Analyzing CDP Questionnaires With Natural Language Processing

By Matt Gargiulo

Problem Statement

CDP

DISCLOSURE INSIGHT ACTION

CDP provides the premier framework for both cities and corporations to disclose environmental and social data, metrics, targets, and initiatives.

Question to Solve:

Can the answers provided by each respondee provide context to where they are located?

Techniques Utilized:

I will incorporate data science techniques such as:

1. Natural language processing through spaCy
2. Dimensional reduction through primary component analysis and isometric mapping
3. Clustering through KMeans

Our Data

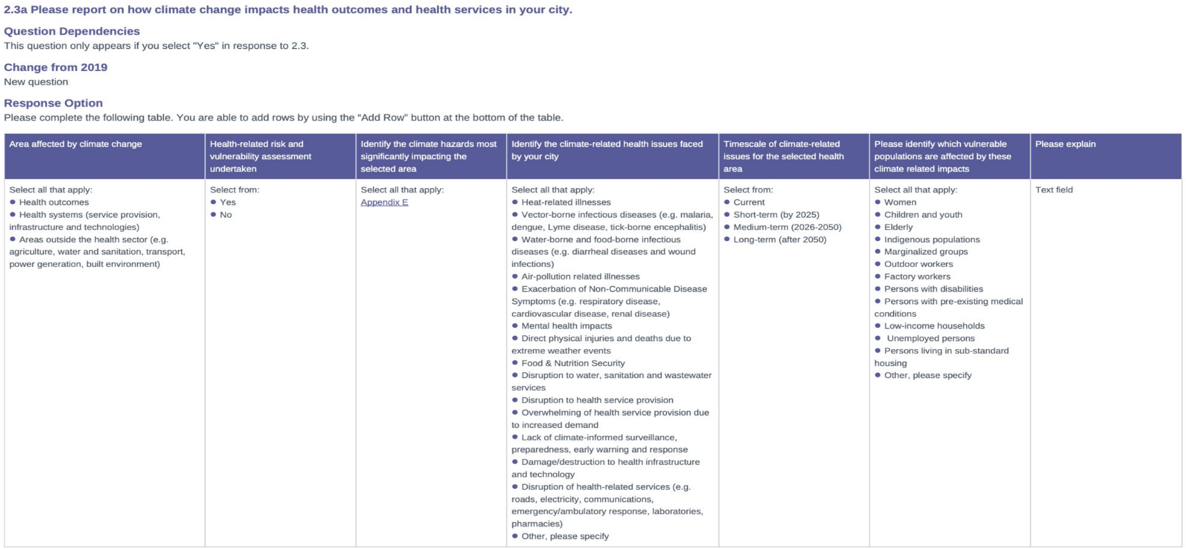
CDP provided data from both corporations and cities from 2018, 2019, 2020.

Each data comes with a ‘disclosures’ file that contains information about the city or company, and a ‘responses’ file that contains responses to the questionnaire.

The data contains over 6,700 respondents that span nearly 100 countries and numerous languages.

Insight Into The Questionnaire

Questionnaire questions may contain sub-questions in the form of a table. In the screenshot below, question 2.3a is composed of seven sub-questions.



2.3a Please report on how climate change impacts health outcomes and health services in your city.

Question Dependencies

This question only appears if you select "Yes" in response to 2.3.

Change from 2019

New question

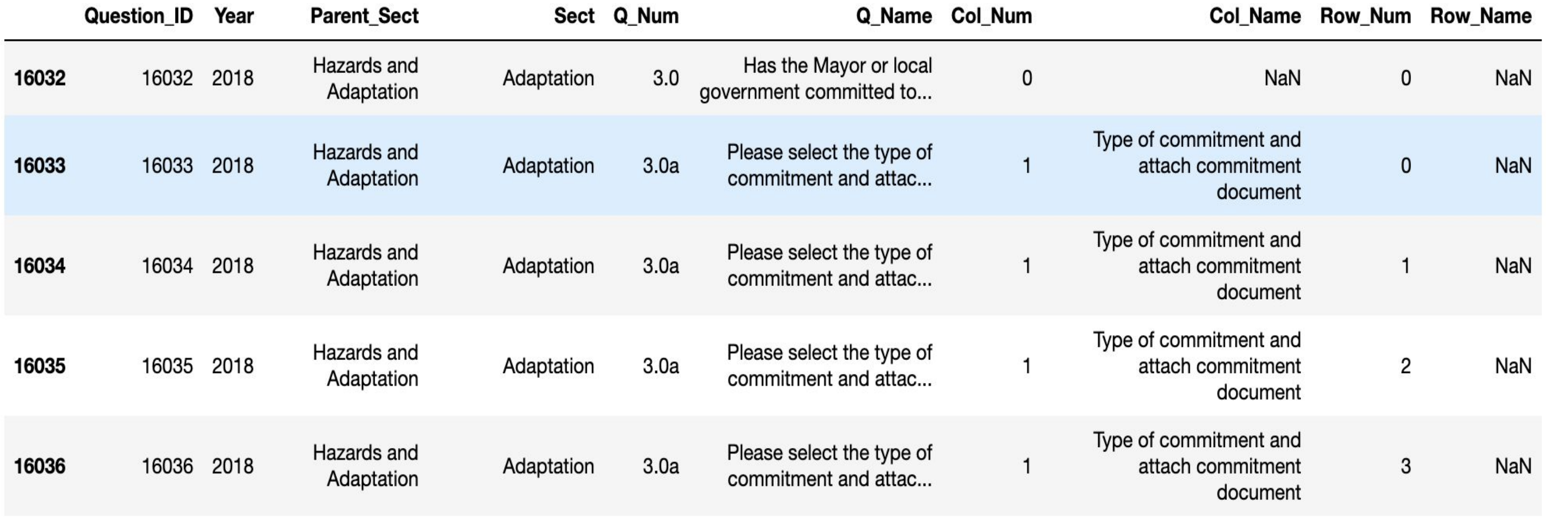
Response Option

Please complete the following table. You are able to add rows by using the "Add Row" button at the bottom of the table.

