University Fitness Center Data Analysis Discovers Interesting Patterns And Allows Prediction

R.M.Noorullah[1], Nihal Singh[2] , Niharika Thumma[3] , Champati Manikanta Varma[4]

[1]Associate Professor, CSE Department, Institute of Aeronautical Engineering, Hyderabad,India.
[2]CSE Department, Institute of Aeronautical Engineering, Hyderabad,India.
[3]CSE Department, Institute of Aeronautical Engineering, Hyderabad,India.
[4]CSE Department, Institute of Aeronautical Engineering, Hyderabad,India.

Abstract– Data is being used more and more to make daily life simpler and safer. Applications like waiting time forecasting, traffic forecast, and parking quest are excellent examples of how data from various sources can be used to make our lives easier. In this analysis, we look at a data base that is underutilised: university ID cards. In several campuses, these cards are used to buy food, gain access to various locations, and even take class attendance. In this post, we assess use of the university fitness centre using data from our university and develop a predictor for potential visit frequency using data from our university. The work contributes in many ways: it highlights the data source's diversity, demonstrates how the data can be used to enhance student services, uncovers fascinating patterns and behaviour, and acts as a case study for the whole data science process.

Keywords: Pattern analysis, modeling and prediction, data mining, machine learning, time series analysis, computer applications miscellaneous.

1. Introduction

Washington State University is a land-grant university in Pullman, Washington, with a student population of about 20,000. The Student Recreation Center (SRC) is one of the college's most often used grounds offices. The Cougar Card, an official college ID card, is used to track entry to the SRC; clients swipe the card at each line. The information gathered from these card swipe activities is extensive, and it can be used to gain valuable insights into daily life and workout habits. Whatever the case may be, this opportunity has received no attention in the past. One of the more broad purposes of this work is to demonstrate how to deal with this possibility by careful consideration and the stimulation of additional examinations in the vicinity of engaged knowledge disclosure. The SRC is a well-known local attraction, and with its popularity comes two distinct types of requirements. Understanding the usage pattern of the offices is critical for SRC managers to have agreeable understudy administrations. Knowing potential visit volumes is important so that you can keep a safe distance from the SRC when it is usually crowded. This project focuses on meeting all of these needs at the same time. The methodologies we use are in accordance with the two requirements. To continue, we use data mining procedures to uncover interesting use trends at the SRC based on verifiable data obtained from card swipes; this will provide useful information to the SRC directors to assist with day-to-day activities such as transfer booking or event planning. Second, we assume whether these data may be used to predict how crowded the SRC would be over a given time period.

Exploratory data analysis (EDA) is a well-established statistical tradition that offers philosophical and analytical methods for identifying trends and refining hypotheses. In confirmatory data processing, these methods and behaviours complement the application of relevance and hypothesis tests. While EDA supplements rather than replaces, it is rarely necessary to use without it. Also where well-defined hypotheses are retained, EDA aids in the interpretation of data and can expose unexpected or deceptive trends. This essay discusses EDA's central heuristics and statistical methods, as well as how it differs from CDA and exploratory statistics in general. Using previously reported psychological evidence, EDA strategies are demonstrated. To apply the tools, changes in statistical teaching and practise are recommended.

Agents based on data mining are often used in applications such as estimating wait times or predicting traffic flow. Huge volumes of data from several sources are often required for such methods, which can be difficult to collect and result in sparse or noisy datasets. Data from university ID cards, on the other hand, is very easy to view and has very little noise. However, the use of these databases has received

little consideration, and few applications to enhance student facilities on campus have been created. The data for this project comes from CougrCard, Washington State University's official ID card, which is used by the majority of students on a regular basis. Our mission is to create an intelligent agent that can anticipate crowding at various campus facilities and thereby increase student service efficiency. This ongoing project aims to improve students' gym opportunities at the University Recreation Center, one of the most common facilities on campus.

