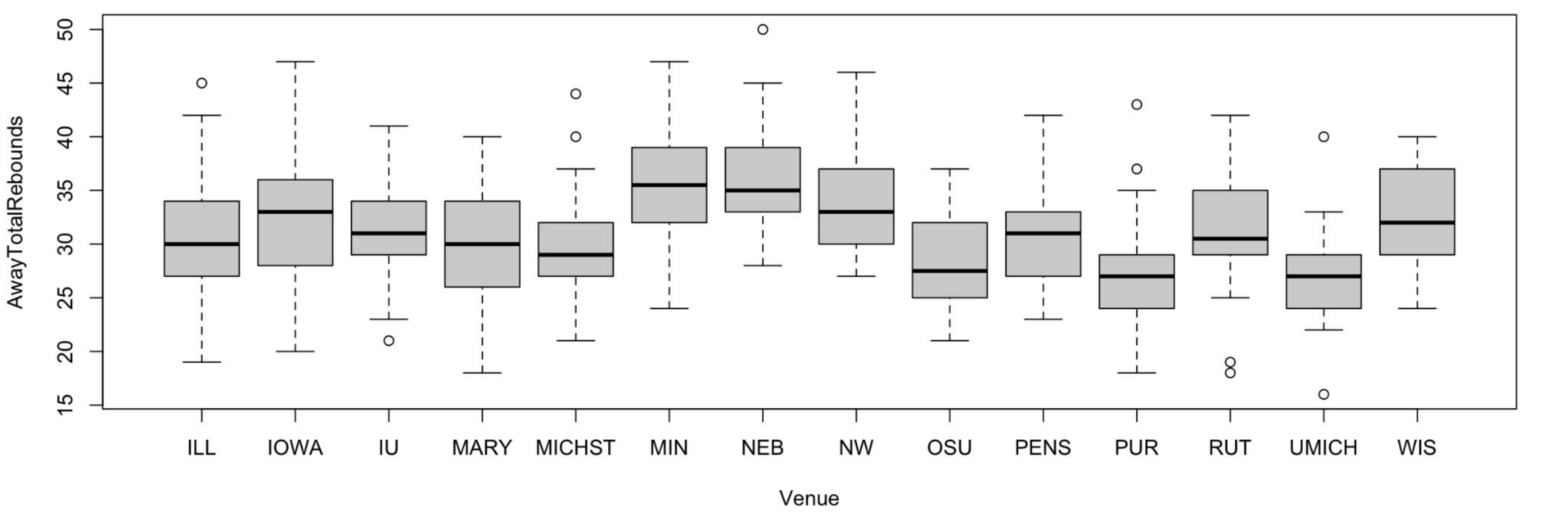
$$\sum_{i=1}^{N} n_i(\bar{y}_{i,a} - \bar{y}_{t,a})(\bar{y}_{i,b} - \bar{y}_{t,b})$$

$$\sum_{i=1}^{N} \sum_{j=1}^{n_i} (Y_{i,j,a} - \bar{y}_{i,a})(Y_{i,j,b} - \bar{y}_{i,b})$$

