DSC 383W Project:
United Way - RMAPI Residents Survey Analysis and Prediction During Covid-19 Pandemic

1. INTRODUCTION

Under the COVID-19 pandemic, the life of the residents in Rochester Monroe region has been impacted. The point in time survey was administered to get a community assessment on how the community is responding to the crisis and gain insight into the community's post lockdown needs & concerns.

On May 1, RMAPI launched a new survey to better understand the impact of COVID- 19 on community member's income and basic needs as well as what community members need to be safe and financially secure. This survey and responses are intended to inform how the Rochester and Monroe County community responds to COVID-19 to support households experiencing poverty. RMAPI intends to lift the community voice as a key metric for data-driven decision making.

With the insight from the analysis result, the United Way may determine which kind of assistance to be provided, and what features of living necessities are more important for community members.

The survey is administered by RMAPI, and most of the responses from the residents provided detailed living needs and requirements. Analyzing the survey responses would provide an overview and get a glance at the basic living conditions of the local community.

2. DATA DESCRIPTION

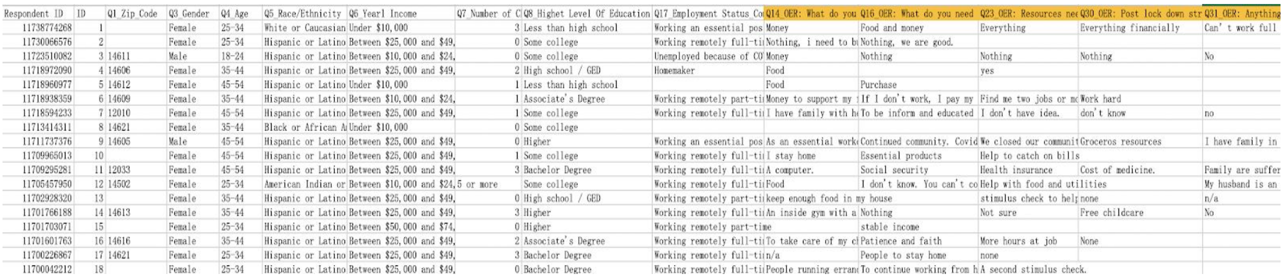
2.1 Data Collection

The dataset of Rochester Community Feedback on Covid-19 Crisis was sent out by Rochester Monroe Anti-Poverty Initiative, and it is available on RMAPI website.

2.2 Data Overview

The dataset contains 553 rows of data entries, with 15 attributes for all entries.

Attributes in this dataset contain basic information such as Respondent_ID, ID, Zip Code, Employment Status, etc. As well as selection questions such as Gender, Age, etc. Followed by Open-ended questions regarding their post-pandemic needs and living conditions. The questions include "What do you need to stay at home" and so forth questions.



Respondent ID	ID	Q1_Zip_Code	Q3_Gender	Q4_Age	Q5_Race/Ethnicity	Q6_Yearl Income	Q7_Number of C	Q8_Highet Level Of Education	Q17_Employment Status_Co	Q14_OER: What do you	Q16_OER: What do you need	Q23_OER: Resources nee	Q30_OER: Post lock down str	Q31_OER: Anything
11738774268	1		Female	25-34	White or Caucasian	Under $10,000	3	Less than high school	Working an essential pos	Money	Food and money	Everything	Everything financially	Can' t work full
11730066576	2		Female	25-34	Hispanic or Latino	Between $25,000 and $49,	0	Some college	Working remotely full-ti	Nothing, i need to b	Nothing, we are good.			
11723510082	3	14611	Male	18-24	Hispanic or Latino	Between $10,000 and $24,	0	Some college	Unemployed because of CO	Money	Nothing	Nothing	Nothing	No
11718972090	4	14606	Female	35-44	Hispanic or Latino	Between $25,000 and $49,	2	High school / GED	Homemaker	Food		yes		
11718960977	5	14612	Female	45-54	Hispanic or Latino	Under $10,000	1	Less than high school		Food	Purchase			
11718938359	6	14609	Female	35-44	Hispanic or Latino	Between $10,000 and $24,	1	Associate's Degree	Working remotely part-ti	Money to support my	If I don't work, I pay my	Find me two jobs or mo	Work hard	
11718594233	7	12010	Female	45-54	Hispanic or Latino	Between $25,000 and $49,	1	Some college	Working remotely full-ti	I have family with h	To be inform and educated	I don't have idea.	don't know	no
11713414311	8	14621	Female	35-44	Black or African A	Under $10,000	0	Some college						
11711737376	9	14605	Male	45-54	Hispanic or Latino	Between $25,000 and $49,	0	Higher	Working an essential pos	As an essential work	Continued community. Covid	We closed our communit	Groceros resources	I have family in
11709965013	10		Female	45-54	Hispanic or Latino	Between $25,000 and $49,	1	Some college	Working remotely full-ti	I stay home	Essential products	Help to catch on bills		
11709295281	11	12033	Female	45-54	Hispanic or Latino	Between $25,000 and $49,	3	Bachelor Degree	Working remotely full-ti	A computer.	Social security	Health insurance	Cost of medicine.	Family are suffer
11705457950	12	14502	Female	25-34	American Indian or	Between $10,000 and $24,	5 or more	Some college	Working remotely full-ti	Food	I don't know. You can't co	Help with food and utilities		My husband is an
11702928320	13		Female	35-44	Hispanic or Latino	Between $25,000 and $49,	0	High school / GED	Working remotely part-ti	keep enough food in my house		stimulus check to hel	none	n/a
11701766188	14	14613	Female	35-44	Hispanic or Latino	Between $25,000 and $49,	3	Higher	Working remotely full-ti	An inside gym with a	Nothing	Not sure	Free childcare	No
11701703071	15		Female	25-34	Hispanic or Latino	Between $50,000 and $74,	0	Higher	Working remotely part-time		stable income			
11701601763	16	14616	Female	35-44	Hispanic or Latino	Between $25,000 and $49,	2	Associate's Degree	Working remotely full-ti	To take care of my c	Patience and faith	More hours at job	None	
11700226867	17	14621	Female	25-34	Hispanic or Latino	Between $25,000 and $49,	3	Bachelor Degree	Working remotely full-ti	n/a	People to stay home	none		
11700042212	18		Female	25-34	Hispanic or Latino	Between $25,000 and $49,	0	Bachelor Degree	Working remotely full-ti	People running errand	To continue working from h	A second stimulus check.		