Many robotics applications, such as map building and coverage, require exploration in unfamiliar conditions. Several studies have suggested exploration techniques that lead a mobile robot to select where to go next in order to map an initially unfamiliar area incrementally. In this article, we look at the worst and average travelled distances needed to explore graph-based ecosystems using certain discovery techniques that take distance and knowledge benefit into account when deciding where to go next.

2. Related Work

It includes potential forecasts of students' temperament and academic performance so that professors can figure out the best way to handle and encourage those students so that their scholarly presentation improves. The Student Recreation Center (SRC) is one of the college's most often used grounds offices. The Cougar Card, an official college ID card, is used to track entry to the SRC; clients swipe the card at each line. The information gleaned from these card swipe activities is extensive, and it can be used to gain valuable insights into everyday life and fitness routines. Whatever the case may be, this opportunity has received no attention in the past. One of the more broad goals of this project is to demonstrate how to deal with this possibility by careful study and the initiation of additional inquiries in the vicinity of active information disclosure.

The Student Recreation Center (SRC) is one of the most commonly used university buildings. The use of an official university ID card, the Cougar Card, is used to track access to the SRC; users swipe the card upon entry. These card swipe exercises yield a wealth of information that can be used to obtain useful insights into student life and fitness habits. However, this opportunity has received no recognition in the past. One of the larger aims of this project is to show how to exploit this ability by careful study and to encourage further research on campus-based information exploration. The SRC is a popular destination on campus, and its success has resulted in two distinct categories of needs. Knowing the use trend of the facilities is essential to the SRC managers in order to provide adequate student services. Knowing potential visit volumes is critical from the students' perspective in order to escape the SRC when it is most crowded. In spirit, our work is close to a few current applications that are based on information gained from the collection and analysis of activity data. One example is the Orlando Undercover Tourist App. This smart phone app collects real-time facility utilisation data at Orlando Disney World and forecasts wait times so visitors can schedule their tours more effectively.

The methods we use are in line with the two requirements. First, we use data mining methods to discover fascinating utilisation trends at the SRC based on historical data obtained from card swipes; this insight can be useful to SRC administrators for everyday operations such as shift management and event preparation. Second, we decide whether these data can be used to forecast how crowded the SRC will be over a given time period.

3. Exploratory Data Analysis

EDA (Exploratory Data Analysis) is a computational method for investigating the properties of a dataset. During EDA, we use tools like plots, diagrams, and summary statistics to see what the data might teach us before we match a formal model or test a hypothesis. We use these approaches at the SRC to help explain student usage habits and speculate about potential reasons. We begin by looking at the timestamp dataset to see whether there are any trends, seasonality, outliers, or other irregularities in student exercise habits. Then we looked at user accounts to see how use differs by gender and social status.

