Crime Behavior Analysis Progress Report

Lauryn Davis, L Dettling, Jessica Malinowski

Our progress

Our implementation and progress of the crime behavior analysis project can be reported upon in consideration with the timeline provided within the proposal. As discussed, by this date our team hoped to have the data cleaned/merged and one multivariate and geospatial technique/visualization produced. Our data analysis has been conducted in a shared Github repository. This repository stores all excel and geographic datasets from 2012 through 2017 that were utilized within the "Real-Time Crime Forecasting Challenge" (*Real-Time*).

We started development by creating a file to read excel data via brute-force, and a main file to run our code (along with the appropriate testing file). Our testing file implements a merging test to verify that our merge contains all appropriate rows of information. We plan to explore more testing options for validation purposes. During this initial release, we used Anaconda to install Geopandas. After some initial releases, we created another file containing the code for importing and implementing Geopandas, a pandas geometrical library (*Documentation*). After the initial release, we organized our python files into executable functions for ease of access. We then implemented functions to handle cross-platform file reading - something our group did not anticipate as a challenge (see 'Challenges' section for more information). We then replaced our excel file-reading from a brute-force approach to an iterative loop implementation using the os library (*File*). Using the same technique, we attempted to replace Geopandas brute-force file reading with os as well, but ran into some compatibility issues with the files (see 'Challenges' section for more information).

Next we created some distributional balloon plot visualizations (Kassambara) within R studio under the file name, "Multivariate.Rmd" within Github. These figures can be seen on the final page of this document. These visualizations allowed for an initial gauge of trends that persist across time within the Portland Police Bureau's data. From Figure 1, the highest and most persistent call group is "disorder". The size in frequency across time for each call group tends to be stationary with dips in 2017, which makes sense because data was not collected during the second half of 2017. We can get a more representative multivariate picture of what is occurring across time by viewing figure 2. We can see that the most frequent "other" call groups are disorder, noncriminal/admin, and property crime.

After initial distributional visualization, we completed correspondence analysis using the package within R titled, "FactoMineR" under the same file name mentioned above. This is a statistical technique that uses contingency tables to map relationships among qualitative variables. As shown in Figure 3, we can interpret this by noting that 82.7% of the variability within the data can be explained by the first two dimensions of the correspondence analysis. Further interpretation can be understood by noting that variables farther away from the origin are considered more discriminatory (Kassambara). Therefore, personal crime, non criminal/admin, and traffic crimes seem to be quite discriminatory within the data. We can use a chi-squared independence test to see if there is a significant relationship between years and call groups. The chi-square statistic is equal to 604 and the p-value is <.0001. Therefore, we have evidence at the .05 alpha level that there is a statistically

significant relationship between the year a crime was committed and the type of call group associated with it.

More advanced multivariate techniques can be achieved when viewing multiple correspondence analysis results. From figure 4, we can represent each variable as a representation of their "squared cosine" values. These values represent the degree of association between variable categories and a particular axis. If a variable category is represented well within the two dimensions, its cosine squared value will be close to one (Kassambara). Although only 21.9% of the variation within the data is being explained within this plot, it's still interesting to note that the categories street crimes and personal crime have an important contribution to the positive pole of the first dimension, while "other" and property crime have an important contribution to the negative pole of the first dimension. We wish to use this information to do geospatial research on the frequencies of various call groups and crime categories. Our progress within this domain can be highlighted within "challenges" and "next steps" sections within this report.