Figure 2.1 Data Overview[1]

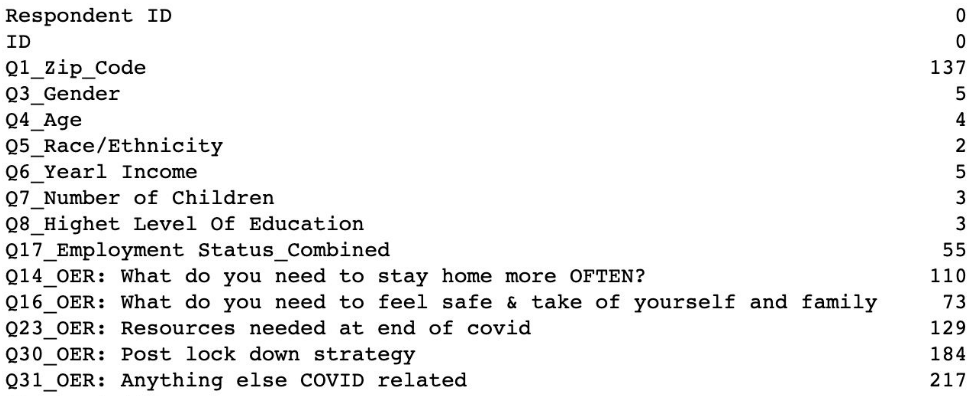
The questions in the format of choices/selections will be the foundation of the analysis, as they will be the features of comparison. Due to the purpose of this dataset and survey is to analyze the needs of the local community for the RMAPI, the Zip_Code will be the feature to be used to visualize the open-ended questions. Regarding the open-ended questions, the sentence formatted answers need to be preprocessed into an analyzable format using NLTK and one-hot encoding, which will be explained in 2.4.

2.3 Data Cleaning

2.3.1 Missing values:

Many entries of answers are left blank, which would be considered missing values. As shown in figure 2.2, the number of the missing values exists in all the questions, except for the first two auto-generated attributes for survey takers.

[1] RMAPI Rochester Community Feedback on Covid-19 Crisis Dataset



```
Respondent ID                                                          0
ID                                                                     0
Q1_Zip_Code                                                          137
Q3_Gender                                                              5
Q4_Age                                                                 4
Q5_Race/Ethnicity                                                      2
Q6_Yearl Income                                                        5
Q7_Number of Children                                                  3
Q8_Highet Level Of Education                                           3
Q17_Employment Status_Combined                                        55
Q14_OER: What do you need to stay home more OFTEN?                   110
Q16_OER: What do you need to feel safe & take of yourself and family  73
Q23_OER: Resources needed at end of covid                            129
Q30_OER: Post lock down strategy                                     184
Q31_OER: Anything else COVID related                                 217
```

Figure 2.2 Number of Missing Values for Each Question

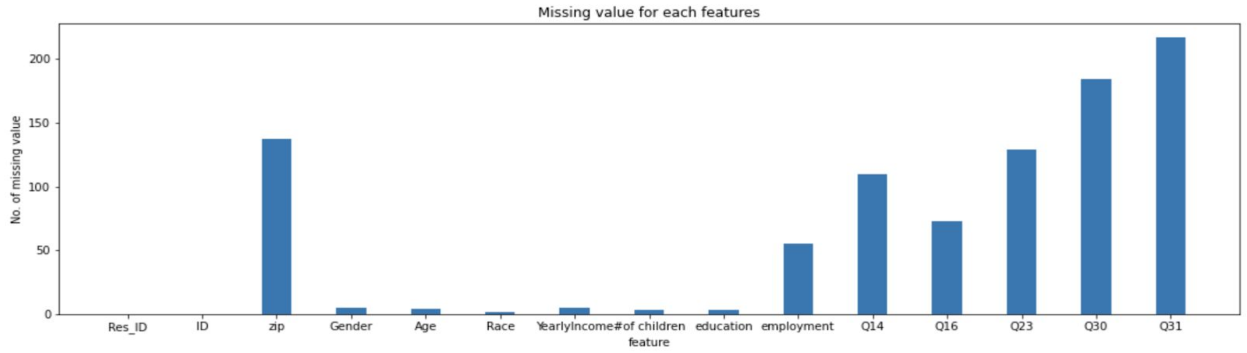


Figure 2.3 Distribution of Missing Values on Questions

As shown in figure 2.3, the most frequent missing values are in Zip_Code and open ended response. This will increase the erroneousness in analyzing the needs(extracted from Open Ended Questions) based on the area(attribute Zip_Code).

