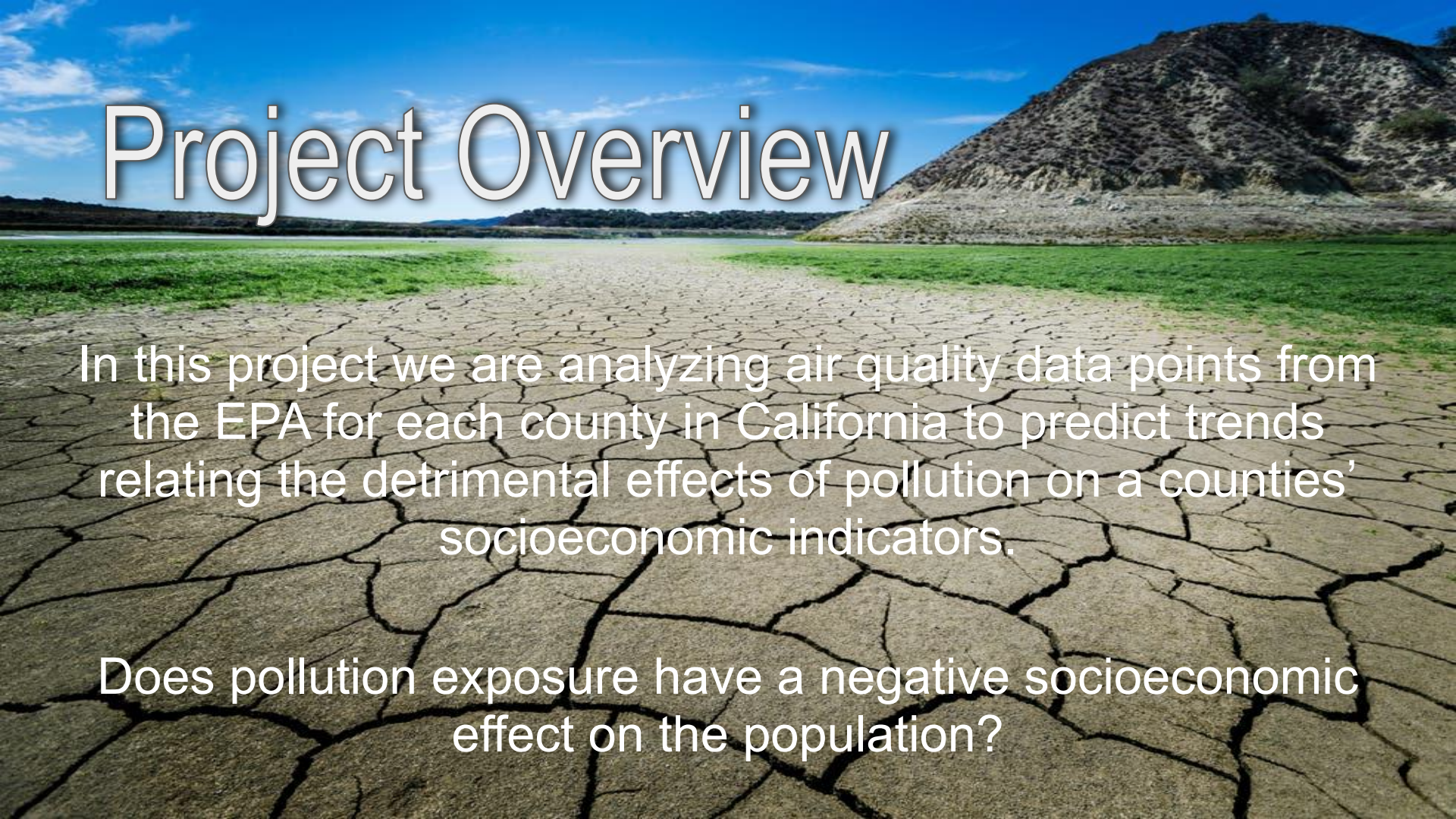

Air Quality Prediction
The Effects of Pollution on California's Socioeconomic Indicators
David Aaronson
Josh Allen
Ashley Burneka
Cynthia Marin



Project Overview

In this project we are analyzing air quality data points from the EPA for each county in California to predict trends relating the detrimental effects of pollution on a counties' socioeconomic indicators.

Does pollution exposure have a negative socioeconomic effect on the population?

Statistical Hypothesis Testing

Hypothesis: The pollution exposure of different regions in California presents a negative socioeconomic effect on the population.

Null Hypothesis: There is no negative correlation between pollution exposure and negative socioeconomic variables.

Alternative Hypothesis: There is a negative correlation between pollution exposure and negative socioeconomic variables.

Location Variables:
Census Tract
Total population
California County
Zip
Longitude
Latitude

Pollution Exposure Variables:
Census_Tract
Ozone Concentrations
PM2.5 Concentrations
Diesel PM Emissions
Pesticides Use
Toxic Release from Facilities
Traffic Density

Socioeconomic Variables:
Census Tract
Asthma
Low Birth Weight
Cardiovascular Diseases
Poverty
Unemployment
Traffic

Air Quality ERD

Machine Learning Method

We intend to use Neural networks (also known as artificial neural networks, or ANN). Neural networks are an advanced form of machine learning that recognizes patterns and features in input data and provides a clear quantitative output. In its simplest form, a neural network contains layers of neurons, which perform individual computations. These computations are connected and weighed against one another until the neurons reach the final layer, which returns a numerical result, or an encoded categorical result.

Advantages:
Effective at detecting complex, nonlinear relationships.

Have greater tolerance for messy data and can learn to ignore noisy characteristics in data.

Disadvantages:
The layers of neurons are often too complex to dissect and understand (creating a black box problem).

Prone to overfitting (characterizing the training data so well that it does not generalize to test data effectively).

Machine Learning Model

Using SK-learn for linear regression prediction model yielded poor accuracy when using all 6 feature parameters without scaling feature data

- Needs: Try scaling X input data for model using Sklearn.preprocessing StandardScaler()
- Tools:
  - Sklearn.preprocessing StandardScaler()
  - Sklearn.linear_model LinearRegression()
  - Numpy reshape()
    - We tried using the KMeans optimization method of building the NN and so far results have shown no prediction accuracy and complete data loss.

Machine Learning Model

Used **ReLu activation** as we are predicting something that is using linear regression

Used **keras-tuner** to find the best KMeans NN model

- Using **adam** optimizer
- **Binary_crossentopy** loss

Tools:

- **KMeans**
- **TensorFlow**
- **matplotlib.pyplot**

Tools for Dashboard

- Tableau will be used to create an interactive dashboard that will display our results and tell a story about the air quality in relation to health concerns in the state of California.
  - A “clean” version of our data will be imported into Tableau
- GEOJson to create custom maps to be used in Tableau.
- Javascript will be used to create a website that will display our results.
  - The website will contain tables that can be filtered to view different results.
- Link to Dashboard/Storyboard in progress.

Interactive Elements

- Layered maps with the ability to select one or more variables.
- We will also create a webpage that will be viewable in the Tableau dashboard as an object.
    - Webpage will also include tables that can be filtered to view different results.
- Use multiple views to filter other views in our dashboard.
- Ability to navigate from one view to another view, dashboard, or story
- We will use the Highlight action to bring attention to specific results.

Visuals on Tableau