CIS 528

**Introduction**

Here 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 analysists for big money as this kind of analysis provides insight on different aspects of college.

**Specification of Data Set.**

Data Set website - <https://www.kaggle.com/kaggle/college-scorecard>

We have data for over 100,000 colleges in the US spanning over 14 years.

Data Size – 1.33 GB

File Format – CSV

**Data Analysis Tools**

* Community Data Bricks – It was used for writing the spark code.

Databricks fully manages Apache Spark clusters in the cloud, giving it the ability  to ingest, analyze and visualize the data.

This platform includes many features like multiuser support, Interactive workspace and more.

**Cluster information**

Community Data Bricks Cluster

Memory – 6GB

Cores – 0.88

CPU Nodes – 1 Master Node

**Query**

Query 1 – results = sqlContext.sql('SELECT INSTNM, mn\_earn\_wne\_p10, STABBR FROM Scorecard\_Project order by (mn\_earn\_wne\_p10) desc')

Query 2 - results = sqlContext.sql('SELECT INSTNM, NPT4\_PUB, Year, STABBR, CITY FROM Scorecard\_Project where City = "Los Angeles" or City = "New York" order by LENGTH(NPT4\_PUB) desc')

Query 3 - results = sqlContext.sql('SELECT INSTNM, SATVR75, SATWR75, SATMT75, mn\_earn\_wne\_p10, CITY FROM Scorecard\_Project where SATMT75>500 order by SATMT75 desc, SATVR75 desc, SATWR75')

Query 4 - results = sqlContext.sql('SELECT INSTNM, CITY, STABBR, UGDS, PCTPELL FROM Scorecard\_Project where UGDS>1000 order by PCTPELL desc')

**Terms and Terminology**

1. Mean Earnings

Mean (mn\_earn\_wne\_p\*) earnings are for the institutional aggregate of all federally aided students who enroll in an institution each year and who are employed but not enrolled.

(mn\_earn\_wne\_p10) – Mean earnings for 10 years

(INSTNM) – Name of the College

1. Average Net Price

There are several elements in the Avg Net Price that are derived from the full cost of attendance (including tuition and fees, books and supplies, and living expenses) minus federal, state, and institutional aid, for undergraduate student. NPT4\_PUB

(NPT4\_PUB) – Average Net Price for Public Colleges

(INSTNM) – Name of the College

(CITY) – City of the College

1. SAT Scores – Verbal Math x College x Mean Earnings

Test scores of enrolled students are not reported for all institutions, but may help students to find a school that is a good academic match. The query includes 75th percentiles of SAT reading (SATVR75), math (SATMT75);

SAT Verbal (SATVR75) – SAT VERBAL.

SAT Math (SATMT75) – SAT Math.

(INSTNM) – Name of the College

1. Percent of Undergraduates Receiving PELL GRANT x College.

This element (PCTPELL), shows the share of undergraduate students who received Pell Grants in a given year. This is an important measure of the access a school provides to low-income students.

(PCTPELL) – Percent of UG Receiving Federal Loans.

(INSTNM) – Name of the College

**Conclusion**

We would like to conclude that 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 analysist for what we just presented.

**GitHUB**

**https://github.com/atinder03/CollegeScorecardAnalysis**

**Reference**

[**https://www.kaggle.com/kaggle/college-scorecard**](https://www.kaggle.com/kaggle/college-scorecard)

**http://heinz.cmu.edu/school-of-public-policy-management/public-policy-management-msppm/msppm-track-options/data-analytics-track/index.aspx**