Area affected by climate change	Health-related risk and vulnerability assessment undertaken	Identify the climate hazards most significantly impacting the selected area	Identify the climate-related health issues faced by your city	Timescale of climate-related issues for the selected health area	Please identify which vulnerable populations are affected by these climate related impacts	Please explain
Select all that apply: ● Health outcomes ● Health systems (service provision, infrastructure and technologies) ● Areas outside the health sector (e.g. agriculture, water and sanitation, transport, power generation, built environment)	Select from: ● Yes ● No	Select all that apply: Appendix E	Select all that apply: ● Heat-related illnesses ● Vector-borne infectious diseases (e.g. malaria, dengue, Lyme disease, tick-borne encephalitis) ● Water-borne and food-borne infectious diseases (e.g. diarrheal diseases and wound infections) ● Air-pollution related illnesses ● Exacerbation of Non-Communicable Disease Symptoms (e.g. respiratory disease, cardiovascular disease, renal disease) ● Mental health impacts ● Direct physical injuries and deaths due to extreme weather events ● Food & Nutrition Security ● Disruption to water, sanitation and wastewater services ● Disruption to health service provision ● Overwhelming of health service provision due to increased demand ● Lack of climate-informed surveillance, preparedness, early warning and response ● Damage/destruction to health infrastructure and technology ● Disruption of health-related services (e.g. roads, electricity, communications, emergency/ambulatory response, laboratories, pharmacies) ● Other, please specify	Select from: ● Current ● Short-term (by 2025) ● Medium-term (2026-2050) ● Long-term (after 2050)	Select all that apply: ● Women ● Children and youth ● Elderly ● Indigenous populations ● Marginalized groups ● Outdoor workers ● Factory workers ● Persons with disabilities ● Persons with pre-existing medical conditions ● Low-income households ● Unemployed persons ● Persons living in sub-standard housing ● Other, please specify	Text field

Translation From Questionnaire To Dataframe

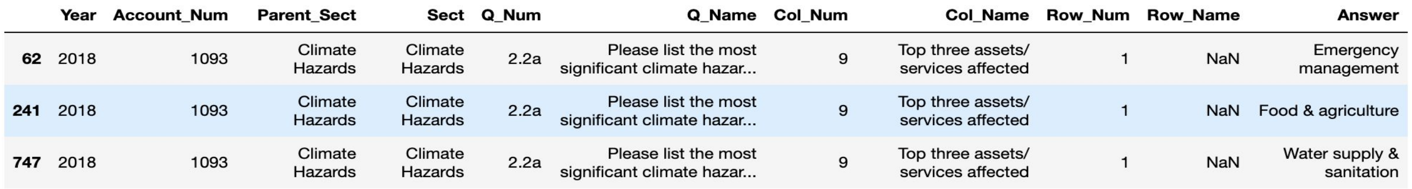
The previous table is arranged in the dataframe as:



	Question_ID	Year	Parent_Sect	Sect	Q_Num	Q_Name	Col_Num	Col_Name	Row_Num	Row_Name
16032	16032	2018	Hazards and Adaptation	Adaptation	3.0	Has the Mayor or local government committed to...	0	NaN	0	NaN
16033	16033	2018	Hazards and Adaptation	Adaptation	3.0a	Please select the type of commitment and attac...	1	Type of commitment and attach commitment document	0	NaN
16034	16034	2018	Hazards and Adaptation	Adaptation	3.0a	Please select the type of commitment and attac...	1	Type of commitment and attach commitment document	1	NaN
16035	16035	2018	Hazards and Adaptation	Adaptation	3.0a	Please select the type of commitment and attac...	1	Type of commitment and attach commitment document	2	NaN
16036	16036	2018	Hazards and Adaptation	Adaptation	3.0a	Please select the type of commitment and attac...	1	Type of commitment and attach commitment document	3	NaN

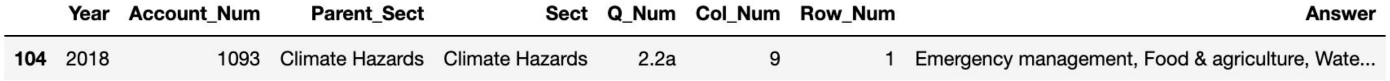
Data Wrangling

‘Climate Hazards’ sector, question 2.2a column number 9, row number 1, has three different, unique answers. In order to combat duplicate issues further in the analysis. I combined these instances into a single row.



	Year	Account_Num	Parent_Sect	Sect	Q_Num	Q_Name	Col_Num	Col_Name	Row_Num	Row_Name	Answer
62	2018	1093	Climate Hazards	Climate Hazards	2.2a	Please list the most significant climate hazar...	9	Top three assets/ services affected	1	NaN	Emergency management
241	2018	1093	Climate Hazards	Climate Hazards	2.2a	Please list the most significant climate hazar...	9	Top three assets/ services affected	1	NaN	Food & agriculture
747	2018	1093	Climate Hazards	Climate Hazards	2.2a	Please list the most significant climate hazar...	9	Top three assets/ services affected	1	NaN	Water supply & sanitation

Turns into:



	Year	Account_Num	Parent_Sect	Sect	Q_Num	Col_Num	Row_Num	Answer
104	2018	1093	Climate Hazards	Climate Hazards	2.2a	9	1	Emergency management, Food & agriculture, Wate...

