CS6220 Final Report: Data Analysis on COVID-19 Vaccine

Alex Yang, Shuaizhen Li, Yiman Liu

I. Introduction

- The problem we're trying to address and why the problem is important:

According to CDC (https://www.cdc.gov/coronavirus/2019-ncov/vaccines), currently available COVID-19 vaccines in the United States have been shown the effectiveness at preventing COVID-19. People hesitate if they should get COVID-19 vaccine. On one hand, vaccination is an important tool to help stop the pandemic, so they can get their normal life back. On the other hand, they are worried about the side effects of the vaccines. Thus, we want to design a micro-service to recommend the most suitable COVID-19 vaccine to users based on their current illnesses, medication, allergies, and medical histories.

Our solution to make the recommendation is that we bring a user's personal information into our training models, one model comes from the data with Moderna vaccine, the other comes from the data with BioNTech vaccine. Then the model will be able to predict a list of symptoms that getting this vaccine could have. Finally, we will compare the two lists that generated from the two models, and recommend the vaccine with less or milder symptoms to this user.

- The reason we use data science:

Data science is a great tool to make this happen. Firstly, we can perform extensive data analysis and visualization by using various tools, such as Pandas and Matlab. Instead of the raw data, this visualization can demonstrate a clear idea of what the data means by providing visual context through graphs or plots. Users can better comprehend what current vaccine progress is and easier to identify trends and patterns. Secondly, by training the dataset, we can get a model to predict a possible outcome given a specific instance. For example, Tom would like to know what possible symptoms he may get after getting Pfizer. He just needs to input some of his personal information, and we will be able to generate the results.

Thus, I think data science can help us mining the data and analyzing the data, and finally find patterns as well as predict certain results. It can also help us visualize huge amounts of data to get insights of the dataset, which could provide more information than raw data.

II. Raw Data

We used three datasets in our analysis:

1. COVID-19 Vaccine Adverse Event Report
   https://www.kaggle.com/ayushggarg/covid19-vaccine-adverse-reactions

2. COVID-19 World Vaccination Progress
   https://www.kaggle.com/gpreda/covid-world-vaccination-progress

3. Data on COVID-19 cases, deaths, hospitalizations and tests
   https://github.com/owid/covid-19-data/tree/master/public/data

COVID-19 Vaccine Adverse Event Report has total 34121 instances and contains three major information:

1. Patients' personal information (35 features included): such as *age*, *sex*, *illnesses at time of vaccination*, *disability*, *allergies,* and *long-standing health conditions.*

2. Vaccine information (8 features included): such as *vaccine name, manufacturer, lot number, route, site,* and *number of previous doses.*

3. Adverse events (11 features included): such as *symptom* and *symptomversion.*

COVID-19 World Vaccination Progress dataset has 8079 instances and 15 features, including *total number of vaccinations in each country, total number of people vaccinated in each country, daily vaccinations, and etc*.

Data on COVID-19 cases, deaths, hospitalizations and tests dataset has 77954 instances and 59 features, including *total COVID-19 cases in each country, number of COVID-19 patients in hospital, and etc*.

III. Data Understanding

To better understand the datasets, we extracted some key features and performed data visualization. Let's first look at how vaccination affects the number of COVID-19 cases in the United States.

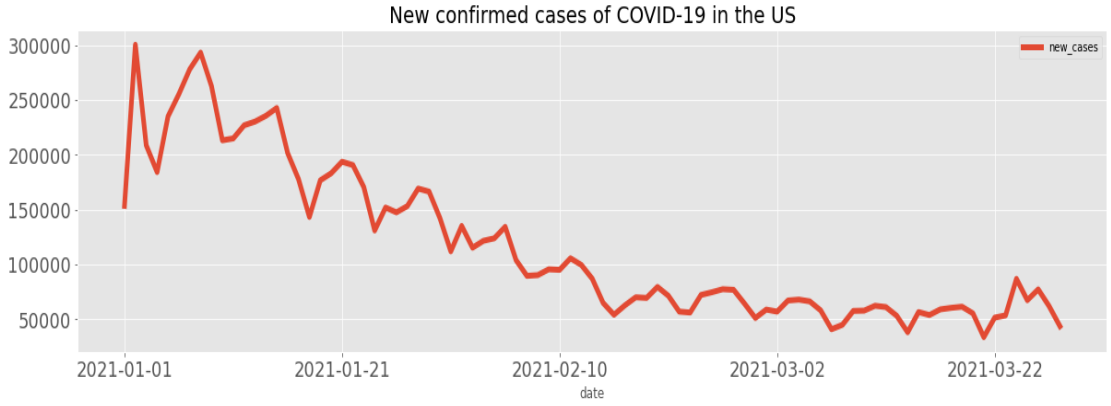


Figure 1: New confirmed cases of COVID-19

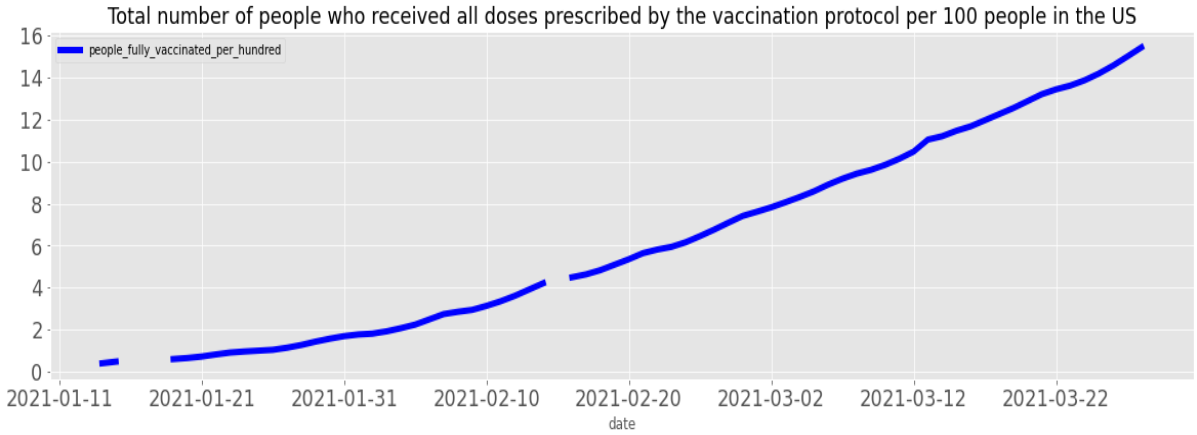


Figure 2: Total number of people fully vaccinated

From Figure 1 and Figure 2, we can see that the number of people fully vaccinated is increasing from January to March, while the number of new confirmed cases is decreasing, which indicates vaccination can help keep people from getting COVID-19, although there could be other reasons that cause the decrease of COVID-19 cases.

