Members:	Kerry Wu	Lab 04	
	David Griswold	Lab 03	**Designated Presenter**
	Yannis Bi	Lab 04	
	Rifayat Uddin	Lab 03	

Description of Dataset: Dataset contains 7 different datasets. We received these datasets from the 2019 Ross Datathon that was held on March 22nd. We do not have exact sources for all the data given. Dataset 1 is about chemical contamination for counties in the United States from 1999 to 2016. It contains ~882,000 rows with 11 variables. We intend to use fips (unique code for state and county), county, state, chemical_species (Type of chemical contaminant or particulate), conntaminant_level (Classification of contaminant measure; 3 types: less or equal/greater than mean concentration, non-detect), unit_measurement (Measurement units of the contaminant), and value (Measure contaminant value, in measurement units). Dataset 2 is about drought severity for counties in the US and contains ~1 million rows & 11 variables. We intend to use fips, county, state, and classification of droughts (% of population affected by severity of droughts denoted by: none = not affected, d0 = abnormally dry, d1 = moderate, d2 = severe, d3 = extreme, d4 = exceptional). We do not intend to use dataset 3 which contains median earning for specific counties. We do not intend to use dataset 4 which contains an breakdown of responses of educational degree for a given county. Dataset 5 contains population breakdowns for specific industries and includes 5.712 rows and 18 variables. We intend to use fips, county, state, year and 14 various estimated working population in a specific industry. Dataset 6 provides various water metrics such as water use for irrigation, agricultural or hydroelectric purposes. It includes 3,224 rows & 117 variables. We intend to use fips, state, county, year, population, and some of the 110 other parameters given. Dataset 7 is a dictionary of the 110 other parameters given in dataset 6.

Issues: Some issues that we may encounter is not enough data for every single county. There is only data for 24 states for dataset 1, but dataset 2 contains data for 50 states + Puerto Rico and DC. We have to make sure that we are matching state for state and county for county, this is why we intend to use fips because fips is unique for a given state and county. Additionally, we have to worry about the year in which the data was collected as cross examining data from different years would skew our outcomes.

Question(s):

- How does industrial water usage affect the water contamination level of a certain county? Are there industries that contaminate water more than others?
- Do counties with lower fresh water levels have lower educational or industrial outputs?
- Which factors allow us to best predict the water contamination level of any given county?

Statistical Tools: We intend to use R to create the models which we interpret (such as KNN, log reg, lin reg, and more as we continue to explore the data) and Python for data manipulation. We will know more about the models once we continue to explore the data and select a question to explore.