Creating My Minimal Viable Product Dataset

I’ll subset the data to 5000 samples.

Containing only 2020 data from cities within the United States.

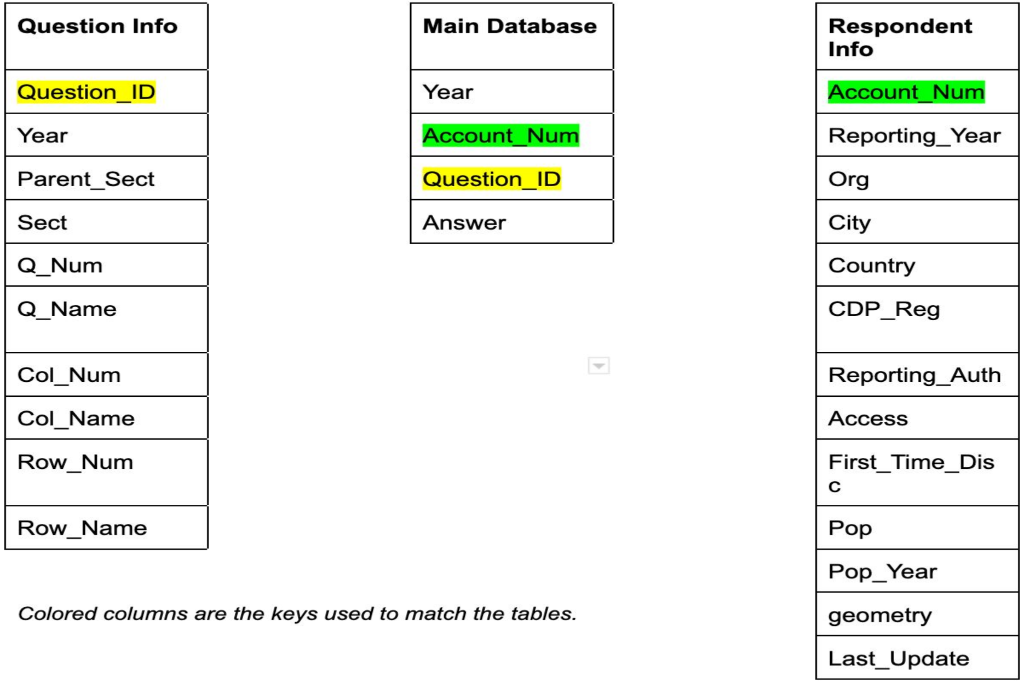
This way I can increase performance for later analysis and language translations won’t inhibit progress.

Data Organization

I spit the two data tables into three unique tables.

This organization is known as Star Schema.

Colored rows are primary-foreign key pairs.



Question Info
Question_ID
Year
Parent_Sect
Sect
Q_Num
Q_Name
Col_Num
Col_Name
Row_Num
Row_Name

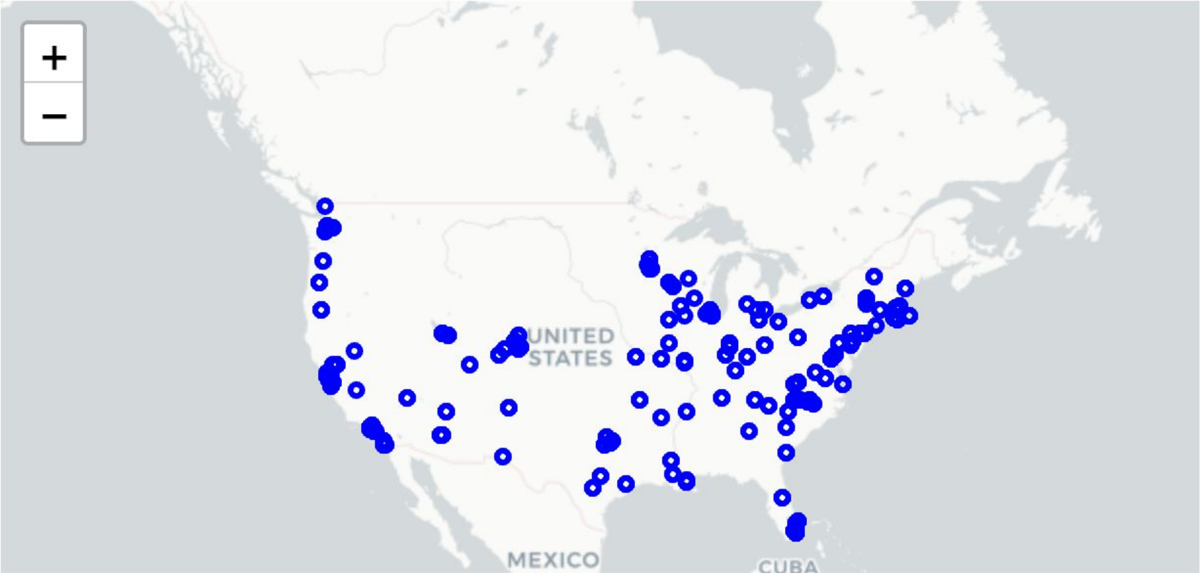
Main Database
Year
Account_Num
Question_ID
Answer

Respondent Info
Account_Num
Reporting_Year
Org
City
Country
CDP_Reg
Reporting_Auth
Access
First_Time_Dis c
Pop
Pop_Year
geometry
Last_Update

Colored columns are the keys used to match the tables.

