Using Big Data To Analyze FAFSA Data For Various Trends

A project by the Honey Badgers

Phase 3

Val Booth
vxb4825@rit.edu

Nicholas Jenis
ngj5017@rit.edu

Mark Petrie
mjp2542@rit.edu

ABSTRACT

The FAFSA, the Free Application for Federal Student Aid, is a form filled out yearly by students who will be attending college in the following year. The FAFSA tries to gauge how eligible the student is for college aid and is based off of many factors. One such factor is dependency status, which in turn is based off of a student's age, type of program (undergraduate vs. doctorate, etc.), and military status. Team Honey Badgers has written an application that uses FAFSA data from 2013-2015 hosted on data.gov. Our application allows the user to view graphs of dependency FAFSA data filtered by geographic region or state, year or school type. Our application allows the user to not only analyze FAFSA dependency status in several different ways, but also to try and find trends among the data. Team Honey Badgers used the application to try and determine whether or not there has been an increase in independent filers versus dependent filers, as some news sources have claimed.

1. OVERVIEW

The team has chosen to analyze data from the FAFSA, the Free Application for Federal Student Aid. This is a form that is typically filled out each year by students planning to attend or already attending college in the US. This form helps students and families determine their eligibility for student aid. The Financial Aid Department for the University of Texas at Austin notes that in almost all cases a student is eligible for at least some financial aid, though the amount of aid varies by need [12]. The Federal Student Aid Office states that their aid programs are based on the concept that it's the responsibility of the student and the family to pay for the student's education [2]. The Federal Student Aid office still gives out many types of financial aid awards to students, with the most common being the Pell Grant, Stafford Loan, and Federal Perkins Loan [10].

We have built an application to view and try and find trends in the numbers of dependent and independent filers using data publicly available from data.gov [9].

Our motivation for starting this project is in section 2. A description of the data domain is in section 3, with details on development of application and its functionality being found in section 4 and section 5 respectively. Finally, lessons learned throughout the project are in section 6.

2. PROJECT MOTIVATION

This paper attempts to use FAFSA data to validate numerous articles [13, 14, 11] and a higher education dataset [4] that indicate that the percentage of new adult students attending college is greater than the percentage of new traditional age populations. The FAFSA considers a filer an independent student if they are 26 years of age or older versus a dependent student who is under the age of 26. Higher education defines a traditional age student as age 17-25 and an adult student over the age of 25. While there is a one year difference in the classification between adult and traditional between the higher education difference and the FAFSA definition, we assumed that the difference was not significant and that therefore the FAFSA data could be used to test this theory.

3. DATA DOMAIN

data.gov currently provides Microsoft Excel spreadsheets of dependent and independent counts per post-secondary institution, for each quarter, from Q1 2013 - Q3 2016. There are 6 quarters in each year.

Filing as an independent or dependent student depends on various factors about the student. This determination is made through a list of questions such as the following, taken from the Federal Student Aid Office website regarding the 2016-2017 FAFSA:

- Were you born before Jan. 1, 1993?
- As of today, are you married?
- At the beginning of the 2016-17 school year, will you be working on a masters or doctorate program?
- Are you currently serving on active duty in the U.S. armed forces for purposes other than training?
- Are you a veteran of the U.S. armed forces?

Answering yes to one or more of these questions indicates that the student is an independent and will not provide information about their parents on the FAFSA. If the student answers no to all of the questions asked, then they are considered a dependent for the purposes of federal student aid

and they must provide information about their parent's income on the FAFSA [2].

A small sample of the data we will be analyzing can be seen in Figure 1.

