Historical College Scorecard Big Data Analysis using in-Memory Processing

Kunal Pritwani [α], Atinder Singh [σ], Dharmesh Soni [ρ], Mounika Vallabhaneni [ω] & Jongwook Woo[¥]

Abstract- Data set is collected for colleges of United States. We would like to analyze different dimensions like SAT scores, ear- ning after graduation, net price and grant financial aids which is a great analyzation for the students. Big Data platform and BI tool such as Spark and tableau are adopted for data analy- zation and visualization. It is found that the top colleges for mean earnings are from medical field, mean earnings with respect to states, detailed comparison of average net price of California and New York, SAT scores for different colleges and also average undergraduates receiving Pell Grant in each colle- ges which will help students to select a college which meets their requirement.

Keywords: big data, spark, hadoop, college scorecard analysis, mean earnings, in-memory.

I. Introduction

We are to analyze the basic fundamentals of college which are important factors in big data analytics. This kind of data is analyzed by big name analyst for big money as this kind of analysis provides insight on differ- rent aspects of college. The outcomes by this analysis will help students to compare between different colleges and can select college according to their own needs and education goals.

Big Data is defined as non-expensive frameworks that can store a large scale data and process it in parallel [7, 8]. A large scale data means really a big data, this data cannot be processed using traditional computing techniques. Data is getting generated everyday through social media, websites, mobile applications etc. To analyze and store data we use Hadoop, which is an open source framework which provides distributed storage on the commodity hardware. Hadoop has two major components which are MapReduce and HDFS (Hadoop Distributed File System).

Apache spark is popular for processing big data using Hadoop architecture. As it's the updated version for map reduce. Apache Spark runs 100 times faster than Hadoop but it doesn't have its own HDFS. So it uses HDFS as its filesystem and runs on top of Hadoop by using memory. Spark uses RDD (Resilient Distributed Datasets) which replaces the MapReduce functionality to write the data to physical storage every time.

Author α σ ρ ω ¥: Department of Computer Information Systems, College of Business and Economics, California State University, Los Angeles. e-mails: kpritwa@calstatela.edu, asingh37@calstatela.edu dsoni3@calstatela.edu, mvallab@calstatela.edu jwoo5@exchange.calstate.edu.

II. Related Works

Nick does the data analysis using statistical techniques to find the correlation between different columns. But, we have used spark to manipulate and visualize the data to get useful insights [10]. Student Responsiveness by Hurwitz et al is analyzed by selecting the earnings using statistical techniques to find the correlation and by using scatter plots for visualization [11]. We simply used geographical visualization to show top earning states. Besides, Spark computation is less time consuming to process the results.

We have used Big Data Spark platform to store and analyze the data and BI tool such as tableau for visualizations. By analyzing the 100,000 colleges data of 14 years, we have different results as we analyzed a very huge dataset. We have the detailed analysis for 100,000 colleges and they have analysis for around 600 colleges. We found the top major which has high paid jobs is in medical field [9]. Spark helps to process the queries and gives the results fast.

III. Methods

First, we collected the data from an online community dedicated to data scientists where the dataset comprises of historical data of 100,000 colleges in the US spanning over 14 years to compare and analyze. Further, by using the Spark technique to find different terminologies like Mean and Median Earnings of the College, Average Net Price of a College, Verbal and Math Sat Score Analysis and Percent of Undergraduates Receiving PELL GRANT. Detailed Analysis of college score card has been performed using data visualization tools.

a) Specification of Data Set

The data is collected from an online community. We have historical data of about 100,000 colleges within United States spanning of 14 years. The data size is 1.33 GB and file is in CSV (Comma Separated Values) format [1].

b) Tools

Data Analysis tools used are Apache Spark cluster on Databricks cloud platform, and visualization tool Tableau 9.2 is used for detailed data analysis for daily and yearly records.

c) Terminology

i. *Mean and Median earnings of the College*

Mean earnings are for the institutional total of all governmentally helped understudies who select in an

organization every year and who are working but not taking any classes.

ii. *Average Net Price of a College*

There are a few components in the Average Net Price that are gotten from the full cost of participation (counting educational cost and charges, books and supplies, and everyday costs) less government, state, and institutional guide, for undergrad understudy.

iii. *Verbal and Math Sat Score Analysis*

Test scores of enrolled students are not reported for all institutions, but rather may help students to discover a school that is a decent scholastic match. The query incorporates 75th percentiles of SAT Verbal (SATVR75), SAT Math (SATMT75).

iv. *Percent of Undergraduates Receiving PELL GRANT*

This column (PCTPELL), reflects the share of undergraduate students who have got Pell Grants in a given year. This has an important measure of the access a school provides to low-income students.

