Data Exploration

The team discussed a variety of different topics and subjects to consider for this project, we also looked at and read various topics on data exploration[1]. As we looked at the data available to us, we started shaping the discussion around our purpose. This gave us a more defined scope from which to focus. After sorting thru retail data, online sales and crises-specific major events, we opted for donation data focused on presidential election years.

Once we agreed on donation data, we focused on presidential elections and ways that the elections may be affected. This led us to narrow our focus on six specific states (AZ, FL, MI, NC, PA, and WI) and the counties therein. Historically, these states have been considered battleground and/or swing states, having a significant impact on the presidential election and electoral votes.

We set out to answer the following;

- Do presidential candidates influence how people donate?
- Can we predict who the president will be based on donation data?
- How much money is donated across candidates? States? Counties?
- Are donations affected by employment/unemployment?
- How do rural and urban neighborhoods affect voting?
- Does FL, MI, PA, NC, WI and AZ have a significant impact on elections?
- Do economics effect donations?
- How do major event effects on donations (9/11, real estate market crash, COVID) and/or presidential elections?
- Do demographics, such as age and education influence these data?
- What significance does health information have?

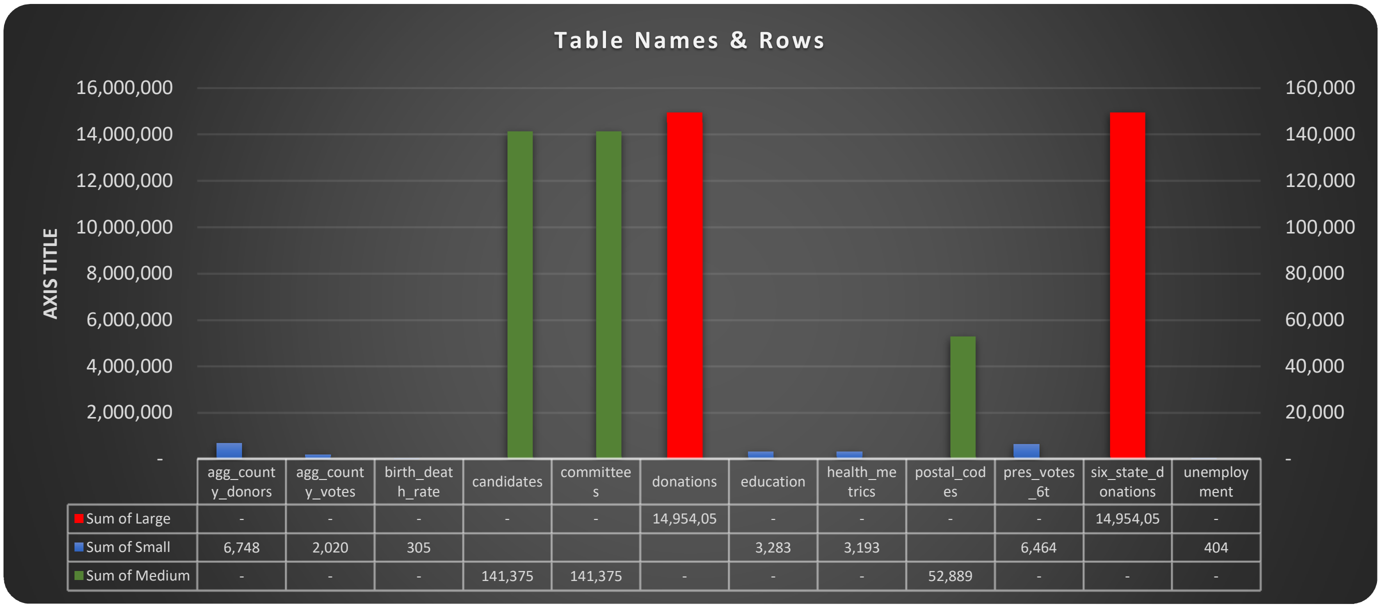
We explored the data and created several visualizations[2], looking at different aspects of the data we sought to explore. We also did some value counts of the datasets to determine the best storage options and how to pull the recall the information to create our models, predictions and perform analysis. We have twelve tables for this analysis. Below is a list and graph with the information.