First, from September 1st, 2012 to December 31st, 2016, we plotted the cumulative number of visits to the SRC for each month. The "time" column of Table 1 was used to aggregate the monthly amount of visits. This time series has a lot of seasonality, as seen in Figure 1. Although each year's trends are identical, the Summer semester (May to August) had slightly less people than the Fall and Spring semesters (September to April). This is unsurprising with how many students leave campus for the summer. Furthermore, we discovered that the overall number of visits in the Spring semesters is consistently higher than the norm, while in the Fall semesters, only September and October surpass the mean, whereas November and December fall below it.
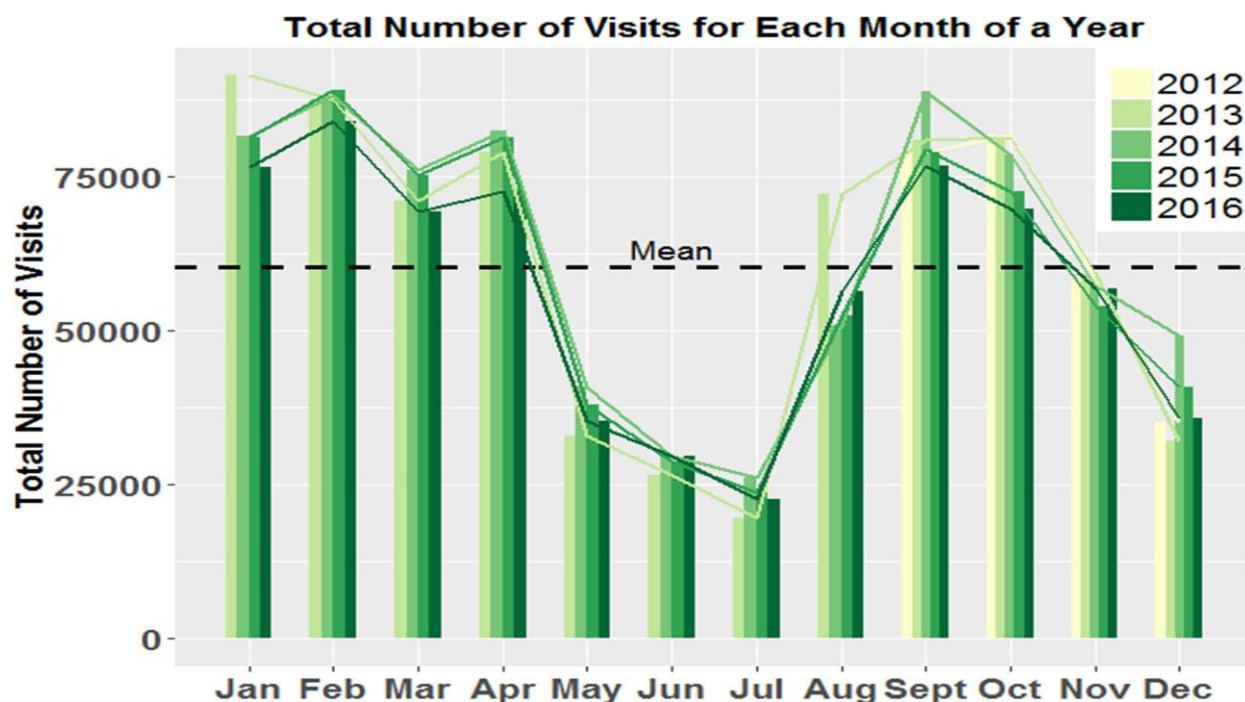
Fig. 1. This figure shows the total number of visits for each month of a year. The dashed line indicates the mean of all years.

To help students determine the best time to visit the SRC, we formulated this time series prediction problem as a regression task (rather than a classification task, as was done previously in [1]), with the aim of predicting the number of people visiting the SRC for a given time period. Based on historical SRC consumption info, several models were developed for forecasting visit volumes. The approach and findings for the three models we built are presented first in this section. Then there's a debate about which model is better for our mission. We compare the suitability of modelling approaches from two fields of analysis, statistics and machine learning, for our dilemma. We started with a seasonal naive (Snaive) model as a benchmark in terms of statistical modelling. This model "naively" predicts future visit volume based on the previous season's observed value. Second, we used an autoregressive integrated moving average (ARIMA) model, one of the benchmark time series models, which outperformed the baseline model in terms of estimation. In terms of machine learning simulation, we used a random forest (RF) model to match the data and found that it outperformed both the baseline and the ARIMA model in our experiment. All of the models were created with the R programming environment's available packages [13]. The models were tested using k-fold cross-validation [14], with k = 4. In specific, as seen in Figure 9, we divided the data into four folds. The I th fold included I years of training data (green block), with August 20th, 2012 as the starting date. For research, we used data from the four months after the conclusion of the training period (yellow block). Fold 1 had one year of training data from August 20th, 2012 to August 19th, 2013, and testing data from August 20th, 2013 to December 31st, 2013, for example; fold 2 had two years of training data from August 20th, 2012 to August 19th, 2014, and testing data from August 20th, 2014 to December 31st, 2014, and so on5. All of the data (grey blocks) has been left out.