Not only in the United States, visualization has been proved as a useful tool to reduce the growth of COVID-19 cases around the world.

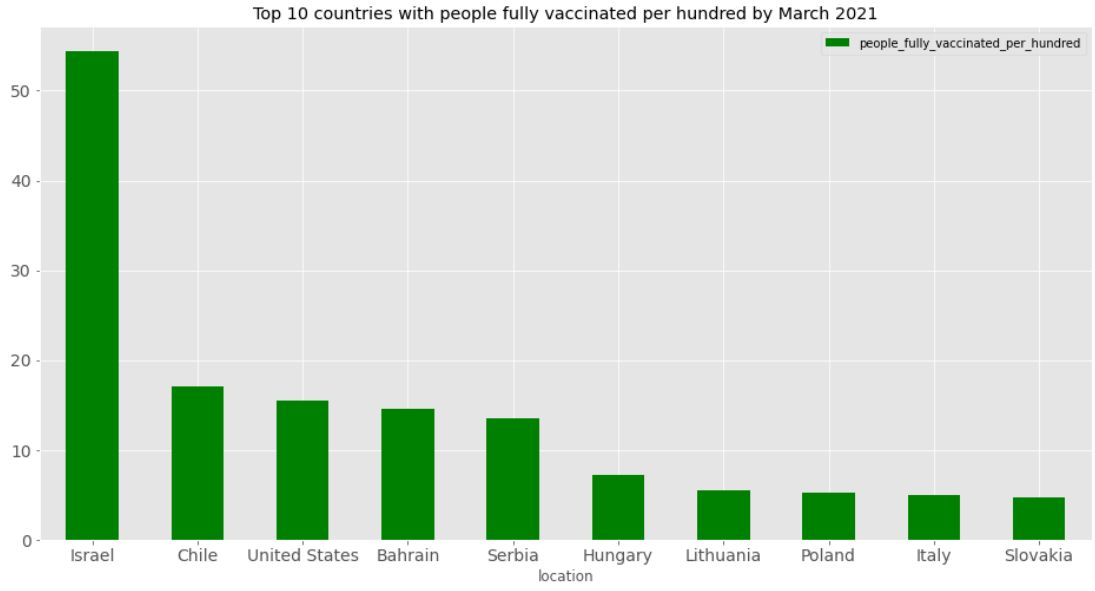


Figure 3: Top 10 countries with people fully vaccinated

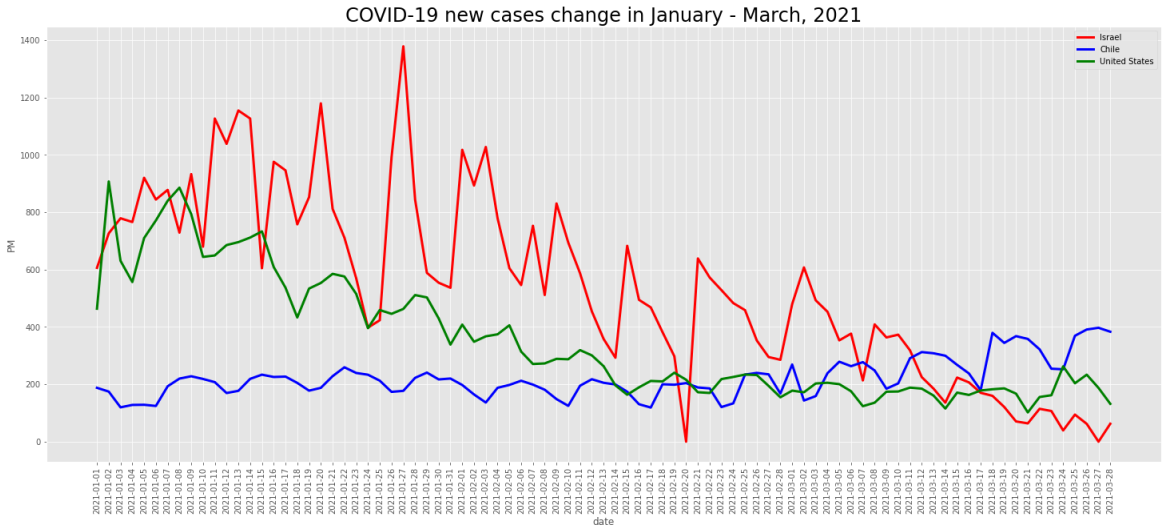


Figure 4: COVID-19 new cases change in Israel, Chile and United States

From Figure 3 and Figure 4, we can see that except Chile, new cases in Israel and United States are dropped very fast from January to March, and Israel has the most number of people fully vaccinated in the world. Thus, we can say visualization should be an efficient way to keep people from getting COVID-19. Also, we can find that the more people were fully vaccinated, the faster the new cases decreased.

Since vaccination can be regarded as a useful tool to help stop the pandemic, now let's look at the distribution of vaccines around the world.

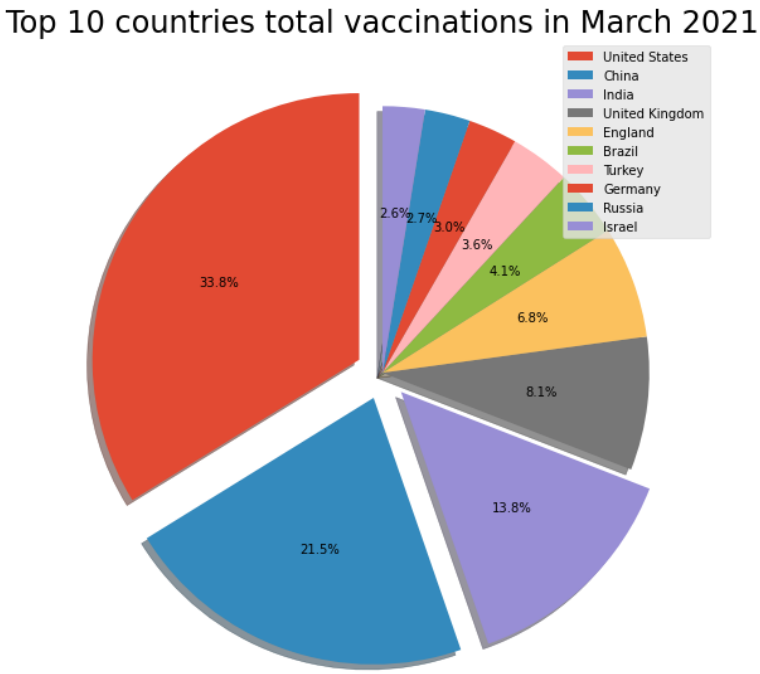


Figure 5: Total vaccinations in March, 2021

From Figure 5, we can see that the United States has the most number of COVID-19 vaccines, so it won't take too long for everyone in the United States to get the vaccine.

