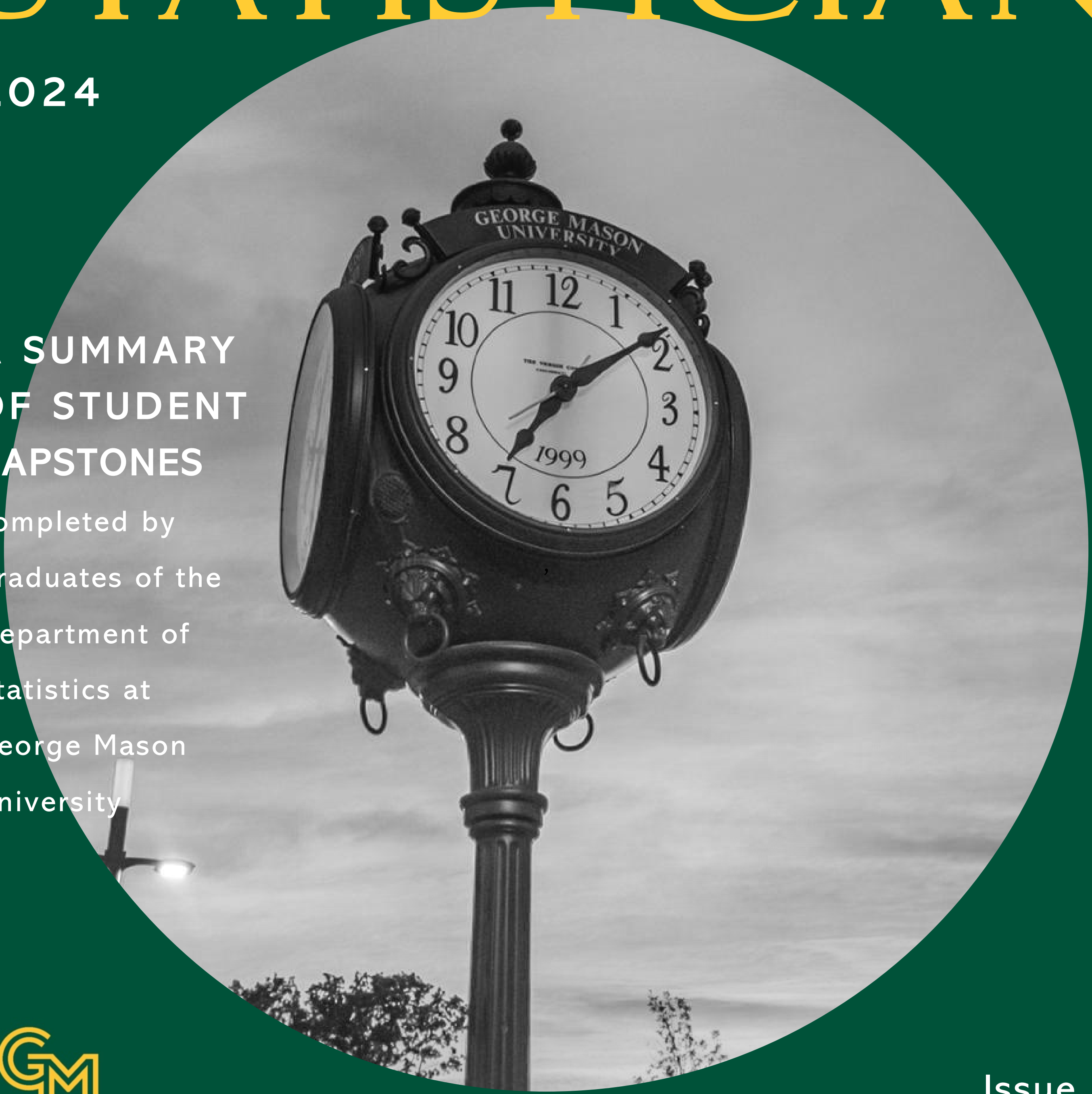
THE MASON STATISTICIAN

2024



A SUMMARY OF STUDENT CAPSTONES

completed by graduates of the Department of Statistics at George Mason University

Issue 1

edited by
Jonathan Auerbach & Charlotte Tihen
Department of Statistics
George Mason University
contact: Amir Hasan, stat@gmu.edu

TABLE OF CONTENTS

MESSAGE FROM THE CHAIR

DEAR GRADUATING CLASS,

Congratulations on completing your capstone projects and preparing for the next chapter of your life! These projects are the seeds of transformation, reflecting your hard work, problem-solving skills, intelligence, and statistical knowledge. They also demonstrate their potential impact on data science and practice, while showcasing your mentors' and professors' dedication to preparing students for the real world.