2.3.2 Unusual values or Outliers:

No noise or outliers for choice responses.

In open ended responses, there is no outliers or noise, as all the open responses are taken in consideration of the feature/label extraction process.

The noises and outliers are generated after the Natural Language Processing step, as the words being mentioned such as “the” or “a” will become noise.

2.3.3 Duplicates: None of the entries are duplicated.

2.4 Data Preprocessing

2.4.1 NLP

The Natural Language Toolkit (NLTK) is a platform used for building programs for text analysis.

To judge whether a word is important in an answer, the first step is the Tokenization. It is the process by which a big quantity of text is divided into smaller parts called tokens. By doing this, the separated tokens are very useful for finding such patterns as well as are considered as a base step for stemming and lemmatization.

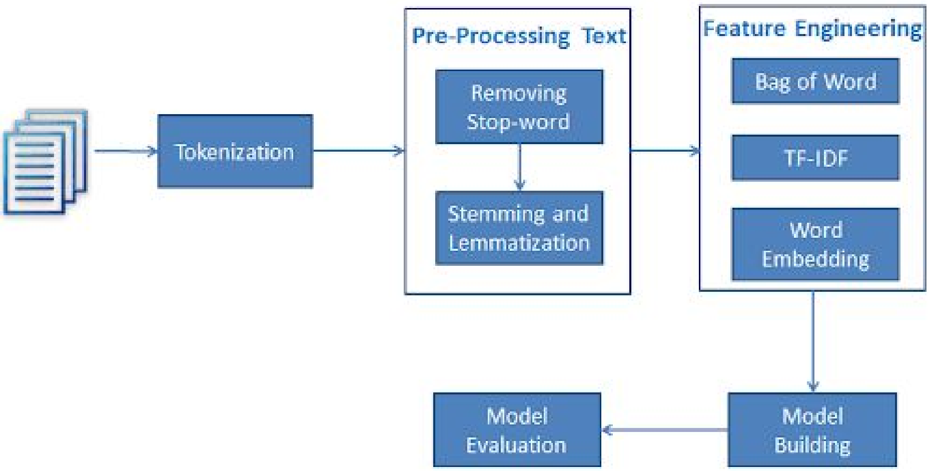


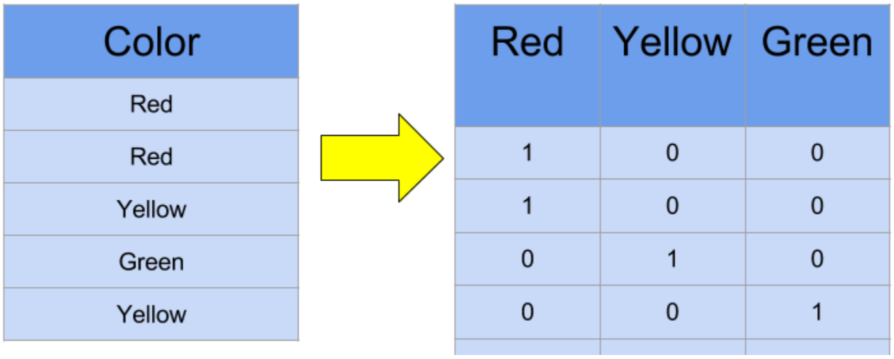
Figure 2.4 Tokenization[2]

Stemming: Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. It is trying to cut off the small tail of inflection that does not affect the part of speech.

Lemmatization: Categorize various types of word inflections into one form.

2.4.2 One-Hot-Encoding

A one hot encoding is a representation of categorical variables as binary vectors. This first requires that the categorical values be mapped to integer values. Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1. During our exploration of the raw data, it will changed the response from attributes.

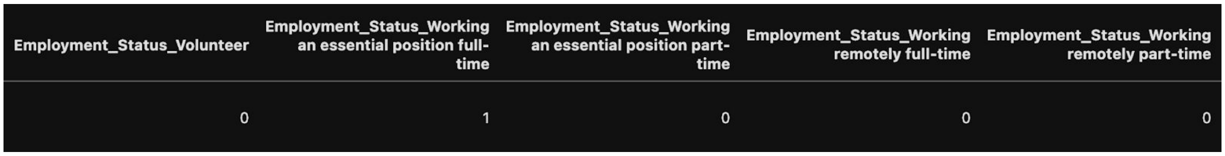


Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1

[2] Reference 1.Dansbecker. (2018, January 22). Using Categorical Data with One Hot Encoding. Retrieved December 09, 2020, from https://www.kaggle.com/dansbecker/using-categorical-data-with-one-hot-encoding

Figure 2.5 One-Hot-Encoding [3]



Employment_Status_Volunteer	Employment_Status_Working an essential position full-time	Employment_Status_Working an essential position part-time	Employment_Status_Working remotely full-time	Employment_Status_Working remotely part-time
0	1	0	0	0

Figure 2.6 One-Hot-Encoding for RMAPI Dataset

All 13 Columns except for 2 Unique ID, have all been transformed into binary vectors by One Hot Encoding.