However, people may concern if vaccination is suitable for them, what kind of symptoms they may have after getting vaccines, and which vaccine they should get. So, now let's look at how many vaccines are available now, what they are, and how they differ from each other.

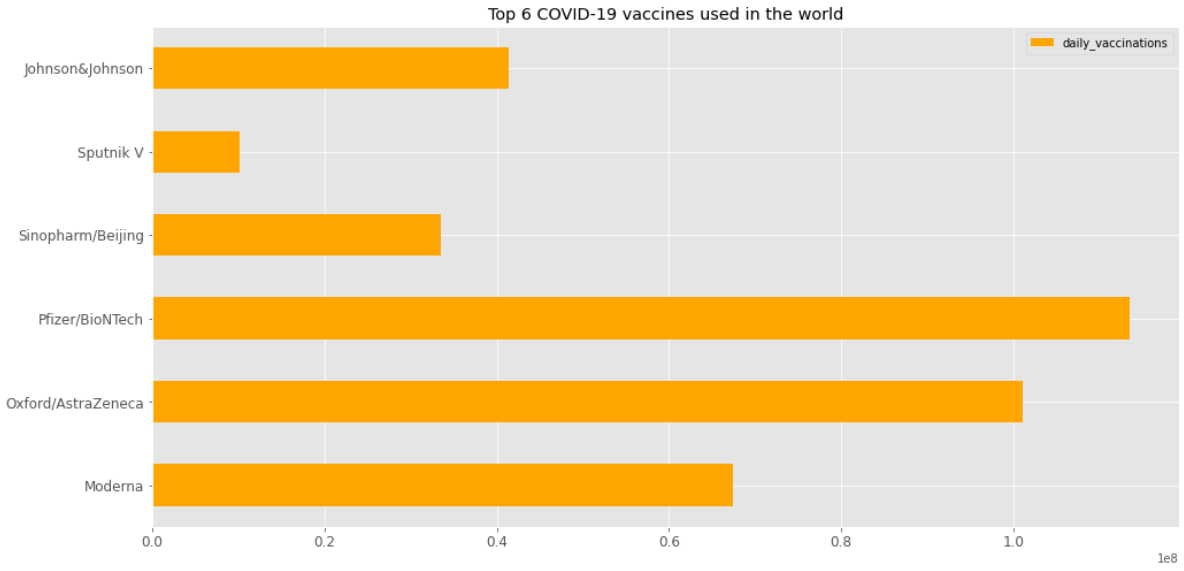


Figure 6: Top 6 COVID-19 vaccines used in the world

From figure 6, we can see Pfizer/BioNTech is the most popular vaccine used in the world. Now, let's look at how effective they are in preventing the COVID-19 virus.

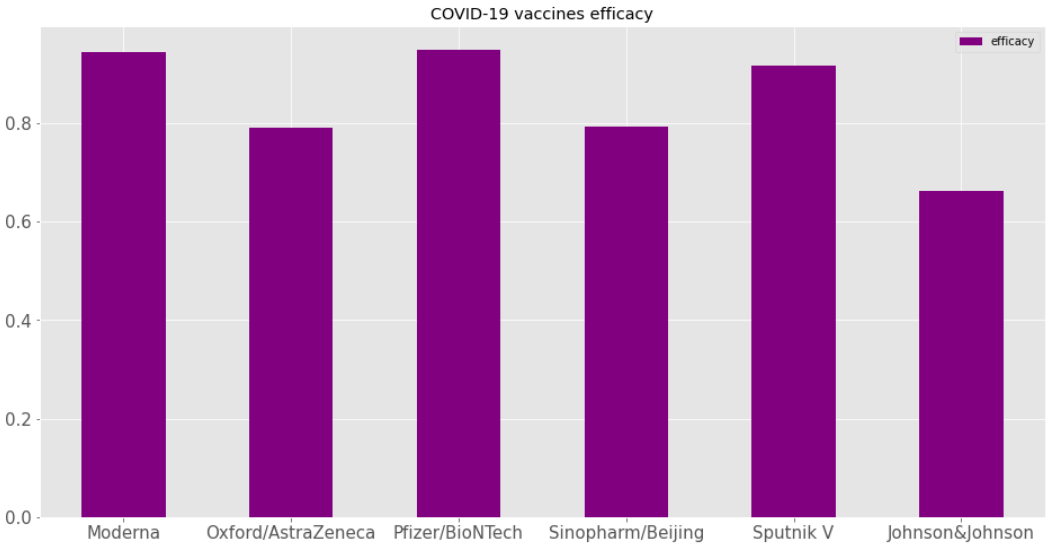


Figure 7: COVID-19 vaccines efficacy

From Figure 7, we can see that Pfizer/BioNTech and Moderna have the maximum efficacy among other COVID-19 vaccines. Thus, next, we will mainly analyze the symptoms of getting Pfizer/BioNTech and Moderna based on people's medical history, current illness and medication. We trained the dataset created by the Food and Drug Administration (FDA) and Centers for Disease Control and Prevention (CDC) to compare the possible side effects of getting those two vaccines, and finally recommend users the best suitable COVID-19 vaccines for him/her.

Let's now look at the most frequent symptoms people had after getting vaccines.