As you step into the world armed with knowledge, skills, and dreams, remember that you are in control of your career path. Mason will continue to support you as your academic family. Your statistical education is adaptable; you can grow and evolve no matter how the world changes—as long as you remain curious and inquisitive about the world around you. Remember: a statistician is defined not only by their training and statistical skills but also by a commitment to scientific inquiry coupled with presenting information with honesty and integrity.

Please take our survey to stay in touch! tinyurl.com/GMU-STAT-Exit123

Jiayang Sun
Professor, Bernard J. Dunn Eminent Scholar, and Chair

CAPSTONE ABSTRACTS

David Buxton

Objectively Grading the Baseball Swing

In my capstone, I analyzed the baseball swing of elite baseball players using objective measurements. Specifically, I wanted to find how changes in two variables, bat speed and attack angle, changed the quality of contact to maximize potential hitting outcomes. To do this I used opensource data from Driveline Baseball's OpenBiomechanics Dataset to predict batted ball outcomes (measured by xwOBAcon) with characteristics of each swing. The data was cleaned and processed by Driveline Baseball, a leader in the field of baseball biomechanics.

I used the statistic 'xwOBAcon' to assess the value of a batted ball, calculating it based on Exit Speed and Launch angle of the baseball. I then used a random forest model trained on seven metrics describing bat path as measured by a Blast bat tracking device and Computer Vision to model xwOBAcon. This model was used to predict the values of xwOBAcon of 70% of the data, and the other 30% was used to analyze the accuracy of the model.

I found that my model was able to accurately predict the value of a batted ball depending on the change in one variable, specifically bat speed and attack angle. These metrics are two of the most straightforward and easily measurable of all metrics in hitting and directly correlate to the effectiveness of a swing. Utilizing these plots, we can see that expected performance goes up as bat speed does to a certain degree and that the ideal attack angle for hitters usually sits around eight degrees but could be different depending on other characteristics.

Why did you choose statistics?

I chose to major in statistics because I wanted to choose a major that applies to all job fields and will help me make decisions and advance in my career in whatever field I choose. I also find that the statistics courses have many opportunities to challenge myself in theory of math and learn coding while also grounding my learning in real world problems.

Acknowledgments

I would like to thank Dr. Kenneth Strazzeri for welcoming me and encouraging my learning when I first arrived as a transfer student to George Mason. I would also like to thank Dr. Hunter, Dr. Izmirli, Dr. Kepplinger and Dr. Auerbach for the support of my learning objectives and making sure I was given every opportunity to succeed and grow in and out of the classroom.
I would ultimately like to thank my wife, Ali, for her support of me and sacrifices she made in my pursuit to be the best student I could be and lay the groundwork for our future.

Julian Christov-bakargiev

Analyzing Soccer Match Results Using Transfer Values

My investigation explores whether players' transfer values predict soccer match outcomes better than traditional Elo ratings. I focused my investigation on World Cup knockout stage matches from 2006 to 2022, and the collected data include the transfer values and ages of starting players. I then combined the data with the match outcomes to construct a series of Bayesian logistic regression models. I used the models to assess the relationship between transfer value, age, and match outcomes, aiming to provide insights into the effectiveness of using transfer values in predicting World Cup match results.

The analysis primarily relied on Bayesian logistic regression models to assess the relationship between players' transfer values, ages, and the outcomes of World Cup knockout stage matches. This method accounted for uncertainty inherent in the data and provided a framework to evaluate the predictive power of transfer values compared to traditional Elo ratings. Additionally, Leave-One-Out

Cross-Validation (LOO-CV) was employed to assess the predictive performance of the models and to ensure robustness by simulating predictions on unseen data points iteratively.

The findings indicate that incorporating players' transfer values into predictive models for World Cup knockout stage matches can significantly enhance predictive accuracy compared to traditional Elo ratings. The Bayesian logistic regression models leveraging transfer values and player ages demonstrated superior performance, achieving a success rate of 92.5% in predicting match outcomes. These results suggest that transfer values offer valuable insights into teams' potential performance, highlighting the importance of considering economic indicators alongside traditional performance metrics in soccer match prediction models.

Why did you choose statistics?

I initially chose statistics because my parents thought it would be a good fit for me. I'm now drawn to statistics for its emphasis on mathematics, problem-solving prowess, real-world relevance, and the ability to approach problems from various angles, much like the flexibility and creativity found in art.