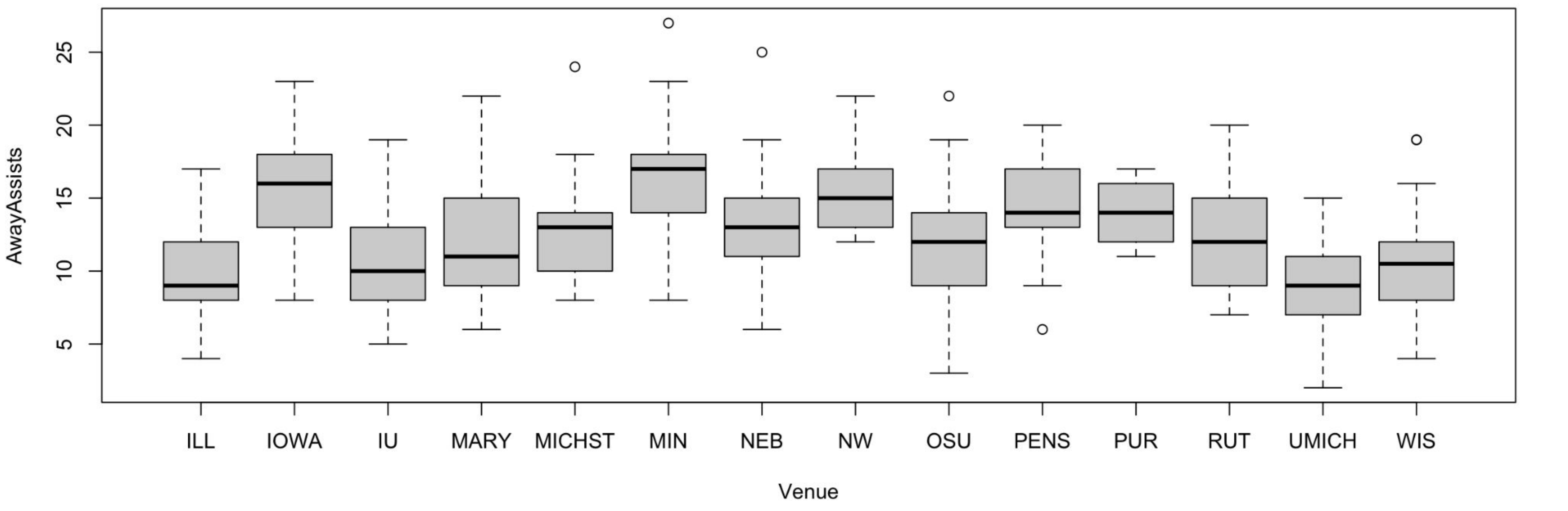
A few boxplots



A few boxplots



A few boxplots



MANOVA Conditions

```
> mshapiro.test(t(matrix_for_test))

        Shapiro-Wilk normality test

data:  Z
W = 0.98451, p-value = 0.0002037
```

Multicollinearity

- Correlation between FG made and FG% was 0.79, all else below 0.54
- Removed FG made

Multivariate normality

- Extension of the Shapiro-Wilk test
  - Null hypothesis is that the data is normally distributed
- Therefore, we reject the null hypothesis, and this condition is NOT satisfied

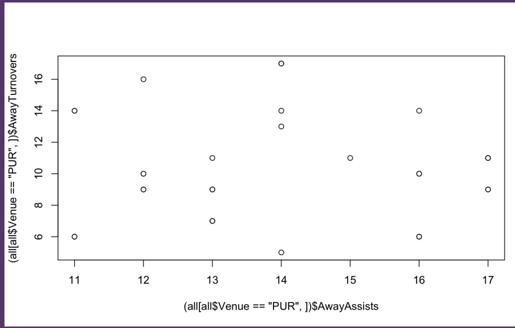
```
> boxM(Y = box_all[, box_indeps], group = box_all$Venue)

        Box's M-test for Homogeneity of Covariance Matrices

data:  box_all[, box_indeps]
Chi-Sq (approx.) = 508.43, df = 468, p-value = 0.0956
```

Homogenous variance-covariance matrices

- Box's M test
  - Null hypothesis is that variance-covariance matrices are equal for all groups
- Therefore, we fail to reject the null hypothesis, and the conditions is satisfied

Linear dependent variables

- Examine scatterplots for each pair of dependent variables in each group:
- Linearity doesn't always look great...



Independence

- Also probably not...