3. Result discussions:

```
In [5]: plt.figure(figsize=[10,8])
        plt.bar(data['Month'],data['Freq'], width = 0.5, color='#0504aa')
        plt.xlabel('Month',fontsize=15)
        plt.ylabel('Total number of visits',fontsize=15)
        plt.title('Total number of visits for each Month',fontsize=15)
        plt.show()
```
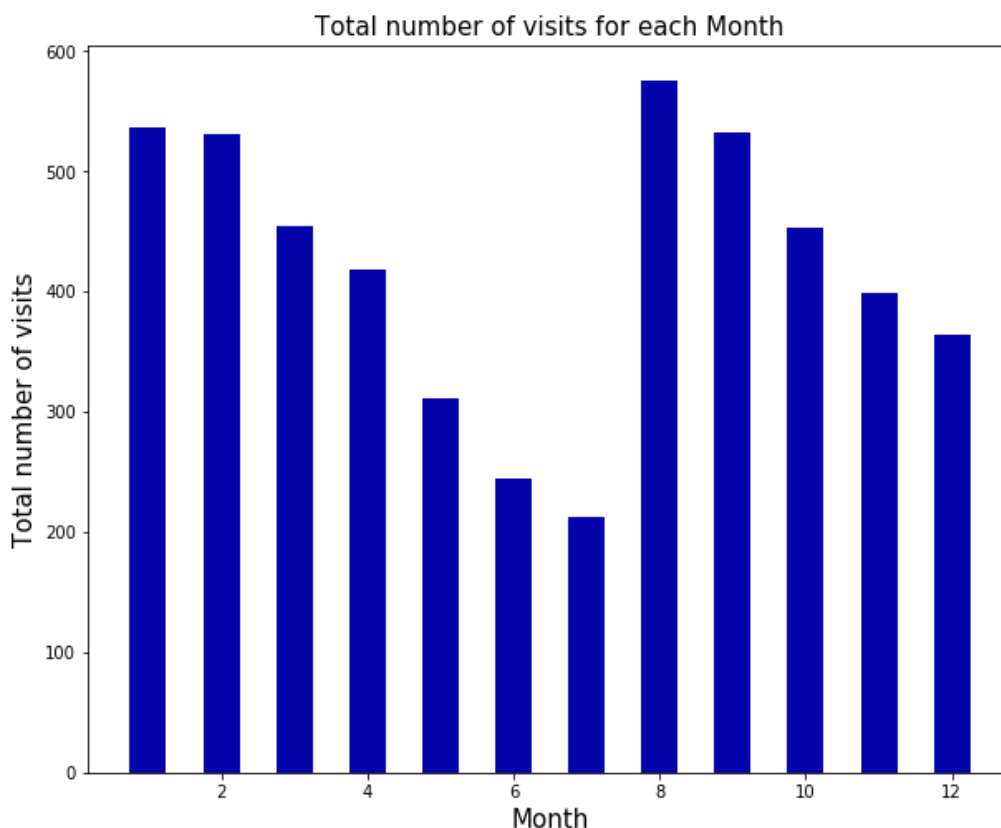