Acknowledgments

I would like to thank my parents and all my teachers.

Victoria Gonzales

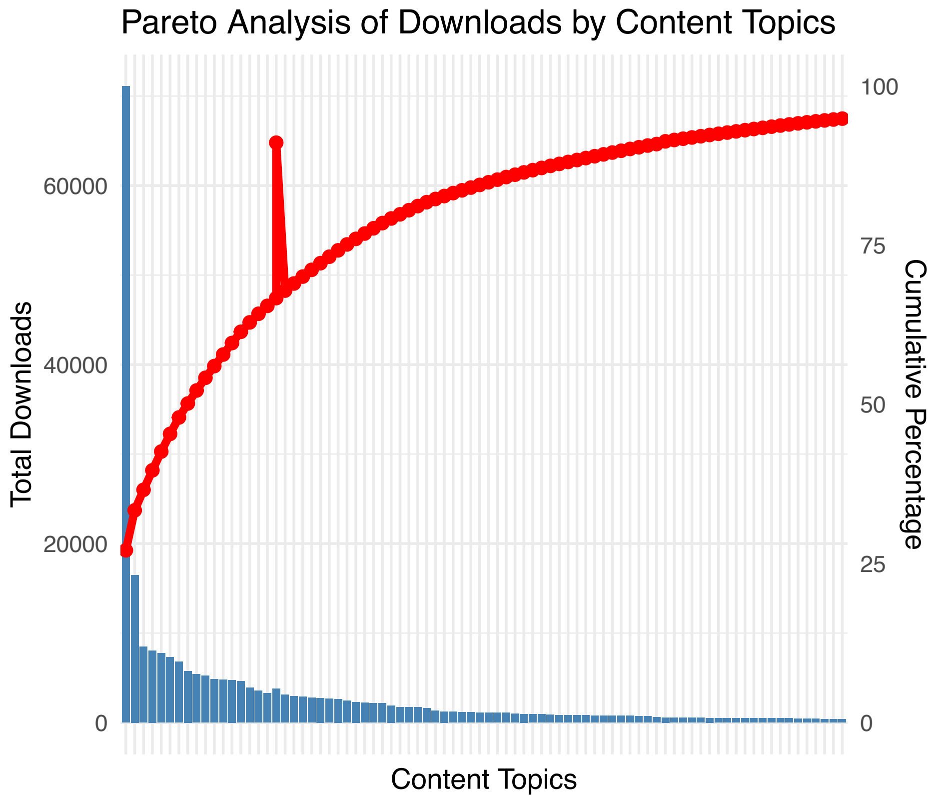
Optimizing the Location of Data on a Statistical Agency Website

In my capstone, I investigated a statistical agency's website which provides downloadable information. The objective was to understand how the content's location on the website affects the accessibility of information and the user's experience.

The data used in the analysis included the number of downloads and the actions taken to a user to navigate to the location of the downloadable content. Pareto analysis was used to determine the top content of interest based on number of downloads. Linear regression was used to find the correlation between downloads and location of content.

I found that 82 out of 287 content topics were downloaded the most. Moreover, the number of downloads and their location on the website were correlated. I conclude that there are opportunities to optimize the accessibility of information by prominently featuring popular content on the agency's website.

Figure: 80% of website downloads are for 20% of content



Why did you choose statistics?

I chose to major in statistics because it is the foundation of modern decision making and research. The application of statistics is endless, and I found the versatility of it fascinating. I like that statistics is an avenue for analytical thinking. Statistics assist to provide clarity when there is uncertainty, and I think that is very beneficial in a world hungry for knowledge.

Acknowledgments

I would like to thank my family for all of their support.