3. EXPLORATORY ANALYSIS

3.1 Data Visualization

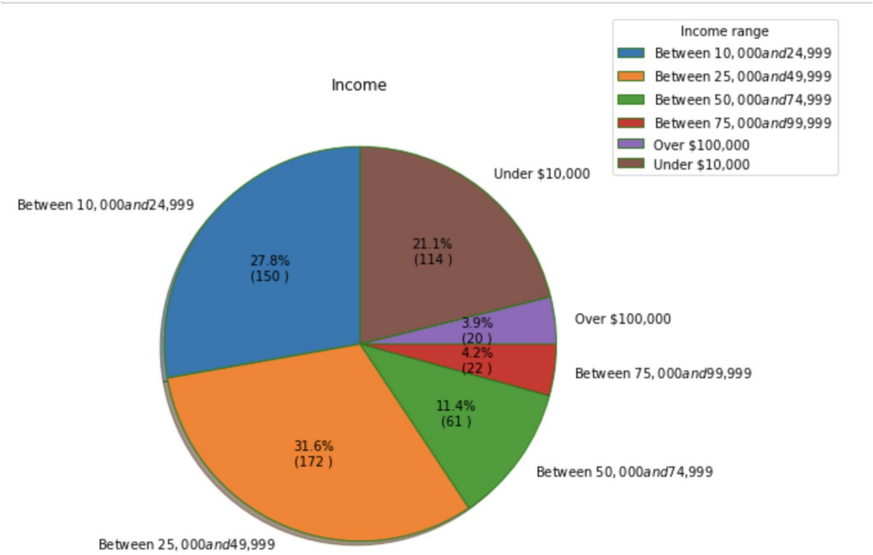


Figure 3.1: Pie Chart for Yearly Income of survey participants

[3] Reference 2. (Tutorial) Text ANALYTICS for Beginners using NLTK. (n.d.). Retrieved December 09, 2020, from https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk

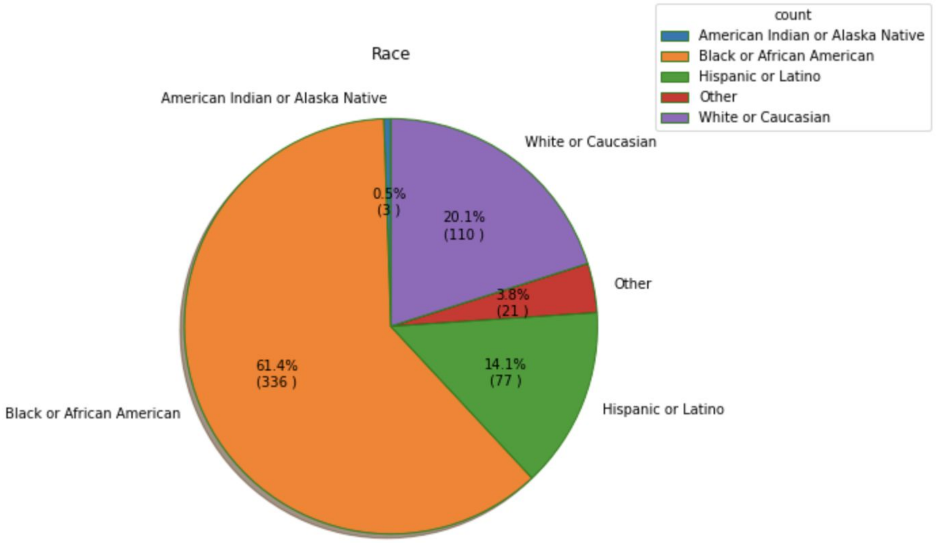


Figure 3.2: Pie Chart for Race distribution of survey participants

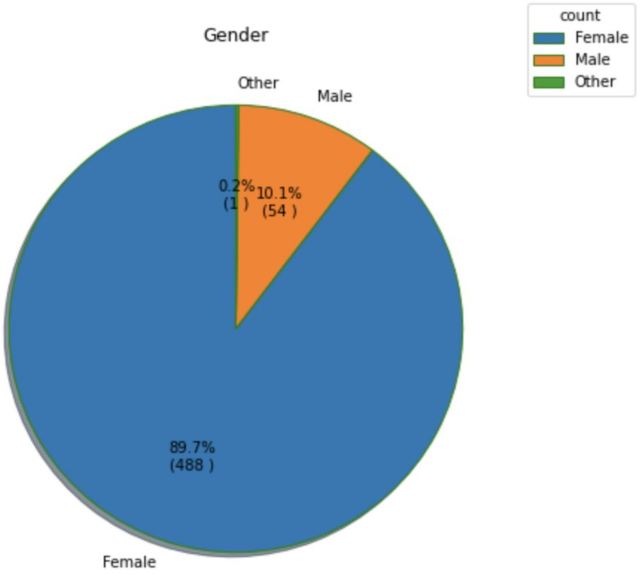


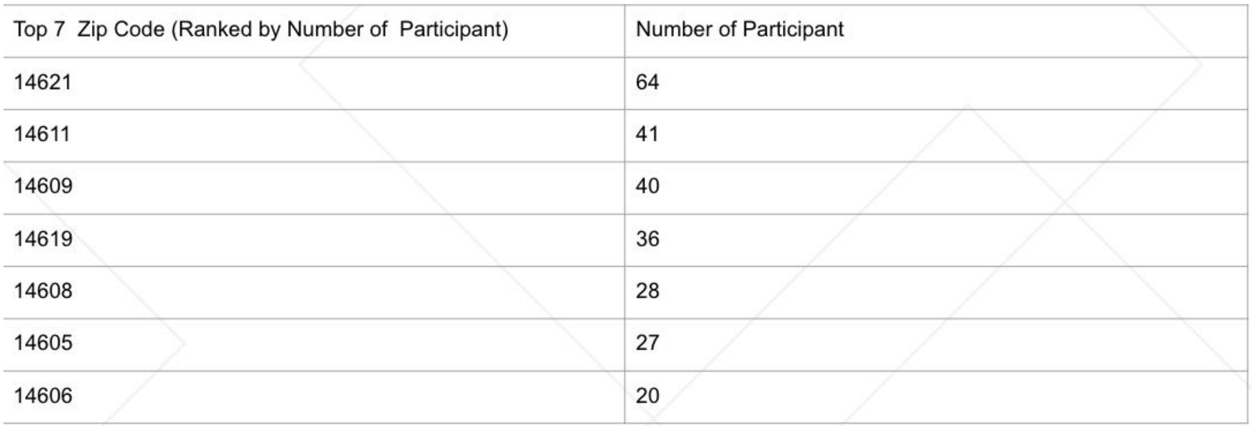
Figure 3.3: Pie Chart for Gender distribution of survey participants

Figure 3.1 plots the percentage distribution of yearly income for all participants. From the pie chart, it is clear that most of the participants are experiencing poverty. Nearly 50% of all participants are below or only slightly above the poverty line of New York state. At the same time, only about 20% of the participants can make above-average income. People who are experiencing financial issues are more likely to respond to the survey. This is, for the sponsor, a good sign, because it means that the survey is directed to the correct group.

Figure 3.2 plots the percentage distribution of race for all participants. Roughly 80% of the people who answered the survey are from minority groups. Minority groups are more willing to answer the survey compared to other participants. Compared to white people, in the United States, minority groups are more likely to suffer poverty. This is another sign that shows RMAPI is receiving data from their targeted group.

Figure 3.3 shows a significant imbalance concerning gender division among participants. 90% of the participants are female, only 10% are male. It is not clear why during this pandemic, females are much more willing to respond to the survey compared to males. But this phenomenon should raise consideration of how accurate, on reflecting the real situation, the survey is. In Monroe County, the division between male and female is about 50 to 50. As a lot of information from the male group is missing, the survey data collected could be highly biased. The survey collected might only be able to represent the female group but not the male group. As the project progresses, the sponsor should take this critical information revealed from the data into consideration.

3.2 Visualization Based on Area(zip code 14621 is used as an example)



Top 7 Zip Code (Ranked by Number of Participant)	Number of Participant
14621	64
14611	41
14609	40
14619	36
14608	28
14605	27
14606	20

Figure 3.4: Top 7 zip code (# of participants)

From Figure 3.4 we can see that 64 participants respond to the survey in area 14621, this is also the area that has the greatest number of participants. It will be used as an example to illustrate the visualization of the data in a specific zip code.

