



# Recommender System & Chatbot AI Studio Final Presentation

Social Science Research Council 1B  
December 4th, 2023



# Introductions



# Meet Our Team!



**Loyd Flores**  
CUNY Queens College



**Labiba Aziz**  
CUNY Queens College



**Fatima Ali**  
SUNY Farmingdale



**Xinyu Huang**  
CUNY Baruch College



**Kirsty Ihenetu**  
Barnard College



# Our AI Studio TA and Challenge Advisors



**Ryan Hardesty Lewis**  
AI Studio TA



**Rebecca Gluskin**  
Challenge Advisor



# Presentation Agenda

1. Overall Project Challenge
  - I. Pannel Challenge and Potential Impact
2. AI Studio Project Overview
  - I. Data Collection
  - II. Data Cleaning
  - III. Feature Engineering
  - IV. Clustering
  - V. Recommender System
  - VI. Chatbot
  - VII. Implementation
  - VIII. Demo
3. Final thoughts and Closing



Pannel Challenge and Potential Impact :

A more efficient way of accessing the Census API, by making DATA2GO.NYC even easier to use. How?

1. A data **recommendation system** for variables, correlations, and infographics based on the key search terms they are already looking for.
2. An interactive **Chatbot/Assistant** where you can have deeper insights



# Our Goal and Potential Impact

- To enhance the accessibility and user experience of DATA2GO.NYC for those with limited technical proficiency, we propose a multifaceted approach. This includes:
  1. **User-Friendly Interface Redesign:** Simplify the interface to make it easy for users of all technical levels to access and understand the data.
  2. **Advanced Analytical Tools:** Add tools for deeper analysis, helping users easily find and understand complex data correlations.
  3. **Comprehensive Data Coverage:** Ensure that all important variables are included, providing a more complete and accurate data set.
  4. **Enhanced User Engagement:** Create interactive features to make the platform more engaging and easier to use, encouraging wider exploration of the data.



# AI Studio Project Overview





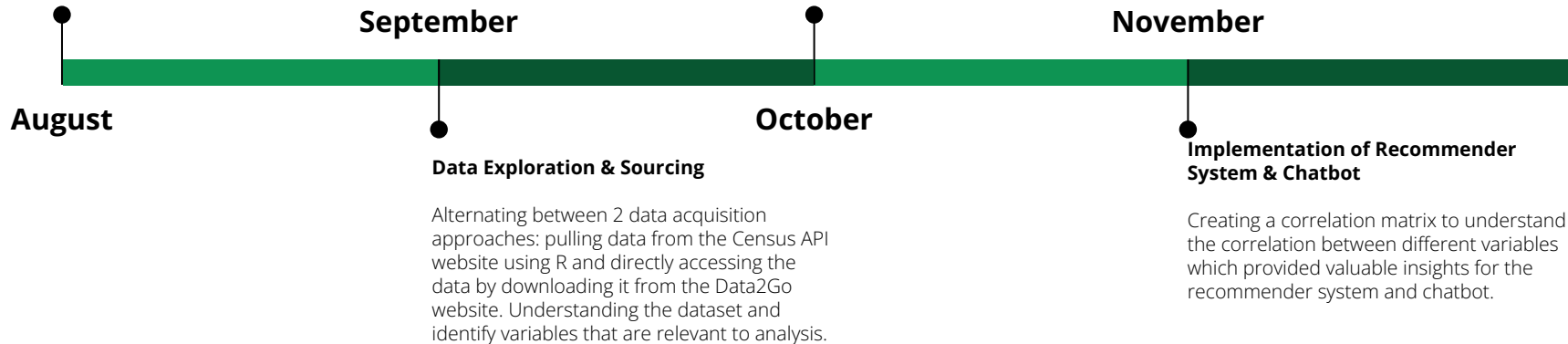
# Our Approach

## Business Understanding / Website Exploration

Going through the Data2Go website and gathering insights on our future ML models. Each member also created data stories based on personal experience and highlighting variables that captures our interests.

## Data Cleaning & Modeling

Removing unnecessary columns/rows and employed feature engineering to enrich the dataset. Applied clustering on the 'Sector' variable to understand patterns and relationships within the dataset.