Distribution of cities

The dataset is well distributed across the United States. There is, however, an observable lack of data points from the northern part of the central time zone.

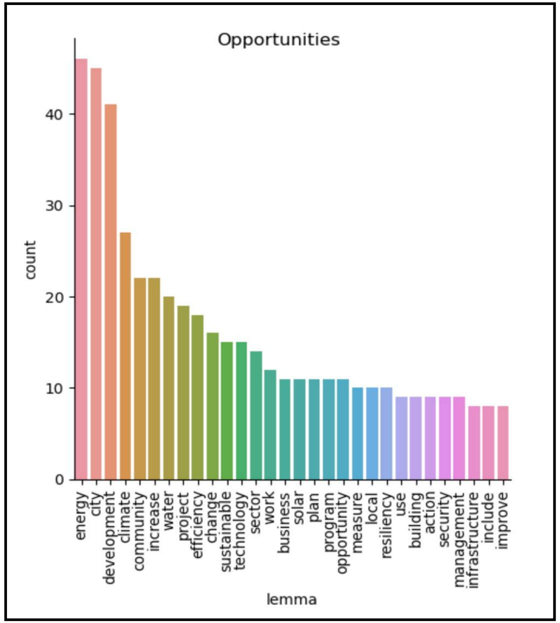


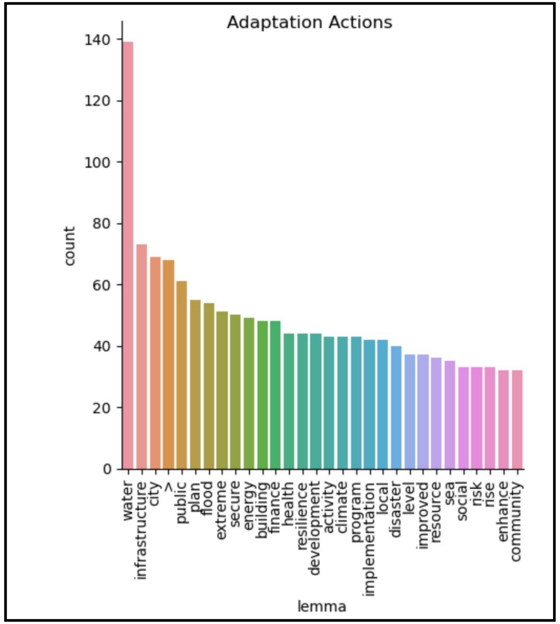
Distribution Of Words Per Sector

Each of the 25 sectors has a unique distribution of most common words. Here are two examples:

The ‘Opportunities’ sector had higher counts of words such as energy, city, development, and climate.

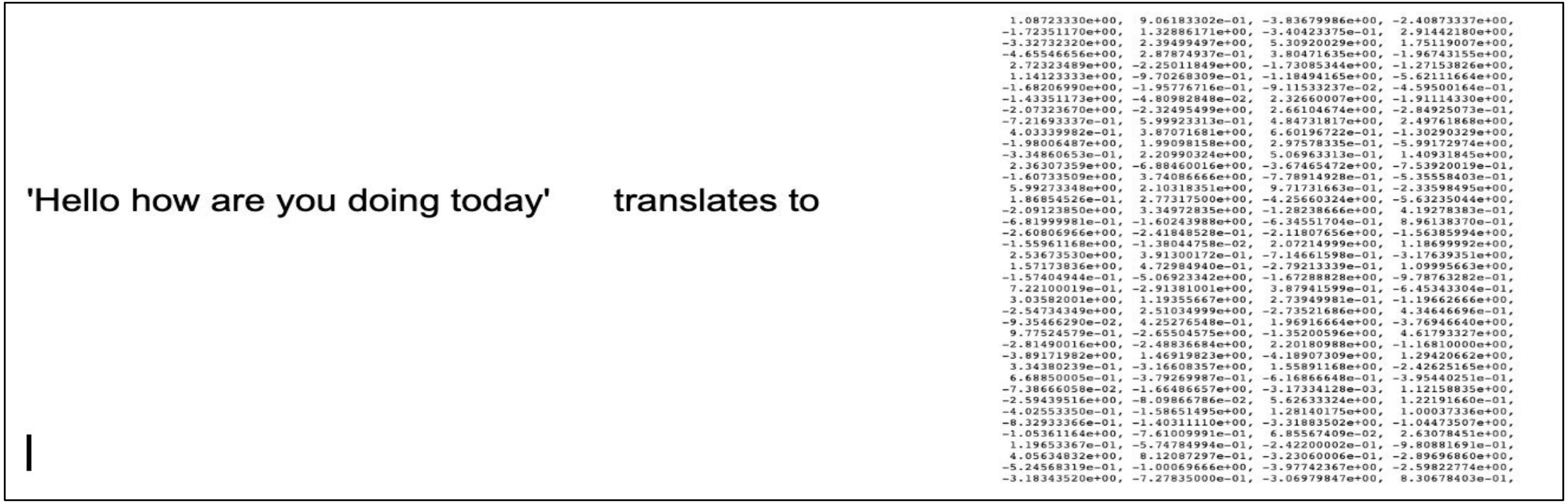
The ‘Adaptation Actions’ had almost 140 different counts of the word water, nearly double of the second most common word, infrastructure.





