Proposal Guidelines (100 points):

– (20 points) Summarize and motivate your proposed project.

My proposed project is to understand what factors such as income levels may impact pollution levels. For instance, one insight I am hoping to understand whether individuals who are richer live in areas that have little to no pollution while individuals who live in poorer areas have greater levels of pollution. It would be interesting to see what impacted what: if areas with high pollution were priced lower so only individuals with lower socioeconomic status live there. This has significant impact because areas that have high pollution are also correlated with higher levels of health problems. This project idea stems from my work in East Africa where I witnessed areas that were poorer were also living in closer proximity to factories, etc… that released pollutants in the air. They also had worst health outcomes such as asthma, heart disease, etc…

From the two datasets that I found, described further below, I am combining pollution level data with income level data in the United States. From this, I hope to better understand if there are factors such as number of households (how crowded an area is) or income level that are correlated with pollution levels.

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

The U.S. Pollution dataset contains the four major pollutants (Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide and Ozone) by mean concentration, air quality index within a given day, maximum value obtained in a given data, and the hour when the maximum concentration was recorded. City, county, state and date of monitoring are also captured in the dataset. More information on the dataset can be found here: https://www.kaggle.com/sogun3/uspollution

The US Income dataset contains mean and median household income data, number of households, square area of land and square area of water at location. City, county, and state are also captured in the dataset. The database contains 32,000 records on US Household Income Statistics & Geo Locations. More information can be found here on the dataset: https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations

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

Since both data sets have county level information, one possible methodology I will use to join the datasets is to join them via County Name. By doing this, I will be able to see both pollution measurements as well as income level measurements by county. I will use a join statement to merge the two datasets together. For counties that are missing, I will give them null values.

– (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.

1. Using mrjob, I will use map and reduce to count the number of cities within each county that are over the maximum pollution levels for Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide and Ozone set by the Environmental Protection Agency (EPA). Conducting this computational task will allow me to see which counties are polluting the greatest.

2. Using sparksql, I will take the mean income level for each county and map it against the number of cities within each county that are polluting over the maximum level set by the (EPA) for Nitrogen Dioxide, Sulphur Dioxide, Carbon Monoxide and Ozone. I will order the data by income level to see if there is a trend between income level and number of cities polluting over the maximum standard and if there are any differences between the pollutants. This analysis will allow me to see if there is any correlation between income level and various pollutants.

3. Using sparksql, I will examine if the number of households in a county have an impact on pollution levels within the same county. I will do this by counting the number of households in each county mapping this against the number of counties that exceed the maximum pollutant levels for each pollutant.

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

One visualization that I hope to create is from the second computational task that I am preforming. I hope to create a bar graph which illustrates income level on the x-axis and number of cities exceeding maximum pollution levels on the y-axis. This will allow individuals to easily see if there is a trend between income and pollution.