# Resources We Leveraged

- Google Collab
- VS Code
- GitHub
- RStudio
- Kaggle
- ChatGPT





# I. Data Collection



# Data Collection

Data Pulled from the **Census** using slightly modified code provided to us


- Had to pull for multiple years

City	age_pyramid_total_nyc	median_household_income_nyc	median_personal_earnings_nyc	no_hs_nyc	at_least_hs_nyc
New York 0 city, New York	8550405	55752	36087	19.097573	80.902427



# Data Collection

Dataset provided within the **DATA2GO.NYC** website



MEASURE OF AMERICA

of the Social Science Research Council

DATA2GO.NYC

2ND EDITION

www.measureofamerica.org

• contact@measureofamerica.org

	SECTOR	DEMOGRAPHICS	DEMOGRAPHICS	DEMOGRAPHICS	DEMOGRAPHICS
	UNIT OF GEOGRAPHY	Community District	Community District	Community District	Community District
	DISPLAY NAME	Adult Stability (% of adu	Black (% of total popul	Black Population (#)	Females 15-24 (#)
	VARIABLE	adult_samehome_cd	aframer_pop_cd	aframer_total_pop_cd	age_pyramid_female_15_t
	SOURCE NAME	Measure of America calcula	Measure of America calcula	Measure of America calcula	Measure of America calcula
	SOURCE URL	https://ssrc.formstack.com/	https://ssrc.formstack.com/	https://ssrc.formstack.com/	https://ssrc.formstack.com/
	DESCRIPTOR	X.#%	X	X	X
	YEAR	2010-2014	2010-2014	2010-2014	2010-2014
	NOTES	Percentage of adults age 18	Black or African American n	Black or African American non-Latino	population
GEO_ID	GEO_LABE	GEO_DISPLAY_NAME			
201	Bronx CD (Mott Haven and Melrose		90.40969371	27.86549304	26565.83859
202	Bronx CD (Hunts Point and Longwood		86.72006375	23.0081972	12558.01405
203	Bronx CD (Morrisania and Crotona		89.76304802	38.13314971	31455.71424
204	Bronx CD (Highbridge and Concourse		89.89152782	30.91387421	45673.54209
205	Bronx CD (Fordham and University Heights		87.034046	26.55404482	32811.18434
206	Bronx CD (Belmont and East Tremont		84.68095697	24.45313865	20945.89002
207	Bronx CD (Kingsbridge Heights and Bedford		86.94871601	17.67692999	24113.55598
208	Bronx CD (Riverdale and Fieldston		88.15644389	12.39578932	12808.77953
209	Bronx CD (Parkchester and Soundview		90.73871125	31.26239495	55426.13598
210	Bronx CD (Throgs Neck and Co-op City		90.16645595	21.76323093	27851.1699
211	Bronx CD (Morris Park and Bronxdale		90.22365152	20.94490669	25288.13527



## II. Data Cleaning



# Data Cleaning - Census Data

Combining multiple years because each year was a separate dataset

	City	age_pyramid_total_nyc	median_household_income_nyc	median_personal_earnings_nyc	no_hs_nyc	at_least_hs_nyc
0	New York city, New York	8550405	55752	36087	19.097573	80.902427
0	New York city, New York	8537673	58856	36871	18.456166	81.543834
0	New York city, New York	8622698	60879	38430	18.114607	81.885393
0	New York city, New York	8398748	63799	40932	17.287630	82.712370
0	New York city, New York	8336817	69407	42326	16.779843	83.220157

5 rows × 199 columns



# Data Cleaning - DATA2GO.NYC data

Extracting necessary rows and columns and pivoting data

- Data originally came as a row, we converted each column into a row entry
- Could only use 5th and 4th version because the older versions had different formatting (Variable names, # of total indicators varied)

Variable/Indicator	Data	Sector
adult_samehome_nyc	89.4606976027601	DEMOGRAPHICS
asian_api_pop_nyc	14.1090314889791	DEMOGRAPHICS
asian_api_total_nyc	1184982	DEMOGRAPHICS
asian_api_pop_nyc_historical	9.83	DEMOGRAPHICS
asian_api_pop_change_nyc	43.5303305084344	DEMOGRAPHICS
aframer_pop_nyc	21.7136292218793	DEMOGRAPHICS
black_pop_nyc	21.7136292218793	DEMOGRAPHICS
aframer_total_pop_nyc	1823673	DEMOGRAPHICS
black_pop_nyc_historical	24.5015719983747	DEMOGRAPHICS
black_pop_change_nyc	-11.3786281822257	DEMOGRAPHICS
child_samehome_nyc	92.7889816079	DEMOGRAPHICS
disabled_nyc	10.4698333264522	DEMOGRAPHICS
divorced_nyc	7.7806099247245	DEMOGRAPHICS
lonelyaged_nyc	11.6407431505645	DEMOGRAPHICS
sixtyfive_nyc	14.8293531369199	DEMOGRAPHICS
familynochild_nyc	31.3456823308932	DEMOGRAPHICS
female_veterans_nyc	11209	DEMOGRAPHICS
female_veterans_percent_nyc	7.80364527492725	DEMOGRAPHICS
age_pyramid_female_under_5_nyc	261720	DEMOGRAPHICS