Erick Guevara

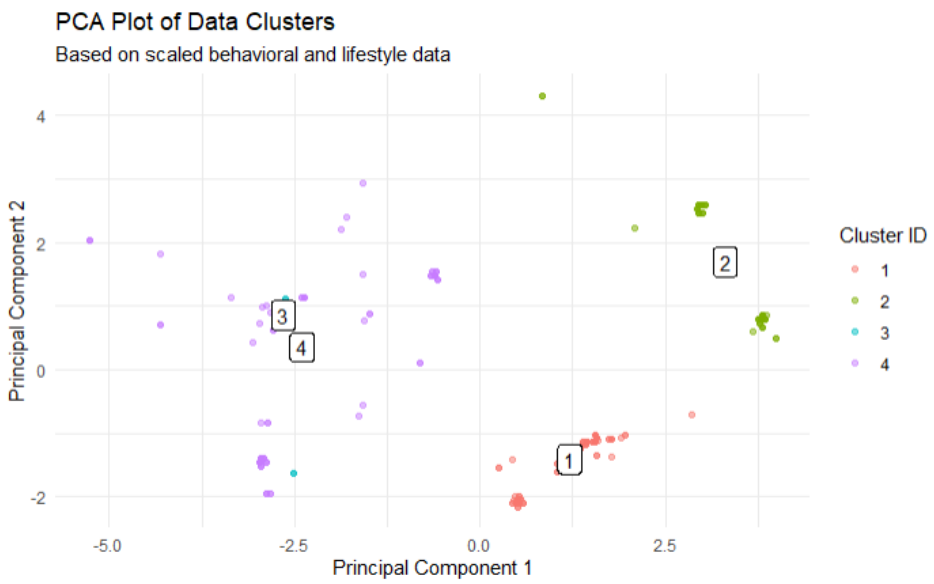
Assessing Risk Factors of Sleep Disorders

I examined whether there was a combination of behavioral risk factors that affected sleep more than other factors. The data is from the behavioral risk factor surveillance system, collected to keep track of risk factors from the population by conducting surveys. The data includes lifestyle choices, behavioral factors, and sleep data from more than 300 people.

I used cluster analysis to group individuals based on their risk profiles and identified which combination of groups are at the highest risk of developing a sleep disorder. I then used multinomial logistic regression analysis and random forest to understand the risk score outcome between factors in groups that might correlate to a specific sleep disorder.

I found that certain occupations like doctors, lawyers, and nurses cluster up mostly with variables like sleep quality, duration, and stress. Those occupations have a positive predictive probability of developing insomnia. There are other occupations where a sleep disorder isn't present; however, their BMI, stress level, physical activity, and sleep duration is presented as negative factors which can have a potential development in a sleep disorder.

Figure: Individuals are grouped based on their risk profiles



Why did you choose statistics?

I enjoy mathematics, particularly applying formulas and other methods I have learned. Observing patterns and analyzing data is one of my strengths, which is why I decided on majoring in statistics. I look forward to applying the tools I have learned in the future.

Acknowledgments

I would like to thank my professors for keeping me on track to graduate over the years and for making statistics enjoyable.

Jenifer Nguyen

Exploring Regional Variations in Mental Health

In my project, I looked at how mental health disorders differ across regions by comparing the percentage of people with these disorders in various countries. I grouped countries by region and analyzed the data to see the patterns of mental health issues in different parts of the world.

I used spatial analysis to determine how anxiety disorders and depression vary across regions worldwide. I analyzed the data by mapping the percentage of people with anxiety disorders and depression across different regions to identify areas with both high and low prevalence rates. This method allowed me to understand the geographic differences of various mental health disorders, explore potential contributing causes, and identify any similarities between them.

My analysis revealed that Africa has the highest rate of people with depression, whereas South America has the highest rate of people with anxiety disorders. This suggests socioeconomic status, cultural background, and access to healthcare are key determinants. Further exploration may yield insight into why certain regions have higher prevalence rates of specific mental health disorders. This insight would enable targeted efforts and resources to address the underlying causes of mental health differences.

Why did you choose statistics?

I chose to major in statistics because I enjoy working with data and want to learn how to analyze various datasets effectively. I was particularly drawn to the field’s ability to reveal valuable insights from a wide range of topics through statistical methods. Statistics provides a solid foundation for understanding uncertainty and making predictions about the topics I'm studying, but it also covers various data visualization methods. I'm drawn to data visualization because it allows me to present complex findings clearly and engagingly.

Acknowledgments

I am extremely grateful to my family for their constant support and encouragement throughout my academic journey. A sincere thank you to my mentor, Jim, whose wisdom and guidance have profoundly influenced my learning experience. And to my statistics buddy, Chris, whose support keeps me on track with attending classes and completing assignments. I'll miss the structured environment that keeps me motivated to learn, as well as the interactions with classmates and professors who share my interests. Additionally, I’ll miss studying at the Fenwick Library, where I can be productive and focused.

Spencer Retcher

Improving Campus Safety through Officer Evaluation

I investigated the reporting time of incidents to Mason Police. Specifically, I aimed to determine whether incidents were reported promptly after they occurred and if there were any significant differences in reporting time among officers. To address these questions, I collected incident reports from 2016 to March 2024 from daily crime and fire logs. I then measured the reporting time for each report by analyzing the duration between report and incident timestamps.

