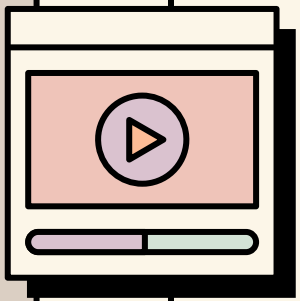
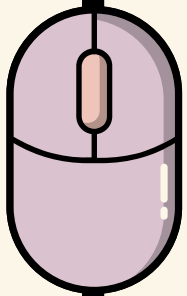
NYU Datathon 2023 Submission





By the Data Divas: Zakaria Arshad and Shubhi Upadhyay

NYU

Dataset

Includes information relating to wellbeing and basic needs of adults in the United States in the year 2020

Examples:

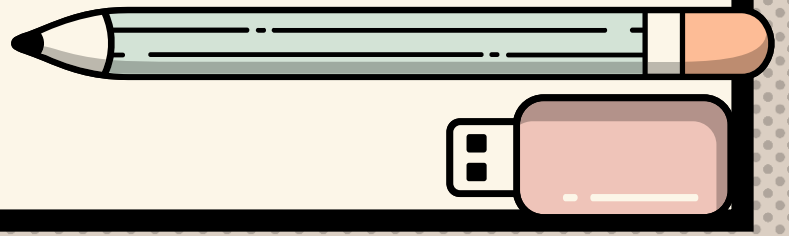
- Income
- Access to healthcare
- Health Conditions
- Food security
- Education
- Perception of COVID-19 and COVID-19 Vaccine
- State of Residence

7737 rows, 392 columns

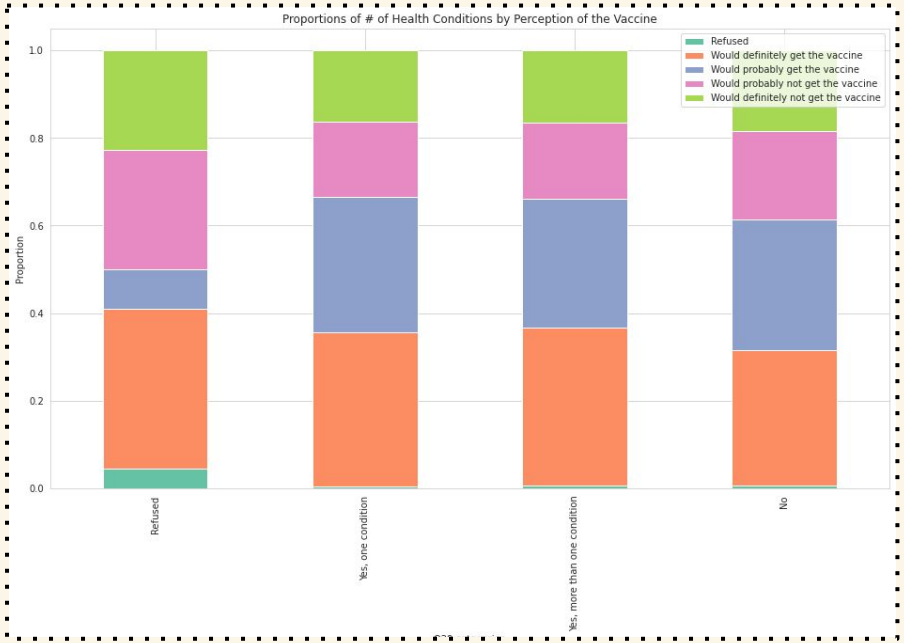
Main Focuses

Visualizing and Modeling of:

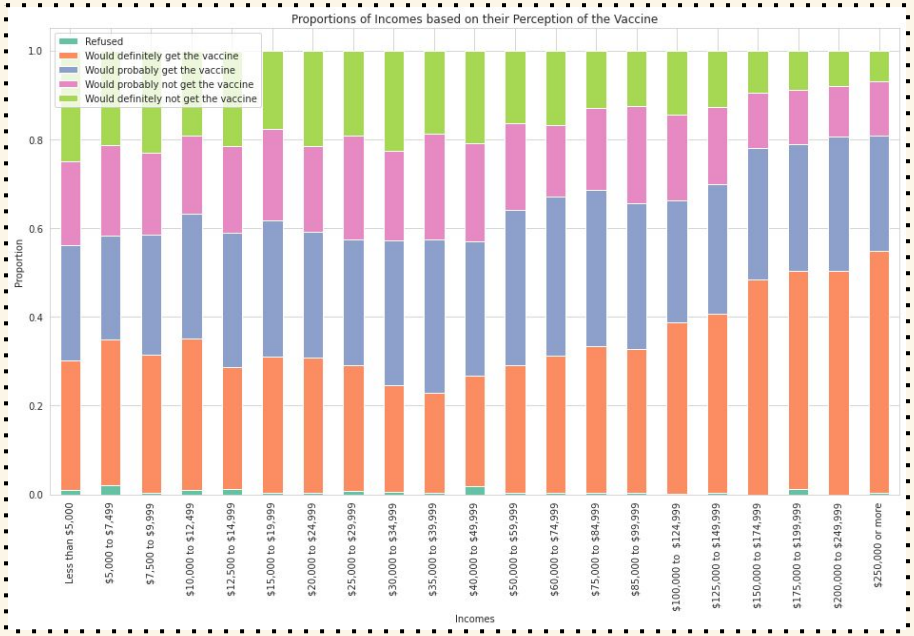
1. Individual's demographics (e.g. health, state of residence, household income, and education) to their perception of COVID
2. Household income to quality of children's diet
3. Correlation between negative emotions