IV. Detail Data Analysis Results

a) Mean and Median earnings of the College

This formula selects columns the institute name (INSTNM), Mean and Median Earnings of the college (mn_earn_wne_p10) and state name(STABBR). Results are stored in 'results' RDD and then displayed using Spark Display command. Spark SQL commands are used for fast processing of SQL context queries. It shortens the query length and gives faster results than SQL.

->results = sqlContext.sql('SELECT INSTNM, mn_earn_wne_p10, STABBR FROM Scorecard_Project order by (mn_earn_wne_p10) desc')
-> display(results)

Figure 1. shows the top colleges with mean earnings, In this case its Medical college of Wisconsin with mean earnings as 250K.

Figure 2. shows the states with highest(Blue - California), medium(Gray - Texas) and lowest(Red - Oregon) mean earnings as for CA it's more than 60 million. The results are listed in Figure 1 and Figure 2 below.

b) Comparing Average Net Price of Two States

This formula selects columns the institute name(INSTNM), Average Net price of state (NPT4_PUB) and CITY(CITY). Results are stored in 'results' RDD and then displayed using Spark Display command. Spark SQL commands are used for fast processing of SQL context queries. It shortens the query length and gives faster results than SQL. Refer the code at Github [5], [6].

Figure 3. shows the average net price with comparison of two states. UCLA has 13,817 and Cal State La has 4,37.

Figure 4. shows the top net prices for public universities like Blue Hills Regional Technical School has 26, 475.

Figure 5 shows the top net prices for private universities like Aerosim Flight Academy has around 87K.Figure 3, Figure 4 and Figure 5 display the results below.

c) SAT Scores in Different Colleges

This formula selects the top institutes where SAT verbal and Mathematics score is maximum. Refer the code at Github [5], [6].

Figure 6. shows the SAT scores and mean earnings like California Institute of Technology has Math's score(Blue) as 800, Verbal score(Orange) as 778.9 and Mean earning(Purple) as 98,700. Figure 6 display the result below.

d) Comparing Average Undergraduates Receiving PELL GRANT

Amounts can change yearly. For the 2016–17 award year (July 1, 2016, to June 30, 2017), the maximum award is $5,815. The amount you get, though, will depend on:

- your financial need
- your cost of attendance
- your status as a full-time or part-time student
- your plans to attend school for a full academic year or less.

You may not receive Federal Pell Grant funds from more than one school at a time.

This formula will select the columns from the database, institute name(INSTNM), Mean, state name (STABBR), Average Undergraduate Students(UGDS) and percentage of Pell grant(PCTPELL) which has UGDS > 1000. Results are stored in 'results' RDD and then displayed using Spark Display command. Spark SQL commands are used for fast processing of SQL context queries. It shortens the query length and gives faster results than SQL. Refer the code at Github [5, 6].

Figure 7. shows Universal Career Community College has the full PELL grant like 1.0 which means 100% scholarship

Figure 8. shows that East Georgia State College has 2,854 Avg. no undergraduate students and also PELL grant percentage is 97.285%. Figure 7 and 8 shows the result below.