### III. Feature Engineering



# Feature Engineering - Census Data

- Removing the city column and replacing the index to the year of which the data was pulled

	age_pyramid_total_nyc	median_household_income_nyc	median_personal_earnings_nyc	no_hs_nyc	at_least_hs_nyc
year					
2015	0.672076	-1.452290	-1.325879	1.434938	-1.434938
2016	0.539949	-0.811021	-1.032473	0.801368	-0.801368
2017	1.422298	-0.393080	-0.449029	0.463983	-0.463983
2018	-0.901748	0.210175	0.487326	-0.352890	0.352890
2019	-1.544438	1.368757	1.009020	-0.854473	0.854473
2021	-0.188138	1.077459	1.311034	-1.492926	1.492926

6 rows × 197 columns



# Feature Engineering - DATA2GO.NYC

- With limited data, we One hot encoded the “sector” column to allow us to make correlations

- 1 : In this sector

- 0 : Not in this sector

Sector_EDUCATION	Sector_ENVIRONMENT	Sector_FOOD SYSTEMS	Sector_HEALTH	Sector_HOUSING & INFRASTRUCTURE	Sector_POLITICAL ENGAGEMENT	Sector_PUBLIC FUNDING & SERVICES	Sector_SAFETY & SECURITY	Sector_WORK, WEALTH & POVERTY
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0



## IV. Clustering

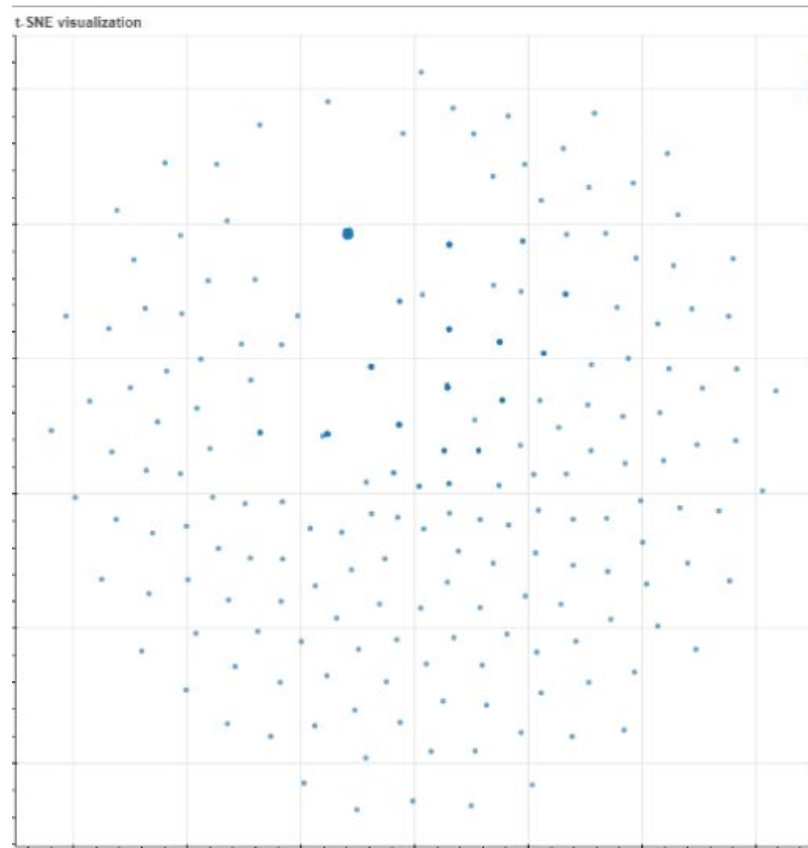


# Clustering with K-means

- Pre clustered data
  - How does our data look?
  - First assumptions

Pre-processing steps :

1. Standardization of values using,  
`from sklearn.preprocessing import StandardScaler`  
because some values were percentages (%), and others were solid or total count values.
2. Optimal Hyperparameter :  
# of clusters





