Capstone 2 Report
Justin Holladay, 2021

Applied Behavior Analysis is a data driven method of behavior change that began in the 1970s.

Problems:

1. Data is collected and examined. Entering data from paper into spreadsheets has been common for many years, but this leads to different methods and templates being used. Collating this data to be inputted into modern collection and examination tools can jumpstart companies' transition from paper to digital data.

Collating this data can:

- Allow examination of a clients historical data by modern software

2. There are many factors that can correlate to a clients' progress, but it is very time consuming to constantly examine all the possible relationships to a clients' behavioral progress. There are relationships that are not obvious as well as relationships that change over time.

Finding these relationships can help discover things like:

- If there are any relationships between bx and a client
- If more bx is observed with a particular staff member
- If a client reacts differently to one therapist than others
- If a therapist observes more of a type of bx between clients

######

Assist in training new therapists, as to prepare them for behaviors found frequently together with a client.

Identify anomalies between students bx as well as interactions of clients and therapists -- one therapist doing something differently than the others is causing something else to happen.

######

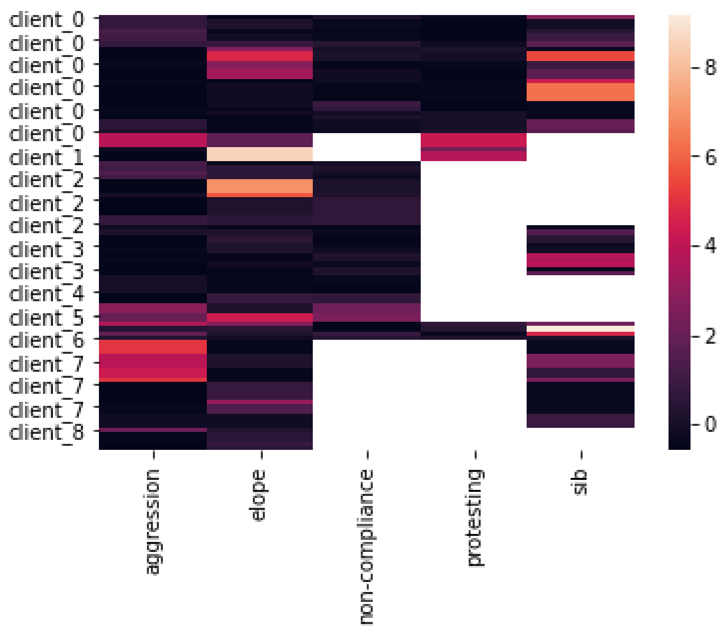
Data Wrangling

Behavior data was extracted from many differently styled spreadsheets of client sessions. Data was replicated from original sources using sessions of clients.

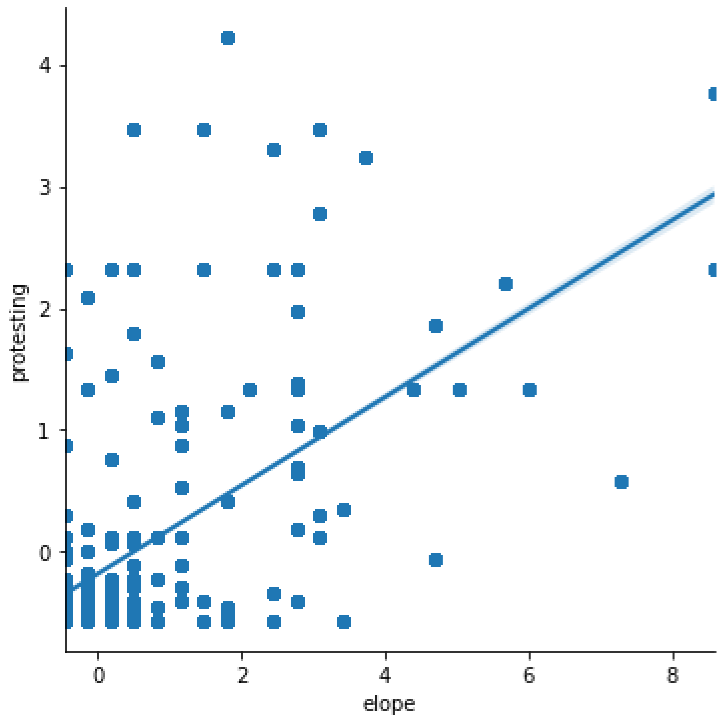
After defining which behaviors I wanted to standardize, the FuzzyWuzzy library in python to pull columns with approximate behavior names from many different spreadsheets into a single standardized frame. Each threshold for the fuzz matching was generated after manual examination of different possible inputs.

<u>EDA</u>

Looking initially through the dataset we can see there are some relationships between our variables. This heatmap shows some at a glance.



Through exploring our newly created dataset we can see an interesting relationship between eloping and protesting have the strongest correlation with overall behaviors. Another strong correlation is between aggression and non-compliance.