MANOVA Output

```r
###MANOVA

#to make MANOVA code easier, we define our dependent variables and our independent variable
depvs <- cbind(all$AwayFGpct,all$Away3ptpct,all$AwayTotalRebounds,all$AwayAssists,
             all$AwaySteals,all$AwayBlocks,all$AwayTurnovers,all$AwayFouls)
indepv <- all$Venue

#MANOVA test
our_manova <- manova(depvs ~ indepv, data = all)

#Summary of output
summary(our_manova)
```

```
> summary(our_manova)
           Df  Pillai approx F num Df den Df    Pr(>F)
indepv     13 0.93906   4.1023    104   3208 < 2.2e-16 ***
Residuals 401
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

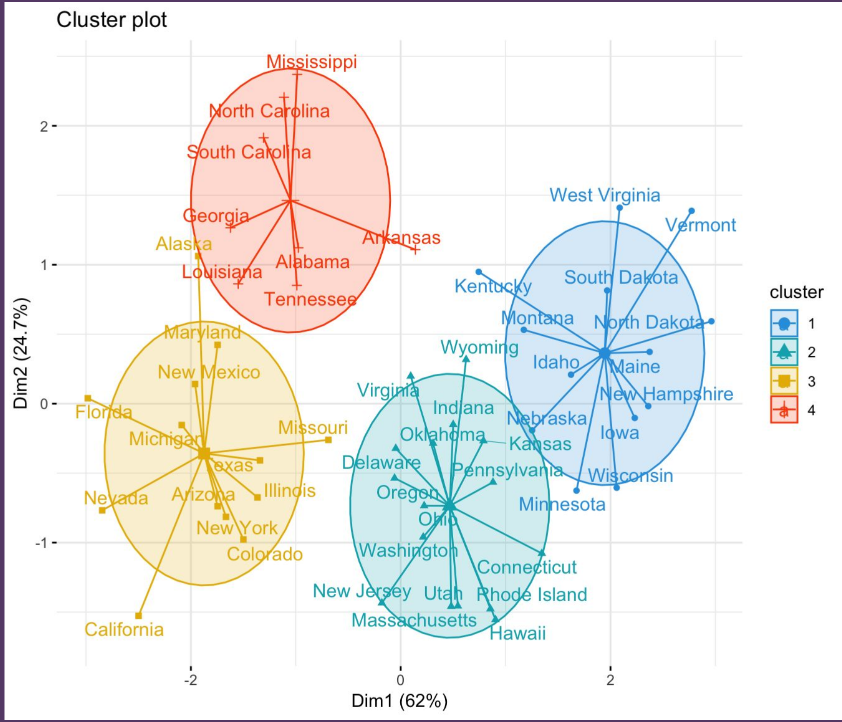
MANOVA Analysis

- We reject the null hypothesis, meaning there is SOME difference in at least one average statistic at at least one different Big Ten venue

- Normally, the next step would be to do Post-Hoc analysis
  - Dimensionality reduction: Linear Discriminant Analysis (LDA)

- However, because of our conditions, we chose another Machine Learning method instead:

K-Means Clustering



(Example diagram)

- Unsupervised non-linear algorithm that clusters data based on their similarity to one another
- Uses a a pre-specified number of clusters
  - We chose 3 (K= sqrt(n/2))
- K-means clustering does not require the same linearity and normality assumptions as MANOVA
- We want to group the venues based on away-team performance
  - Can venues be grouped by their specific difficulty to play in?

Cluster Classification

Cluster	1	2	3
Illinois	0	0	3
Iowa	0	1	2
Indiana	0	0	3
Maryland	0	0	3
Michigan State	0	0	3
Minnesota	0	3	0
Nebraska	0	3	0
Northwestern	0	3	0
Ohio State	3	0	0
Penn State	3	0	0
Purdue	3	0	0
Rutgers	3	0	0
Michigan	3	0	0
Wisconsin	3	0	0

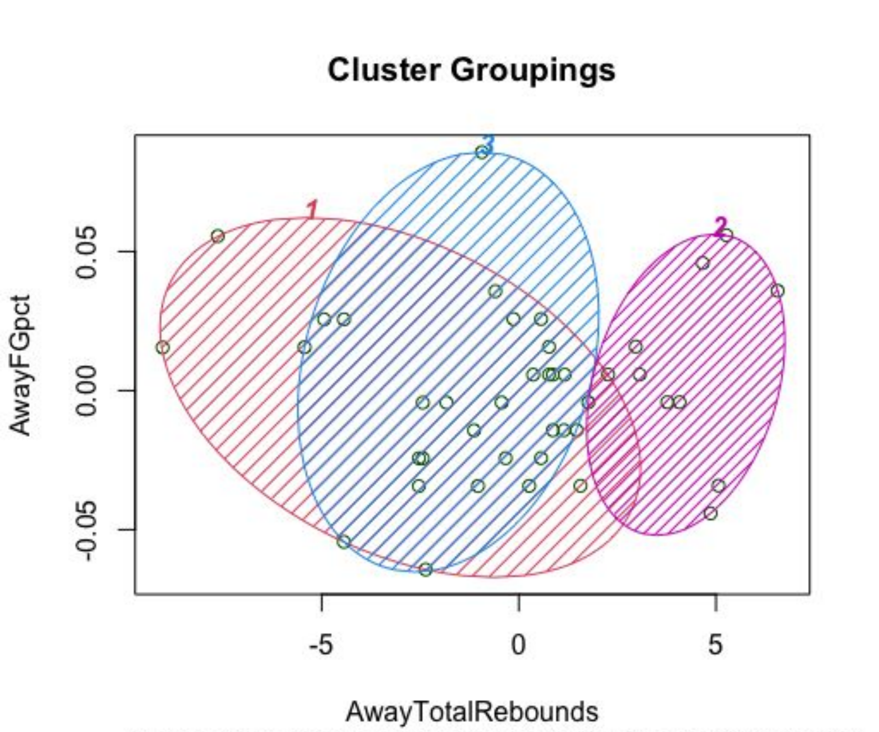
Cluster 1: Worse offense, least aggressive defensive playstyle - least steals, least rebounds, less fouls

Cluster 2: Least challenging - More assists, better shooting, better defense (blocks), least fouls

Cluster 3: Similar offensive to cluster 1 but more aggressive playstyle - Significantly more steals, rebounds and fouls