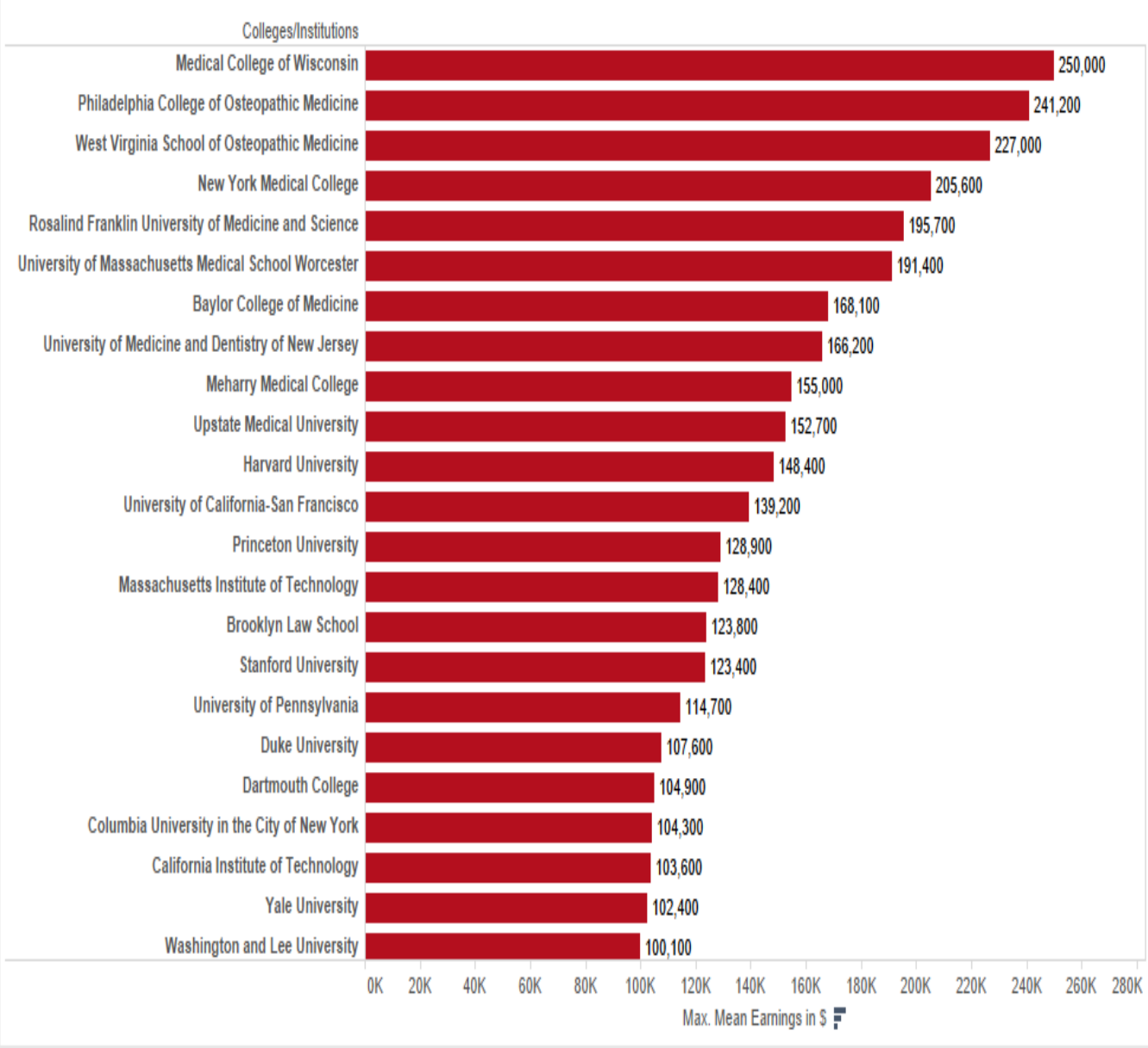


Figure 1: Top Mean and Median earnings of the College in USD.