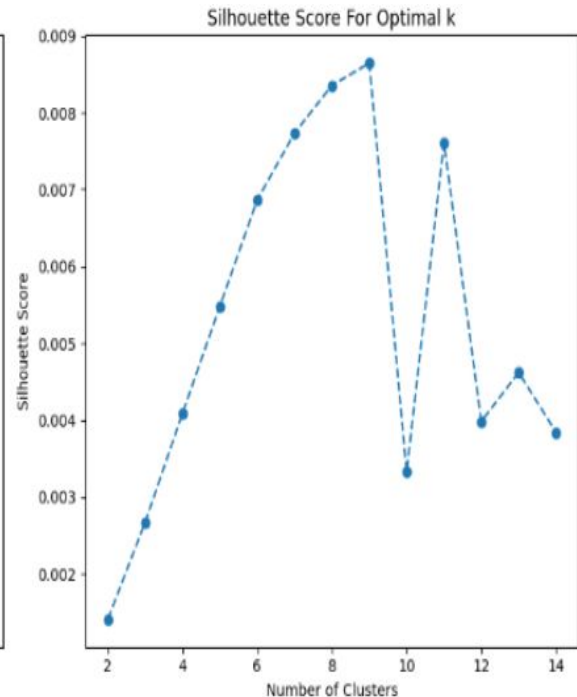
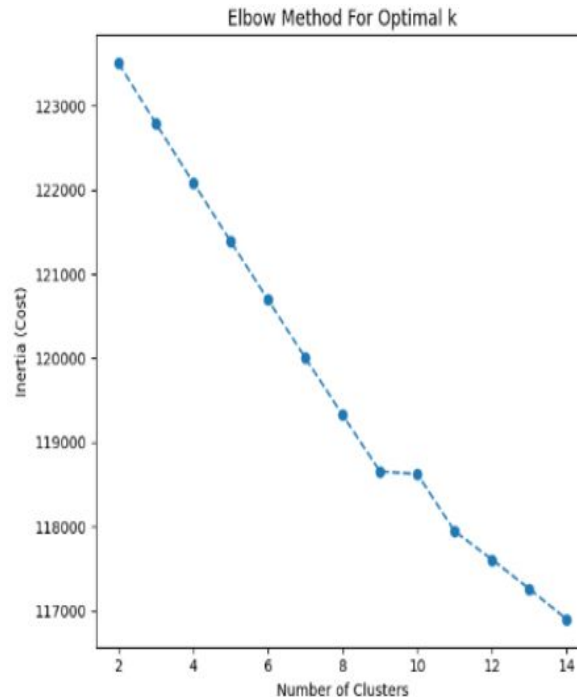
# Clustering with K-means

- K-means requires a hyperparameter : **# of clusters**
- We used 2 methods to verify the best number of clusters

How Best Hyperparameter is determined:

1. Elbow method - Value before plateau

2. Silhouette Score - Value before the steepest drop





# Optimal number of hyperparameters : 6

**We started out with 10 sectors (Education, Health, Environment,...), and ended up with 6? Why?**

1. If we use too few clusters it may oversimplify the data, if we use too much it may overfit and not accurately cluster. In summary they might not be accurate.
2. Clustering removes 1 sector, and implicitly if your data point isn't clustered to any explicit cluster implicitly it is part of the unstated or removed cluster.
3. Some sectors had limited entries, making that sector not as impactful as others,  
*Ex : Food Systems only had 4 entries. This sector could be merged.*

159	bodega_supermarket_nyc	13	FOOD SYSTEMS
160	food_insecure_nyc	12.9	FOOD SYSTEMS
161	meal_gap_nyc	241956200	FOOD SYSTEMS
162	households_receiving_snap_benefits_nyc	20.2806346750004	FOOD SYSTEMS



## V. Correlation Analysis & Recommender System



# Correlation Analysis using **Correlation Matrix**

`correlation_matrix = df.corr()`



## Problems Faced as we attempted to correlate:

*There were clusters being formed in the previous step, but no correlation could be extracted out of the DATA2GO.NYC data since it was only 2 years.*

## Solution :

We were stuck and opted to attempt to cluster using the cleaned data from the census, despite losing some variables it was a tradeoff for a more accurate model that could correlate our data