AwayAssists	AwayFGpct	Away3ptpct	AwayTotalRebounds	AwaySteals	AwayBlocks	AwayTurnovers	AwayFouls
11.91389	0.4327778	0.3383333	29.21167	5.078333	3.088889	10.22778	16.77000
15.11300	0.4410000	0.3420000	35.24100	5.624000	3.687000	11.34300	16.60600
11.88071	0.4314286	0.3357143	30.37714	5.672143	3.043571	10.51286	17.53214

Conclusions from Clustering



- Classification of clusters had decent accuracy
- Distinguished between different venues
- Differences in team performance and play-style were clearly recognized
- Team strength was a clear limitation
- Could definitely be used by schools to identify more challenging arenas
- Next step: Working to separate team strength from the effects of each venue

*Each axis scale shows distance from mean

Opportunities for Future Work

- Look into Linear Discriminant Analysis (LDA) to understand MANOVA outputs
- Consider more advanced stats, such as defensive and offensive efficiency
- Consider variables such as attendance, stadium capacity, day of the week, etc.
- Include more years in the analysis
- Look into different conferences outside of the Big Ten
- Take into account stadium changes (Nebraska built a new stadium in 2013, Maryland in 2002, etc.)
- Try to find a way to separate home team strength from the impacts of the Venue

Sources

- https://online.stat.psu.edu/stat505/book/export/html/762
- https://rpubs.com/KyleRuaya/1038546#:~:text=The%20Theory%20of%20MANOVA%20in,factors)%20of%20the%20independent%20variable
- https://www.r-bloggers.com/2021/11/manovamultivariate-analysis-of-variance-using-r/#google_vignette
- http://www.sthda.com/english/wiki/manova-test-in-r-multivariate-analysis-of-variance#google_vignette
- https://www.jstor.org/stable/2333709?seq=17
- https://cran.r-project.org/web/packages/mvnormtest/mvnormtest.pdf
- https://www.sports-reference.com/cbb/schools/ohio-state/men/2023-gamelogs.html
- https://bradleyboehmke.github.io/HOML/
- https://rua.ua.es/dspace/bitstream/10045/130341/6/JHSE_18-2_13.pdf
- https://content.iospress.com/articles/journal-of-sports-analytics/jsa200450