	Description	Table Name	Rows	Columns
1.	Donors on selected states	agg_county_donors	6,748	12
2.	Voters on selected states	agg_county_votes	2,020	14
3.	Births and deaths	birth_death_rate	305	6
4.	Presidential candidates since 2000	candidates	141,375	17

[1] Toward Data Science: 15 Steps to Data Exploration https://towardsdatascience.com/15-data-exploration-techniques-to-go-from-data-to-insights-93f66e6805df

[2] https://www.sisense.com/glossary/data-exploration/

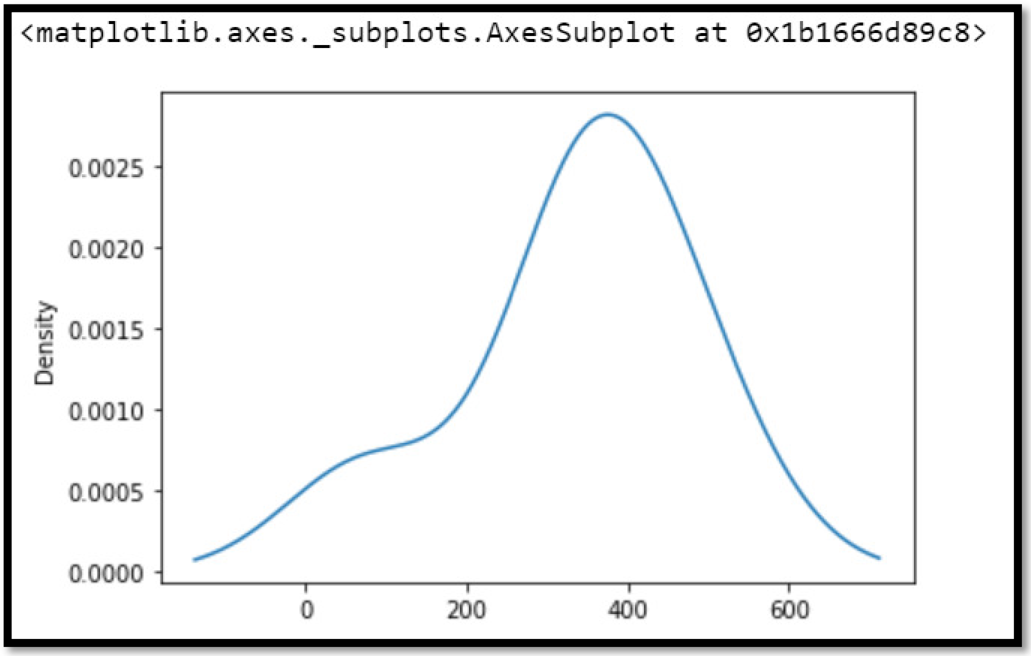
5. Committee presidential donations	committees	141,375	15
6. Donations for presidential votes	donations	14,954,054	21
7. Educational levels by state	education	3,283	47
8. Health information	health_metrics	3,193	507
9. County postal codes	postal_codes	52,889	5
10. Presidential votes on selected states	pres_votes_6T	6,464	11
11. Donations for presidential votes	six_state_donations	14,954,054	21
12. Unemployment rates	unemployment	404	114



	agg_county_donors	agg_county_votes	birth_death_rate	candidates	committees	donations	education	health_metrics	postal_codes	pres_votes_6t	six_state_donations	unemployment
🟥 Sum of Large	-	-	-	-	-	14,954,05	-	-	-	-	14,954,05	-
🟦 Sum of Small	6,748	2,020	305				3,283	3,193		6,464		404
🟩 Sum of Medium	-	-	-	141,375	141,375	-	-	-	52,889	-	-	-

We created various charts and graphs while exploring the data.

