Client:

- **Who is your client and how will that client benefits from exploring this question or building this model/system?**

Washington State Department of Health. They'll benefit by being able to better identify schools at risk of lead contamination in drinking water and provide resources to schools for prevention.

Question/Need:

- What is the question behind your analysis or model and what practical impact will your work have?

Can we identify the common factors in schools with high lead contamination in water?
The government can work with the schools to provide resources to reduce and prevent lead contamination in water to increase the health of the students.

- What data science opportunity and potential application do you see?

Exposure to lead contamination can affect children's' growth, behavior and learning problems. We can use data to identify the common factors in school with lead contamination and use this to better address and prevent the problem.

- What is the purpose of the EDA you will do and the data science model you will propose?

The purpose is to answer questions about the factors behind unsafe lead contamination levels such as:
Which schools are at risk of lead contamination and what is the risk level?
When are lead samples taken (day of the week and time)?
Where are the schools located (city, rural or urban)?
Are there areas that are prone to lead contaminations (located near farms or factories)?
What are the demographics of the student body and the median household income of the location?

Impact

- What impact are you trying to achieve? In other words, what will your work do for your client

The government will be able to address high lead contamination in drinking water in the short and long terms. By both identifying the schools with high lead contamination and understanding the common factors, the government can identify the steps to create safe drinking water for students.

- What is your impact hypothesis?

By identifying the factors of lead contaminations in water and school that are high risk, the government can work with the schools to improve the water system and implement frequent testing to ensure that the water is safe.

Data Description:

- What dataset(s) do you plan to use, and how will you obtain the data? Please include a link! (The link can be to the dataset you're downloading, the site you're scraping, etc.)

1. WA School Water Lead Test Data (school year 2018-2019)

Source: https://data.wa.gov/Health/WA-School-Water-Lead-Test-Data/i3jn-y8vx

2. List of public schools in Washington state (county, city, and district information)

Source 1: https://nces.ed.gov/ccd/schoolsearch/
Source 2: https://www.k12.wa.us/data-reporting/data-portal

3. K-12 enrollment by race/ethnicity, public schools in Washington (2018-2019)

Source: https://datacenter.kidscount.org/

4. Median household income estimates (WA)

https://ofm.wa.gov/washington-data-research/economy-and-labor-force/median-household-income-estimates

- What is an individual sample/unit of analysis in this project? In other words, what does one row or observation of the data represent?

One row of data represents the school's name, school district, county, city, date and time of lead sample collection, and lead concentration. Additional data sets provide demographics of student body at the school and median household income at the county level.

- What characteristics/features do you expect to work with? In other words, what are your columns of interest?

Lead concentration, date and time of analysis (day of the week and time of collection), demographics and income.

Solution Path

- What is your data science solution path?

Regression model: to identify the features that contribute to lead contamination in water and to predict the risk level of lead contaminations.

- What other solution paths might be viable?

Unsupervised clustering: to better understand the group of schools at risk of lead contaminations and the underlying factors that contribute to lead contaminations in water.

- How else could you accomplish your goals if not through data science? (Think practically; you might not even need advanced methods. If this is the case, make a justification for why advanced methods are the preferred solution path.)

The simplest way to address the lead contamination is to test more and try replacing pipes / installing water filters when the lead levels are high. However, data science is useful for a few reasons.

- Identifying underlying patterns: are pipes costly to replace, which means schools that don't have funding are more likely to have problems? Are lead levels a bigger problem in agricultural and/or industrial regions?
- Is there a proper amount of testing for this or is the testing too infrequent?

- Is this problem treated with the proper importance? Are there schools with high levels of contamination for a long period of time?
- Are there any outliers? Is the problem individual sinks rather than the entire school?
- Replacing the water pipes and installing water filters can be costly, so getting a better understanding of the extent and nature of the problem will be beneficial.
- Checking the water quality of other places in the city (city water quality report) will help identify whether everyone has access to clean drinking water.

Criteria for Success

- What does success look like?

Identifying the factors that cause lead contaminations and schools that are at high risk of contamination currently or in the future.

- How specifically will you measure your results in order to determine whether your work is successful? Think about the practical goals of your work.
- Test water samples for lead at school and the city.
- Potentially sample drinking water at home and in the community to check if the drinking water elsewhere is contaminated or not.
- Take blood samples of students before and after implementations of water contamination prevention (such as replacing pipes or installing water filters) and see if lead is found in their system.

Assumptions and Risks:

- Given what you've said about your desired impact and the solution(s) you will use to achieve it, what are some assumptions you're making?

Assuming that:

- lead testing was done correctly.
- lead contamination is not a seasonal problem, since the data of water testing was collected in the winter only.
- students live in the city of their schools
- schools do not take current prevention methods such as water filters installations and replacements of old pipes and don't report them.
- students drink water at school and not from bottled water.

- What are the risks to the approach you've identified?
- Frequent water testing is costly.
- Schools lead risk level is based on one year sample, so we could potentially misidentify schools that may or may not be at risk.
- Not accounting for schools that were not included in the state wide water testing.

Tools:

- How do you intend to meet the tools requirement of the project?

Excel for exploratory data analysis and Tableau for visualizations

- Are you planning in advance to need or use additional tools beyond those required?

Web scarping for additional data that not available as an excel file.

MVP Goal:

- What would a minimum viable product (MVP) look like for this project?

Graphs showing the top 10 schools with the highest lead contamination and the location of the schools.