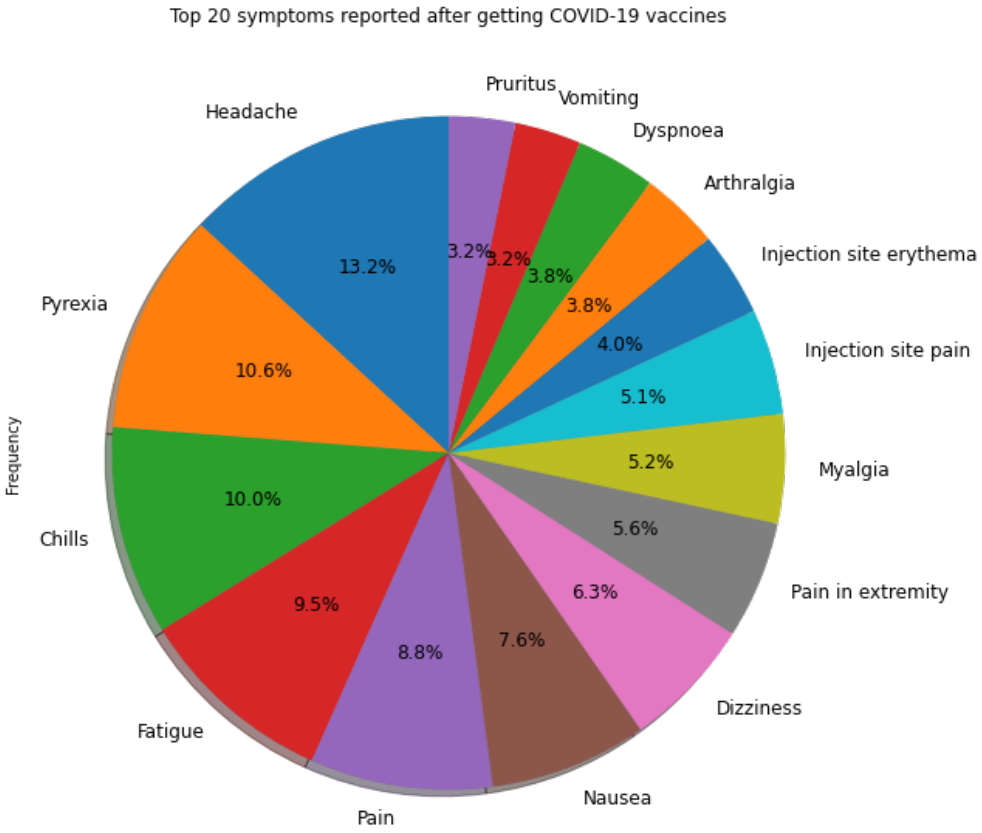


Figure 8: Top 20 symptoms reported after getting COVID-19 vaccines

We can see that 5.3% people in the dataset reported they got xanthopsia, zoster, or vomit. 5.2% people reported having vertigo or vasculitis.

We would like to know if a symptom is related to a certain health condition, so once we have a patient's personal medical data, we will be able to predict what kind of symptoms he/she may have, and recommend his/her the best vaccine. Now, let's look at what common long-term health conditions people have in the report.

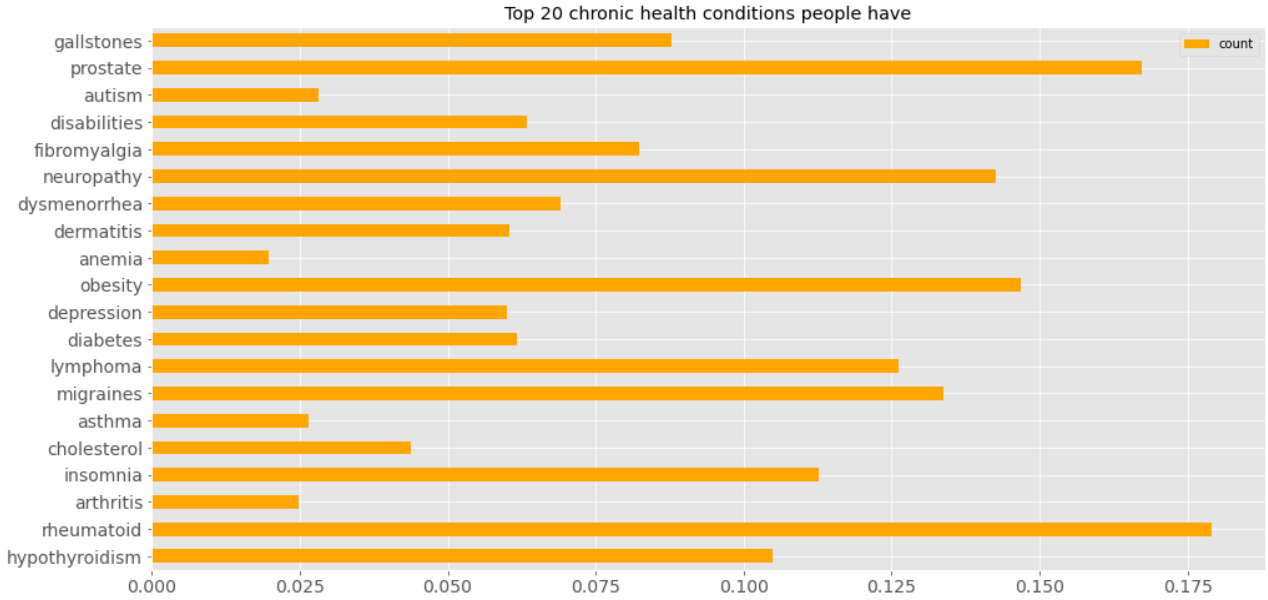


Figure 9: Top 20 chronic health conditions people have

In the *COVID-19 Vaccine Adverse Event Report*, all the sample instances had adverse symptoms. There are 18% people in the report having rheumatoid, 16% people having prostate, and 14% people having obesity. Next, we want to find out if the symptoms are related to people's current medications as well.

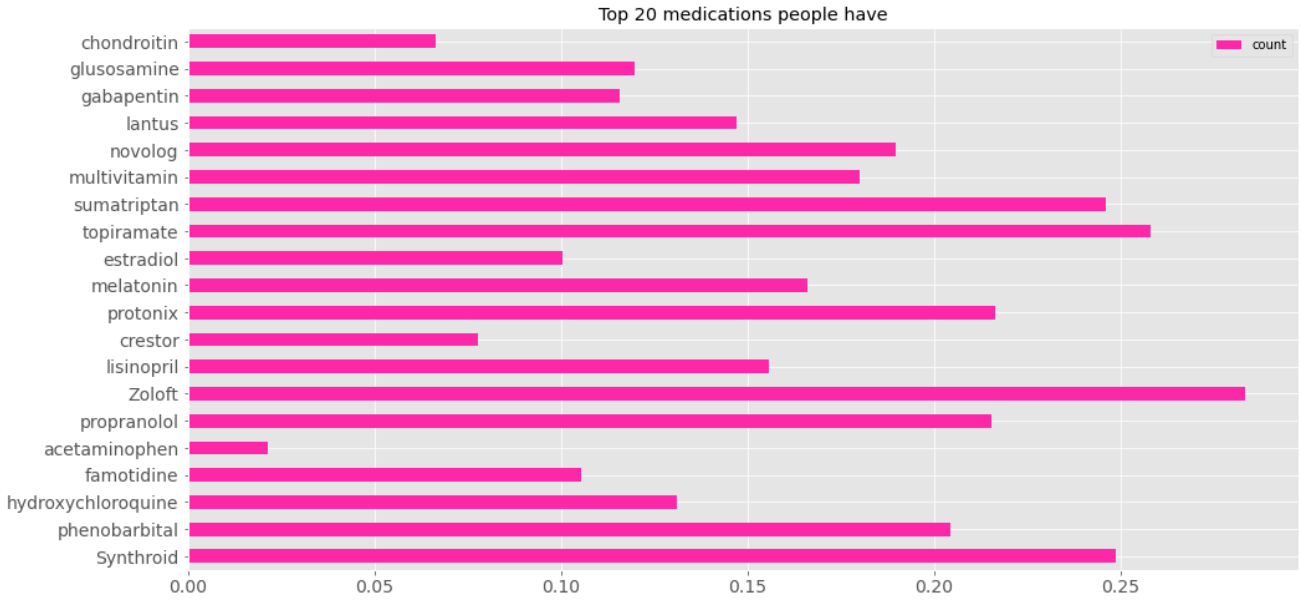


Figure 10: Top 20 medications people have

As we can see, there are 28% people currently taking Zoloft, 26% people currently taking topiramate, and 25% people currently taking Synthroid.