Word Embedding

While the above analysis is great for initial exploration, text data needs to be analyzed in relation to their larger context. SpaCy provides word embedding to help achieve this. By utilizing neural networks, spaCy is able to take a sentence and turn it into a vector 300 features long. As an example:

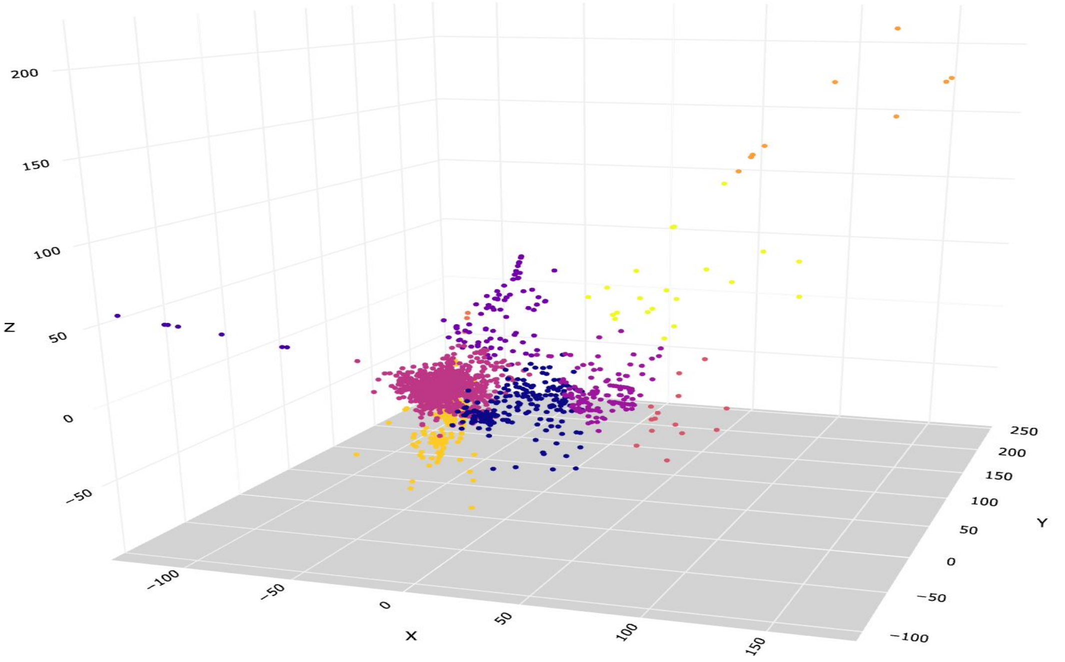


'Hello how are you doing today' translates to

```
 1.08723330e+00,  9.06183302e-01, -3.83679986e+00, -2.40873337e+00,
-1.72351170e+00,  1.32886171e+00, -3.40423375e-01,  2.91442180e+00,
-3.32732320e+00,  2.39499497e+00,  5.30920029e+00,  1.75119007e+00,
-4.65546656e+00,  2.87874937e-01,  3.80471635e+00, -1.96743155e+00,
 2.72323489e+00, -2.25011849e+00, -1.73085344e+00, -1.27153826e+00,
 1.14123333e+00, -9.70268309e-01, -1.18494165e+00, -5.62111664e+00,
-1.68206990e+00, -1.95776716e-01, -9.11533237e-02, -4.59500164e-01,
-1.43351173e+00, -4.80982848e-02,  2.32660007e+00, -1.91114330e+00,
-2.07323670e+00, -2.32495499e+00,  2.66104674e+00, -2.84925073e-01,
-7.21693337e-01,  5.99923313e-01,  4.84731817e+00,  2.49761868e+00,
 4.03339982e-01,  3.87071681e+00,  6.60196722e-01, -1.30290329e+00,
-1.98006487e+00,  1.99098158e+00,  2.97578335e-01, -5.99172974e+00,
-3.34860653e-01,  2.20990324e+00,  5.06963313e-01,  1.40931845e+00,
 2.36307359e+00, -6.88460016e+00, -3.67465472e+00, -7.53920019e-01,
-1.60733509e+00,  3.74086666e+00, -7.78914928e-01, -5.35558403e-01,
 5.99273348e+00,  2.10318351e+00,  9.71731663e-01, -2.33598495e+00,
 1.86854526e-01,  2.77317500e+00, -4.25660324e+00, -5.63235044e+00,
-2.09123850e+00,  3.34972835e+00, -1.28238666e+00,  4.19278383e-01,
-6.81999981e-01, -1.60243988e+00, -6.34551704e-01,  8.96138370e-01,
-2.60806966e+00, -2.41848528e-01, -2.11807656e+00, -1.56385994e+00,
-1.55961168e+00, -1.38044758e-02,  2.07214999e+00,  1.18699992e+00,
 2.53673530e+00,  3.91300172e-01, -7.14661598e-01, -3.17639351e+00,
 1.57173836e+00,  4.72984940e-01, -2.79213339e-01,  1.09995663e+00,
-1.57404944e-01, -5.06923342e+00, -1.67288828e+00, -9.78763282e-01,
 7.22100019e-01, -2.91381001e+00,  3.87941599e-01, -6.45343304e-01,
 3.03582001e+00,  1.19355667e+00,  2.73949981e-01, -1.19662666e+00,
-2.54734349e+00,  2.51034999e+00, -2.73521686e+00,  4.34646696e-01,
-9.35466290e-02,  4.25276548e-01,  1.96916664e+00, -3.76946640e+00,
 9.77524579e-01, -2.65504575e+00, -1.35200596e+00,  4.61793327e+00,
-2.81490016e+00, -2.48836684e+00,  2.20180988e+00, -1.16810000e+00,
-3.89171982e+00,  1.46919823e+00, -4.18907309e+00,  1.29420662e+00,
 3.34380239e-01, -3.16608357e+00,  1.55891168e+00, -2.42625165e+00,
 6.68850005e-01, -3.79269987e-01, -6.16866648e-01, -3.95440251e-01,
-7.38666058e-02, -1.66486657e+00, -3.17334128e-03,  1.12158835e+00,
-2.59439516e+00, -8.09866786e-02,  5.62633324e+00,  1.22191660e-01,
-4.02553350e-01, -1.58651495e+00,  1.28140175e+00,  1.00037336e+00,
-8.32933366e-01, -1.40311110e+00, -3.31883502e+00, -1.04473507e+00,
-1.05361164e+00, -7.61009991e-01,  6.85567409e-02,  2.63078451e+00,
 1.19653367e-01, -5.74784994e-01, -2.42200002e-01, -9.80881691e-01,
 4.05634832e+00,  8.12087297e-01, -3.23060006e-01, -2.89696860e+00,
-5.24568319e-01, -1.00069666e+00, -3.97742367e+00, -2.59822774e+00,
-3.18343520e+00, -7.27835000e-01, -3.06979847e+00,  8.30678403e-01,
```

