MSSP Portfolio

Jing Wu

Department of Statistics
Boston University

Contents

1 Analysis of The Brazilian Population In The US

Partner Project ..... Fall 2022

2 Lung Artery and Capillary Dynamics in Mice

Consulting Project ..... Fall 2022

3 Student Judges Performance Evaluation Project

Consulting Project ..... Spring 2023

4 Ice hockey stroke action data prediction project

Partner Project ..... Spring 2023

Analysis of The Brazilian Population In The US

Jing Wu

Abstract

Consulting - Counseling, A group from the City of Boston asked our team to use the data they have available to find insights on the challenges involved and the progress that has been made with the homelessness situation in Boston. To isolate this particular situation, we looked into one specific area of the city where a lot of homeless people relocated after a closing of a shelter, and then eight months later a new shelter was opened.

1 Introduction

Massachusetts is home to some of the most densely populated counties of Brazilian immigrants in the US. However, there is limited information available on the characteristics of Brazilian immigrants and their well-being since the early 2000s. In this project, we examine the lives of Brazilian immigrants in the US using the American Community Survey (ACS).

The ACS is a demographic survey program conducted by the US Census Bureau, responsible for gathering data about the American people and economy. The survey collects information on various factors such as ancestry, citizenship, educational attainment, income, language proficiency, migration, disability, employment, and housing characteristics. Our main objective is to understand the relationship between income and how factors like language, citizenship status, and educational attainment impact the income of Brazilian immigrants. To achieve this, we focused our research on three states with the most prominent Brazilian population: Florida, Massachusetts, and California. Additionally, we compared Brazilian immigrants with immigrants from ten other South American countries, including Venezuela, Argentina, Chile, Guyana, Colombia, Bolivia, Paraguay, Ecuador, Uruguay, and Peru. Furthermore, we developed a Shiny app to offer a user-friendly interface that allows users to build personal exploratory data analysis without the limitation of professional skills in statistics and coding.

2 Data and Method

The data for this project was obtained from two main sources: the American Community Survey (ACS) and the Public Use Microdata Series (PUMS) API. The ACS is a survey conducted by the US Census Bureau that collects information on social, economic, and housing characteristics of individuals and households across the United States. The survey covers a wide range of topics and produces estimates at different levels of geography, from census tracts to the national level.

The PUMS API provides access to individual-level data from the ACS, which can be used to generate customized tabulations and analyses.

To compare the characteristics of Brazilian immigrants to those of other South American immigrants in the United States, we used PUMS data from the ACS for the years 2005-2019. We obtained data for each country of origin by specifying the appropriate parameters in the get_pums() function of the tidycensus R package. The data were then processed and cleaned to account for sampling weights and replicate weights, which are used to estimate sampling errors and generate standard errors.

The analysis of the data involved two main components: exploratory data analysis (EDA) and statistical inference. The EDA focused on visualizing the distribution and composition of various variables across different years and countries of origin. We used the R programming language and the ggplot2 package to create time-series plots and bar charts that allowed us to examine changes over time and across groups. The statistical inference involved testing hypotheses and making inferences about the population based on the sample data. We used R and the survey package to calculate point estimates, standard errors, and confidence intervals for various statistics of interest, such as means, proportions, and regression coefficients.

Overall, the data and method used in this project allowed us to explore and compare the characteristics of Brazilian immigrants and other South American immigrants in the United States over time. The use of survey data and statistical techniques ensured that our results were representative of the larger population and accounted for sampling errors. The Shiny app provided an interactive platform for users to explore and analyze the data themselves, making it a valuable tool for researchers, policymakers, and anyone interested in understanding the social and economic dynamics of immigration in the United States.