I used the Wilcoxon Signed Rank Test to determine whether the median reporting time was less than one day, indicating a prompt response to incidents. Following that, I employed the Kruskal Wallis Test to explore potential differences in reporting time among officers, which could provide insights into variations in performance. The Wilcoxon Signed Rank Test results ($W=1925739$, $p=\sim 0$) indicate that there is enough evidence to conclude that the overall median reporting time is less than one day. Furthermore, the Kruskal Wallis Test results ($H=85$, $p=0.48$) revealed that there are no significant differences in reporting time between Mason police officers. This suggests that Mason police officers tend to report incidents promptly, typically reporting incidents on the same day they occur. This guarantees that incidents are promptly addressed, fostering safety and adherence to the Clery Act.

Figure: Summary of data collected

Summary Statistics for Reporting Time (Days)	
Number of Reports	3984.000000
Mean	5.114373
Std	38.287602
Min	0.000000
25%	0.000000
50%	0.000000
75%	0.093924
Max	836.221181

Median Reporting Times by Officer		
Median Reporting Time	Officers	Total Reports
5+ days	Jackson, Unknown	143
1-2 days	Green, Surber, Tarad	6
0-1 days	Baucom, Cruz, Dean, Johnson, Ly, Spano	65
0 days	75 Officers	3,770

Wilcoxon Signed Rank Test	
W Statistic	1925739.0
P-Value	~0
Alpha	0.05

Kruskal Wallis Test	
H Statistic	85.0
P-Value	0.48
Alpha	0.05

Why did you choose statistics?

I chose to major in statistics because I've always been fascinated by the power of data. Statistics provides the tools and techniques to analyze and interpret data, allowing me to understand patterns, make predictions, and solve real-world problems.

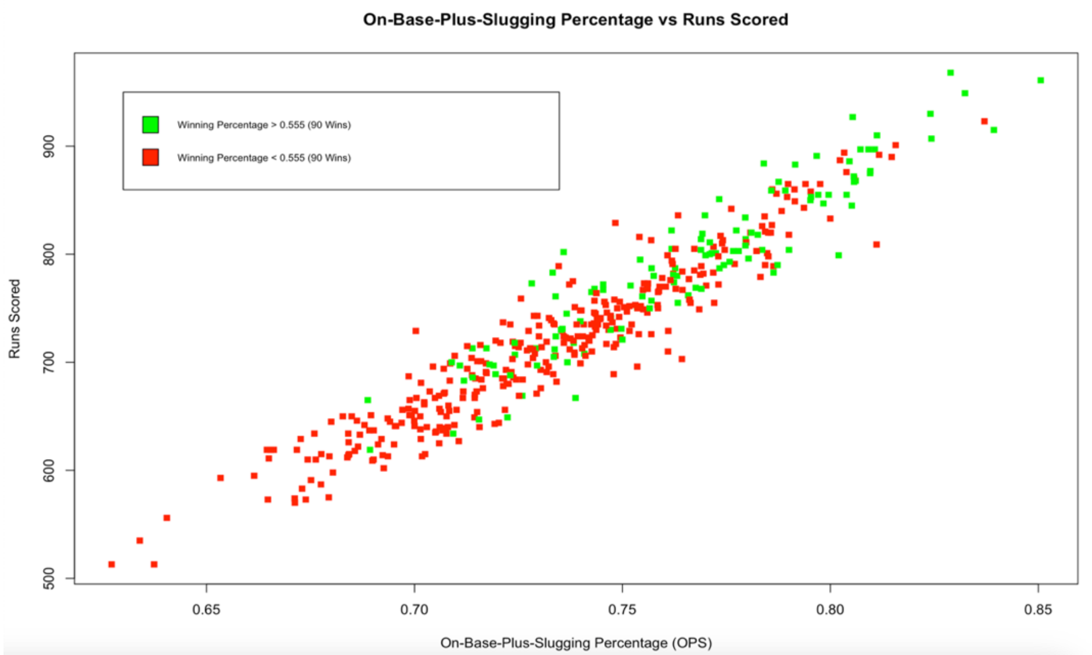
Constantine Tzoumas

Different Statistics Correlated to Scoring Runs in Major League Baseball

I investigated different hitting metrics and how they correlated to scoring runs in baseball. To conduct my investigation, I obtained publicly accessible data on baseball-reference and baseball savant to investigate. I then used random forest to determine which hitting statistic was most important in predicting how many runs a team would score during the season. I also used random forest to find the most important individual hitting statistics that contribute to On-Base-Plus-Slugging (OPS) percentage. Finally, I examined whether the previous two seasons' individual hitting statistics could predict the next season's hitting statistics.