A number of different datasets were generated including lists of extreme behavior observations per client and reported by staff.

In [23]:

```
########################################################################
# Purpose: Produce list of how many extreme observations per client #
########################################################################


client_extreme = pd.DataFrame(columns=['client','entries']) # add bx col

for c in extreme_obs.client.unique():#extreme_obs.index.unique():

    aa = pd.DataFrame(extreme_obs.loc[extreme_obs.client == c])
    l = len(extreme_obs.loc[extreme_obs.client == c])


    client_extreme = client_extreme.append({'client' : c, 'entries' : l}, ignore_index=True).sort_values(by='entries

client_extreme
```

Out[23]:

	client	entries
0	client_0	26
1	client_7	19
2	client_2	14
3	client_3	11
4	client_1	6
5	client_4	6
8	client_8	5
6	client_5	4
7	client_6	4

Preprocessing

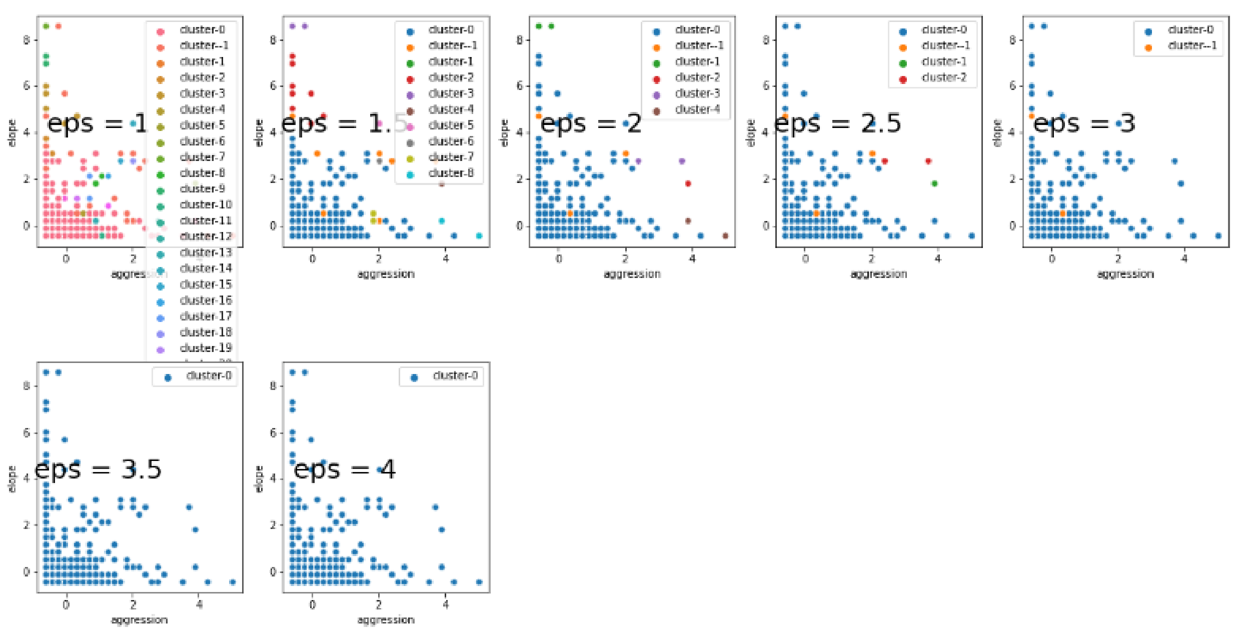
Dummy variables were created for categorical variables and merged into the datasets. The bx data was standardized using StandardScaler(). Using this standardized data, a manual anomaly detection method was created to define outliers within a session.

First, a threshold was generated for each client and each behavior observed in their sessions. Sessions were extracted where at least one extreme behavior was observed into extreme_df. After manual review, this will serve as true values.