Challenges

There have been some noteworthy challenges within this project so far. One challenge our group did not anticipate was file-reading between different operating systems. Because our group uses both Mac and pc, we have had to make sure that file reading is integrated for cross-platform implementation. This has been resolved with the os library in python, and using a file_name/directory variable for our excel data. This can be viewed in our Github repository under the file name, "Macxlsxdata.py"**.**

As stated, a goal we wished to achieve was geospatial analysis. A common library that conducts such work is known as Geopandas within Python. We have encountered a number of issues in attempting to successfully install and use a Geopandas environment. Two members had issues installing Geopandas from Conda, which is a known bug. To get around this issue we issued a multi line command that utilizes a common Conda recipe. The working installation instructions are located in the ReadMe file on the repository. Once all members successfully downloaded Geopandas we have still encountered issues with implementation. We have tried multiple approaches to reading the shapefiles. Two group members encountered an error with the Fiona file reader component of Geopandas, and another group member encountered an error with an unsupported filetype. The Fiona error could have been caused by a corrupted Conda installation or be an indication of a problem in the shapefiles themselves. Since the third member of our group also encountered an unsupported filetype error on some of the shapefiles, we believe there may be an issue with the shapefiles themselves. Remaining remediation efforts are detailed in the next steps section.

Collaboration

Our team meets one to two times per week on zoom or after class to discuss the crime analysis project. We have found a meaningful way to communicate with one another in order to define what deliverables and computer science platforms are needed within this project. For example, we have downloaded Anaconda and HomeBrew in order to successfully install packages that are

needed to complete this project. As of today, there have not been any issues in contribution of any team member. We continue to hold each other accountable in our responsibilities within this domain.

Next Steps

The plan to get Geopandas working for all users is multifold. The users with Fiona errors could consider a Conda reinstall or using Conda to reinstall Geopandas. The user with the filetype errors will try re-downloading from source to see if that alleviates that error. A third member is going to set up a docker container to try installing and running the project in that environment. If that works the docker image will be added to the repository, and the members can all develop in that environment. If none of these efforts are successful then the group will look at alternatives to Geopandas.

Our group would like to proceed with cleaning the *census_tract* column of our data set using Geopandas. Our current challenges with Geopandas have prevented us from using the *x_coordniate* and *y_coordniate* attributes to clean the *census_tract* data. We hope after the remediation efforts outlined above, we will be able to complete this step. Once Geopandas is implemented, and we can clean the census_tract data, we hope to move forward with a geospatial analysis technique in python.

Finally, we'd like to implement further multivariate techniques to tailor our project around the research questions asked within the project proposal. Many multivariate techniques must be conducted on quantitative data. A source discussing this says that there are methods that can be utilized to transform qualitative data into quantitative. This is a goal we wish to perform upon the qualitative crime data. We would then be able to use other classification and/or clustering methods such as principal component analysis, K-nearest neighbor, and non-hierarchical clustering (Donaires, 5295).

References

Documentation — GeoPandas 0+untagged.50.G9a9f097
https://geopandas.org/en/stable/docs.html. Accessed 18 Nov. 2023.

Donaires, Omar S. *Multivariate data analysis of categorical data: taking advantage of the rhetorical power of numbers in qualitative research*. Springer Nature, 2023, p. 5295.

File and Directory Access. Python Documentation, https://docs.python.org/3/library/filesys.html. Accessed 18 Nov. 2023.

Kassambara, Alboukadel. *Practical Guide To Principal Component Methods in R*. 1st ed., Statistical tools for high-throughput data analysis, 2017, www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/113-ca-correspondence-analysis-in-r-essentials/.

Kassambara, Alboukadel. *Practical Guide To Principal Component Methods in R*. 1st ed., Statistical tools for high-throughput data analysis, 2017, www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/114-mca-multiple-correspondence-analysis-in-r-essentials/.

Kassambara, Alboukadel. *R Graphics Essentials for Great Data Visualizations*. 1st ed., Statistical

tools for high-throughput data analysis, 2017, www.sthda.com/english/articles/32-r-graphics-essentials/129-visualizing-multivariate-categorical-data/.