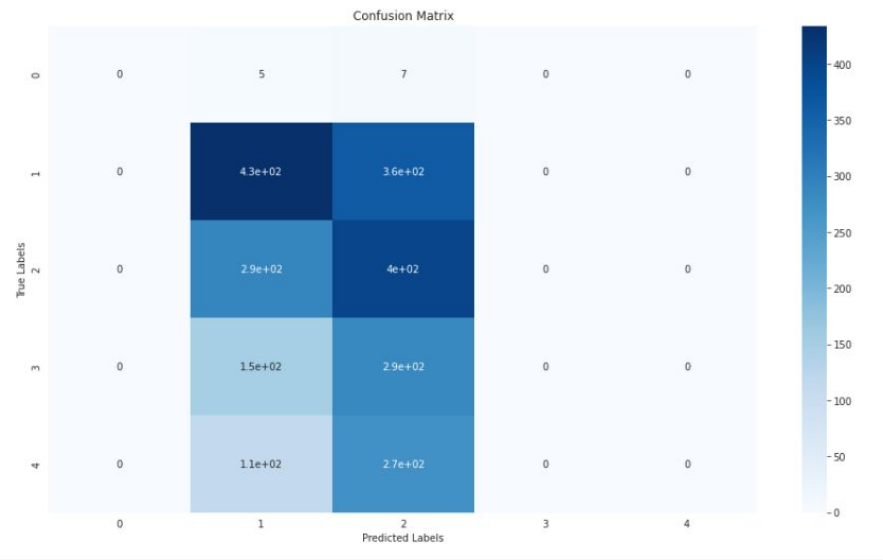
Proportions of # of Health Conditions by Perception of the Vaccine



Proportions of Incomes by Perceptions of the Vaccine



Modeling Education to Perception of the Vaccine



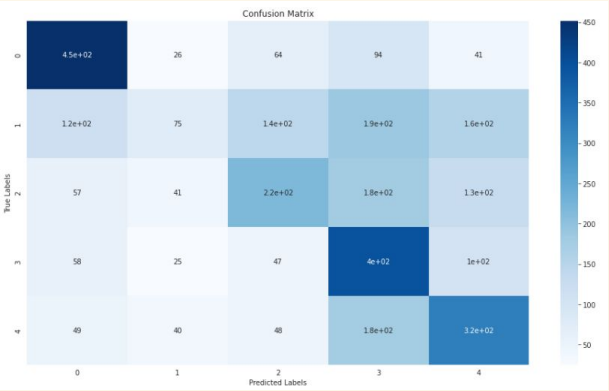
K-nearest neighbors model classifying only labels 1 and 2, which is “Would definitely get the vaccine” and “Would probably get the vaccine”.

Evidence of classifier being biased to majority classes

SMOTE Technique for Previous Model

```
{-1: 2578, 1: 2578, 2: 2578, 3: 2578, 4: 2578}
```

Using SMOTE (oversampling), all classes have the same number of instances. Improved confusion matrix.