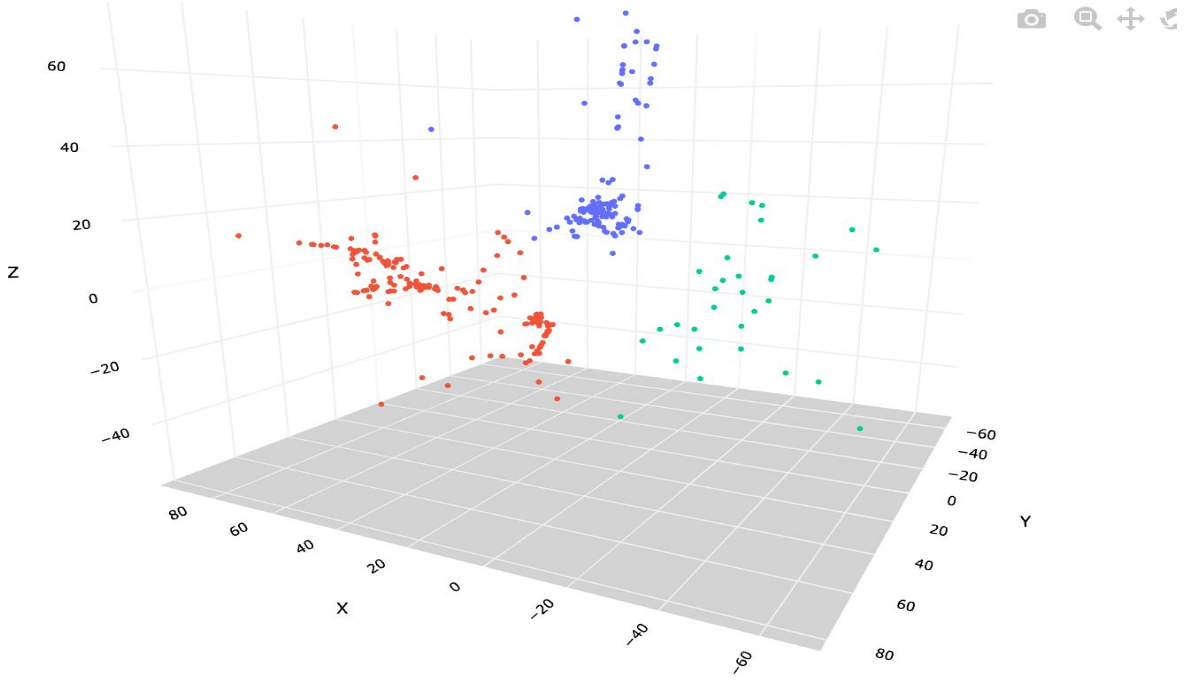
Mapping With Natural Language Processing

Once I embedded the answers, I visualized them with a 3 dimensional scatter plot.