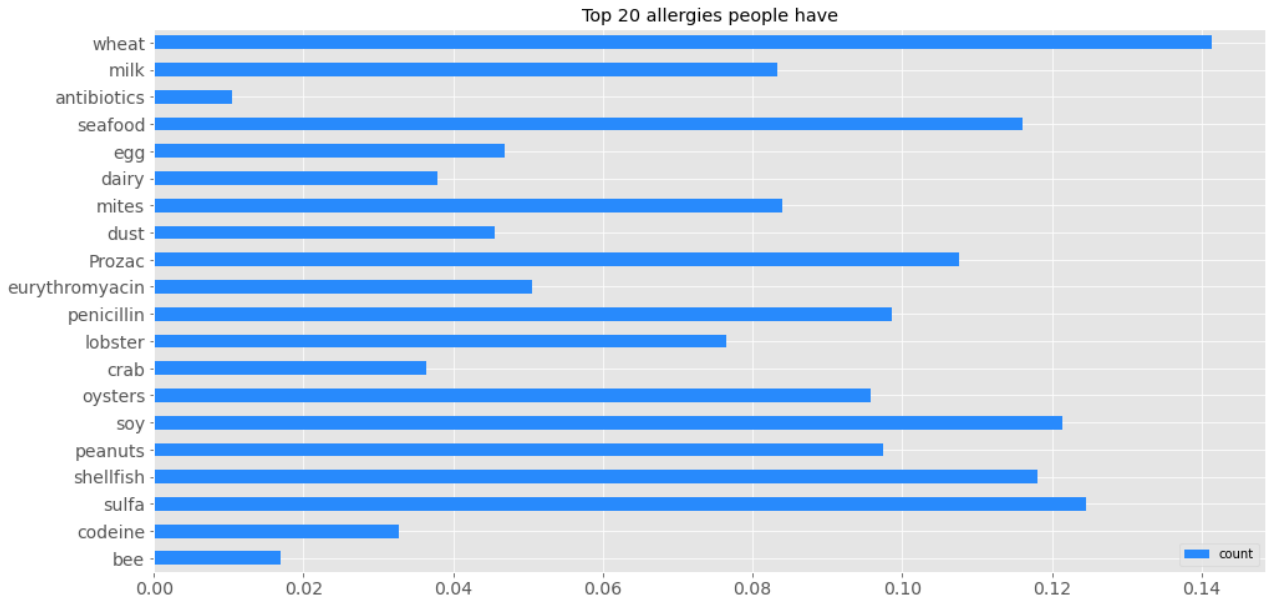


Figure 11: Top 20 allergies people have

Allergies may also have effects on the symptoms people have after getting the vaccines. As we can see from Figure 11, there are 15% people allergic to wheat. 12% people are allergic to sulfa, and 11.5% people are allergic to soy.

From above analysis, we can see that there are common symptoms reported and common health conditions reported, so there is a chance that we can associate people's health conditions with the symptoms they get after getting COVID-19 vaccines, and establish some kind of relations between them, so we can predict possible symptoms according to people's medical history, current illness and medications.

IV. Data Modeling

1. Model selection: K-means

Our initial approach using multilabel classification models gave us very low accuracies in predicting the vaccine side effects. The reason is that each patient in the dataset is associated with a subset of 4238 vaccine side effects. With such a wide range of symptoms, it is too difficult for the classification model to produce accurate predictions.

However, After trying different learning models, we think that using the k-means clustering model will give us better accuracy in predicting the vaccine side effects. Our idea is to cluster the patients based on their age, gender, and medical conditions. If their characteristics and vaccine side effects are correlated, the patients in the same cluster should have similar vaccine side effects. Thus, we can use each cluster's frequent side effects as a prediction, such as when a new patient is clustered in a particular group, we can output the group's frequent side effects as the prediction. If 3 out of 4 symptoms are matching, the model is considered 75% accurate.

2. Data preprocessing

The data frame we are using:

	VAERS_ID	AGE_YRS	SEX	CUR_ILL	HISTORY	ALLERGIES	VAX_NAME	SYMPTOMS	AGE_CAT	HEALTH_CONDITION
0	916600	33.0	F			pcn and bee venom	MODERNA	[Dysphagia, Epiglottitis]	20-39 yr	,,pcn and bee venom
1	916601	73.0	F	patient residing at nursing facility. see pati...	patient residing at nursing facility. see pati...	"dairy"	MODERNA	[Anxiety, Dyspnoea]	over_60 yr	patient residing at nursing facility. see pati...
2	916602	23.0	F			shellfish	PFIZER	[Chest discomfort, Dysphagia, Pain in extremit...	20-39 yr	,,shellfish
3	916603	58.0	F	kidney infection	diverticulitis, mitral valve prolapse, osteoar...	diclofec, novacaine, lidocaine, pickles, tomat...	MODERNA	[Dizziness, Fatigue, Mobility decreased]	40-59 yr	kidney infection,diverticulitis, mitral valve ...
4	916604	47.0	F				MODERNA	[Injection site erythema, Injection site pruri...	40-59 yr	,,

- AGE_CAT : we converted the age column from numeric to categorical by grouping them to buckets of “0-2 yr” , “3-12yr”, “13-19yr” , “20-39yr” , “40-59” and “over_60yr”.

- HEALTH_CONDITION: the values in this column are the concatenation of the text features “cur_ill” , “History”, and “Allergies”. Merging these columns allows us to apply data cleanings on one column instead of three.

Columns for training the model:

Features	Data type		Target	Data type
AGE_CAT	Nominal		SYMPTOMS	List [“symp1”,” symp2” ...]
SEX	Nominal			
HEALTH_CONDITION	Free-form text			

To facilitate the K-Means Model to group patients with similar health conditions, we need to extract meaningful keywords related to the patients’ health from the text in the health_condition column. To do so, we performed the following step to clean the data. There are lots of inconsistencies in the text, such as the same word in different forms and misspellings.