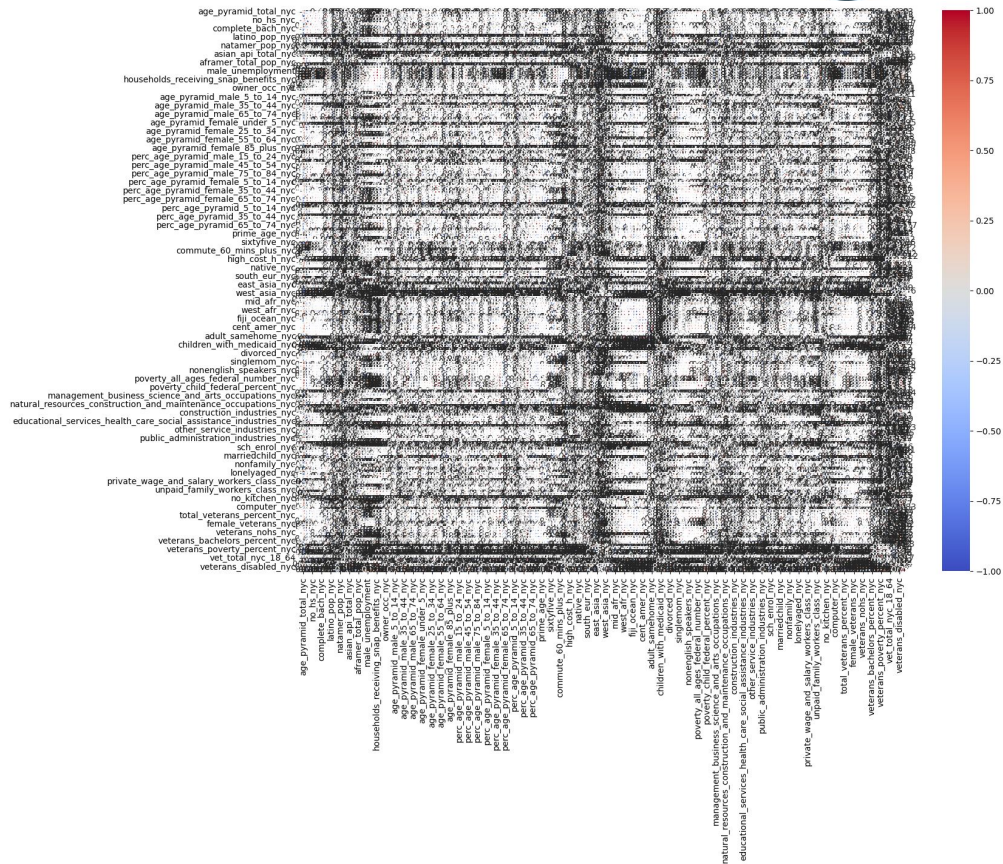


# Correlation Matrix (All variables)



It worked! Correlations were formed with the Census Data!

This graph is unreadable due to the 197 variables compared to each other.



# Correlation Matrix (Top 5 variables)



Zoom in to just 5 variables compared.

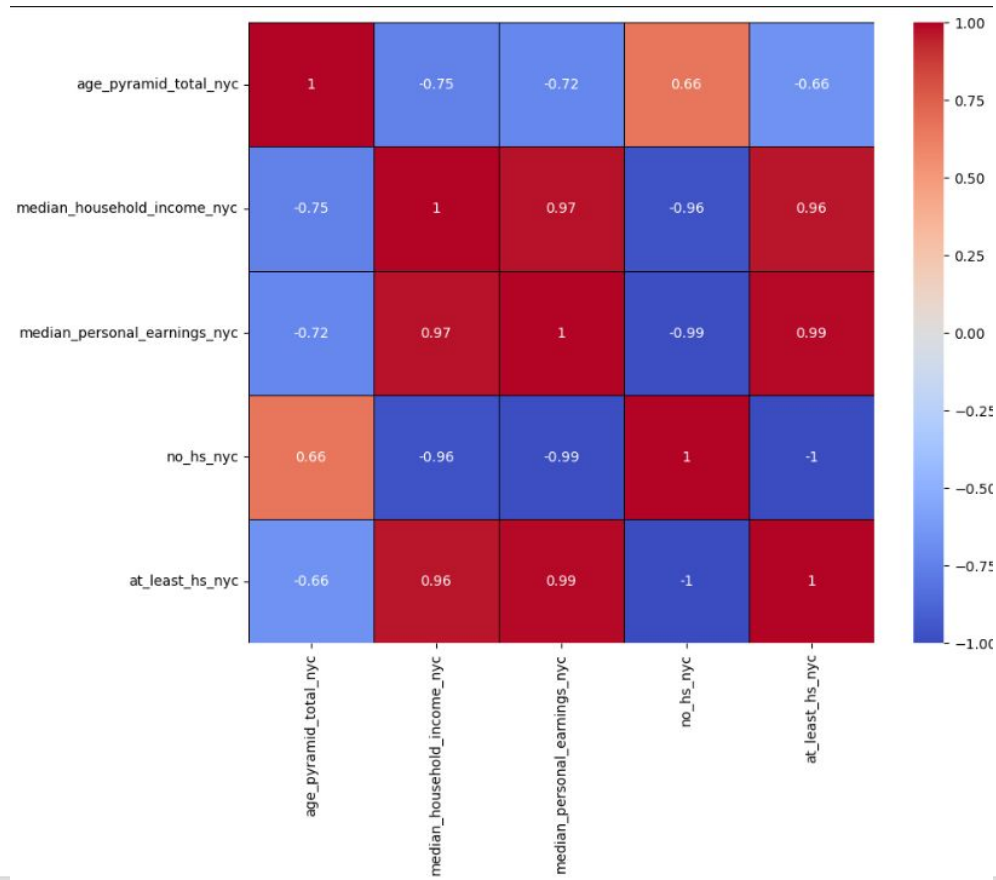
## 1 - Strong Positive Correlation :

The closer to 1 or the more red it is the stronger the correlation

## -1 - Strong Negative Correlation :

The closer to -1 or the more blue, the stronger the negative effects are.