I found that OPS was the most important factor in determining how many runs a team scores throughout the season. The most important statistics in predicting OPS were On-Base Percentage, Slugging Percentage, Batting Average, Barrel Rate, and Walk Percentage. It is important for teams to look for players who excel in these statistics when building a roster. I also found that the previous seasons' statistics did not predict next season's statistics.

Figure: An example relationship between baseball metrics



Why did you choose statistics?

I chose statistics as my major because I have always had a skill in math and a passion for sports. I have enjoyed the challenging yet rewarding coursework of the statistics major. The real world investigations have been my favorite part of my major.

Acknowledgments

I would like to thank my teachers for giving me a great education and always pushing me to succeed. I would also like to thank my parents for always supporting me and being great models of hard work and success.

William Gene Warren

College Basketball: The Effect of Team Conference on Performance

I investigated whether a team's particular conference had an effect on a team's individual statistic. To answer this question, I gathered data from the RealGM website from 2014 to 2024 using a web scraper. The data contained all of the major basketball statistics for each of the 350 teams. (See table on opposite page.) I then used MANOVA, ANOVA, and linear regression to test if a team's conference is a significant predictor of their individual statistic.

Specifically, I first ran the MANOVA test to see if conference was a significant predictor compared to all of the other variables contained in the study. Next, I ran individual ANOVA tests to identify every variable for which conference was a significant predictor. Lastly, I ran individual linear regression tests to see which particular conferences were significant for each variable.

I found that conference is overall a significant predictor of team's individual statistics (p-value 2.2e-16). I also found conference to be an individually significant predictor of all of the variables contained in this study at the 0.001 significance level. Lastly, linear regression revealed that useful insight can be attained from an individual conference's play style. Overall, I conclude that conference is a significant predictor of team statistics and that the playing field in men's college basketball may not be balanced.