Application Data by School

2015-2016 Application Cycle

Quarter 3 (07/01/15 - 09/30/15)

					NUMBER OF FAFSA APPLICATIONS PROCESSED IN Q3		
OPE ID	School	State	Zip Code	School Type	Dependent Students	Independent Students	Quarterly Total
00266500	VAUGHN COLLEGE OF AERONAUTICS AND TECHNO	NY	11369-9997	Private	220	252	472
00266600	ADELPHI UNIVERSITY	NY	11530-4299	Private	386	789	1,175
00266700	DOWLING COLLEGE	NY	11769-1999	Private	135	203	338
00266800	ALFRED UNIVERSITY	NY	14802-1205	Private	126	189	315
00266900	BANK STREET COLLEGE OF EDUCATION	NY	10025-0000	Private	2	82	84
00267100	BARD COLLEGE	NY	12504-5000	Private	38	30	68
00267400	NEW YORK THEOLOGICAL SEMINARY	NY	10115-0002	Private	1	95	96
00267700	BROOKLYN LAW SCHOOL	NY	11201-0000	Private	14	203	217
00267800	BRYANT & STRATTON COLLEGE	NY	14203-1713	Proprietary	584	3,304	3,888

Figure 1: FAFSA data for NYS institutions retrieved from data.gov

4. IMPLEMENTATION

An application has be created which allows the user to filter the data set by given parameters such as location, year, or type of institution. The user is able to see the data, filtered by their chosen parameters, in a graph form inside an interactive GUI.

The application has been written in Python with an external Python library being used to generate the graphs in the application. Our team chose this language as we are all fairly comfortable with it, it suited our needs well by being easy to develop, and also allowed us to use external libraries to connect to a database and create data graphs. A more detailed description of the application is available in section 5.1.2.

4.1 Design

4.1.1 Data Design

The data pulled from data.gov was already fairly normalized. The data was split into two tables, `School` and `FAFSA_Data`. The ER Diagram for the two tables created in sqlite can be seen in Figure 2.

In order to more efficiently query the database, `Type` and `Region` were stored as integers, being reconverted to strings in the application. For school type, "Public", "Private" and "Proprietary" are represented as the integers 1, 2, and 3 respectively in the database. `Region` was a field not originally in the data, but added by us in order to further analyze the data and possibly find trends. `Region` is determined by the school's state and is either: "Northeast", "Midwest", "South", "West", "Pacific", or "Other". "Other" represents territories outside of the US ("PR" for Puerto Rico, etc.), while all of the other regions are defined by the United States Census Bureau [3]. Regions are represented by the integers 1 through 6 in the respective order above within the database.

The database had 4 indices, split up below by table:

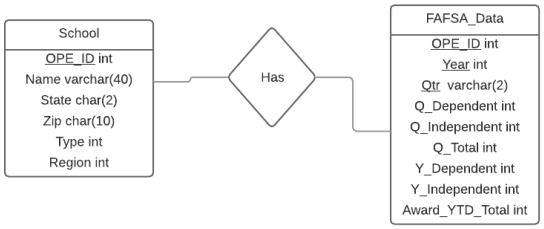

Figure 2: ER Diagram for our sqlite3 database

- `School`:
  - `OPE_ID` (*Note: Unique index.*)
  - `State`, `Type`
  - `State`, `Region`, `Type`
- `FAFSA_Data`: `OPE_ID`, `Y_Dependent`, `Y_Independent`, `Qtr`, `Year`

`OPE_ID` is used to join the two tables together, `State`, `Type`, `Qtr` and `Year` are used to group the data together, and `Y_Dependent` and `Y_Independent` are summed and then selected to be visualized by the graphing application. More information about the database can be found in section 5.1.1.

4.1.2 Application Design

We split our application into two pieces in order to more easily split the work up among the team and also in order to have a more modular final product. All of the application source code files are in the `code` folder in the code submission. The front-end was a GUI that allowed the user to pick which mode, or parameters, to filter the data by. Depending on the filtering selected, the GUI called the corresponding function in the back-end, `filter.py`. `filter.py` connected to the database, composed the appropriate SQL query depending on the function, and sent the row data to functions in `formatData.py` before returning it. More detail on the application can be found in section 5.1.2.