We ignore values exactly 1 because they are correlated to themselves.





# Recommender System - Implementation

- Implemented using a function our team developed. The code was written in Python
- Goes through all the correlations of the selected variable and returns the ones with the highest correlation depending on how many the user wants

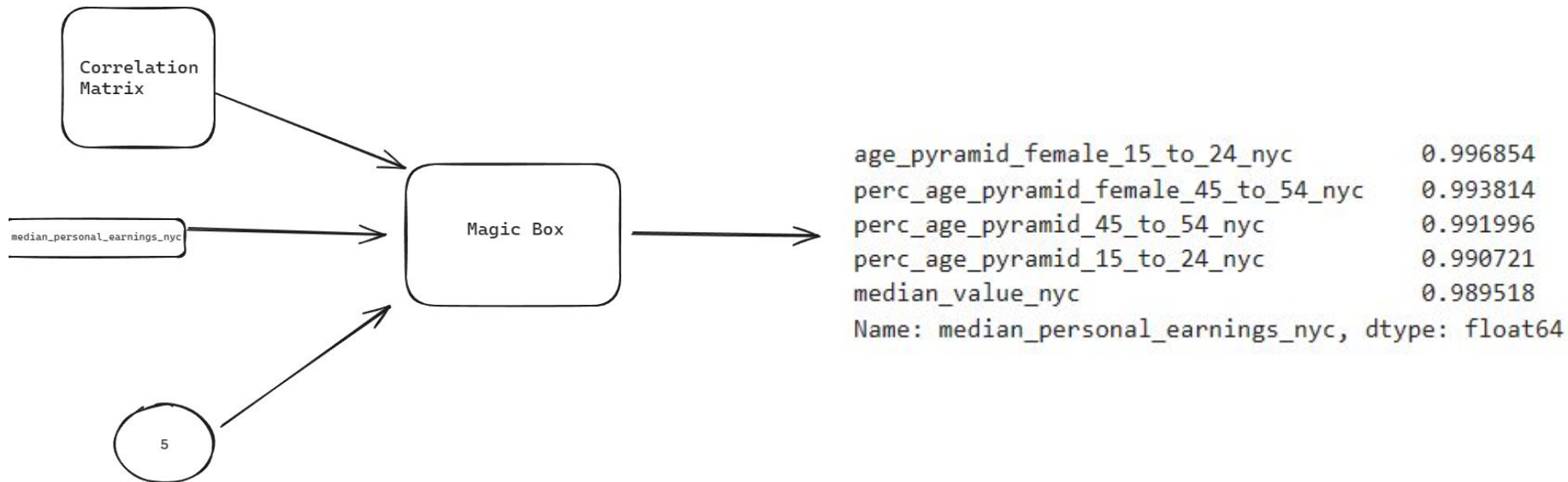


# Recommender System - How it works

```
print(get_top_correlated_variables(correlation_matrix, "median_personal_earnings_nyc", 5))
```

Function\_Call

(Correlation Matrix, Target Variable, # of entries to be returned)





## VI. Chatbot



# Chatbot Implementation - Overview

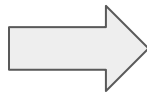
- Due to time constraint our team opted to utilize **OPEN AI's GPT-3.5-Turbo to implement a correlational analyst**. Another option was to implement our own RAG model (LLM like Chat GPT), but we were short on time.
- Open AI's models requires a few things. It needs our data retrieved to be converted into something it understands.
- From Correlation\_Matrix.csv -> knowledge.txt, using a python function implemented by the team.