Table: An example of data scraped from the RealGM website

	Team	GP	MPG	PPG	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	ORB	DRB	RPG	APG	SPG	BPG	TOV	PF	Conf	Year
1	Xavier	37	40.1	80.9	30	61	0.492	7.4	19	0.39	13.6	19.1	0.71	9.2	26.5	35.6	19.1	6.4	3.2	12.2	16	59	
2	Marquette	36	40.5	79.3	29.3	60.4	0.486	8.9	25.3	0.353	11.7	16.3	0.718	8	22	29.9	17.3	9.2	3	10.7	16.2	59	
3	UConn	39	39.8	78.6	27.7	59.8	0.464	9.1	24.9	0.363	14	18.4	0.761	12.2	25.6	37.7	17.5	6.3	4.9	12.6	17.7	59	
4	St. John's	33	40.5	77.3	29.4	65.4	0.449	5.8	17.7	0.331	12.7	18.2	0.699	12.2	26.3	38.5	15.2	8.2	3.7	13.7	16.2	59	
5	Providence	33	40.7	77.3	27.3	59.8	0.456	6.6	19.1	0.346	16.1	21.7	0.742	11.1	25	36.1	14.6	6.8	4.5	11.3	15	59	
6	Creighton	37	40.2	76.4	27.4	58.5	0.469	8.6	24.4	0.354	13	16.6	0.783	7.6	27.5	35.1	15.7	5	4.1	11.2	13.1	59	
7	DePaul	33	40.6	71.2	25.4	59.4	0.427	8.3	22.5	0.368	12.1	16.4	0.739	8.2	22.3	30.4	13.8	6.4	4.7	12.7	18.3	59	
8	Villanova	34	40.1	69.6	23.7	54.5	0.436	8.7	26	0.334	13.4	16.4	0.82	6.4	23.6	29.9	11.1	5.7	1.7	9.7	15.8	59	
9	Georgetow	32	40.1	69.5	25.6	60.4	0.423	5.8	18.7	0.309	12.6	17.6	0.718	10	24.1	34.1	11.9	6.2	4.6	12	14.4	59	
10	Seton Hall	33	39.7	68.4	23.8	54.1	0.439	5.8	17.7	0.325	15.1	22	0.686	9.5	23.6	33	12.2	8.2	4.3	13.3	17.8	59	
11	Butler	32	39.8	65.3	24.1	55.2	0.436	6.5	19.8	0.327	10.6	14.3	0.739	6	22.5	28.5	12.6	6.8	4.3	11.5	13.4	59	
1	Marshall	32	40.4	81.8	30.8	66.3	0.465	8.3	24.4	0.338	11.9	17.1	0.693	11.4	25.3	36.8	17.5	9	5.3	11.5	16.4	33	
2	James Mad	33	40.5	80.5	28.3	61.2	0.462	8	22.8	0.35	16	22.7	0.707	11	24.9	35.9	13.5	9	2.6	12.9	17.7	33	
3	Louisiana	34	40.1	77.4	27.9	58	0.482	7	18.7	0.376	14.4	21.2	0.682	10.1	24.5	34.6	14.4	7.2	3	12	17	33	
4	Southern M	33	39.8	73.8	26.7	58.7	0.455	7	21.6	0.322	13.5	18.8	0.714	10	24.4	34.4	14.8	8.4	2.6	12.1	16.5	33	
5	Coastal Ca	31	40.3	73.3	26.3	60.6	0.434	7.2	22.7	0.316	13.5	17.2	0.784	10.3	24.7	35	12.4	6.1	3.2	12	16	33	
6	Troy	33	40.3	73.2	26	58.9	0.441	7.7	24.2	0.32	13.4	19.2	0.701	9.9	23.2	33.1	13.5	8.7	2.6	12.6	16.9	33	
7	South Alab	35	40.1	70.9	27.3	58.2	0.469	7.6	22	0.345	8.8	12.8	0.687	6.2	24.6	30.7	11.4	6.1	4.3	9.5	14.6	33	
8	Appalachia	32	40.2	70.3	25.6	57.9	0.442	7.6	22.7	0.335	11.5	17	0.676	8.1	25.5	33.6	13	6.4	4.6	10.5	13.8	33	
9	Georgia So	33	40.1	68.4	25.5	58.4	0.436	6.3	20	0.314	11.2	16.3	0.686	8.8	24.3	33.1	10.4	7	2.8	11	16.7	33	
10	Louisiana-M	32	40.3	67.4	23.7	58.1	0.408	7.6	21.6	0.351	12.4	17.4	0.713	9.8	23.4	33.2	12.6	6.9	3.6	12.2	15.1	33	
11	Old Domini	31	39.9	66.9	23.7	56.8	0.418	5.3	16.9	0.314	14.2	20.8	0.68	11	24.6	35.6	10.6	6.4	3.8	11.1	16.7	33	
12	Georgia Sta	31	40.1	66.7	24	57.7	0.416	5.3	17.9	0.295	13.4	18.7	0.713	10.2	22.5	32.7	11.1	5.8	3.6	12.6	17	33	
13	Texas State	35	40	66	24.2	54.8	0.442	4.4	14.3	0.307	13.1	18	0.729	9.4	21.3	30.8	9.9	6.3	3.5	11.2	18	33	
14	Arkansas St	33	40.1	65.7	23.5	54.5	0.432	6.3	18	0.352	12.3	18.2	0.677	9.3	23.1	32.4	13.1	6.7	3.3	11.3	16.8	33	
1	Oral Robert	35	40	83.3	30.2	63.7	0.474	10.7	29.3	0.365	12.1	15.7	0.774	8.2	27.1	35.3	14.7	6.8	5.3	9.2	12.9	32	
2	St. Thomas	33	39.9	74.2	25.7	57.4	0.447	9	25.6	0.352	13.8	17.6	0.785	7.6	22.5	30	12.9	5.9	2.3	9.8	17.4	32	
3	Western Illi	30	40.2	73	26.5	58.5	0.453	7	21.7	0.321	13.1	18.7	0.7	8.7	23.7	32.4	13.7	5.8	3.6	10.4	14.5	32	
4	North Dako	33	39.8	72.8	26.2	57.7	0.455	7.5	22.3	0.336	12.9	18.1	0.712	7.9	26.6	34.5	11.2	3.9	3	11.4	15.4	32	
5	Denver	32	40.6	72.7	26.1	53.6	0.487	4.5	13.8	0.326	16	22.3	0.717	8.4	25.3	33.6	11.3	5.5	3.1	14.7	17	32	
6	North Dako	33	39.7	72.1	26	58.9	0.441	9.2	26.3	0.35	10.9	16	0.683	8.4	23.4	31.8	11.9	4.4	2.7	11.3	15.8	32	
7	South Dakc	32	40.2	70.9	25.4	56.4	0.45	7.7	21.9	0.351	12.5	15.8	0.792	6.1	24.5	30.6	10.7	5.2	2.9	11.3	14.3	32	
8	South Dakc	31	39.9	69.3	24.4	56.5	0.431	8.9	22.8	0.39	11.7	15.4	0.765	7.7	24.2	31.9	12.4	4.5	1.5	11.5	17.4	32	
9	Omaha	32	40	68.5	24.9	56.4	0.442	5.7	16.9	0.339	13	16.7	0.776	7.5	23.4	30.9	11.9	5.4	2.9	12.3	18.3	32	
10	UMKC	32	40.4	64.4	22.6	57.6	0.392	5.9	20.8	0.283	13.3	18.7	0.712	12.2	22.7	34.9	9.4	6	3.7	13	18.3	32	
1	Southern	32	40.3	71.9	25.1	58.5	0.429	7.6	23	0.331	14	19.6	0.715	8	23.2	31.2	14.1	8.7	2.3	14.2	20	31	
2	Alabama A&	33	40.2	69.6	24.9	56.6	0.441	6.8	18.5	0.37	12.9	19.7	0.655	9.3	21.8	31.2	12.7	8	4	14.2	18.9	31	
3	Texas Soutl	35	40.7	69.2	25.5	59.9	0.427	4.7	16.8	0.277	13.5	20.3	0.664	10.7	24.7	35.3	11.8	6.7	3.3	14.3	18.8	31	

