MSDS 694 Distributed Computing

Task 1 - Data Selection

Group: #8, DurianCandy

Members: Sunny Kwong, Charles Siu, Sean Tey, Andrew Young

Topic Proposals

Student Name	Data Titles	Data Source (URL)	Size	Reason	Possible Analytic Goals
Charles Siu	NYC Parking Tickets NYC Parking Violation Codes	https://www.kaggle.com/new-york-city/nyc-parking-tickets https://data.cityofnewyork.us/Transportation/DOF-Parking-Violation-Codes/ncbg-6agr	NYC Parking Tickets: **8GB** (4 CSV files) NYC Parking Violation Codes: **13KB**	As a commuter by car, I am a strong opponent against parking tickets. Meanwhile, I'm annoyed by how often people are plagued by parking tickets everyday. Therefore, I wish to explore more insights about this lucrative source of income for city governments, and perhaps ways to avoid them for good.	Which type of violation takes places the most? Are there any seasonality associated with parking tickets? How much income do parking tickets generate for NYC?
Sunny Kwong	US Dept of Education: College Scorecard	https://www.kaggle.com/kaggle/college-scorecard	US Dept Data: **4 GB**	The expenses of college is becoming more of a problem in America, but we usually only provided a very basic understanding of the cost though tuition. I wanted to take a closer look at how bad this problem is, and also, is it even worth it.	Which schools are the toughest to repay loans (is it necessarily the most expensive schools)? Which schools are giving their students the most/ least help in paying for their education?
Andrew Young	Death in the United States U.S. Education Datasets: Unification Project	https://www.kaggle.com/cdc/mortality https://www.kaggle.com/noriuk/us-education-datasets-unification-project	**4GB** **238MB**	Death is an inevitable part of life. Each death record represents somebody's loved one, often connected with a lifetime of memories and sometimes tragically too short. Going through the cause of death and the correlation of the education background would be interesting. There could be some unknown insight discovered.	Found the correlation of education level and death type. Also, there could be some insight from age and location
Sean Tey	Health Insurance Marketplace U.S. Census Data (State Level) Zillow Rent Data	https://www.kaggle.com/hhs/health-insurance-marketplace https://www.census.gov/data.html https://www.zillow.com/howto/api/APIOverview.htm	**11GB** **TBD** **TBD**	As an international student, I have always been terrified about healthcare cost in the US. Out of curiosity, I would like to explore this data and see if it will give me some insights and maybe it will help me conquer my fear. "Know thyself, know thy enemy. A thousand battles, a thousand victories."	See how much insurance cost varies by state and does it share patterns with rent data etc. See if we can predict insurance cost for each state over the years 2013 to 2015 based on healthcare data but also government census data and zillow rent data as features.

Final Dataset: We decided to choose the **Health Insurance Marketplace** dataset because we agreed that rising costs in health care is the most important issue out of all of the issues we brought up, as health care is a resource that everyone needs, and health care being too expensive can have detrimental effects on society. Therefore, we all decided that we wanted to explore this more.