**Proposal Guidelines (100 points):**

**(20 points) Summarize and motivate your proposed project**

My proposed project is looking to see how energy consumption, production and expenditure affect homelessness in the United States. I would like to see if there seems to be more homeless people in regions of the United States that consume a lot of energy. I plan on establishing this relationship by looking at the median and mean income of each person for each state. By joining together my energy data set and my income dataset, I can then look and see if there is a relationship between energy consumption, homelessness, and income. I also would like to see the difference between the relationship of homelessness and consumption, homelessness and expenditure, and homelessness and production. The year I will be looking into will be specifically 2014 as it is recent enough to make a prediction of current day, and all my datasets include the year as most recent. The more energy a certain geographic region consumes and produces, the more they will expend for energy.

The motivation behind this proposal is the expensive electric bill around Ann Arbor and seeing all the homeless people pan-handling. If there is a relationship, this could explain why there are a lot of homeless people around in Ann Arbor.

**– (20 points) Choose and describe (at least) two different datasets.**

I have gotten two data sets from Kaggle and will be downloading the csv file to gather, join, and analyze the information. The first dataset involves homelessness from 2007-2016.

<https://www.kaggle.com/adamschroeder/homelessness>

This dataset shows every state (categorized by state abbreviations), the count of homeless people in that state, and the year of which this number has been recorded. This also includes a description of their status in homelessness.

My second dataset is called United States energy, census, and gdp 2010-2014.

<https://www.kaggle.com/lislejoem/us_energy_census_gdp_10-14>

This dataset shows each state (categorized by state abbreviations), the total number of consumption, production and expenditure in the millions of BTU and dollars per year.

My third dataset is called US household income statistics. I will be grabbing this data from the US census bureau for the year of 2014.

<https://apps.bea.gov/itable/iTable.cfm?ReqID=70&step=1>

**– (20 points) Describe how you might manipulate and join the two datasets.**

I will be joining my energy dataset with my income dataset by a category of states. Because they both are grouped by actual state names, and then I will join the homeless dataset to my energy dataset by state name abbreviations. For all the cells in the datasets that are empty and not in the year of 2014, I will be removing these cells as they are not relevant.

**– (30 points) Describe at least three large-scale computation tasks you will perform to gain insights from the datasets (e.g. mrjob, spark, and sparksql). Each task should result in one meaningful analysis.**

I will be using mrjob to map, reduce, and combine a lot of my data in each dataset. Since each of the state data goes by county, I need to aggregate all the data together before joining them. An example of this will be getting all my Alabama data together by adding the counties together that fall into the state of Alabama. I will do this for my energy dataset as well.

I will join the data after that and then look at the count of homeless people to the state, to the total consumption of energy in millions BTU, total production of energy in millions BTU, and total expenditure in millions of dollars.

To run all my data quicker, I will be using spark to run most of the datasets because they are in the thousands to millions for rows.

**– (10 points) Describe at least one visualization you might create that highlights insights you hope to gain.**

I will be making a scatter plot to show if there is a positive relationship between energy and income. After this, I will be making a scatterplot showing the relationship between the joined data of energy and income with energy.