1) Tokenize the texts using NLTK.tokenizer(). So that individual words are stored in lists and we can decide which words to keep and remove.
2) Make all words lowercase and remove any punctuations.
3) Keep only the noun words in the texts by using NLTK.pos_tag() to identify the nouns
4) Use TextBlob library to auto-correct the misspellings
5) Remove stops words and words unrelated to patients' health.
6) Normalize the words to their basic form by using NLTK Lemmatization

7) Use difflib.sequenceMatcher to group similar words together, Manually replace words with the same meaning.

```
In [28]: simmilar_words = [{'abnormal': ['abnormal', 'abnormality']}, {'acetaminophen': ['acetaminophen', 'acetiminophen']}, {'acyclovir':
             'hypoglycemia': ['hipoglicemia', 'hypoglycemia']}, {'horse': ['horse', 'horses']}, {'hydronephrosis': ['hydonephrosis', 'hydr

In [29]: simmilar_words2 = [{'hypogammaglobulinemia': ['ypogammaglobuliemia', 'hypogammaglobulinemia', 'ypogammaglobulonemia']}, {'spironc
             'candida', 'candidate', 'candidiasis']}, {'cantalope': ['cantalope', 'cantaloupe', 'canteloupe']}, {'capsaicin': ['capsaicin'

In [30]: simmilar_words3 = [{'fibrosing': ['fibrosing', 'fibrosis']}, {'flaygel': ['flaygel', 'flaygyll']}, {'flexaril': ['flexaril', 'fle
```

After the data cleaning, we were able to reduce the number of keywords from 16361 to 4517.

Separate the data by vaccine :

Our service is designed to allow users to see the respective side effects of the Moderna and Pfizer vaccines. So they can evaluate the risk and determine which vaccine to take. As such, we have to separate the dataset by the two vaccines and train two K-means models.

Split the data for training and testing:

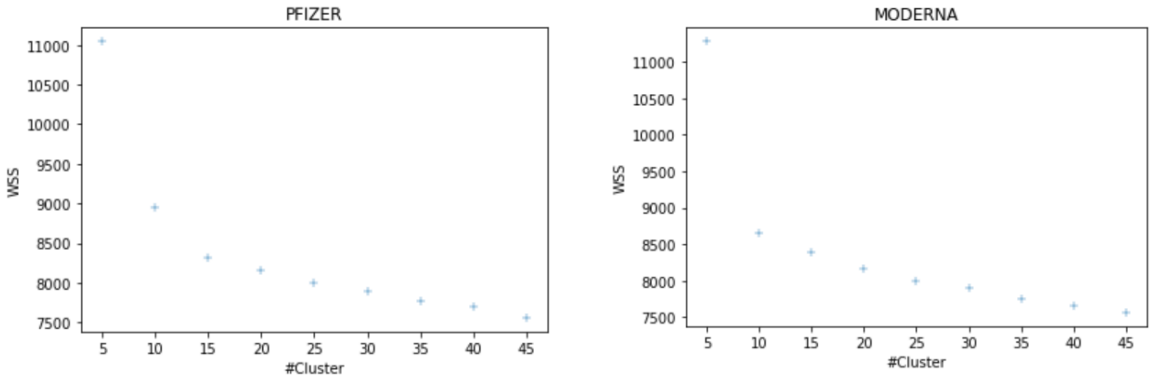
We split the dataset into 95% for training and 5% for testing. The split is stratified by age_cat. So that the testing set has instances from all age categories.

Vectorize the features:

For the K-Means model to calculate the Euclidean distance, we need to vectorize the categorical features. We applied OneHotEncoding to transform the gender and age_cat features to a binary matrix and applied IF-IDF on the health_condition feature to transform the keywords to a term frequency matrix. We then merged the two matrices into one data frame for training the model.

3. Model training

Elbow method:



The Elbow method shows that the optimum number of clusters (K) for Pfizer is 15 and Moderna is 10. However, we found that using such a small number of clusters, the models are only grouping the patients by gender and age categories but not by patient's health conditions.

VAERS_ID	AGE_YRS	SEX	VAX_NAME	SYMPTOMS	AGE_CAT	HEALTH_CONDITION	CLUSTER
1017412	55.0	F	MODERNA	[Adrenal adenoma, COVID-19, Computerised tomog...	40-59 yr	miggraine amox	3
928101	56.0	F	MODERNA	[Back pain, Dyspnoea, Headache, Pyrexia, Respi...	40-59 yr	penicillin laytex	3
919171	57.0	F	MODERNA	[Blood pressure decreased, Dizziness, Electroc...	40-59 yr	hypothridism sulfa drug	3
927997	57.0	F	MODERNA	[Headache, Injection site pain, Myalgia, Retch...	40-59 yr		3
995477	42.0	F	MODERNA	[Limb discomfort, Muscular weakness, Nausea, P...	40-59 yr	morphine proxen shellfish	3
...	...	...	...	...	...	...	...
921620	53.0	F	MODERNA	[Chills, Dyspnoea, Pain, Pyrexia]	40-59 yr	chloesterol	3
919674	50.0	F	MODERNA	[Dizziness, Hot flush, Immediate post-injectio...	40-59 yr	pollen dander	3
1047763	43.0	F	MODERNA	[Abdominal pain upper, Arthralgia, Bone pain, ...	40-59 yr	mast cell activation histamine intolerance epi...	3
916774	50.0	F	MODERNA	[Chills, Pain, Pyrexia]	40-59 yr		3
930097	41.0	F	MODERNA	[Cough, Headache, Hypoaesthesia, Injection sit...	40-59 yr	hypothridism imitrex intolerance	3

By increasing the number of clusters to 100, we start seeing some grouping by health conditions. As you can see below, cluster 83 is grouping the patients who are over 60 , male , and have hypertension. And cluster 84 is grouping the patients who are between the ages of 20 to 29, female and have asthma.

Cluster 83:

VAERS_ID	AGE_YRS	SEX	VAX_NAME	SYMPTOMS	AGE_CAT	HEALTH_CONDITION	CLUSTER
1071396	65.0	M	MODERNA	[Angiogram, Deep vein thrombosis, Echocardiogr...	over_60 yr	hypertenision hyperlidipdemia	83
932817	66.0	M	MODERNA	[Blood test, Macule, Pruritus, SARS-CoV-2 test...	over_60 yr	hypertenision chloesterol penicillin	83
920604	73.0	M	MODERNA	[Diarrhoea, Fatigue, Pyrexia]	over_60 yr	hypertenision obesity	83
1041211	63.0	M	MODERNA	[Alanine aminotransferase increased, Aspartate...	over_60 yr	hypertenision	83
1024627	77.0	M	MODERNA	[Atrial fibrillation, COVID-19, Cough, Death, ...	over_60 yr	dvt popliteal hypertenision hyperlidipdemia	83
1089057	90.0	M	MODERNA	[Computerised tomogram thorax, Deep vein throm...	over_60 yr	hypertenision hyperlidipdemia amox cephalaxin ...	83
1024889	67.0	M	MODERNA	[Injection site erythema, Injection site pain,...	over_60 yr	type hypertenision	83
987374	78.0	M	MODERNA	[Cerebrovascular accident]	over_60 yr	hypertenision hypertenision	83
994502	78.0	M	MODERNA	[Blood test normal, Cardiac function test norm...	over_60 yr	hypertenision afib arteriole	83