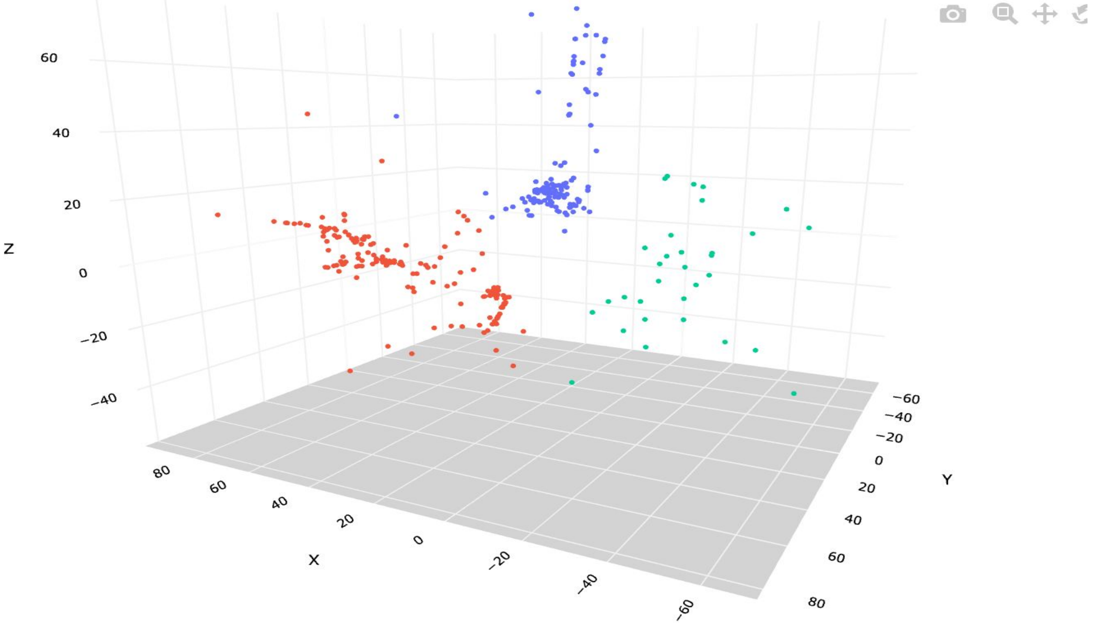
Natural Language Processing

And here is that same image filtered on a single sector - Climate Hazards.



How Are These Clusters Formed?

These 3D visuals utilize Kmeans clustering models. The data is split into three distinct clusters because I choose a hyperparameter value of 3. Tools such as elbow plots or silhouette analysis can help determine the ideal number of clusters.



What Is Causing This Clustering?

Was the clustering caused by cities in different parts of the country preferencing certain words over others?

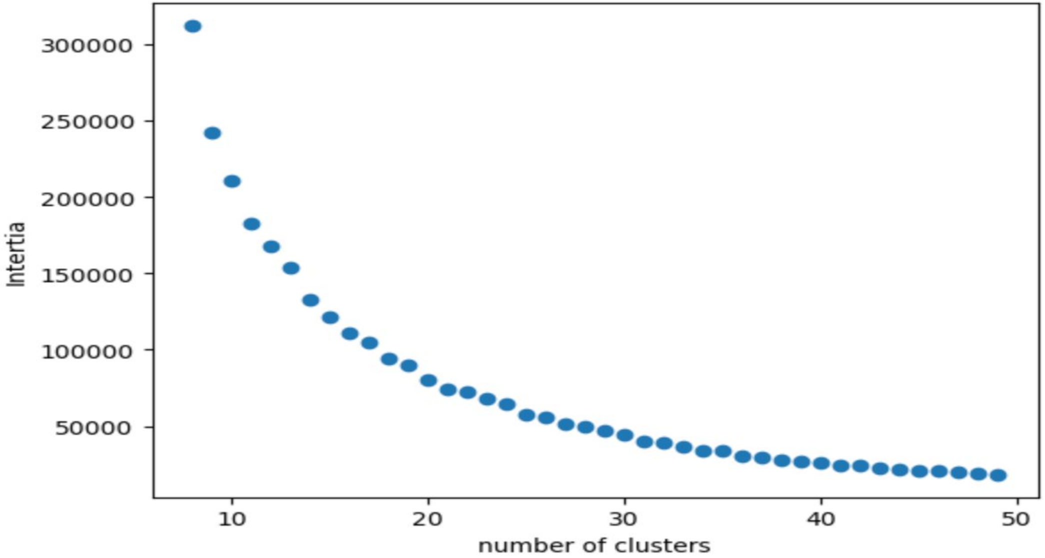
OR

Did the nature of each question within each sector lead to clustering?

Clustering By Sector And Question

These questions were answered by subsetting the data by unique sectors and questions.

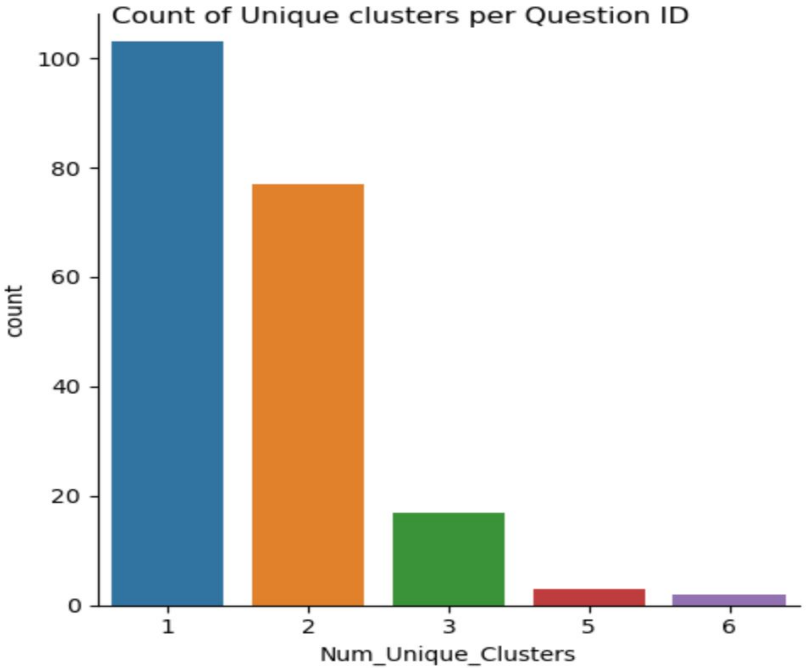
As seen with the elbow plot below, I chose 20 clusters as my hyperparameter value because that is the point the plot becomes linear.



How Unique Is Each Question?

‘Climate Hazards’ contains 202 unique questions squeezed into 20 clusters.

However, As seen by the figure, roughly 180 of the question only had 1 or 2 clusters signifying most of the answers were highly similar to one another.



Model Creation

A random forest and KNearestNeighbors model were selected due to performance and were evaluated on f1 scores.

Both of these models returned low scores.