Total Votes graph

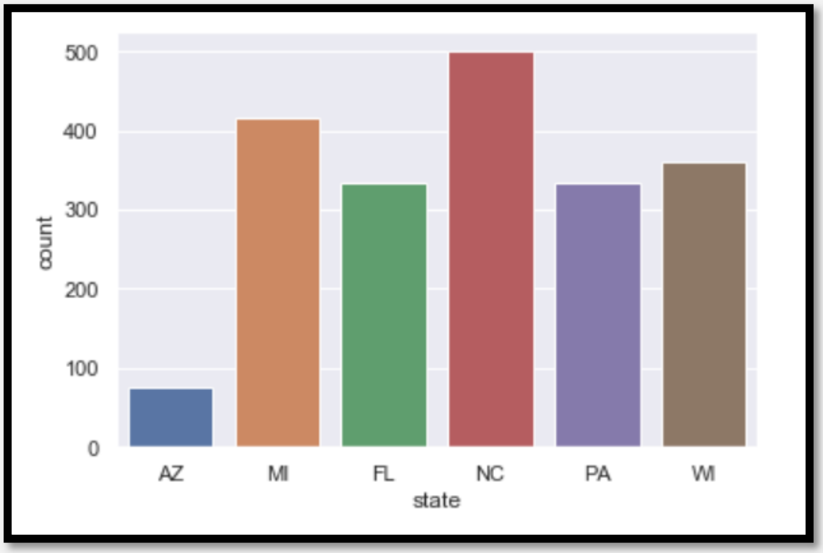


```
<matplotlib.axes._subplots.AxesSubplot at 0x1b1666d89c8>
```

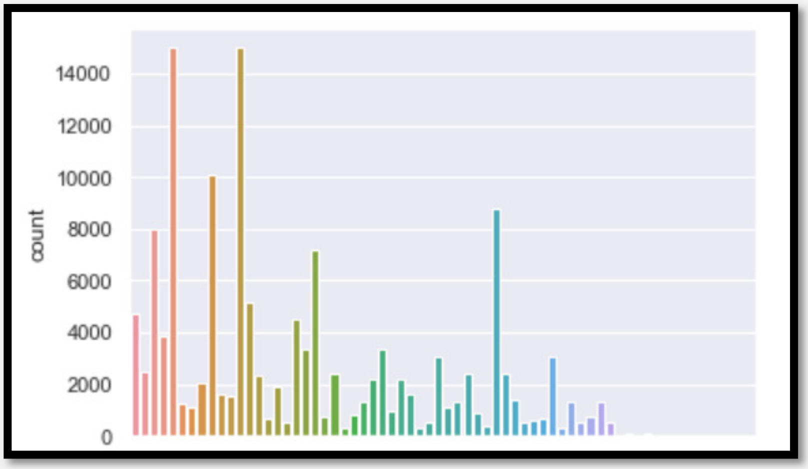
We used several steps out of the 15-step data exploration method[3] to describe our process. We also obtained data exploration information from Medium[4]

1. Unique value counts: examples of votes by county and candidates by state

Aggregated County Votes by State = 2,020



Candidates by State = 141,375

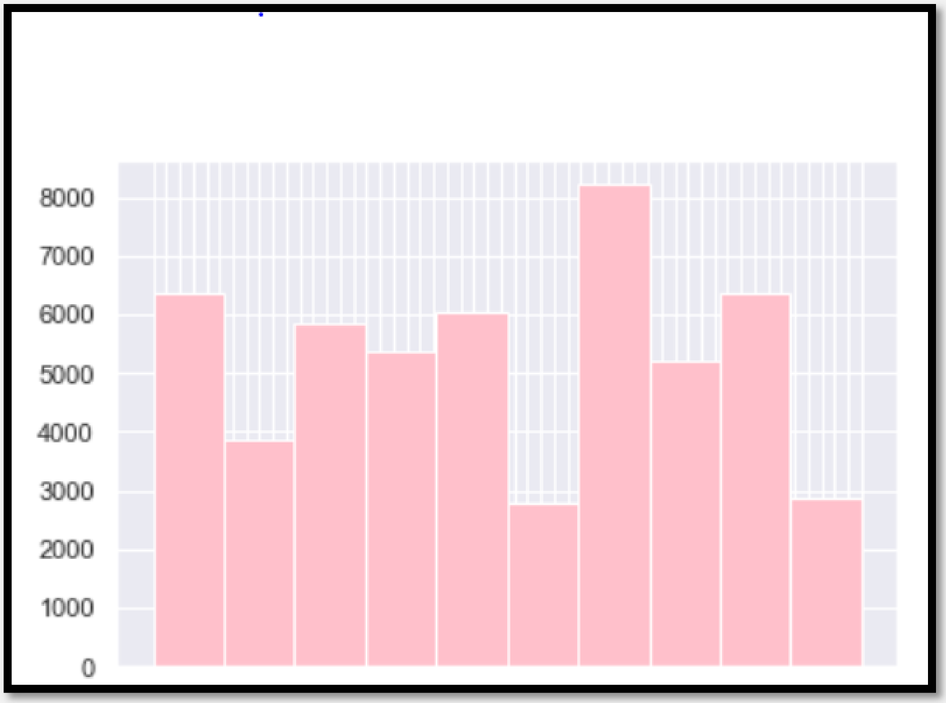


[3] Towards Data Science: "An Extensive Step by Step Guide to Exploratory Data Analysis" https://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-ddd99a03199e