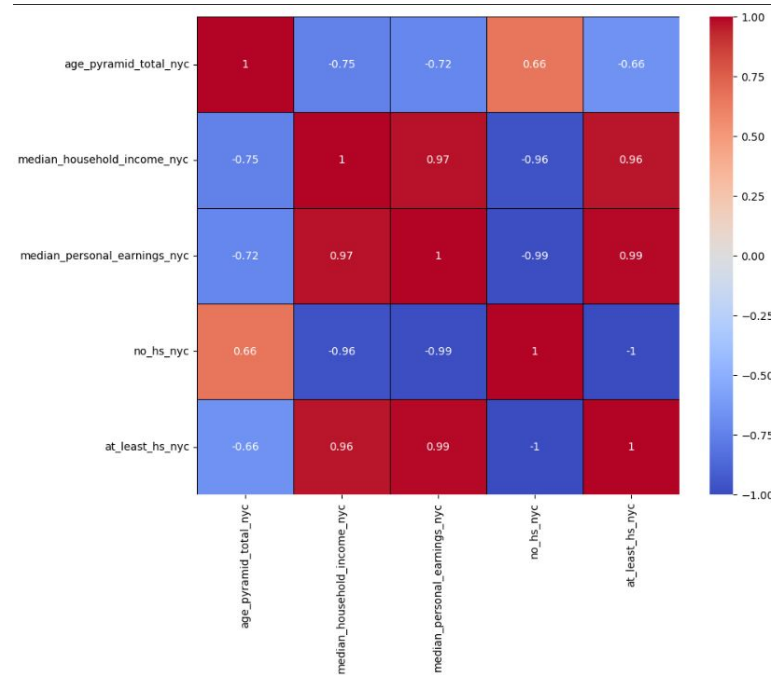
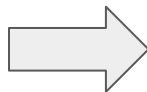


# Chatbot Implementation - Data Conversion

Correlation\_matrix.csv



knowledge.txt



```
1 Significant Correlations:
2 Between 'latino_total_pop_nyc' and 'age_pyramid_total_nyc': Positive Correlation (0.99)
3 Between 'age_pyramid_male_35_to_44_nyc' and 'age_pyramid_total_nyc': Positive Correlation (0.81)
4 Between 'age_pyramid_male_45_to_54_nyc' and 'age_pyramid_total_nyc': Positive Correlation (0.88)
5 Between 'age_pyramid_female_35_to_44_nyc' and 'age_pyramid_total_nyc': Positive Correlation (0.94)
6 Between 'age_pyramid_female_45_to_54_nyc' and 'age_pyramid_total_nyc': Positive Correlation (0.84)
7 Between 'nonfammore1_nyc' and 'age_pyramid_total_nyc': Negative Correlation (-0.82)
8 Between 'other_service_industries_nyc' and 'age_pyramid_total_nyc': Positive Correlation (0.81)
9 Between 'retail_trade_industries_nyc' and 'age_pyramid_total_nyc': Positive Correlation (0.82)
10 Between 'total_18_plus' and 'age_pyramid_total_nyc': Positive Correlation (0.98)
11 Between 'veterans_unemployed_nyc' and 'age_pyramid_total_nyc': Positive Correlation (0.85)
12 Between 'median_personal_earnings_nyc' and 'median_household_income_nyc': Positive Correlation (0.97)
13 Between 'no_hs_nyc' and 'median_household_income_nyc': Negative Correlation (-0.96)
14 Between 'at_least_hs_nyc' and 'median_household_income_nyc': Positive Correlation (0.96)
15 Between 'complete_bach_nyc' and 'median_household_income_nyc': Positive Correlation (0.81)
16 Between 'at_least_bachelors_nyc' and 'median_household_income_nyc': Positive Correlation (0.88)
17 Between 'grad_degree_nyc' and 'median_household_income_nyc': Positive Correlation (0.91)
18 Between 'white_total_pop_nyc' and 'median_household_income_nyc': Negative Correlation (-0.90)
19 Between 'aframer_total_pop_nyc' and 'median_household_income_nyc': Negative Correlation (-0.80)
20 Between 'median_rent_nyc' and 'median_household_income_nyc': Positive Correlation (0.89)
```