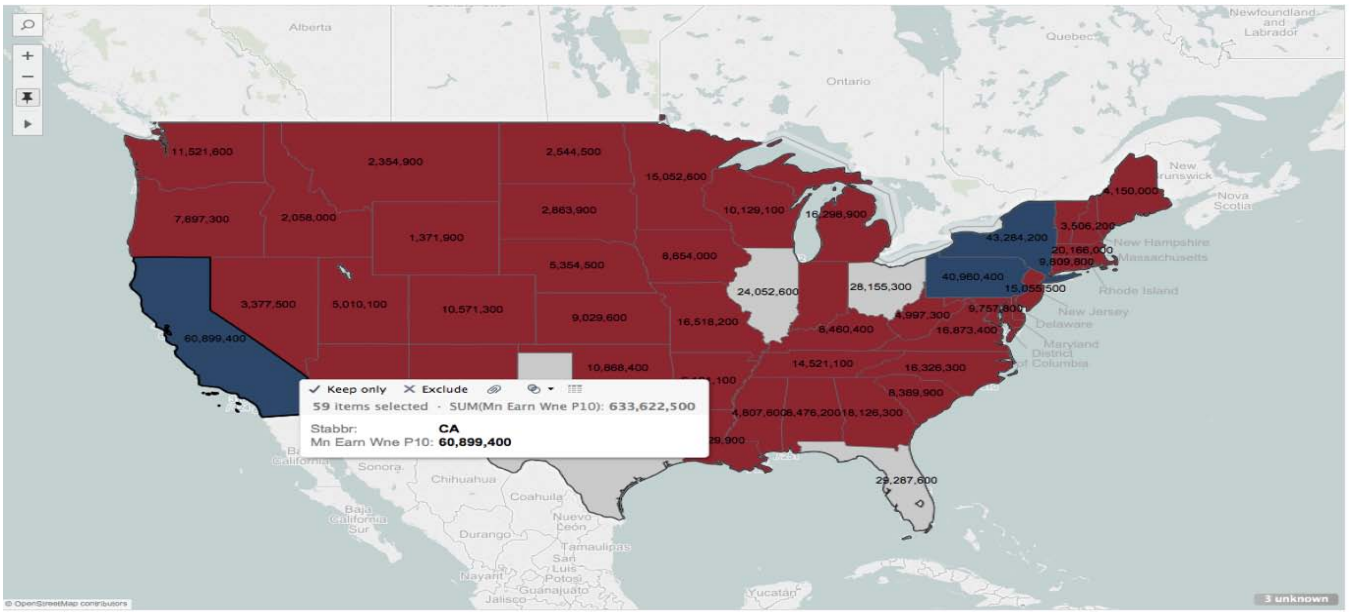


Figure 2: Top Mean Earnings with Respect to states (Blue – High, Gray – medium, Red- less) in USD.

Figure 3: Comparing Average Net Price of Two States in USD.

Public Universities

Figure 4: Net Price comparison of Public Institutions in USD.

