Yasmin Shaker

Oct 24, 2023

Project Plan

For my data capstone project, I wanted to focus on several factors that alter the environment. The main data set for my project is the illegal dumping data set from Data Montgomery. This data set shows reported illegal dumping activity from individuals. It has over 7000 rows of data and 12 columns. The columns that hold more importance to me are; the year of the cases, case description, case subtype (which shows the type of land where the activity occurred), longitude and latitude, to help map the locations, and city. This data set is updated daily and is collected from the Department of Environmental Protection. One of the challenges that I’ve already run into with this data set is trying to map all incidents using the longitude and latitude, there are over 2000 cases where these measurements are not available, and I would be losing over 25% of the data.

Another data set I’ll be using for this project is the Annual air quality index of the county by the Environmental Protection Agency. As I was looking for a data set to show the air quality for the county separated by year, I was not successful until I found this [page](https://aqs.epa.gov/aqsweb/airdata/download_files.html). It has an annual summary for all counties in the US and shows the year for the data, the air quality index (AQI), the median and max AQI, the number of unhealthy days and unhealthy days for each year, and other data for specific pollutants. I collected that data for Montgomery County for the past 20 years. This set could help with showing significant changes in the air quality of Montgomery County.

Another data set is the weather data set from the National Centers for Environmental Information. This set shows the weather for all 12 months of the years from 2000 to 2022. It shows the average temperature, minimum and maximum temperature, precipitation, and heating and cooling degree days. This data set shows the weather patterns for each month and is updated monthly. This helps with seeing the weather patterns of Montgomery County. The last data set is from the Census Bureau and it shows the population demographics for Montgomery County, I think I mainly want to see the population changes in Montgomery County, mainly to see if illegal dumping cases increased as the population increased.

There are several questions I have in mind that I’d like to answer, and they mainly focus on how clean Montgomery County is. First, I want to see the hotspots for illegal dumping activity to see if there are any concentrations in certain towns. I’d also to see if I can create a bar graph to show what types of illegal dumping cases are most common. This could be done using R to map the longitude and latitude of each case. Since the illegal data dumping set has a column that shows the years, I want to see if the number of cases has increased or decreased over the years, and I think graphing it in a line graph could show that. Another question I had was to see if there are significant air quality changes over the years, this also could be done with a lone graph. Using a scatter plot, I’d also like to see if there are any patterns between weather data and air quality data.

Since my main data set is the illegal dumping data set, I’m curious to see if some years with high numbers of cases have any noticeable weather patterns or air quality levels. Through that, I'd like to focus on comparing the years of the highest and the lowest illegal dumping incidents and see if there are major differences in population and air quality. Lastly using a regression model, I’d like to see if I can build a model to forecast future illegal dumping cases in the future.

There are a few main tools I am planning on using for this project, they are Excel, R, and Python. I’ve already begun using excel to create a separate spreadsheet with longitude and latitude to help in mapping the illegal dumping cases. And I think Excel will be helpful in simple format changes in data sets. I am more familiar with R, and I think it will be useful in creating certain graphs and maps. I’d like to use matplotlib in Python to create some of the graphs and I’d like to use Python to create the regression model.

Lastly, there are many challenges in this project. My first unknown is whether or not I will be able to manage my time efficiently to finish each step of this project. I also think I’ll be facing some challenges in creating the regression model because of the lack of expertise in Python. I also have some unknowns when working with the data, and I need to make sure I am correctly understanding and using the information.

Data sets and links

* Illegal dumping is from data Montgomery - <https://data.montgomerycountymd.gov/Environment/Illegal-Dumping-Activity/d985-d2ak>
* Demographics from the census bureau site, for 2000, 2010, and 2020
* Air quality data from <https://aqs.epa.gov/aqsweb/airdata/download_files.html>
* Weather from <https://datacentral.press-citizen.com/weather-data/montgomery-county/24031/2023-09-01/?syear=1895&eyear=2023#summary> provided by National Centers for Environmental Information (NCEI)