Random Forest F1 Score:

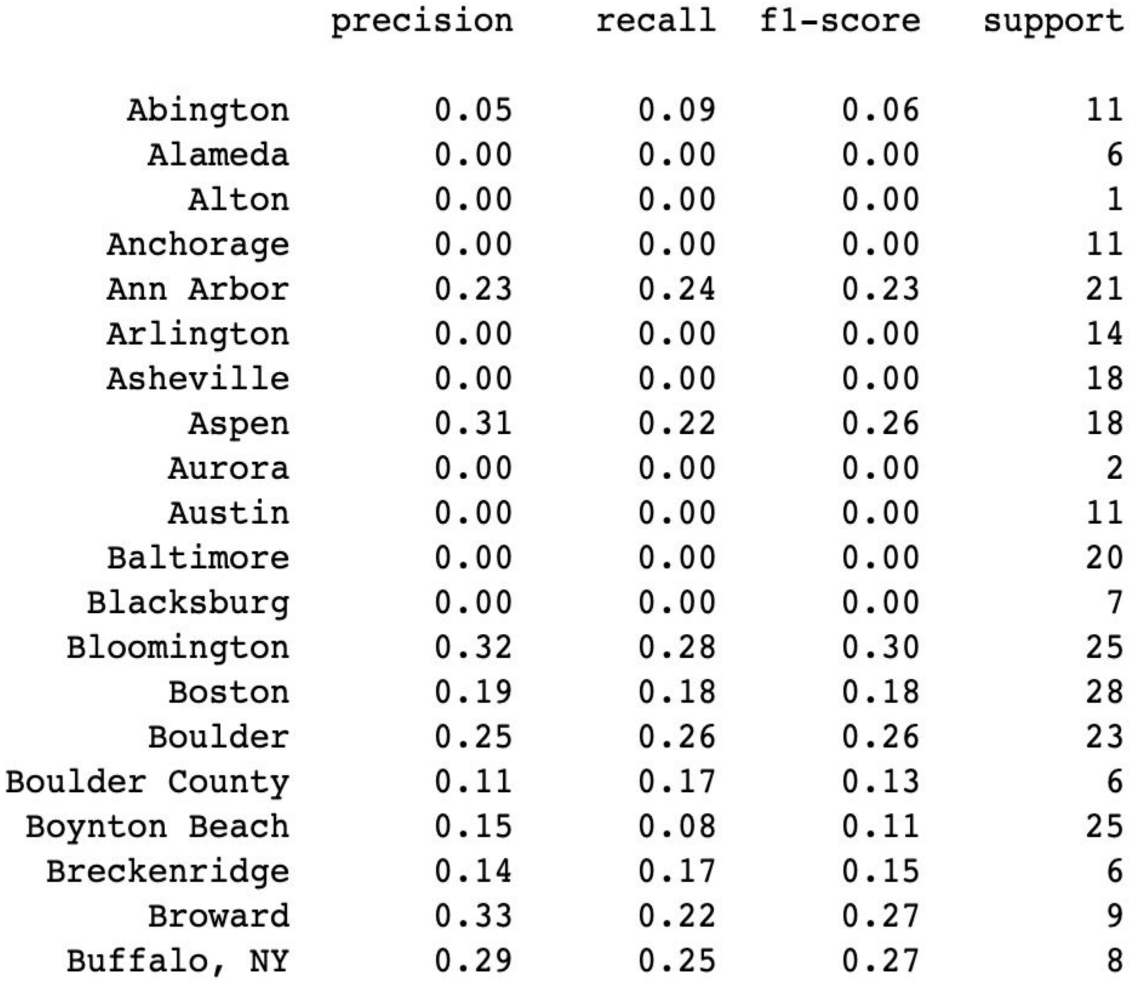
11.5%

KNearestNeighbors F1 Score:

3.5%

Evaluation Of The Model

The table is a sample of the evaluation report card that shows how well the model predicted each city.



```
                precision    recall  f1-score   support

      Abington       0.05      0.09      0.06        11
       Alameda       0.00      0.00      0.00         6
         Alton       0.00      0.00      0.00         1
     Anchorage       0.00      0.00      0.00        11
     Ann Arbor       0.23      0.24      0.23        21
     Arlington       0.00      0.00      0.00        14
     Asheville       0.00      0.00      0.00        18
         Aspen       0.31      0.22      0.26        18
        Aurora       0.00      0.00      0.00         2
        Austin       0.00      0.00      0.00        11
     Baltimore       0.00      0.00      0.00        20
    Blacksburg       0.00      0.00      0.00         7
   Bloomington       0.32      0.28      0.30        25
        Boston       0.19      0.18      0.18        28
       Boulder       0.25      0.26      0.26        23
Boulder County       0.11      0.17      0.13         6
 Boynton Beach       0.15      0.08      0.11        25
  Breckenridge       0.14      0.17      0.15         6
       Broward       0.33      0.22      0.27         9
   Buffalo, NY       0.29      0.25      0.27         8
```

Conclusion And Next Steps

As an MVP, this report does a great job at analyzing the data and creating a basic model to predict city location based on answers to a questionnaire. However, there is a lot of room for improvement.

Next Steps:
Next steps would be to group the cities into larger sections whether that is by state or geo-regions. Unfortunately the dataset did not have states associated with the data and while we may assume, for example, that the City of Austin is referring to Austin, Texas, there is an Austin in seven different states that are represented in the dataset. I'd also like to apply this model to every city in the dataset which would require a translation library such as google translate; and lastly, I'd like to implement the other half of the dataset which are corporate answers to a separate questionnaire.