Real-Time Crime Forecasting Challenge Posting, National Institute of Justice, 1 Aug. 2017, nij.ojp.gov/funding/real-time-crime-forecasting-challenge-posting#data

Visualizations

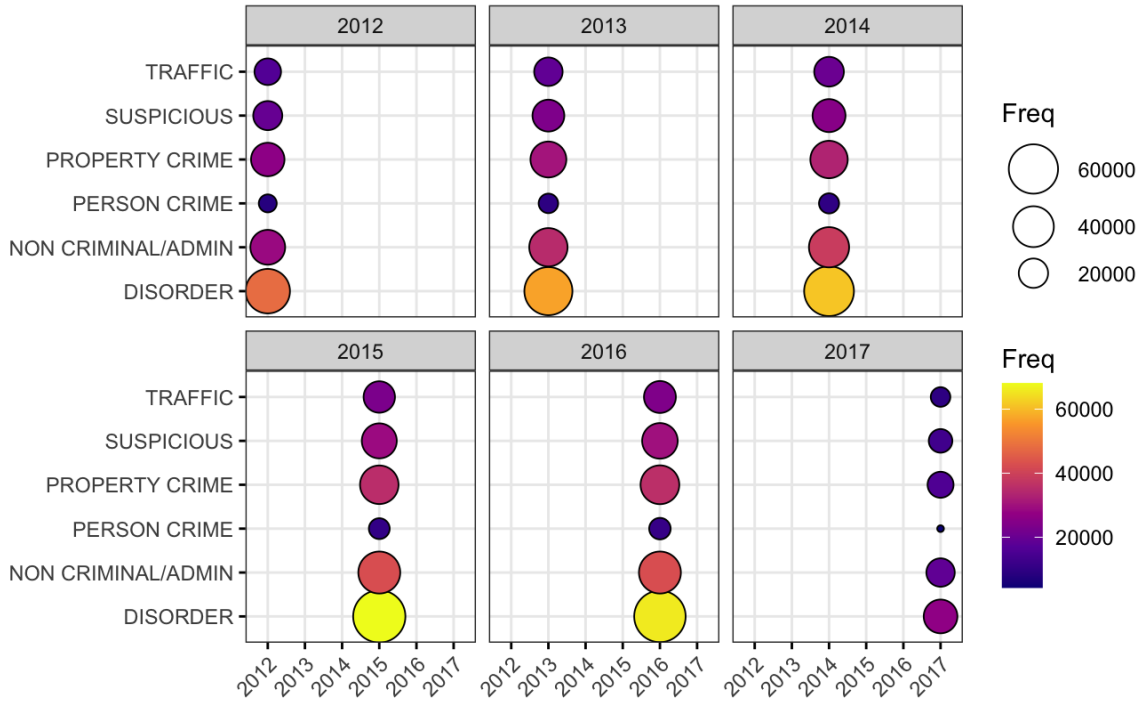

Figure 1: Call Group Distributions Across Time

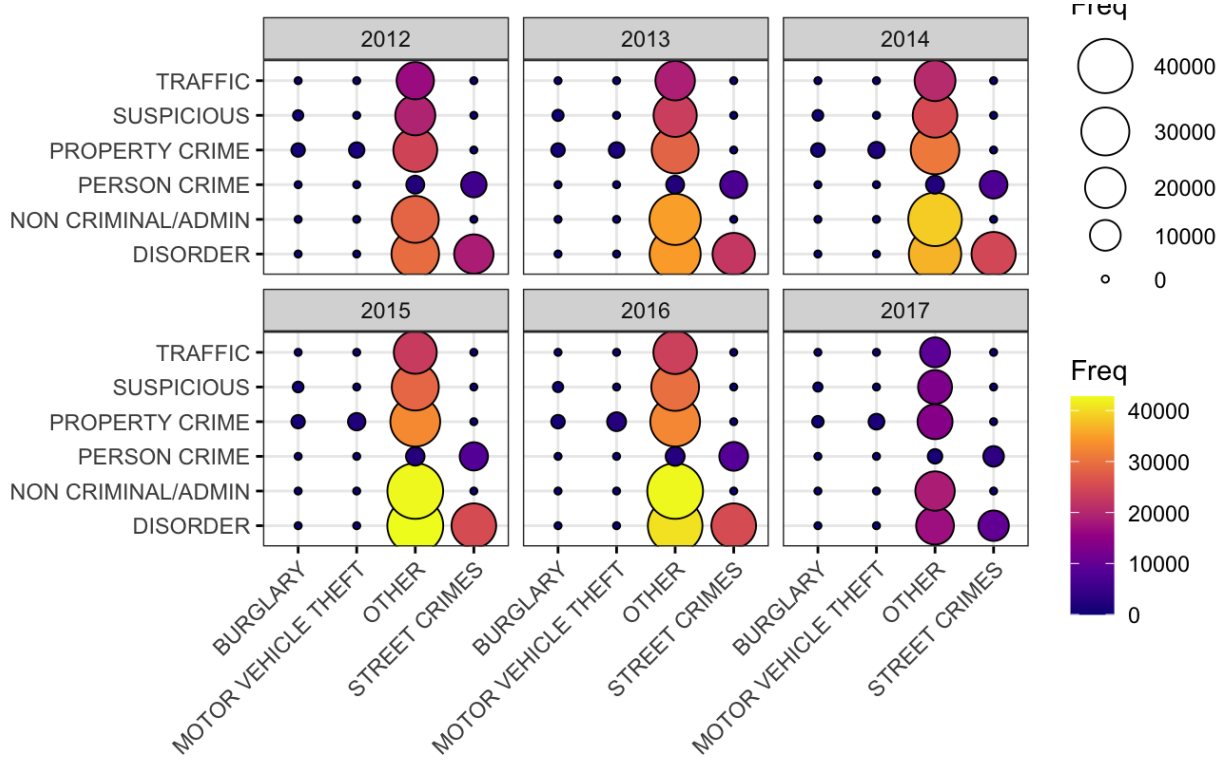

Figure 2: Multivariate distributions across time

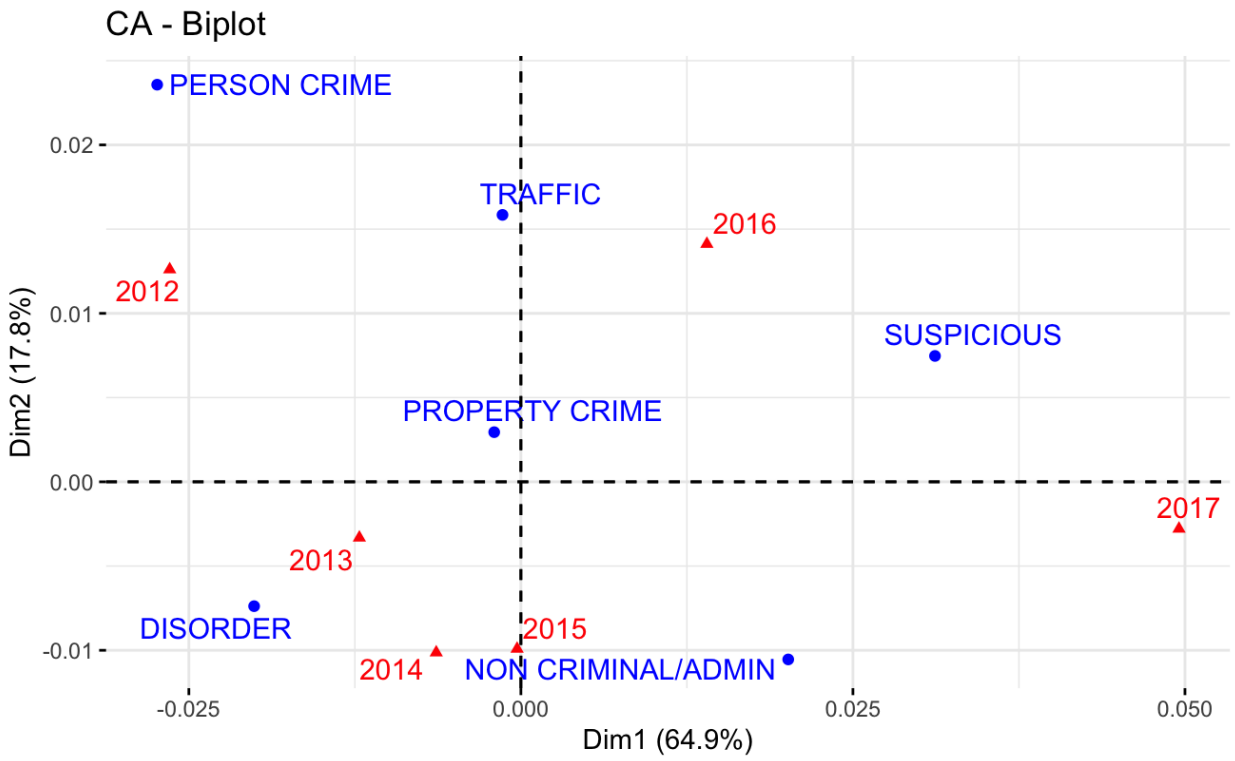