5. CURRENT STATUS AND FUTURE WORK

5.1 What Has Been Done

5.1.1 Database

From the data.gov website we have pulled together 3 years worth of data regarding school, dependent and independent rates all split by quarter. The data was cleaned up to remove extraneous columns, put into a usable Microsoft Excel format, then converted into a sqlite3 [6] database. The database can be explored in the `code` folder in the `HoneyBadgers.db` file and has a total of 7887 rows in the `School` table along with 110057 rows in `FAFSA_Data`.

5.1.2 Graphing Application

We have written a GUI application in Python that allows the user to filter and graph the data by state, larger geographic region, school type, quarter, and year. The user can choose to view all of the 50 states and other territories

individually, in one large graph, or to select two states to compare. To use the application, please view the instructions in `README.txt` of the code submission. The source code for the program can all be found in the `code` folder. The GUI, mainly with code in `gui.py`, takes input from the user on how to filter the data, gets corresponding X and Y axis data from a backend Python script `filter.py`, sends that data to the external Python plotting library, plotly [5], and then renders the graph inside a GUI generated with PyQt5 [1]. One sample graph from the application can be seen in Figure 3. We've bundled the scripts, database, and

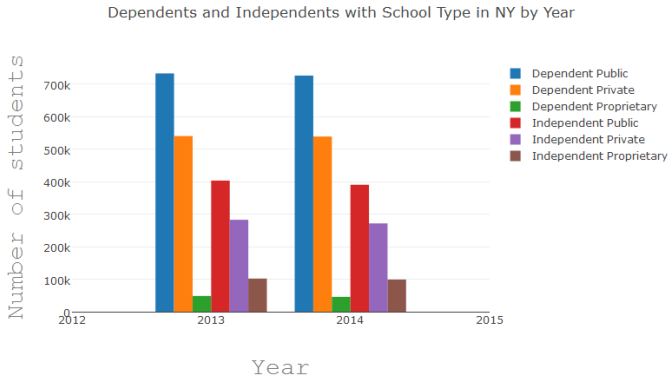

Figure 3: Our application graphing a comparison of dependent vs independent filers in NYS, grouped by school type

external library as a Windows executable, for ease of use, using PyInstaller [15].

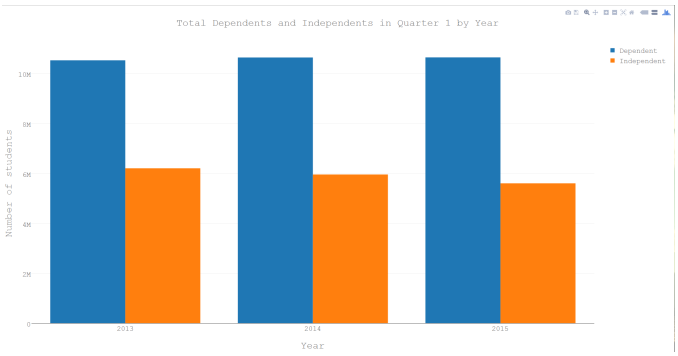

Figure 4: Number of dependent and independent students for quarter 1 by year

5.1.3 Data Analysis

We began this project believing that the the adult population (25+ years old) were enrolling at colleges and universities at a greater rate than the traditional age population (17-24 years old). A brief look at the data using our application contradicts this idea; total filers and dependent and independent filers respectively were lower in 2014 as compared to 2013. Looking on a quarter by quarter basis per year we saw that dependent and independent filers decreased in quarter 1 of each year, with independent filers decreasing at a greater percentage than dependents (see Figure 3). Pivoting to look at breakdown by school type, we observed that proprietary dependent filers were the only type that decreased at a greater rate than their independent counterparts (see Figure 4). Either the assumption that more independents were enrolling is false or it cannot be verified by FAFSA data; many independents may not fill out a FAFSA. They may not believe they are eligible for federal money, or their tuition may be paid by the military or their place of work. We did descriptive analytics on the FAFSA data to determine if the percentage of independent filers is greater than the percentage of dependent filers for the 2 years of data collected. The FAFSA data did not validate our theory. In all but one scenario, pattern recognition revealed a greater percentage decrease in dependent versus independent filers in both 2013 and 2014. The only subcategory where independent filers decreased at a lesser rate than dependent filers was for proprietary institutions. For further data analysis, see section 6.1.