[4] Medium: "Data Exploration" https://medium.com/@abhinavr8/data-preparation-and-exploration-5e09b92cf00e

2. Frequency Count

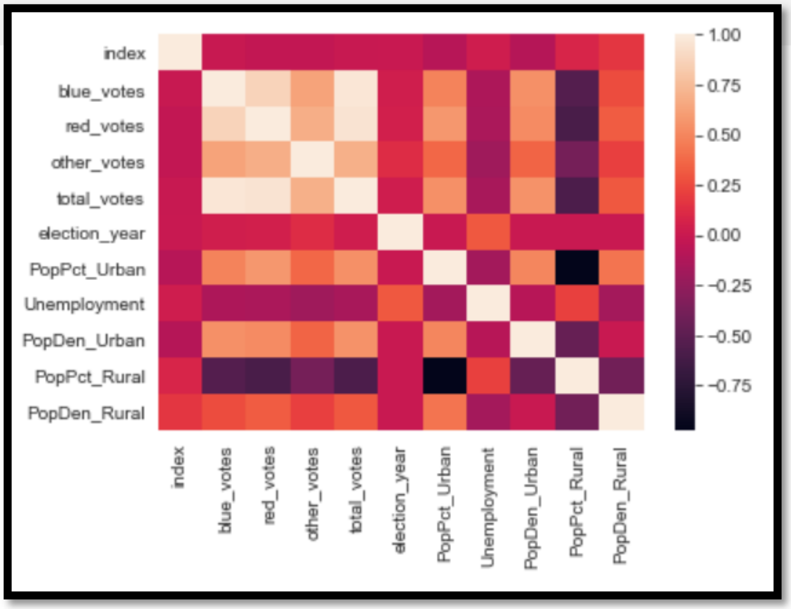
Histogram of Postal Codes by State



3. Correlation

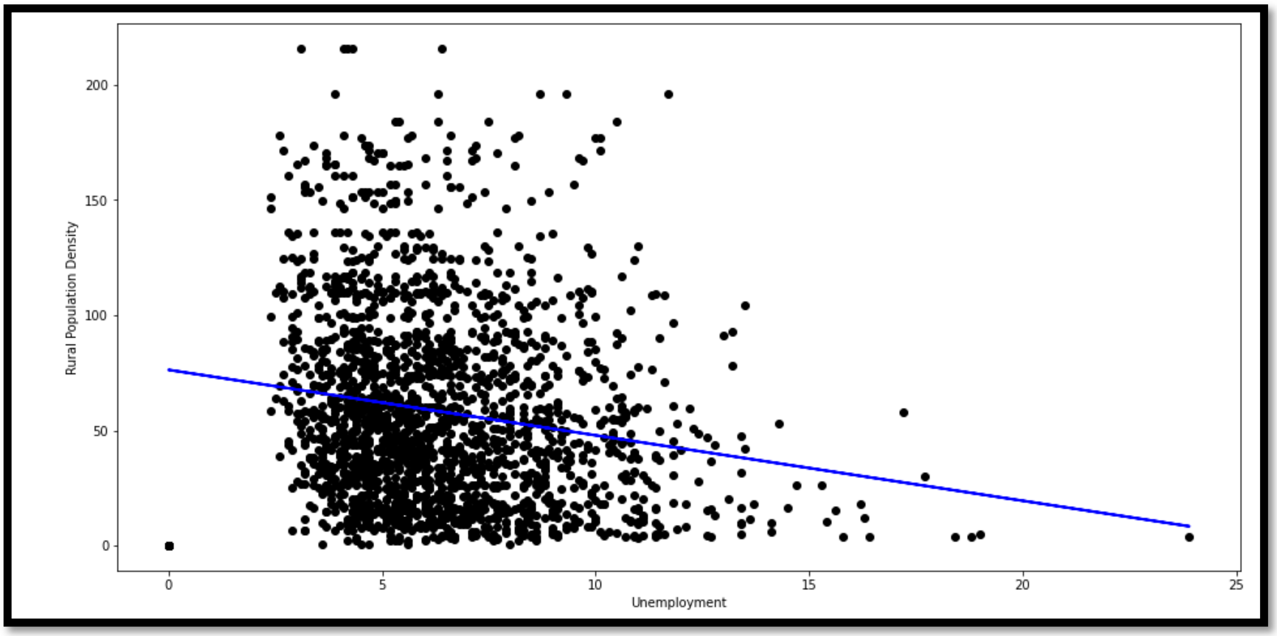
When it comes to analyzing numeric values, some basic information such as minimum, maximum and variance are very useful. Variance gives a good indication how the values are spread.

