Queries and Important Terms for the Project.

**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

**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**