5.2 Deliverables

Our final deliverables are:

- Graphing application executable with source code written in Python, connected to a sqlite3 database. The application is the `app_main.exe` file within the zipped `app_main` directory of the code and poster submission.
- Copies of the source code used to create the graphing application. Source code can be found in the `code` directory of the the code and poster submission.
- This ACM style write up.

6. LESSONS LEARNED

Based on the FAFSA data collected and pattern recognition discovered, we were surprised that the FAFSA data did not validate our theory that adult students as a percentage increase did not outpace traditional age students.

6.1 Additional Observations

We acknowledge that there is a one year difference between the FAFSA definition of an adult versus the standard definition within higher education. Thinking critically about the findings, we stand by our original assumption that the difference in definition is not singularly significant to invalidate our theory.

NCES projected that from 2012 to 2023, the rate of increase for students under age 25 would be 12 percent, compared with 20 percent for students age 25 and over [4]. The Center reports declines in overall enrollment for all Title IV institutions in 2010 and 2011 after 12 years of consecutive increases [7]. It's possible that the Center's projections for 2012 to 2023 [8] are not accurate for 2013 and 2014 given our findings.

The articles by Fuller [13], Soares [14], and Buckner [4] all reference the NCES data as a primary source. For future work, we would like to collect additional data to understand the number of college-going students, dependent and independent, that did not submit a FAFSA. Having the 2013 and 2014 data of non-filers added to our existing FAFSA data would provide a more complete picture of the college-going population during that time period.

7. REFERENCES

[1] https://www.riverbankcomputing.com/software/pyqt/intro, 2015.

[2] Federal student aid: Dependency status. https://studentaid.ed.gov/sa/fafsa/filling-out/dependency, 2015.

[3] Census regions and divisions of the united states. https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf, 2016.

[4] Fast facts. http://nces.ed.gov/fastfacts/display.asp?id=98, 2016.

[5] plotly | make charts and dashboards online. https://plot.ly/, 2016.

[6] Sqlite. https://www.sqlite.org/, 2016.

[7] Total fall enrollment in all postsecondary institutions participating in title iv programs and annual percentage change in enrollment, by degree-granting status and control of institution: 1995 through 2012. http://nces.ed.gov/programs/digest/d13/tables/dt13_303.20.asp, 2016.

[8] Total fall enrollment in degree-granting postsecondary institutions, by attendance status, sex, and age: Selected years, 1970 through 2023. http://nces.ed.gov/programs/digest/d13/tables/dt13_303.40.asp, 2016.

[9] F. S. Aid. Fafsa data by postsecondary school and state of legal residence. https://studentaid.ed.gov/sa/about/data-center/student/application-volume/fafsa-school-state, 2012.

[10] J. Baker. Ifap - dear college letters. https://ifap.ed.gov/dpcletters/GEN1502.html, 2015.

[11] J. J. Buckner. What i'm reading: Post-traditional learners and the transformation of postsecondary education, 2016.

[12] Finaid.utexas.edu. Financial aid: Eligibility for aid faq. http://finaid.utexas.edu/faqs/eligibility.html#Q1, 2015.

[13] A. Fuller. New report projects increased enrollments of women and nontraditional students. http://chronicle.com/blogs/ticker/new-report-projects-increased-enrollments-of-women-and-nontraditional-students/36523, 2016.

[14] L. Soares. Post-traditional learners and the transformation of postsecondary education: A manifesto for college leaders, 2013.

[15] M. Zibricky, H. Goebel, D. Cortesi, and D. Vierra. Welcome to pyinstaller official website. http://www.pyinstaller.org, 2016.