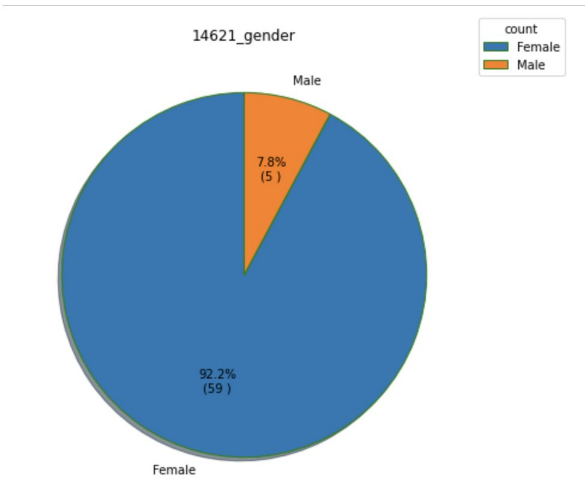


Figure 3.5

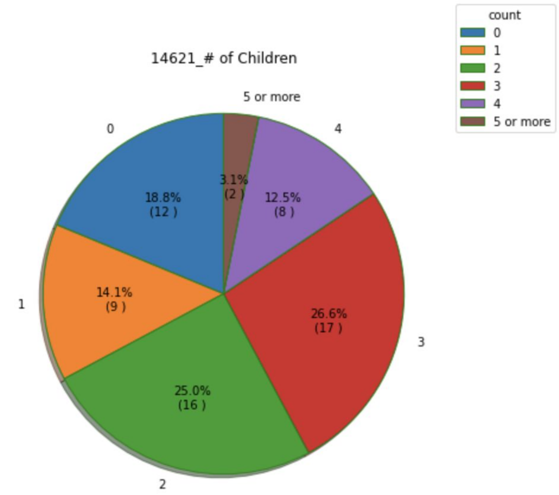


Figure 3.6

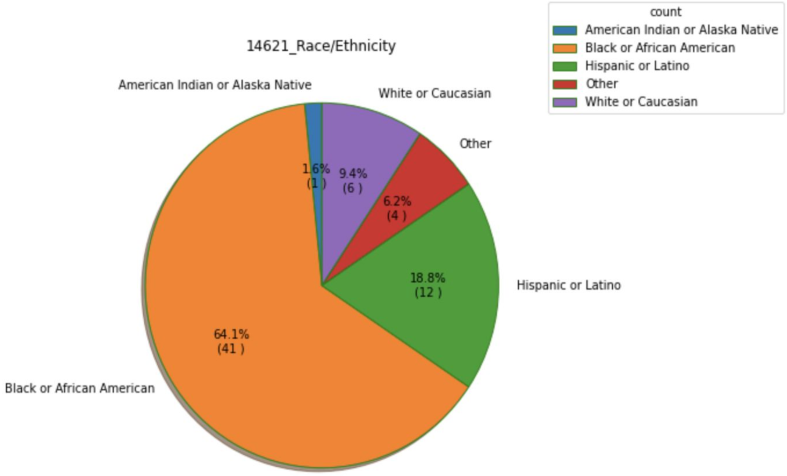


Figure 3.7 Race/Ethnicity

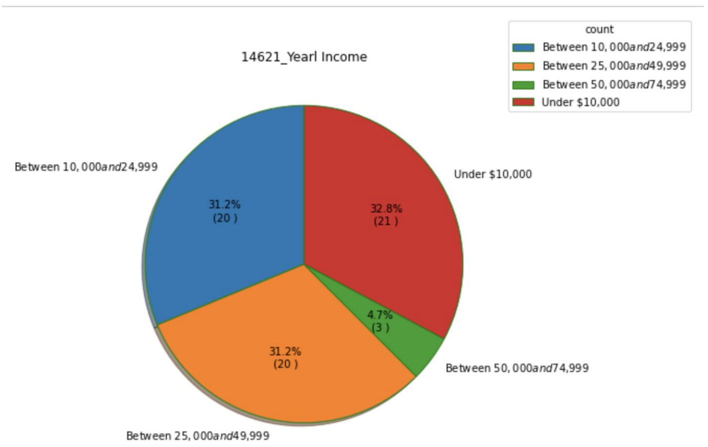


Figure 3.8 Yearly Income

Figure 3.5-3.8 illustrated the statistics for general information in area 14621. In this area, the females are still dominated in number, 92% of the participants are female. This discovery further proved that the gender imbalance mentioned in 3.1 is a common phenomenon in Monroe County. At the same time, the participants in this area are suffering more from poverty compared to the participants in other areas. Only 5% of all participants could make an above-average income, while the rest 95% are either experiencing a harsh financial crisis or barely away from it. Most of the families have more than one child and this would increase their pressure during the pandemic and make the situation they are facing worse. Finally, coinciding with Monroe County as a whole, the participants in this area are mainly made of minority groups.