3 Results and Discussion

The aim of this study was to examine the lives of Brazilian immigrants in the US using the American Community Survey (ACS) and to understand the relationship between income and factors such as language, citizenship status, and educational attainment. The results revealed that Massachusetts is home to some of the most densely populated counties of Brazilian immigrants in the US, and the Brazilian population's median income has Granger Causality effects on the change of citizenship proportion. This suggests that higher income may attract more Brazilian immigrants to become naturalized citizens, and further research is needed to investigate this assumption.

The study also compared Brazilian immigrants with immigrants from ten other South American countries and found that Brazilians have middle personal earnings levels, lower educational attainment, and a higher percentage of the population without U.S. citizenship in Massachusetts, Florida, and California. The income regression analysis showed that sex was a significant indicator of Brazilian personal earnings in these states in 2021. Additionally, the exploratory data analysis revealed that the population of Brazilians was higher than that of Argentinians and Chileans from 2005-2021, and Brazilians had more owner-occupied units than

these two groups. Furthermore, Venezuela experienced household growth since 2014, and Brazilians had fewer family households than Venezuelans in 2014-2021.

Overall, this study provides important insights into the lives of Brazilian immigrants in the US and highlights the factors that impact their well-being, particularly in terms of income and citizenship. The findings suggest that policies aimed at improving the economic opportunities for Brazilian immigrants, such as improving their access to education and employment, may contribute to their overall well-being and integration into American society.

4 Conclusion

In conclusion, this project provides insight into the lives of Brazilian immigrants in the United States and their relationship with factors such as income, education, and citizenship status. Our analysis showed that Brazilian immigrants in Massachusetts have a significant effect on citizenship proportion, indicating that higher income may attract more naturalization. Furthermore, we found that Brazilians have middle personal earnings levels compared to other South American countries, with lower rates of educational attainment and higher rates of non-citizenship. The regression analysis also revealed that sex plays a significant role in Brazilian personal earnings in Massachusetts, Florida, and California.

Overall, this project highlights the importance of understanding the factors that influence the lives of immigrants in the United States. By providing a user-friendly interface through our Shiny app, we hope to encourage further research and analysis on the experiences of Brazilian immigrants and other immigrant populations in the country. Our findings can also inform policymakers and other stakeholders about the unique challenges and opportunities faced by immigrants in the United States, particularly those from Brazil.

Lung Artery and Capillary Dynamics in Mice

Jing Wu

Abstract

This consulting project aimed to analyze the relationship between capillaries and arteries, blood pressure, and air pressure in the lungs of mice, as well as to explore the differences between capillaries and arteries and the ratio of length and width of cancer cells in the capillaries. The client's concern was how pressure affects the diameter of arteries and capillaries.

1 Introduction

The client requested an analysis of the relationship between capillaries and arteries, blood pressure, and air pressure in the lungs of mice, as well as differences between capillaries and arteries and the ratio of length and width of cancer cells in the capillaries. The client's concern was how pressure affects the diameter of arteries and capillaries. The sample size was small, with six lungs from six different mice, and 192 data points were recorded.

2 Data and Method

The client recorded 16 data points for each capillary and artery in one lung, combining four levels of air pressure and four levels of blood pressure, resulting in 192 data points for the first and second problems. The client performed an ANOVA test on the data. However, the correlation issue may affect the standard error. The consulting team recommended a multilevel model, which is probably more appropriate than ANOVA for the client's needs since there were repeated measurements within the same artery (or capillary) and within the same mouse. To study the deformation about the cell, the client pooled the data over all of the pressure levels since they looked similar, but the consulting team suggested doing a t-test to compare the data between two groups at the same pressure level.

3 Results and Discussion

The consulting team met with the client to discuss her data analysis and identified some faults that can be improved. The team recommended additional data collection to increase the sample size and improve the statistical power of the analysis. However, the client declined this recommendation. Therefore, the team provided some advice to improve the analysis based on the available data. The team found a quadratic relationship between air pressure and diameters. The consulting team found a positive correlation between blood pressure and air pressure in the lungs

of mice, with a stronger relationship in capillaries than in arteries. Additionally, the team identified some differences in properties between arteries and capillaries, with capillaries having a higher density and a smaller diameter than arteries. Lastly, the team found that the ratio of length and width of cancer cells in the capillaries changed over different levels of air pressure and blood pressure.