Correlation Heat Map



We looked at the relationship between different aspects of the data, to determine whether the correlation was positive or negative. This example shows a negative correlation.

Correlation: Rural Population w/Unemployment



Data Analysis

Our team used a variety or tools and tables to analyze the data.

- Jupyter Notebook
  This has been our main form of data analysis. We read in the tables from our SQL and Cloud connections to write simple to complex coding, focused on understanding and displaying the necessary data. To measure presidential data, we've pulled in donation and voter information, along with unemployment, education and demographics. We have discovered that these pieces of information have been useful in helping us transform and mold the data into information that is useful for looking at presidential elections.

- Spreadsheets
  Everyone on the team participated in data gathering and located information to assist in our project. This information was located online, using data from the FEC, Kaggle and the Census Bureau. These data were analyzed and saved as CSV files that were uploaded to GitHub, PostgreSQL PGAdmin and Google Cloud for easy access and data sharing.

- Google Cloud & PostgreSQL
  We place large dataset and bucketed information onto this platform. This allows to perform the ETL and join tables. Including the use of Colab and PySpark to load the data. The size of the database is over 15M and using 68GB.

- PG Admin
  We used PGAdmin to perform SQL queries and analyze datasets, as well as review table details.

Feature Engineering & Preprocessing

Feature engineering, also known as data preparation, was done during the first two weeks of the project. The team rallied together to select all the necessary types of data from various online options. We took various tables and processed the data in Jupyter Notebook and Google Cloud. This gave us a better understanding of each data element, along with what information we need to group, focus on and what we needed to remove. Then, we sorted thru the data and optimized it for our project. We believe, from the data, that looking at previous voter information, coupled with donation info and demographics might provide insights on our focus which is donations affecting presidential elections in swing states. We're testing to see if there is a correlation between the money people donate and who is elected.

We did the following joins:

- County
  - Linking with other datasets
    - Unemployment
    - Urban population density
    - Rural population density
  - Running aggregate
- Donations
  - Pull 4-year segment
  - Processed
  - Summation
  - Went from city level to county view
  - Processed for every county, within 6 states
  - Committee ID identifies with party the committee donates to. Mapping the committee IDs to Republican or Democratic. This is what we're using to filter all the donations into either red or blue.
- Voter
  - Keying donations from County, State and Election Year
- ETL
  - FEC dataframe was filtered using Colab to write and perform the scripts
    - Removed approximately 10 columns that were not relevant
    - Kept items that were needed for the analysis

Decision-making Process

After discussing this project with our instructor and teacher's assistants, we determined that we had to do more research to make sure we had enough data elements, feature and information to use a predictive model for our project. Our advisors thought that we should use some classification and regression models to see what more we can learn these data. This led us to work on supervised and unsupervised machine learning models.

Following this process, we realize that we still have some information and interest in using a predictive model. We ran a neural network on the data and the results were .98 in the deep learning model. This led us to believe that the model is likely overfitting.[5] We will look at other portions of the data to determine the next steps.

Peer Review

In order to verify the steps and assist one another in our code, we reviewed the various notebooks and provided updates and enhancement to the info. All our updates are tracked in GitHub.