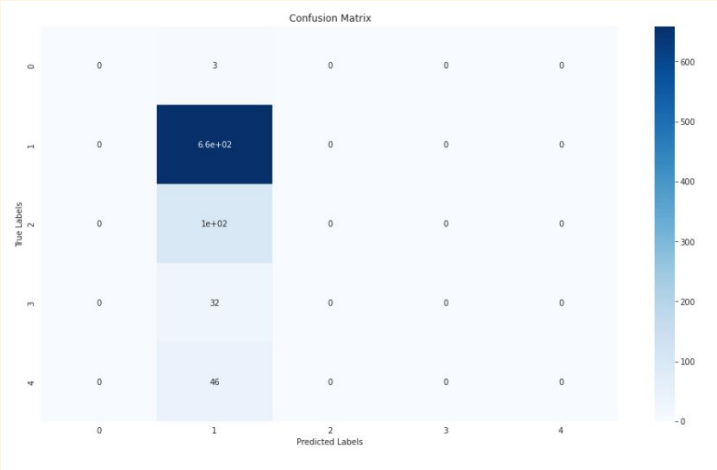

```
Accuracy for class 'Refused': 0.844
Recall for class 'Refused': 0.668
Accuracy for class 'Very Important': 0.772
Recall for class 'Very Important': 0.110
Accuracy for class 'Somewhat Important': 0.782
Recall for class 'Somewhat Important': 0.349
Accuracy for class 'Not Too Important': 0.731
Recall for class 'Not Too Important': 0.629
Accuracy for class 'Not At All Important': 0.771
Recall for class 'Not At All Important': 0.505
```

Accuracy measures the proportion of correctly classified instances (both true positives and true negatives) over the total number of instances. It provides an overall assessment of how well the model predicts the class labels.

Recall measures the proportion of true positives (correctly predicted positives) over the total number of actual positives in the data. It indicates the model's ability to identify positive instances correctly and is especially important when the cost of false negatives is high.

Modeling "Education" and "Income" to the population who would responded “probably/definitely not” get a coronavirus vaccine.

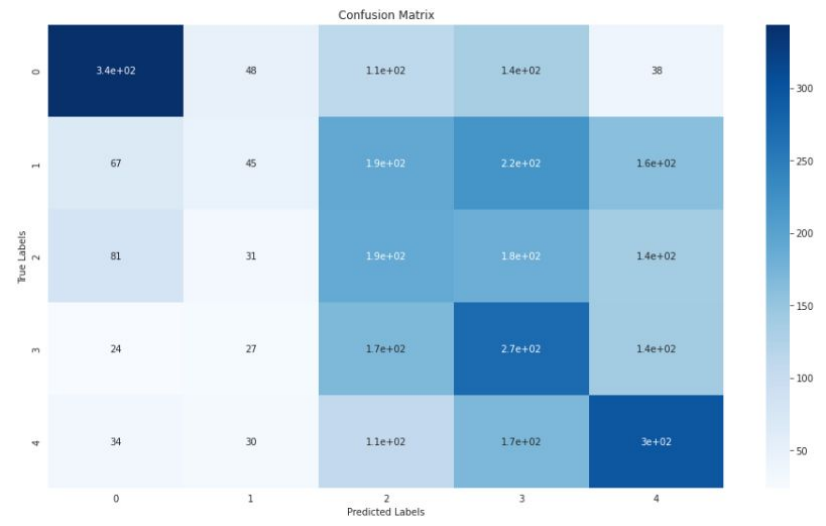
Specifically, their response to "You want to know more about how well the vaccine works"



K-nearest neighbors model

Classifier only classifies to label 1, which is approximately ⅔ of the data. Data is imbalanced.

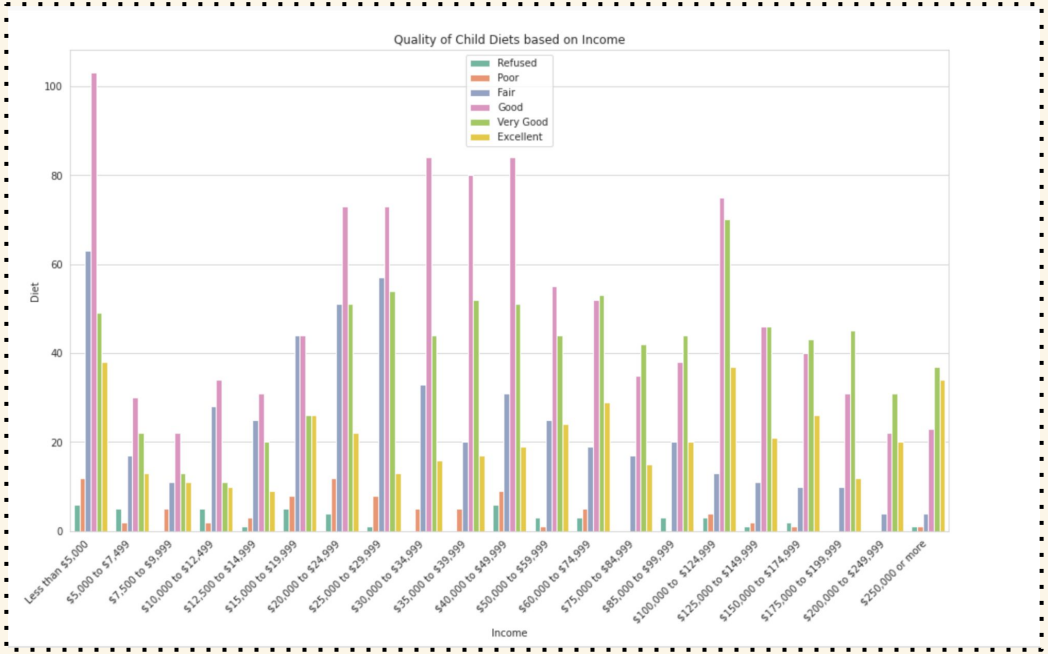
SMOTE Technique for Previous Model




```
Accuracy for class 'Refused': 0.834
Recall for class 'Refused': 0.508
Accuracy for class 'Very Important': 0.762
Recall for class 'Very Important': 0.066
Accuracy for class 'Somewhat Important': 0.688
Recall for class 'Somewhat Important': 0.304
Accuracy for class 'Not Too Important': 0.672
Recall for class 'Not Too Important': 0.432
Accuracy for class 'Not At All Important': 0.749
Recall for class 'Not At All Important': 0.465
```