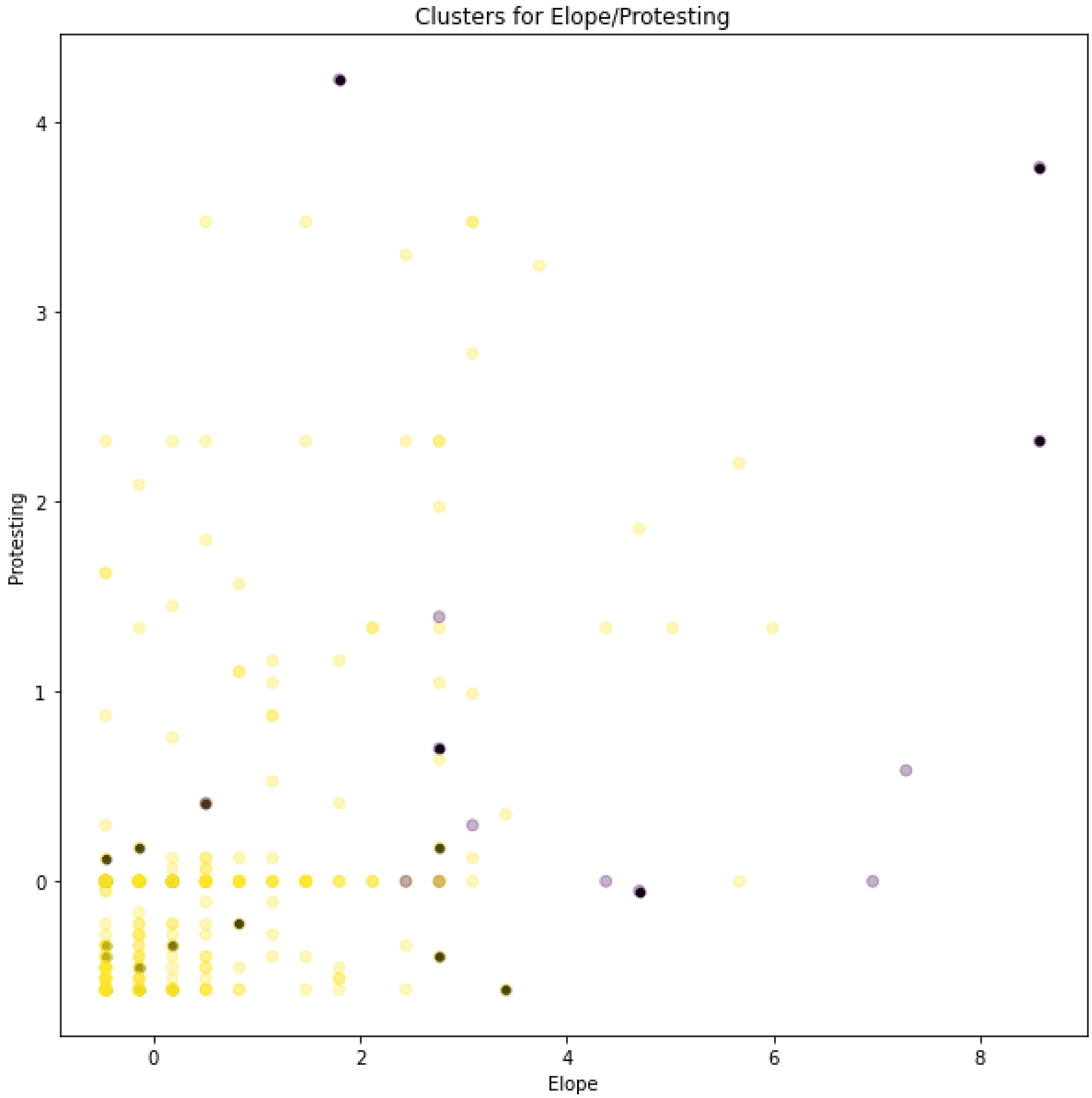
	Problem	Date	Day	Staff Member	Duration	aggression	elope	non-compliance	protesting	sib	client
client_3	elope	2021-06-17	3	staff_17	1.50	-0.585974	0.508188	-0.495307	NaN	0.405570	client_3
client_2	aggression	2021-04-05	0	staff_14	4.33	1.096018	0.508188	0.172368	NaN	NaN	client_2
client_6	na	2021-01-22	4	staff_7	7.00	0.348466	0.508188	-0.495307	0.411591	9.176212	client_6
client_3	na	2021-03-08	0	staff_21	4.00	-0.585974	-0.137430	-0.295005	NaN	3.860672	client_3
client_3	elope	2021-06-17	3	staff_19	4.00	-0.585974	0.508188	-0.495307	NaN	0.405570	client_3

Modeling

Next DBSCAN and KMeans models were created to cluster the original dataframe. DBSCAN hyperparameters were tuned by generating many possible values for epsilon and 2.5 eps with minimum 20 samples. An elbow plot was generated for KMeans, though 2 clusters were chosen for out wanted categories, not-extreme and extreme.



The result from DBSCAN and KMeans methods was similar with a silhouette score of 0.599 and 0.0.494 respectively. Final Project report as PDF. The report should clearly explain the problem, your approach, and your findings. Include ideas for further research, as well as up to 3 concrete recommendations on how your client can use your findings. Neither of these were better than the manual method of filtering.



Conclusion:

Outlier detection in ABA can quickly bring light to anything unusual for a client and discover unseen relationships in a number of ways.

Exploring the data lead to trying 3 different models (Manual, DBSCAN, KMeans) for outlier detection on a dataset of behaviors. The manual model is chosen over the others due to its clear inner workings and ease of making changes. The DBSCAN model performed at nearly 60% and could be investigated further when more features of this project are added.

Future work:

- Extend detection to include skill program targets in addition to behaviors.
- Collect information on gender, observation location and time of session.
- Define more relationships between independent variables.