Cluster = 94 :

VAERS_ID	AGE_YRS	SEX	VAX_NAME	SYMPTOMS	AGE_CAT	HEALTH_CONDITION	CLUSTER
927863	34.0	F	MODERNA	[Injection site erythema, Injection site pain,...	20-39 yr	aasthma zithromax	94
924342	28.0	F	MODERNA	[Injection site erythema, Injection site pruri...	20-39 yr	aasthma	94
920800	39.0	F	MODERNA	[Injection site erythema, Injection site pain,...	20-39 yr	aasthma gerd psoriasis sulfa	94
1089123	23.0	F	MODERNA	[Blood test normal, Chest X-ray normal, Painfu...	20-39 yr	intensity aasthma	94
1075653	34.0	F	MODERNA	[Asthenia, Chest pain, Computerised tomogram c...	20-39 yr	aasthma	94
924667	26.0	F	MODERNA	[Chills, Myalgia, Night sweats, Pyrexia]	20-39 yr	aasthma	94
969685	34.0	F	MODERNA	[Arthralgia, Hypoaesthesia, Lethargy, Nausea, ...	20-39 yr	aasthma neosporin bee	94
926265	22.0	F	MODERNA	[SARS-CoV-2 test negative, Streptococcus test ...	20-39 yr	aasthma	94
930781	24.0	F	MODERNA	[Chills, Headache, Pain, Pyrexia, SARS-CoV-2 t...	20-39 yr	aasthma	94
931248	39.0	F	MODERNA	[Hypoaesthesia, Palpitations]	20-39 yr	aasthma aasthma	94
1046614	36.0	F	MODERNA	[Inflammation, Injection site erythema, Inject...	20-39 yr	aasthma aasthma	94

Cluster 83 - top 10 vaccine symptoms

	Frequency
Death	17
Pyrexia	15
Vomiting	8
Fatigue	8
Headache	7
Myalgia	7
Pain in extremity	7
Dyspnoea	7
Pain	7
Injection site erythema	6

Cluster 94 - top 10 vaccine symptoms

	Frequency
Injection site pain	17
Headache	17
Pain in extremity	14
Pyrexia	13
Injection site erythema	13
Nausea	12
Chills	12
Injection site pruritus	10
Pain	10
Injection site swelling	10

V. Validation and interpretation

1. Model evaluation

We applied the following steps on the test set (1700 samples) to evaluate the accuracy of our models.

1) Use the same OneHotEncoding and the TF-IDF object we have used on the training set to vectorize the test data.
2) Use the models' predict function to get the cluster labels for each instance in the test set
3) Base on the cluster labels, get the top 10 vaccine symptoms from the same cluster in the training set
4) For each instance in the test set, compare the actual symptom with the top 10 symptoms.
5) If 3 out of 4 symptoms are matching with the top 10 symptoms. The prediction is 75% accurate.
6) By averaging the prediction accuracies in the test set give us the overall prediction accuracy of the models.

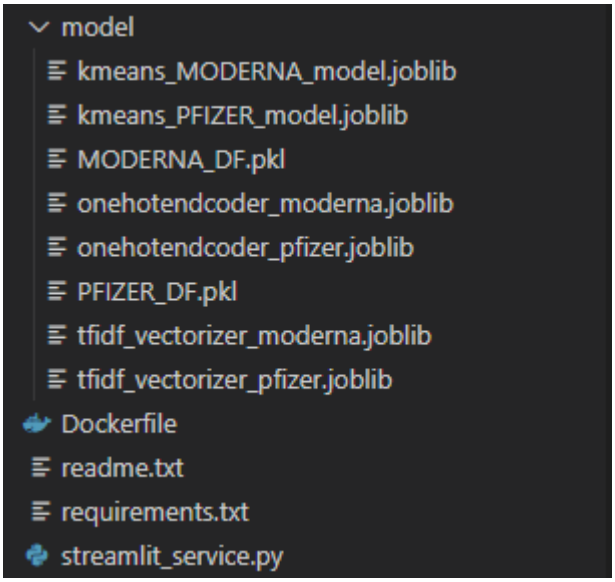
AGE_CAT	SEX	VAX_NAME	AGE_CAT	HEALTH_CONDITION	SYMPTOMS	PREDICTION	MATCHING	TOP10_SYMP
40-59 yr	F	PFIZER	40-59 yr	cough congestion	[Dizziness, Vertigo]	3	0.500000	[Headache, Pyrexia, Chills, Fatigue, Pain, Nau...
40-59 yr	F	PFIZER	40-59 yr	incisor root resorption diabertes incontinence...	[Activated partial thromboplastin time, Blood ...	60	0.000000	[Headache, Injection site pain, Pain, Fatigue,...
20-39 yr	M	PFIZER	20-39 yr	feline	[Arthralgia, Lymphadenitis, Malaise]	6	0.000000	[Headache, Pyrexia, Chills, Fatigue, Pain, Mya...
unknown_age	F	PFIZER	unknown_age		[COVID-19, Feeling abnormal, Malaise, SARS-CoV...	4	0.500000	[SARS-CoV-2 test positive, COVID-19, Headache,...
over_60 yr	M	PFIZER	over_60 yr		[Arthralgia, Vaccination site pain]	0	0.000000	[Death, Pyrexia, Headache, Fatigue, Dyspnoea, ...
over_60 yr	M	PFIZER	over_60 yr	comment	[Facial paralysis]	55	0.000000	[Pyrexia, Chills, Body temperature, Pain, Head...
20-39 yr	M	PFIZER	20-39 yr	seasol alergies	[Arthralgia, Chills, Dizziness, Headache, Myal...	94	0.571429	[Headache, Chills, Pyrexia, Fatigue, Pain, Pai...