Private Universities

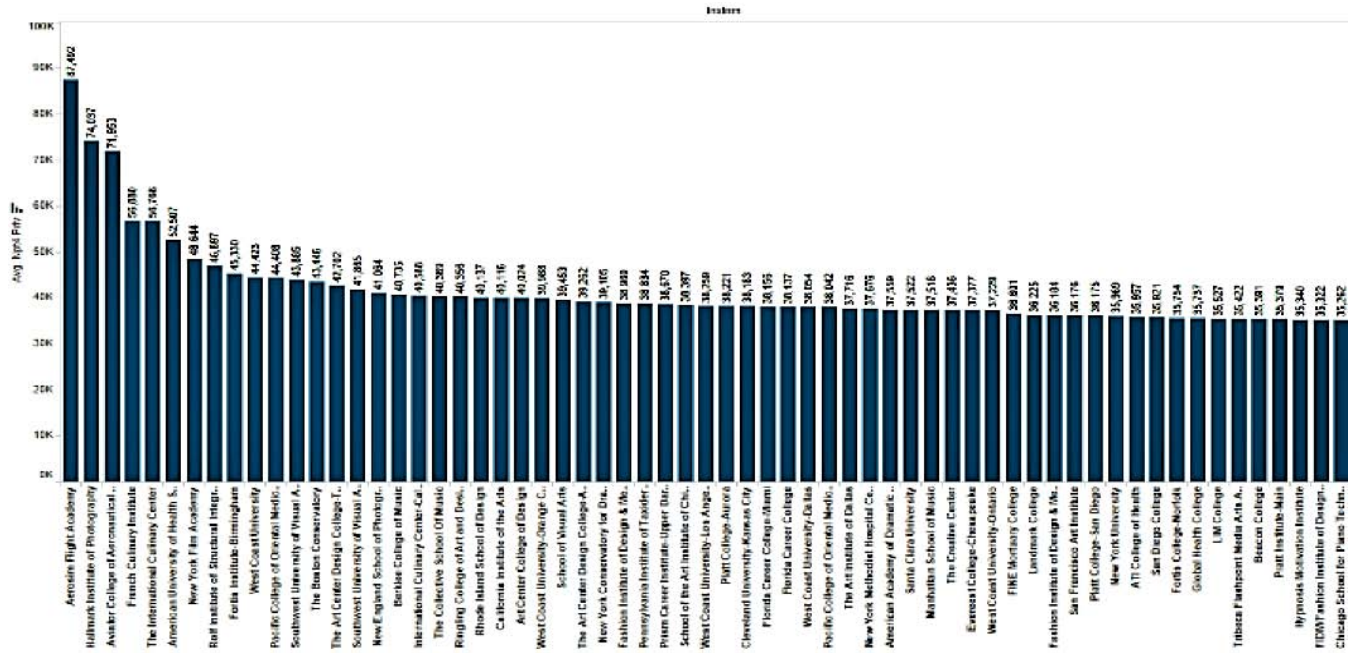


Figure 5: Net Price comparison of Private Institutions in USD.