Fig. 2. This graph depicts the average number of visitors per day over the course of a month. It's worth noting that the standard error of the mean is very small . During the month, there has been a downward trend.
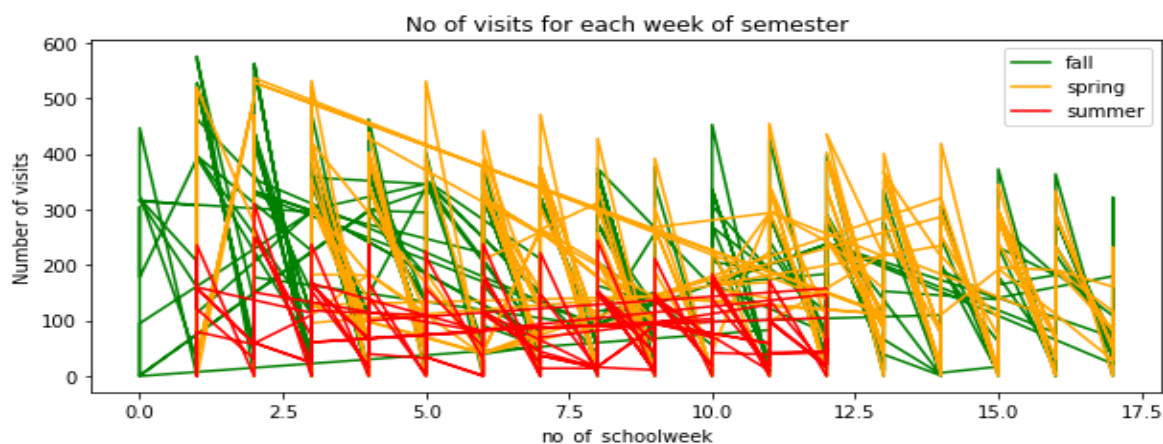
Fig. 3. For the four academic years 2012-2013, 2013-2014, 2014-2015, and 2015-2016, this graph depicts the mean number of visits for each week of a semester. During the Fall and Spring semesters, there are 17 weeks, and during the Summer semesters, there are 12 weeks. The black bars represent the standard error of the mean measured over four academic years.
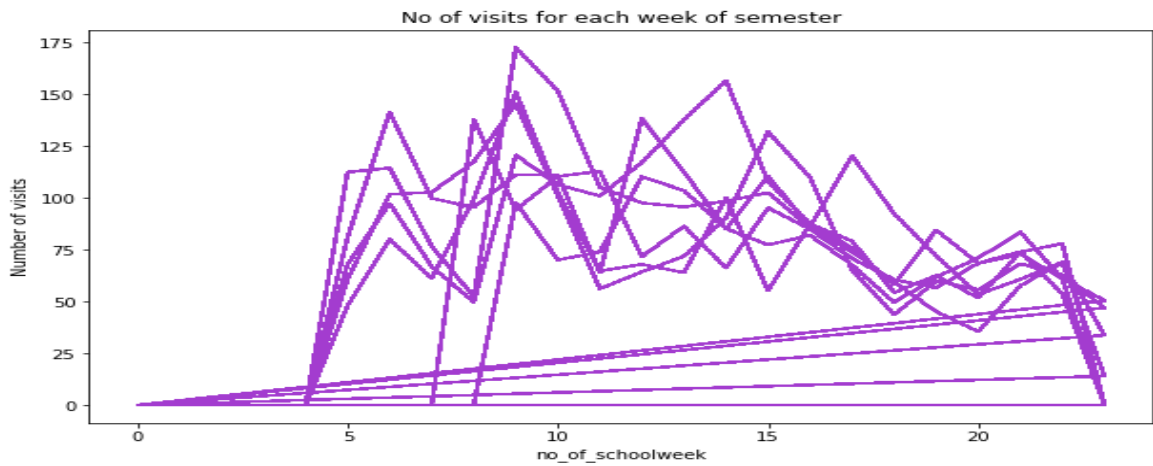
Fig.4.This graph depicts a linear regression between the average number of visits and the number of school weeks. The number of students in each semester is diminishing, with Spring semesters still being more crowded than Fall and Summer semesters.
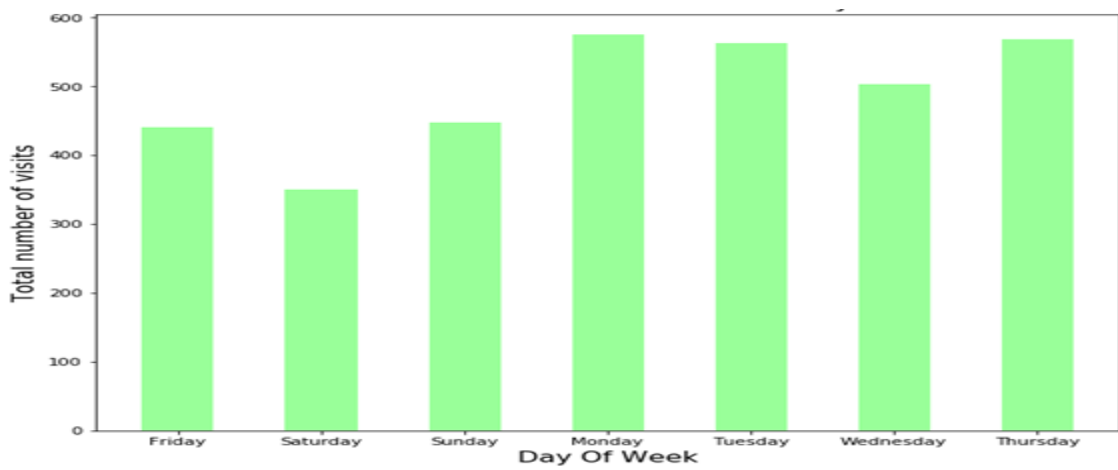