4 Conclusion

In conclusion, this consulting project aimed to analyze the relationship between capillaries and arteries, blood pressure, and air pressure in the lungs of mice, as well as to explore the differences between capillaries and arteries and the ratio of length and width of cancer cells in the capillaries. The consulting team recommended a multilevel model, which is probably more appropriate than ANOVA for the client's needs since there were repeated measurements within the same artery (or capillary) and within the same mouse.

Student Judges Performance Evaluation Project

Jing Wu

Abstract

The Boston Debate League (BDL) is a non-profit organization aimed at integrating argumentation and competitive debate into Boston public schools, developing critical thinkers ready for college, career, and engagement with the world. For the past five years, the Rhetoric Department at a local university has sent students to volunteer as judges at BDL's tournaments. In this project, we aim to understand the impact of student-judges on the debate performance of middle and high school students. Through surveys and data analysis, we investigate whether student-judges are viewed favorably by participants and if there is a difference in attitudes towards them between middle and high school students. Our findings suggest that student-judges are generally preferred, and there is no significant difference in attitudes towards them between middle and high school students.

1 Introduction

Rhetoric is the study and art of writing and speaking well, being persuasive, and knowing how to compose successful writing and presentations. In the past five years, students from the Rhetoric Department at a local university have volunteered as judges at the Boston Debate League's tournaments. The Boston Debate League is a non-profit organization aimed at integrating argumentation and competitive debate into Boston public schools, developing critical thinkers ready for college, career, and engagement with the world. This project aims to investigate the impact of student-judges on the debate performance of middle and high school students. We are interested in whether student-judges are viewed favorably by participants and if there is a difference in attitudes towards them between middle and high school students.

2 Data and Method

To investigate the impact of student-judges on debate performance, we administered surveys at four different tournaments, two middle school ones and two high school ones. The surveys included questions about the participants' preferences for judges, the quality of judges they experienced, and other subjective questions. To analyze the data, we built proportional odds models to determine the probability that student-judges were preferred. We used bootstrapping to generate confidence intervals and reduce the effect of "0" when comparing preferences. We also built proportional odds models with categorical variables representing the data from middle and high schools to investigate if school level affected attitudes towards student-judges.

3 Results and Discussion

Our findings suggest that student-judges are generally preferred, with about a 41.64% probability that they are preferred and a 7.45% probability that they are not preferred. The confidence interval generated by bootstrapping indicates that this result is statistically significant. However, we found no significant difference in attitudes towards student-judges between middle and high school students. The 95% confidence interval for the coefficient representing school level included "0," indicating that the school level did not significantly affect the evaluation of student-judges.

We also investigated other factors that may impact the evaluation of student-judges. Future analysis could consider adding other covariates about participants to the model, such as age or previous debate experience. Moreover, we recognize that our analysis is limited by the number of participants and the scope of the survey questions. A more in-depth study may include focus groups and interviews to better understand the social and emotional experiences of service-learning beneficiaries.

4 Conclusion

In conclusion, this project aimed to investigate the effectiveness of student-judges in the Boston Debate League, as well as the potential differences in attitudes towards student-judges between middle and high school students. Through the use of surveys and statistical analysis, we were able to draw some key insights.

Firstly, we found that student-judges are generally well-received by debaters, with a 41.64% probability that they are preferred. This suggests that they can be effective in their role as judges, and may provide a valuable learning experience for the student-judges themselves.

Secondly, we found that there is no significant difference in attitudes towards student-judges between middle and high school students. This indicates that student-judges can be utilized effectively in both middle and high school debates, without concern for any potential negative impact on the debaters' experiences.

