Project Data Options

ALL first and last group names here

2024-09-08

Your final project is a substantial part of your grade in this class. We will have a series of check ins:

- Project data selection
- Exploratory data analysis
- Model results OR Extended Visualization Poster draft

This will culminate in a final presentation.

For this project check in, complete the following for three candidate datasets. I will then meet with you to collectively decide which one to pursue.

For each dataset write:

- Where you got it. Include any applicable link and embed it in the markdown document. Also link to any accompanying documentation.
- Tell me what the dimensions are. How many rows, how many columns?
- What (if any) data is missing? How many observations?
- What questions interest you with this data set?

Project Pitch 1

- The Austin Animal Center Intakes dataset comes from the Austin, Texas open data portal.
- The dimensions change daily, but are currently 166,888 by 12 (row by col).
- Some entries are lacking an associated name, and some are lacking an associated address beyond the city. All entries are lacking an associated zip code (will need to use zip code tabulation shapefiles found on the US Census Website.
- Questions:
  - How does any particular point or period of time influence the quantity of intakes?
  - What area/regions of the Austin Animal Center's jurisdiction have higher quantities of intaken animals?
  - What breeds and ages of dogs/cats are most common to be intaken?

Project Pitch 2

- The Austin Animal Center Outcomes dataset comes from the Austin, Texas open data portal.
- The dimensions change daily, but are currently 166,889 by 12 (row by col).
- Some entries are lacking a name, and many are lacking an outcome subtype.
- Questions:
  - How does any particular point or period of time influence the quantity of outcomes?
  - What is the frequency of an intact male/female rather than spayed/neutered animal leaving the shelter?
  - Does outcome type change based upon time of year? If so, by how much?

Side note: I think it would be fun and interesting to build and analyze an expanded dataset (column-wise) by merging the intake and outcome datasets by the Animal ID. Some would be lacking connections, but it would *likely* be <1000 out of the 165k+ entries in both datasets.

Project Pitch 3

- The traffic incident report dataset comes from the Austin, Texas open data portal.
- The dimensions change daily, but are currently 389,629 by 10 (row by col).
- Most entries (>350k rows) are lacking an agency entry.
- Questions:
  - In what time frame do traffic incidents occur the most? (12-4a, 4-8a, or another range like 12-2a, 2-4a, etc..)
  - What area of Austin has the highest traffic incident report count?
  - What agency is assigned to the most reports? (will only have around ~33k rows for this question, but still plenty)