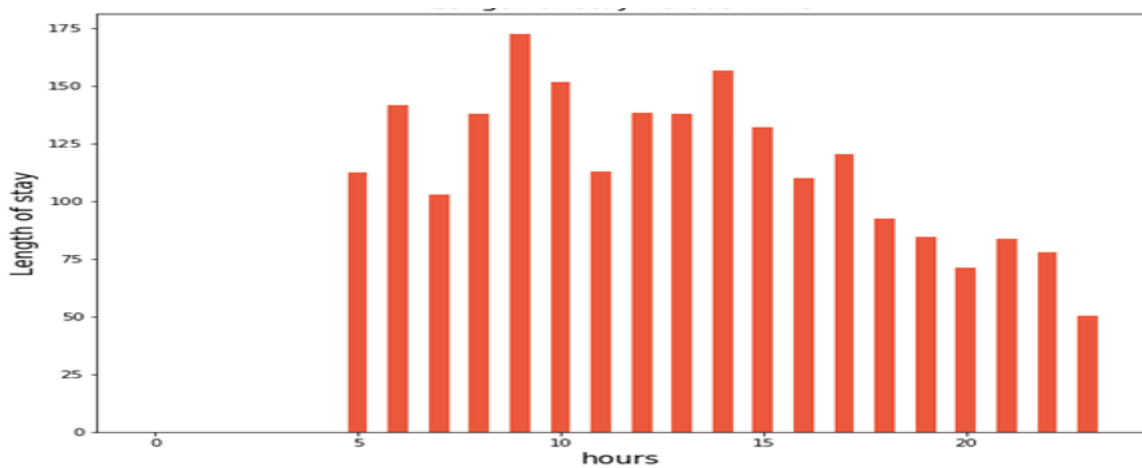Fig.5:Total number of visits for each day