Figure 3: Correspondence Analysis Between Group Calls and Year

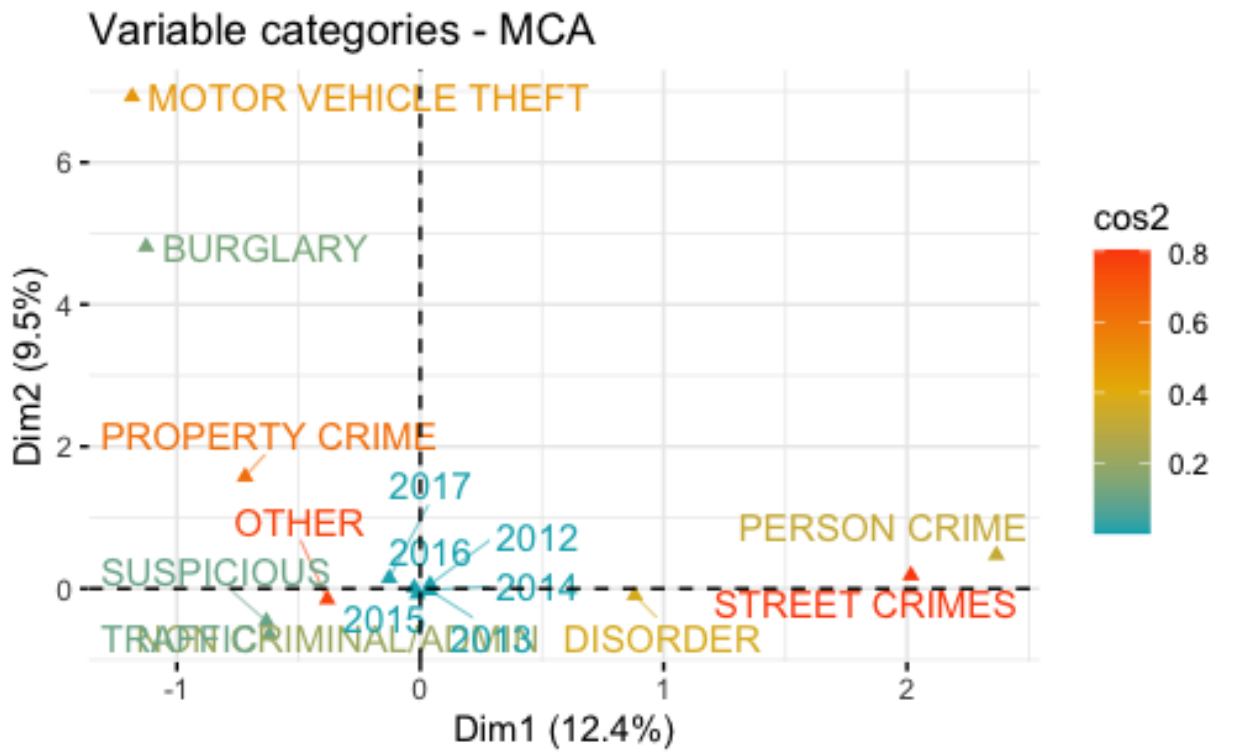

Figure 4: Multiple Correspondence Analysis Between Category, Group Calls, and Year