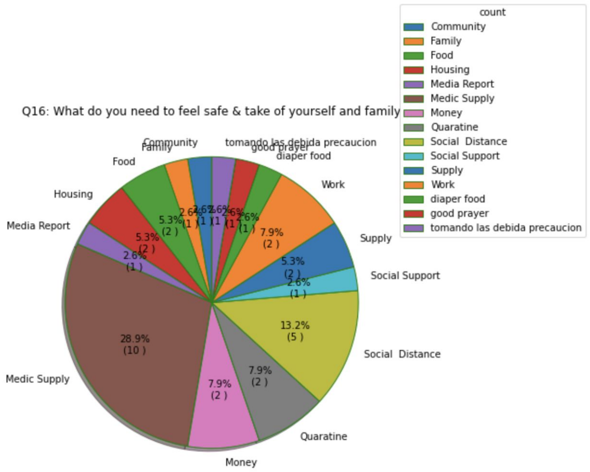


Figure 3.9

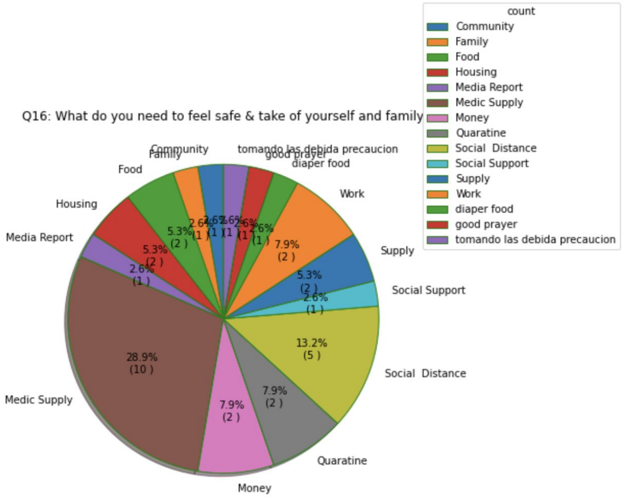


Figure 3.10

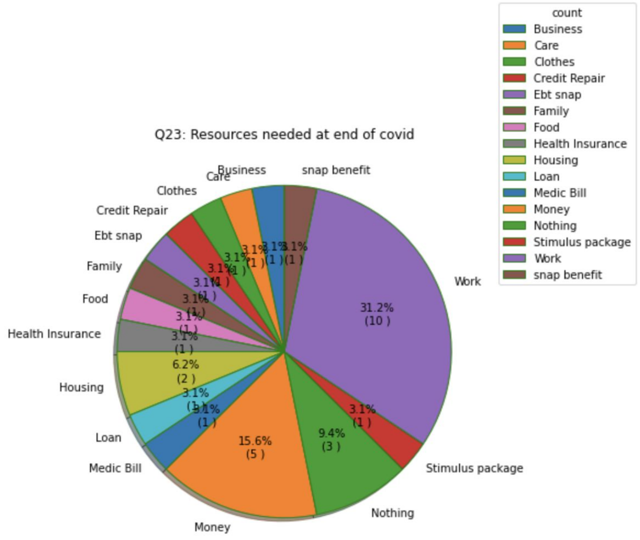


Figure 3.11

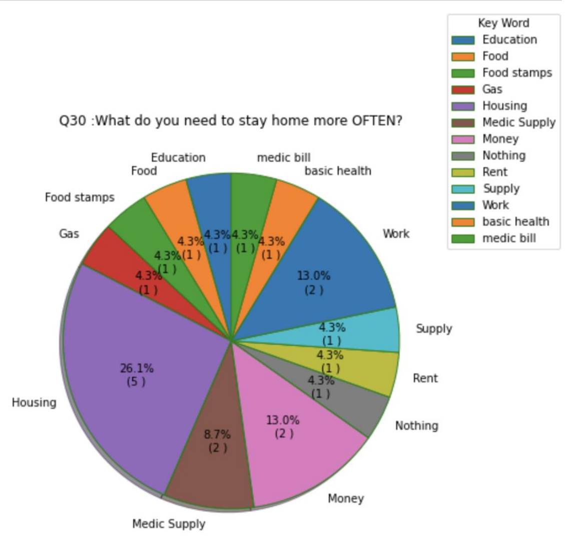


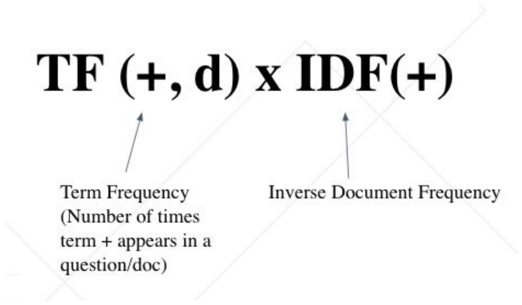
Figure 3.12

Figure 3.9 -3.12 illustrates the frequency of keywords extracted from four open-ended features. Those keywords are generated from the implementation of Natural Language Processing (mentioned in section 2.4) upon open-ended features. From graphs above, it can be viewed that during the pandemic, people are concerned about and experiencing difficulties from many factors. During Covid-19, in 14621, people were mainly concerned about Financial support, Medical supply, Housing, and Food supply. For most of the participants engaged in this survey in this area, they do not have the option to work remotely. Covid-19 raised a serious dilemma - take the risk and go to work or stay at home and have no income.

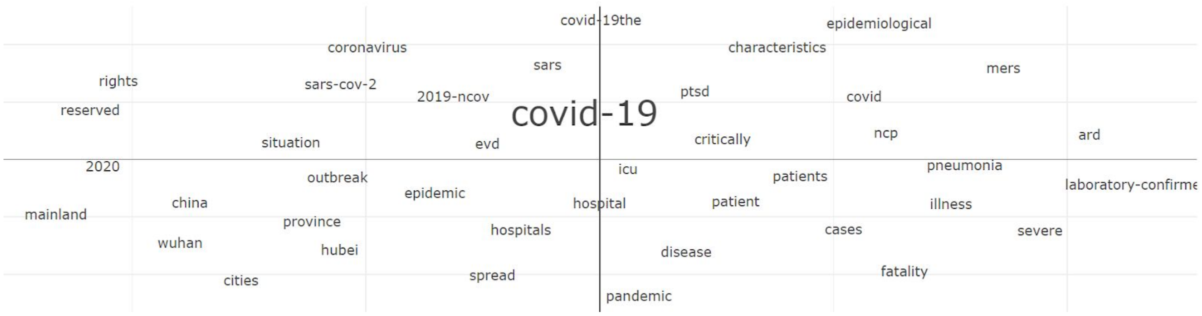
4. MODEL DEVELOPMENT

4.1 TF-IDF

But on the other hand, Some words appear frequently in various responses, and their importance is definitely not as important as those that appear frequently in a certain article. From a statistical point of view, it is to give that Uncommon words are given a larger weight, while the weight of common words is reduced. So here we would use TF-IDF, Term Frequency–Inverse Document Frequency, a commonly used weighting technique for information retrieval and data mining. It is often used to mine keywords in articles, and the algorithm is simple and efficient.


TF (+, d) x IDF(+)
Term Frequency (Number of times term + appears in a question/doc)
Inverse Document Frequency

TF-IDF has two meanings, one is "Term Frequency" (abbreviated as TF), and the other is "Inverse Document Frequency" (abbreviated as IDF). Below is a picture example of the word cloud generated based on TF-IDF model about Covid-19