Why did you choose statistics?

I chose statistics because I was always very interested in data, in particular sports data. I took my first statistics class my senior year of high school, and I fell in love with the subject. Once I met Dr. Johnson my freshman year, I knew that this is what I wanted to pursue.

Acknowledgments

I would like to thank my parents for their support throughout my time in college. I would also like to thank Dr. Elizabeth Johnson for helping convince me to pursue a career in statistics and take up the major.

PREDICTION COMPETITION

When will the cherry trees bloom?

Mason students developed statistical models to predict the day that the cherry trees would first bloom in Spring 2024 at five locations around the world: Washington D.C., USA; Kyoto, Japan; Vancouver, Canada; and Liestal-Weideli, Switzerland.

Students share findings with the press

In the image below, students explain their models to a news reporter

while the cherry trees at Mason Pond bloom in the background. They also share their thoughts on the impact of climate change. The story was featured by several news outlets, including WUSA9, Washington DC's local news channel.

AWARDS & RECOGNITION

Statistics Faculty Award presented to William Gene Warren

for outstanding achievement as an undergraduate student in statistics as determined by vote of the Department of Statistics faculty.

Senior Leadership and Service Award presented to Nafeez Chowdhury

for leadership and service to the Department and wider community as determined by vote of the Department of Statistics faculty.

OSCAR Student Excellence Award presented to Victoria Gonzales

in recognition of outstanding undergraduate students who participate in research and creative activities at George Mason University as determined by the Office of Student Creative Activities and Research from nominations made by Mason students, faculty, and staff.

STATISTICS AT MASON

UNDERGRADUATE PROGRAMS AT MASON

Undergraduate students majoring in statistics can pursue different concentrations as they work towards completing their degrees.

Statistical Analytics

This concentration blends computer science and statistics at the undergraduate level, two core disciplines of data science.

Applied Statistics

This concentration focuses on analytical methods applicable to a specific discipline of the student's choosing.

Mathematical Statistics

This concentration focuses on the theoretical underpinnings of statistics, preparing students for research and graduate study.

Sports Analytics

This concentration prepares students to work with sports teams and related, where data skills are increasingly in demand.

GRADUATE PROGRAMS AT MASON

There are several graduate degrees offered by the Department of Statistics. These programs can be supplemented by concentrations and certificates to create a unique graduate experience.

MS in Statistical Science

This degree program prepares students for statistics and data science occupations in industry and government. Qualified undergraduate students may obtain an Accelerated Statistical Science MS.

MS in Biostatistics

This degree program provides a similar training to the MS in Statistical Science, but it allows students to specialize in the design and analysis of health-related and biological studies.

Statistical Data Science Concentration

This concentration blends computer science and statistics at the graduate level, two core disciplines of data science, preparing students for the analysis of complex data sets.

Federal Statistics Graduate Certificate

This certificate is designed for current practitioners who wish to update their skills in statistics, survey methods, and data analysis, including graphics and data visualization.

PhD in Statistical Science

This degree program is a hybrid of probability, computation, and data analysis.

Department of Statistics
College of Engineering and Computing
Nguyen Engineering Building, Suite 1705
Fairfax, Virginia 22030
703-993-3645