Model is improved and fairly accurate for all classes.
The model correctly identifies the majority of the negative instances as negative, but it fails to identify many positive instances as positive.

Relationship between Quality of Child Diets and Household Income

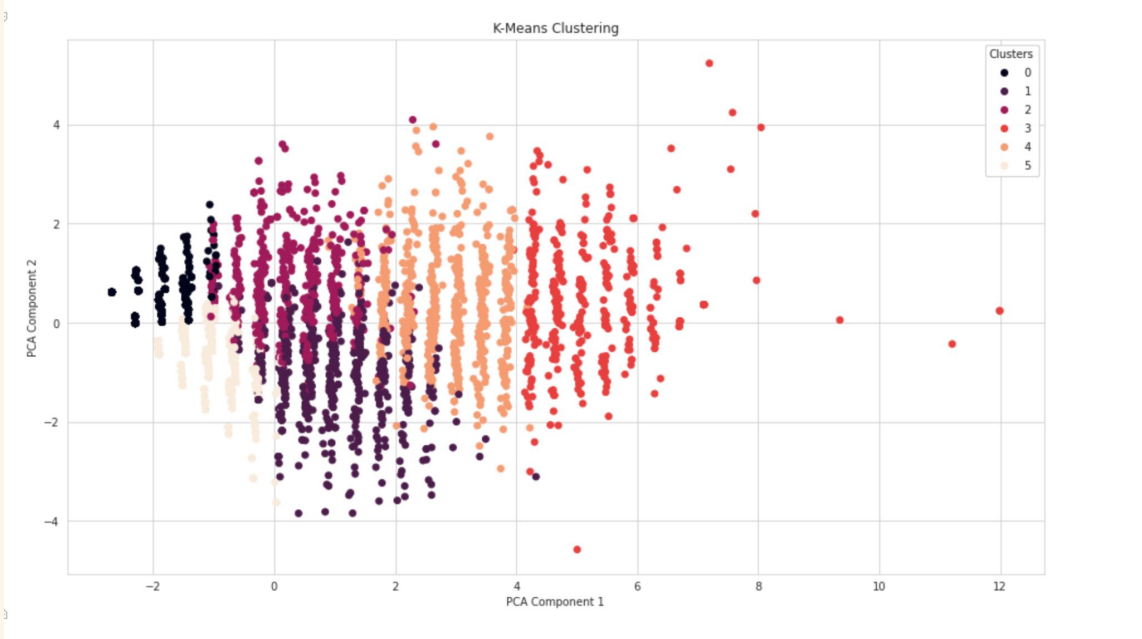


Using Clustering to compare levels of negative emotions between responders

Specific questions:
During the past 30 days, about how often did you feel:

1. ...nervous?
2. ...hopeless?
3. ...restless or fidgety?
4. ...so sad that nothing could cheer you up?
5. ...that everything was an effort?
6. ...worthless?

K-means Clustering



Clustering has revealed four fairly distinct clusters in 0, 3, 4, and 5.

There is some overlap in between clusters.

Clustering seems meaningful, but more analysis is needed to determine the relationship between responses.

Past 30 Days Emotion Correlation

There is a moderately strong, positive correlation amongst all variables.

For our categorical variables, it can be interpreted as if someone reports a lower frequency of a certain negative emotion, their responses to the other questions would also be lower.