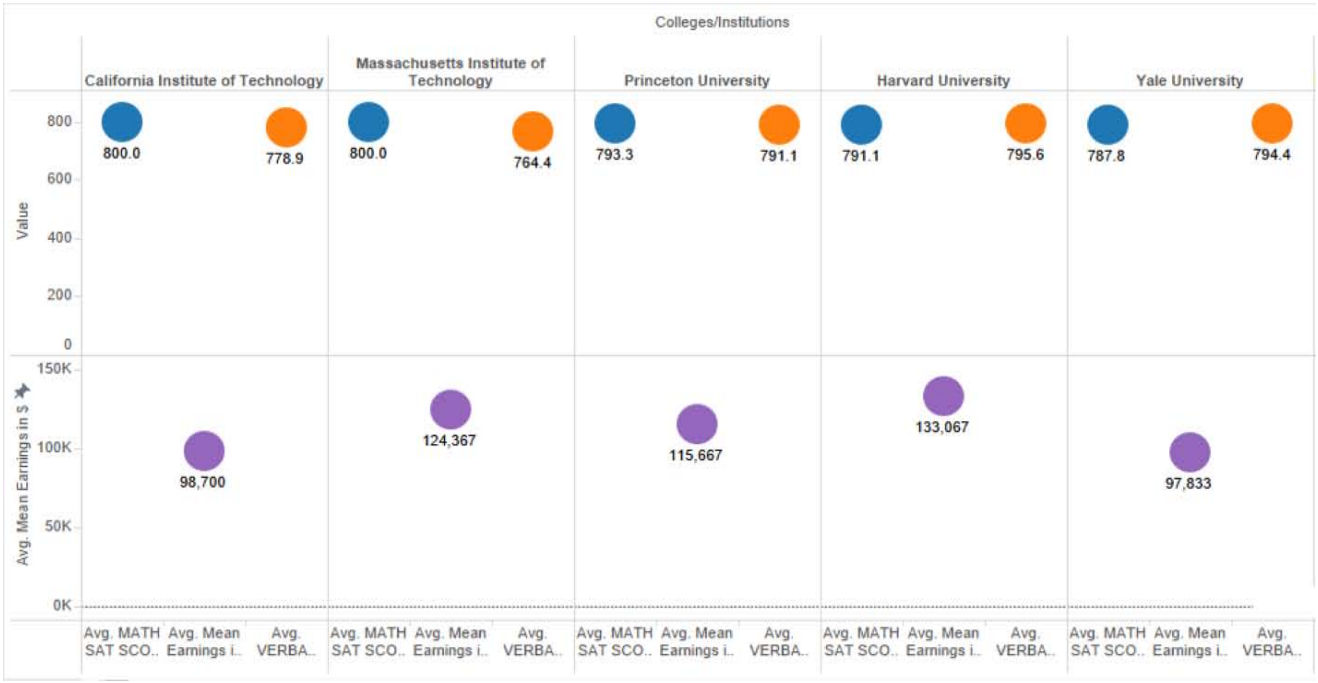
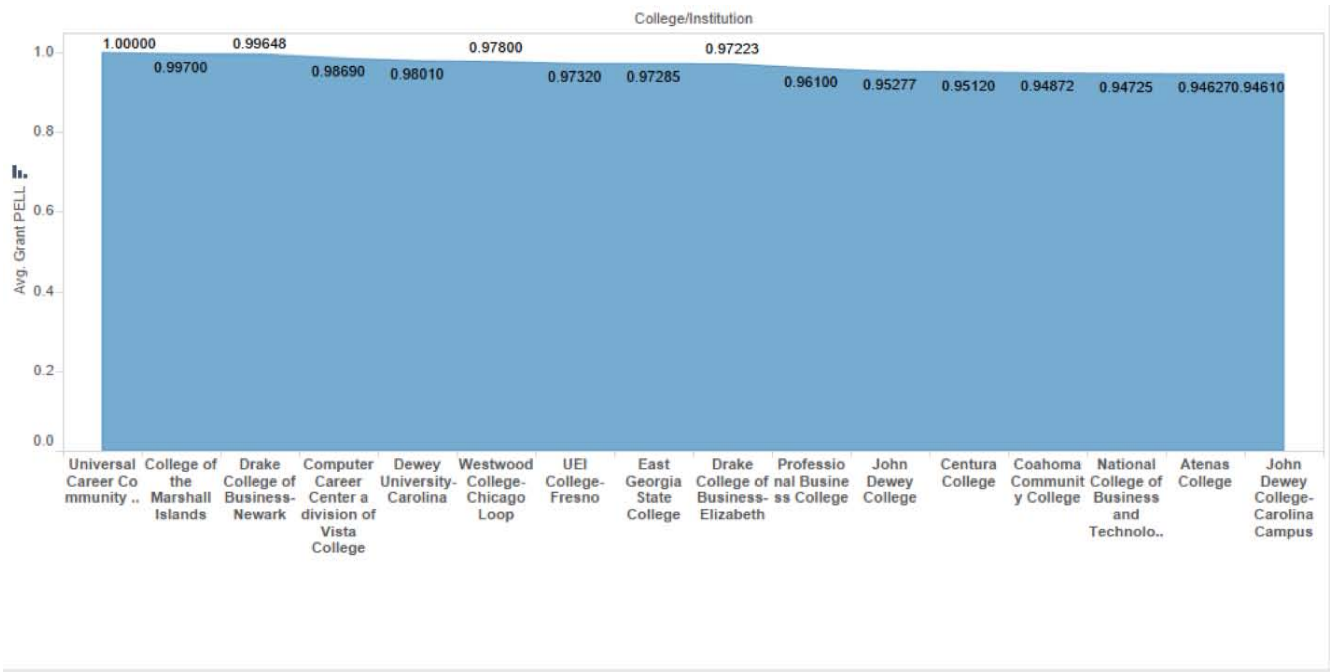