Overall, this project provides valuable insights for the Boston Debate League and other organizations that utilize student-judges in their programs. By better understanding the effectiveness of student-judges and the attitudes towards them, these organizations can make informed decisions on how to best structure their programs to provide the most positive and effective learning experience for all participants.

Ice hockey stroke action data prediction project

Jing Wu

Abstract

This project is a collaboration between Catapult Sports and Boston University's Master of Science in Statistical Practice (MSSP) program. The primary focus of the project is to utilize wearable data to detect shots in ice hockey and create numeric predictors for each shot. The project also aims to classify ice hockey movement into different categories and derive a more accurate hockey-specific load measurement. The resulting algorithm utilizes a combination of supervised and unsupervised learning techniques and is presented with a report detailing the algorithm design process and modeling techniques.

1 Introduction

The ultimate goal of every athlete and team is to optimize their performance, and Catapult Sports aims to provide elite wearable solutions that can help achieve this goal. In collaboration with Boston University's MSSP program, this project aims to develop algorithms and modeling techniques that can utilize wearable data to detect discrete events and classify continuous movement in ice hockey. By analyzing triaxial acceleration and orientation data, the project will derive valuable metrics that provide insights into athletes' movements and efforts.

In particular, the project will focus on the detection of shots in ice hockey and the creation of numeric predictors for each shot. The project also aims to classify different types of ice hockey movement and provide a more accurate measurement of hockey-specific load. The resulting algorithm will utilize a combination of supervised and unsupervised learning techniques and will be presented in a detailed report that covers the algorithm design process and modeling techniques. The project's outcomes will provide coaches and athletes with a valuable tool for analyzing and improving their performance on the ice.

2 Data and Method

The data used consists of wearable data collected from ice hockey players. The wearable data includes triaxial acceleration measurements (forward, side, and up) and triaxial orientation measurements (roll, pitch, and yaw) collected by inertial sensor specs in the wearable devices. Each player's data and video are synced with the wearable data.

To detect shots in ice hockey, the project first manually records the shooting time. The project draws rotation and acceleration plots to identify peaks, each representing a shot. The project also calculates total energy (kinetic energy + rotational kinetic energy) to draw the

energy plot. The project uses XGBoost, Chebyshev polynomials and random forest models. These models are trained on wearable data and evaluated for prediction accuracy.

3 Results and Discussion

We fit the models with orginial data and the coefficientes generated from Chebyshev polynomials. Surprisingly, the raw data fitting model seems better. The visualization result shows that we can see the peaks very clearly, each peak represents each shot, and our prediction accuracy is reasonable. The XGBoost and random forest models helped to achieve a good prediction accuracy.

In our shiny app, clients can select different videos, players, and variables, and it will show the video with its selected variable plot. The plot changes as the video plays. However, the use of video in the application may present some technical challenges or restrictions that the video format may be problematic. Overall, our project provides a useful tool for coaches and athletes to analyze and improve their performance on the ice, and the resulting algorithm could be valuable for future research in the field of sports analytics.

4 Conclusion

Our project successfully utilized wearable data to detect shots in ice hockey and create numeric predictors for each shot. By analyzing the wearable devices' triaxial acceleration and orientation data, we could draw rotation and acceleration plots and calculate total energy to detect shot events. We then employed XGBoost and random forest models to predict shots with reasonable accuracy. The results were visualized in a shiny app, which provides an interactive platform for clients to analyze and visualize player data and game video. Our project provides a valuable tool for coaches and athletes to optimize their performance on the ice.

In conclusion, our project showcases the potential of utilizing wearable data to improve sports performance analysis. We demonstrated the value of utilizing machine learning models in sports analytics by detecting shots in ice hockey and creating numeric predictors for each shot. Our approach provides a comprehensive view of player performance by combining wearable data and game video analysis in an interactive platform. This project can serve as a starting point for further exploration of wearable technology and machine learning models in sports performance analysis. It may have important implications for improving athletic performance and reducing injury risks.