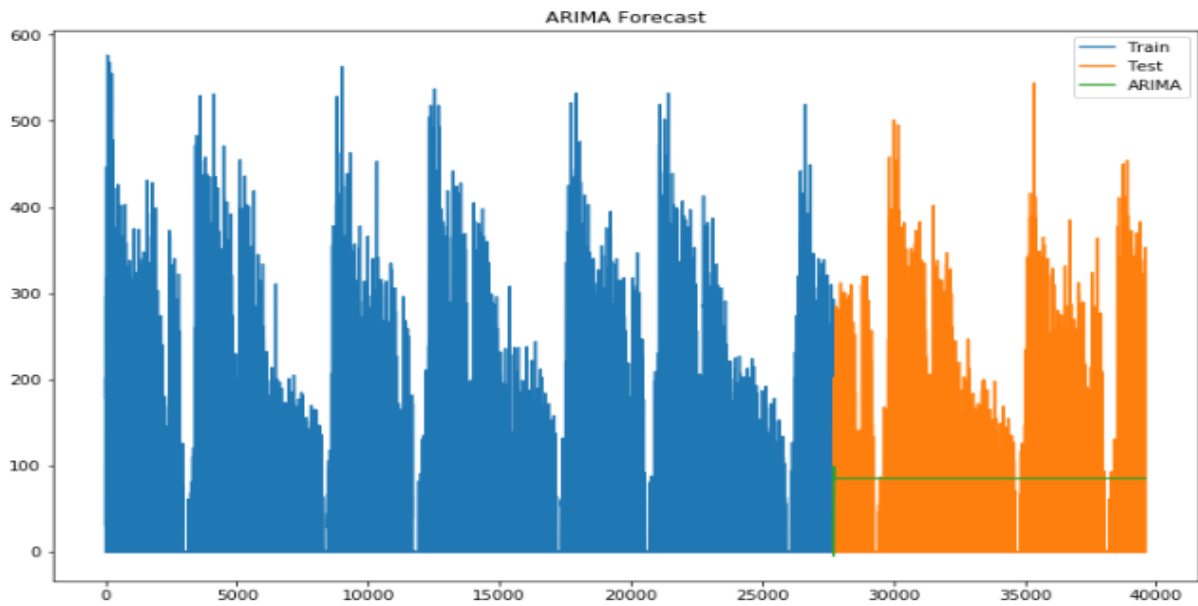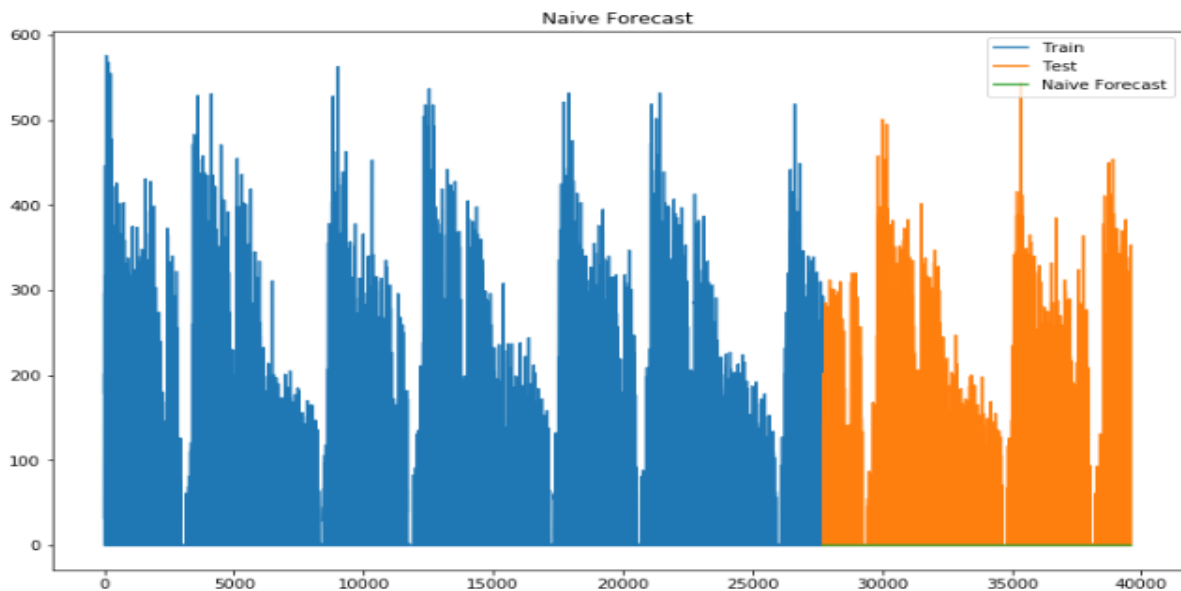