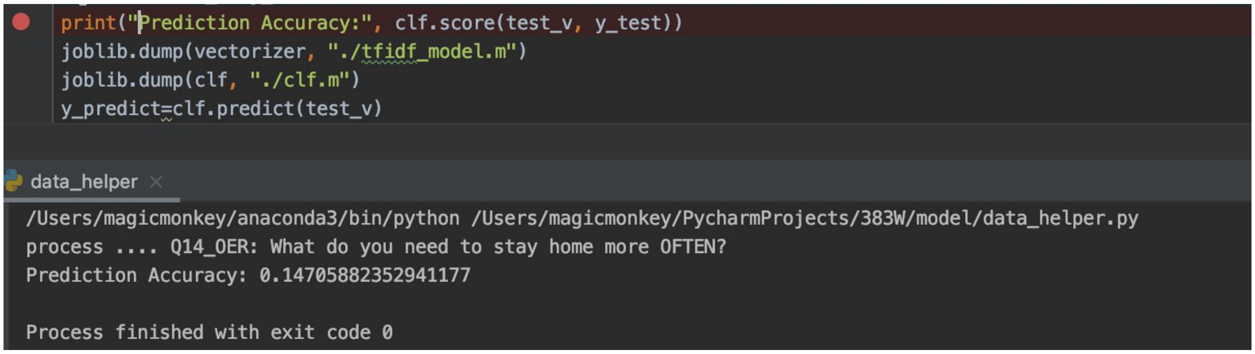


4.2 Model Building Process

5. PERFORMANCE AND RESULTS

5.1 Performance

For Zipcode and other open ended response prediction, We establish a supervised learning model, use Tidif to extract features from question sentences, and then use Bayesian model to classify text and zip code, build a classification model. By giving into the attribute info, the model can predict the zip codes and other open ended responses in case that they haven't filled in the columns during the survey. We do understand the accuracy now is 14.7% and still need to be improved.



```
print("Prediction Accuracy:", clf.score(test_v, y_test))
joblib.dump(vectorizer, "./tfidf_model.m")
joblib.dump(clf, "./clf.m")
y_predict=clf.predict(test_v)
```

```
/Users/magicmonkey/anaconda3/bin/python /Users/magicmonkey/PycharmProjects/383W/model/data_helper.py
process .... Q14_OER: What do you need to stay home more OFTEN?
Prediction Accuracy: 0.14705882352941177

Process finished with exit code 0
```

5.2 Result Analysis

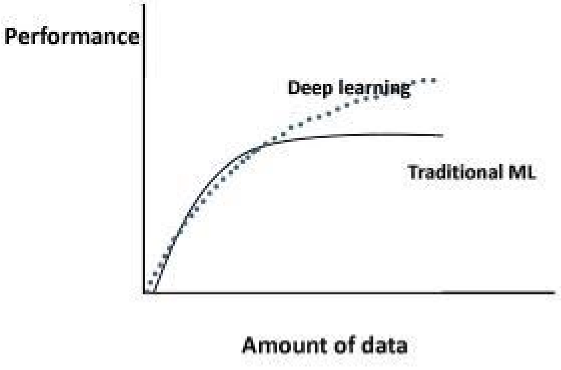
Since we are facing lots of challenges during the data exploration: we may only have up to 560 rows of respondents information and for many of the questions: we receive lots of invalid responses. Such as Question 30, up to 184 people haven't filled in anything, which is one third of the whole group of respondents. We may try to improve the accuracy together with the model by implementing the voice of Social Media: We may integrate with the information derived from social media (such as Facebook and twitter). Since we talked with the sponsor, acknowledging the answers towards questionnaires is really hard, doing this would definitely get access to more data.Also, setting up a website or interface that would help both the sponsor and other users in deriving the result and visualized data directly.

6. CONCLUSION AND FUTURE WORK

This paper and analysis employ the data from RMAPI's ROCHESTER COMMUNITY FEEDBACK ON COVID-19 CRISIS, with the two main goals. The first goal is to analyze the frequently mentioned living necessities from the open-ended responses from the residents. The frequent feature will help the initiative to better understand the community in which supplies to be provided. The second goal is to make predictions based on the area, which will provide necessities that each area will most probably need from the trained model and past data.

The first goal of using NLTK, the result from the frequency mining, provides the living necessities of each area with low noise, which does not contain repeated key features nor noise values such as "the" or "a". This result is satisfying and useful for future analysis.

Regarding the second goal, the prediction model can predict the features based on the Zip code. However, since the amount of data in the training process is not enough to create a high accuracy. As shown in figure 6.1, the current result in low precision has resulted from the lack of amount of data. With more data in the future surveys, combining with the current data, the accuracy will be exponentially improved.



[4]

Figure 6.1

[4] Mitsa, T. (2019, April 23). How Do You Know You Have Enough Training Data? Retrieved December 15, 2020, from https://towardsdatascience.com/how-do-you-know-you-have-enough-training-data-ad9b1fd679ee

Acknowledgements

For this project, we greatly appreciate and thank Mr. Zaman, Ms. Angela, Mr. Zhao and Professor Ajay Anand, and Professor Pedro Fernandez for helpful suggestions and feedback.

References:

1. " Rochester Community Feedback On COVID-19 Crisis. (n.d.). Retrieved December 15, 2020, from https://endingpovertynow.org/rochester-covid-19-feedback/
2. Mitsa, T. (2019, April 23). How Do You Know You Have Enough Training Data? Retrieved December 15, 2020, from https://towardsdatascience.com/how-do-you-know-you-have-enough-training-data-ad9b1fd679ee
3. Natural Language Toolkit¶. (n.d.). Retrieved December 16, 2020, from https://www.nltk.org/
4. Tokenization. (n.d.). Retrieved December 16, 2020, from https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html
5. Li, S. (2019, August 02). Hands On Bayesian Statistics with Python, PyMC3 & ArviZ. Retrieved December 16, 2020, from https://towardsdatascience.com/hands-on-bayesian-statistics-with-python-pymc3-arviz-499db9a59501
6. Lavin, M. (2019, May 13). Analyzing Documents with TF-IDF. Retrieved December 16, 2020, from https://programminghistorian.org/en/lessons/analyzing-documents-with-tfidf
7. Keyword extraction from text using nlp and machine learning. (2019, December 27). Retrieved December 16, 2020, from https://www.einfochips.com/blog/how-to-extract-keywords-from-text-using-nlp-and-machine-learning/
8. DelSole, M. (2018, April 24). What is One Hot Encoding and How to Do It. Retrieved December 16, 2020, from https://medium.com/@michaeldelsole/what-is-one-hot-encoding-and-how-to-do-it-f0ae272f1179