Final Dashboard – Tool Description

- Tableau with aggregate CSVs and donation
- Google Slides for main presentation with link to the Tableau substory
- Flask app, interactive dashboard to display results of the model and run statistics
    - Angular
    - HTML/CSS
    - MySQL
    - Python
    - Flask
    - JavaScript
    - JQuery

Dashboard Interactive Elements – Description

- Tableau – for story and data
- Interactive – displaying data based on the user input

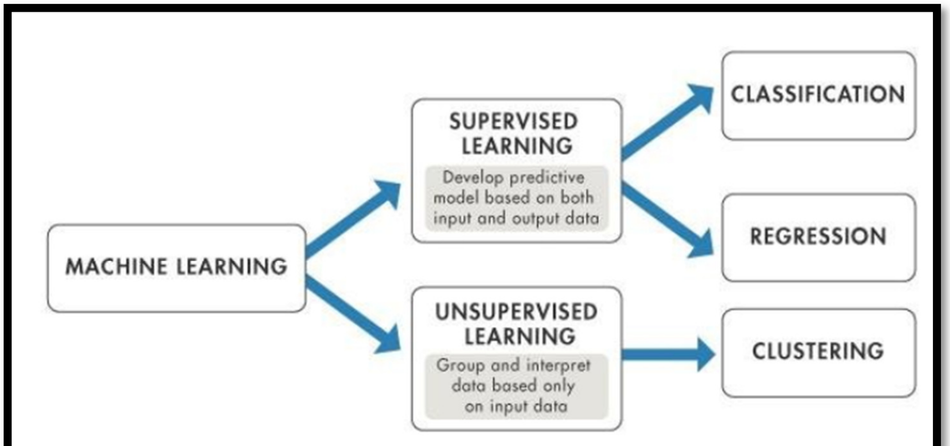
Data Sources

- Donation
    - https://www.fec.gov/campaign-finance-data/party-code-descriptions/
    - https://www.fec.gov/data/browse-data/?tab=bulk-data
- Voter
    - https://www.theguardian.com/news/datablog/2012/nov/07/us-2012-election-county-results-download#data
- Health Metrics
    - http://ghdx.healthdata.org/gbd-results-tool
- Education/Unemployment
    - https://www.census.gov/programs-surveys/popest/guidance.html
- Electoral College Votes

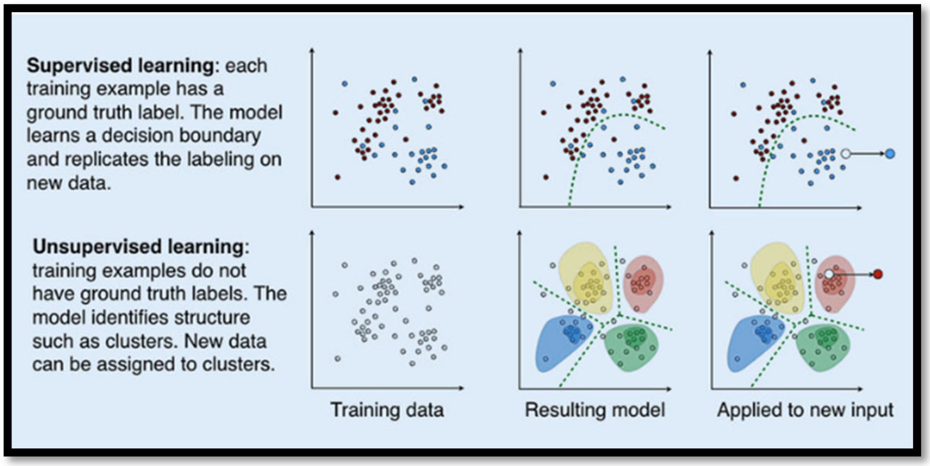
[5] Recognising Traffic Signs With 98% Accuracy Using Deep Learning https://towardsdatascience.com/recognizing-traffic-signs-with-over-98-accuracy-using-deep-learning-86737aedc2ab

- https://data.world/government/us-election-results
- https://worldpopulationreview.com/state-rankings/electoral-votes-by-state

Machine Learning Diagram


MACHINE LEARNING
SUPERVISED LEARNING
Develop predictive model based on both input and output data
UNSUPERVISED LEARNING
Group and interpret data based only on input data
CLASSIFICATION
REGRESSION
CLUSTERING

Supervised Learning versus Unsupervised Learning (Mathworks, n.d.)



Supervised learning: each training example has a ground truth label. The model learns a decision boundary and replicates the labeling on new data.

Unsupervised learning: training examples do not have ground truth labels. The model identifies structure such as clusters. New data can be assigned to clusters.

Presentation Slide Deck

https://docs.google.com/presentation/d/1ijhyfkdBBYox_7o6rQUraLtBufkcBuDwIpVaxm5wSqs/edit?usp=sharing