Fig 6:Length of stay versus time

Fig 7: ARIMA

Fig 8:Naïve forecast

Dataset :

	TRANDATE	Freq	Month	DayofWeek	Hour	Semester	NumSchoolWeek	IsHolidays	IsVacations	IsOpen	PrevFreq	LenStay
0	17-08-2012 05:00	32	8	0	5	0	0	0	0	1	0	68.000000
1	17-08-2012 06:00	36	8	0	6	0	0	0	0	1	32	97.424658
2	17-08-2012 07:00	65	8	0	7	0	0	0	0	1	36	66.923611
3	17-08-2012 08:00	76	8	0	8	0	0	0	0	1	65	49.882883
4	17-08-2012 09:00	132	8	0	9	0	0	0	0	1	76	120.717391
5	17-08-2012 10:00	125	8	0	10	0	0	0	0	1	132	106.461539
6	17-08-2012 11:00	127	8	0	11	0	0	0	0	1	125	100.968000
7	17-08-2012 12:00	164	8	0	12	0	0	0	0	1	127	117.211765
8	17-08-2012 13:00	202	8	0	13	0	0	0	0	1	164	137.835714
9	17-08-2012 14:00	178	8	0	14	0	0	0	0	1	202	156.684849

4. Conclusion

We investigated and visualised trends of student fitness behaviour in terms of time and demographics in this report. We measured and contrasted the predictive precision of a seasonal naive model, an ARIMA

model, and a random forest model, and discovered that the random forest model better suited our dataset. The random forest model correctly estimated the visit volume at the SRC for a given time period. The SRC workers have benefited from the deployed webpage in terms of regular activities such as personnel schedules. We hope that this work will act as a case study for the whole data science process, revealing how data can be gathered, analysed, examined, and used to produce a user-facing data app.

REFERENCES

[1] Y. Du and M. E. Taylor, "Work in-progress: Mining the studentdata for fitness," presented at the 12th Int. Workshop Agents Data Mining Interaction, Singapore, May 2016.

[2] InsiderGuide Inc., "Disney world wait times, touring plans free by undercover tourist," 2015. [Online]. Available: https://www. undercovertourist.com/apps/, Accessed: Jan. 30, 2016.

[3] Waze Inc., "Waze - GPS, maps and social traffic," 2016. [Online]. Available: https://www.waze.com, Accessed: Jan. 30, 2016.

[4] Google Maps, "Popular times," 2016. [Online]. Available: https:// support.google.com/business/answer/6263531?hl=en, Accessed: Jan. 30, 2016.

[5] E. Davami and G. Sukthankar, "Improving the performance of mobile phone crowd sourcing applications," in Proc. Int. Conf. Auton. Agents Multiagent Syst., 2015, pp. 145–153.

[6] R. Schutt and C. O'Neil, Doing Data Science: Straight Talk from the Frontline. Sebastopol, CA, USA: O'Reilly Media, Inc., 2013.

[7] R. Hyndman, G. Athanasopoulos, C. Bergmeir, G. Caceres, L. Chhay, M. O'Hara-Wild, F. Petropoulos, S. Razbash, E. Wang, and F. Yasmeen, Forecast: Forecasting Functions for Time Series and Linear Models, R package version 8.4. 2018. [Online]. Available: http://pkg.robjhyndman.com/forecast

[8] R. J. Hyndman and G. Athanasopoulos, "ARIMA models," in Forecasting: Principles and Practice. OTexts, 2014. [Online]. Available: https://otexts.org/fpp2/

[9l] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[10] R. J. Hyndman and G. Athanasopoulos, "The forecasters toolbox," in Forecasting: Principles and Practice. OTexts, 2014. [Online]. Available: https://otexts.org/fpp2/