We ran the models using cluster (k) = 100 , 150 ,and 200, the accuracy of our model stays around 30%.

k = 100	Moderna	0.3234430636
	Pfizer	0.3162084521
k = 150	Moderna	0.3242052988
	Pfizer	0.3169193292
k = 200	Moderna	0.319561228
	Pfizer	0.3114315175

2. Service and User Interaction

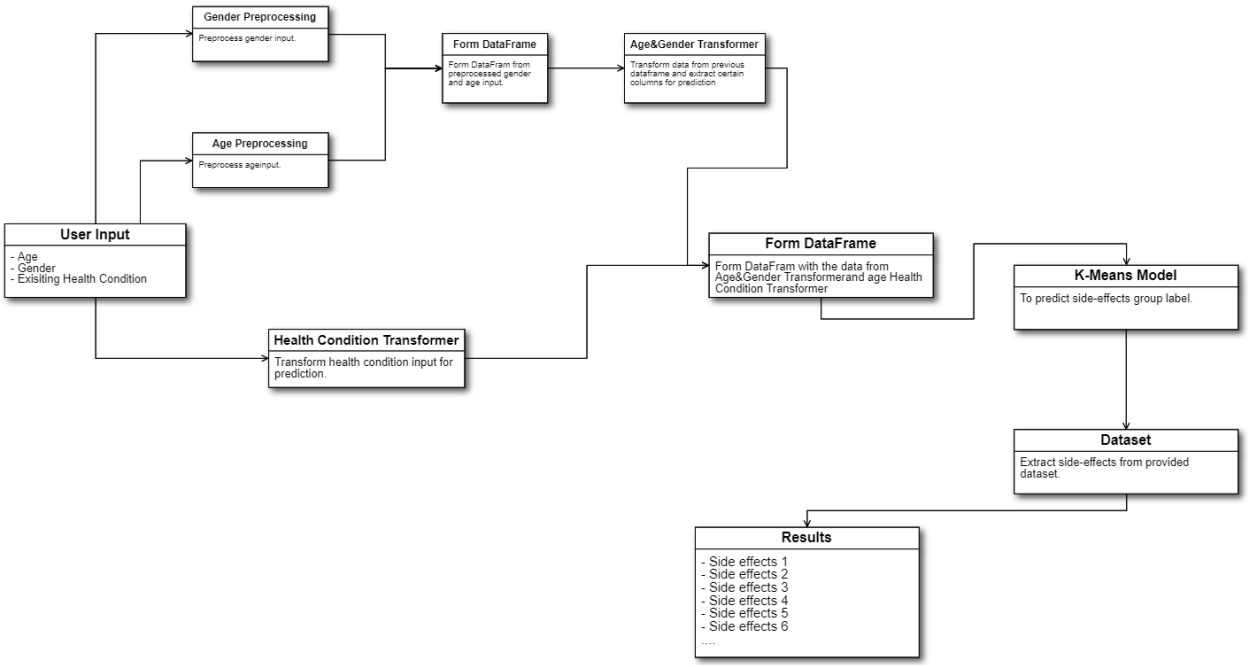
1) Service Architecture and Flow.



As can be seen in the above figure, the architecture of this service is clear. With transformers “onehotendcoder_moderna.joblib” and “onehotendcoder_pfizer.joblib” to process age input, “tfidf_vectorizer_pfizer.joblib” and “tfidf_vectorizer_moderna.joblib” to process existing health condition input for different vaccine brand.

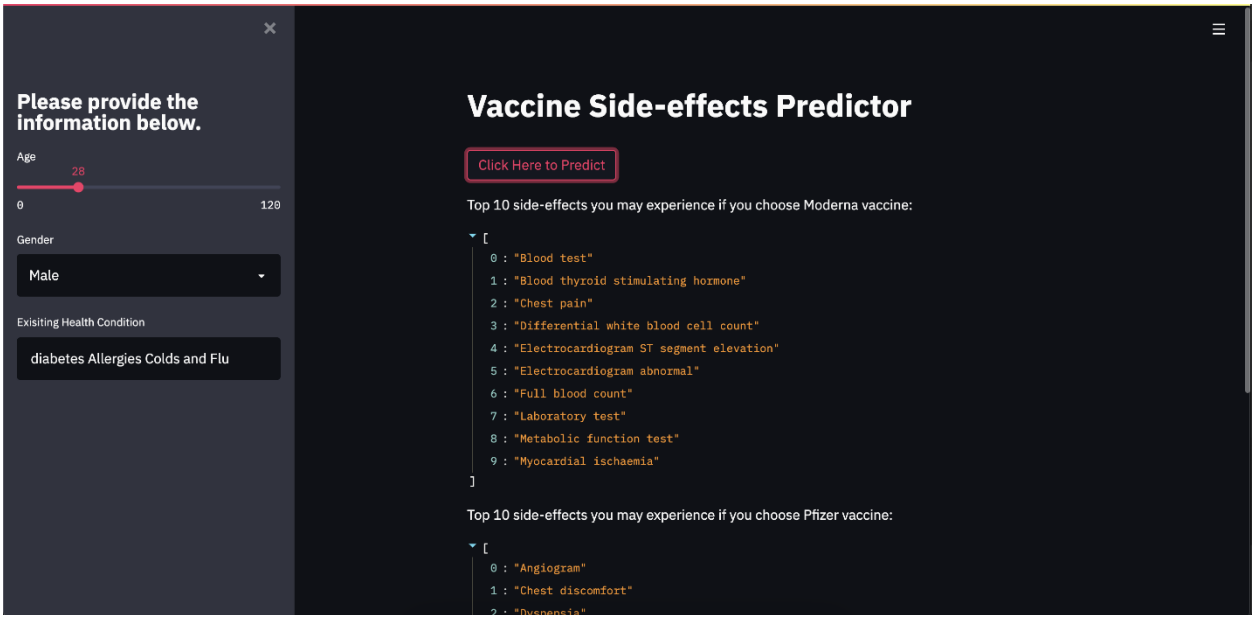
With models “kmeans_PFIZER_model.joblib” and “kmeans_MODERNA_model.joblib”, preprocessed and transformed input can be used to predict results. Then with “MODERNA_DF.pkl” and “PFIZER_DF.pkl” we can use predicted results to extract side-effects groups.

Specific flow:



2) User Interface and User Interaction.

The models and transformers from Model part are used in the service. The service uses Streamlit as the framework and create a single page application for models to transform inputs from users and predict results.



As can be seen in above image, there is a sidebar of the application. With three inputs, Age(slider), Gender(dropdown), and Existing Health Condition(string box) for user to input data. And there is a prediction button on the main interface. By clicking “Click Here to Predict” to get side-effects results.

VI. Knowledge

- Benefits of our solution provides:

Our service can help people make decisions if they should get COVID-19 vaccines and which vaccine is better for them by providing them the possible symptoms after getting COVID-19 vaccines. Many people can benefit from this as COVID-19 vaccines are now available for all people in the U.S. and they can prepare for possible side effects they may have.

- Final thoughts:

Although this is not a real medical research, it is a good practice using data science to analyze COVID-19 vaccines. The biggest challenge for our data modeling is how we delt with categorical data, such as extracting meaningful words from a text, validating mis-spelling words, and converting strings into numerical data so that we can apply it into clustering models. If we can have a well-recorded health conditions and symptoms, our prediction will be more accurate. In addition, there are only data with patients who had side effects in the dataset we used, not including people who didn't have symptoms after getting COVID-19. If we can find more data point that also include people who don't have side effects, our model will be better.