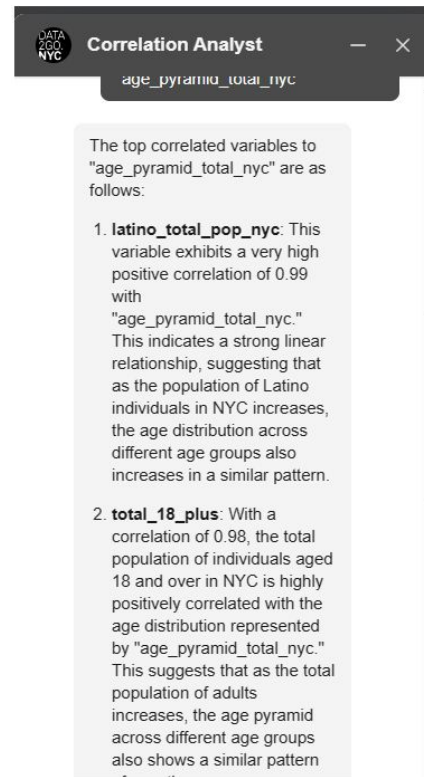
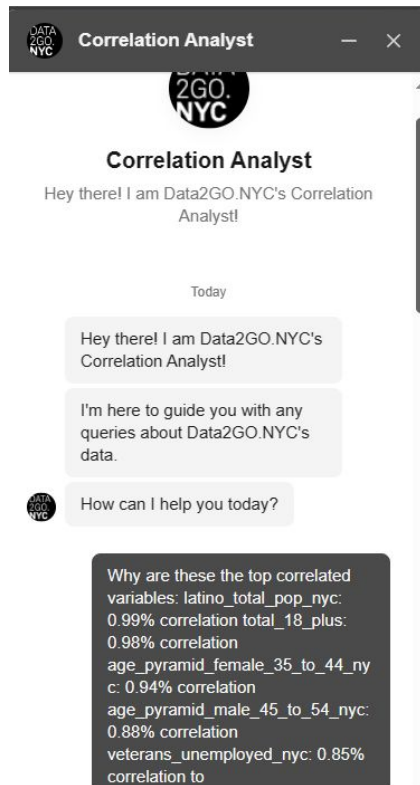
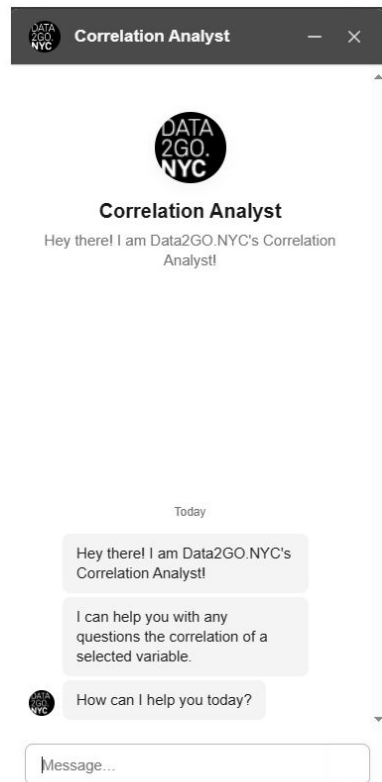


# Chatbot Implementation - Creation of our Model

- Written in python, code was written to host our model that was connected to Chat-GPT's api on the back end.
- We trained our model using the `knowledge.txt` file that was generated previously
- In Open AI's API it indicates that we have to give it specific instructions on how it will analyze our data. After tinkering we finally ended up with a working model that fits specifically for our use case.



# Chatbot Implementation - How it looks





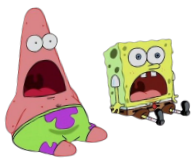
## VII. Implementation - Putting it all together



# How do we implement and tie everything together?

We wanted to tie everything together and for it to work as a unit.

- Webpage was written in REACT/Js
- Backend of the Chatbot was written in Python
- Recommender System was originally written in python but was converted to javascript
- Note, Backend of the webpage is completely separate from the backend of the chatbot
- This webpage could be implemented on DATA2GO.NYC when people are looking for variables, the input could be the user's click.



DATA  
2GO.  
NYC

Home

About

## Top Correlated Variables for `age_pyramid_total_nyc`

Sort By:

Absolute Value

Number of Results:

5

Select an Indicator:

`age_pyramid_total_nyc`

`latino_total_pop_nyc`: 0.99% correlation

`total_18_plus`: 0.98% correlation

`age_pyramid_female_35_to_44_nyc`: 0.94% correlation

`age_pyramid_male_45_to_54_nyc`: 0.88% correlation

`veterans_unemployed_nyc`: 0.85% correlation

BREAK  
THROUGH  
TECH



MEASURE OF AMERICA  
of the Social Science Research Council

HELMSELEY

DATA  
2GO.  
NYC

Correlation Analyst



### Correlation Analyst

Hey there! I am Data2GO.NYC's Correlation Analyst!

Today

Hey there! I am Data2GO.NYC's Correlation Analyst!

I'm here to guide you with any queries about Data2GO.NYC's data.



How can I help you today?

Why are these the results for  
`age_pyramid_total_nyc`  
`latino_total_pop_nyc`: 0.99  
`total_18_plus`: 0.98  
`age_pyramid_female_35_to_44_nyc`: 0.94  
`age_pyramid_male_45_to_54_nyc`: 0.88  
`veterans_unemployed_nyc`:

Message...



## VIII. Demo



Final Thoughts



# What We Learned

- How to work with RStudio
- How stressful it is to look for data and to clean up outdated datasets
- Machine learning cycle from data exploration to implementation
- It is not linear. Most of the time you think you have something figured out but you don't
- The importance of being able to learn on the fly





# Potential Next Steps

- Implementation on DATA2GO.NYC
  - Changing the recommender system to dynamically recommend when a user clicks on a variable
  - Chatbot will be available on the website



Questions?