Figure 6: SAT Scores in Different Colleges on the scale of 800



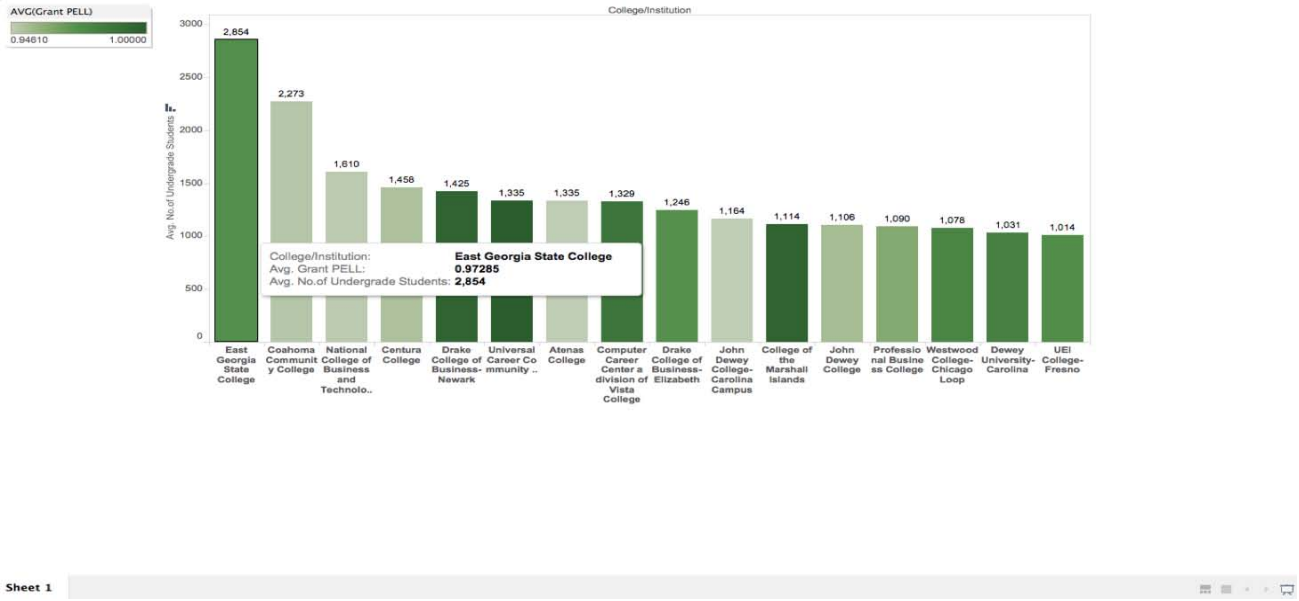


Figure 8: Average Undergraduates Receiving PELLGRANT in Each College in USD.

V. Conclusion

We adopt Spark Big Data platform to analyze college score card and display the insights. Choosing a college for your undergrad right after high school is every child's nightmare and insights like these give you a clear picture of the where about of the college. This kind of insight will be charged huge sum by data analyst for what we just presented. We have found out different colleges have different values in terms of earnings after degree. Two states have California and New York has the maximum average earnings after graduation. Also PELL grant is high in community colleges. These analysis is helpful for students to select the colleges based on their interest.

References Références Referencias

1. Kaggle. "US Dept of Education: College Scorecard Kaggle, n.d. Web. May 2016.
2. Federal Pell Grants." Federal Student Aid. U.S. Department of Eductaion, 2016. Web. May 2016.
3. "Data Analytics Track." Better Policy Decisions. Carnegie Mellon University's Heinz College, 2016. Web. May 2016.
4. "Highest Paying Bachelor Degrees by Salary Potential." Highest Paying Bachelors Degrees Pay Scale. Payscale, n.d. Web. May 2016.
5. Singh,Atinder. "CollegeScorecardAnalysis." Github - Atinder03. Github, n.d. Web. May 2016.
6. Pritwani, Kunal. "College-Historical-Analysis." Pritwanikunal/College-Historical-Analysis. Github, n.d. Web. May 2016.
7. "Market Basket Analysis Algorithm with Map/Reduce of Cloud Computing", Jongwook Woo and Yuhang Xu, The 2011 international Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2011), Las Vegas (July 18-21, 2011).
8. Market Basket Analysis Algorithms with Map Reduce", Jongwook Woo, DMKD-00150, Wiley Inter disciplinary Reviews Data Mining and Knowledge Discovery, Oct 28 2013, Volume 3, Issue 6, pp445-452, ISSN 1942-4795.
9. "Best Graduate Schools by Salary Potential", Pay Scale. (2016). Retrieved November 04, 2016, from http://www.payscale.com/college-salary-report/grad
10. "Association for Public Policy Analysis & Management", Nick Huntington-Klein Retrieved November 14, 2016, from https://appam.confex.com/appam/2016/webprogram/Paper16913.html.
11. Hurwitz, Michael and Smith, Jonathan Student Responsiveness to Earnings Data in the college Scorecard (October 1, 2016). Available at SSRN: https://ssrn.corn/abstract=2768157.

This page is intentionally left blank