**R Final Project**

**Abstract**

1. **Dataset**

Dataset – Death Due to Air Pollution:

<https://www.kaggle.com/datasets/akshat0giri/death-due-to-air-pollution-19902017/data>

There have been a few people that analyzed this dataset and posted it on Kaggle, however,  
the analysis mostly consists of making graphs and only surface level analysis, such as  
identifying which countries had the most mortalities. The code with the most analysis,  
generally only focused on total mortalities, and trends regarding that, without making  
much distinction between the different types of air pollution, so we’d like to go more in  
depth.

1. **Research Question/Problem**
2. **Research Methods**

With this dataset, we plan on using regression techniques to uncover long-term

trends in mortality rates in countries and their correspondence with specific types of

pollution. Additionally, we plan on adding a new column, mapping the country code to

specific regions, so we can not only do country-specific analysis, but also aggregate data

by region and find insights within a broader geographical scope. One way we plan on doing

this regional analysis is by using clustering algorithms on a country and regional level, then

comparing the results for the two. Through this, we could identify whether there’s regions

that tend to be at higher risk, or if these risks tend to be more localized

**4. Project Plan**

**5. Analyze Data and Results**

**6. Summary or Conclusions**

**References**