COVID Long Haulers
Final Submission
BMI 6016-001 Biomed Data Wrangling

Shad Morton, shad.morton@utah.edu, u1328277
Logan Greydanus, u1324767@utah.edu, u1324767
John Arnn, john.arnn@utah.edu, u1328576
Kevin Yang, kevin.yang@hsc.utah.edu, u0630183

https://github.com/jwarnn/BMI-6016-Group-Covid-Long-Haulers

The Wrangling Problem

The novelty of SARS-CoV-2 and the wide range of reported symptoms makes it difficult to predict the long-term complications that may follow the short-term effects. In the future it will be crucial to determine which patients are at risk of developing these complications and to understand the underlying mechanisms that trigger their development. This information will help identify at-risk populations, triaging patients and developing new care plans as variants are discovered. In this approach; population health, clinical and translational informaticians will need to work together to parse and apply these new data. Unfortunately the long-term complications and symptoms of SARS-CoV-2 are not well defined in the medical literature which makes identifying meaningful data from health records beyond the capabilities of this project. Instead we will use the medical records of patients who have had confirmed cases of SARS-CoV-2 to create a dataset to be used for a clustering analysis to explore what symptoms 14 days after that diagnosis are related to which demographics, previous diagnoses and medication use. We had to harmonize, integrate, and transform large datasets in order to accomplish our goal.

Data quality assessment (descriptive statistics) of the dataset(s). Use graphs were applicable

The primary goal of the medications dataset was to create a binary array using the codes attached to the patient ids. So the assessment done on the medications dataset was to check for 'Unknown' values and NaN values:

```
df[df['code'].str.contains('Unknown')]
```

patient_id	encounter_id	code_system	code	start_date	route	brand	strength	derived_by_TriNetX

It appears there are no 'Unknown' code values.

```
print('any NaN values?', df.isnull().values.any())
```

```
any NaN values? True
```

There are NaN values that need to be dropped with df.dropna().

The patient csv contained seven rows of data for > 1.6 million unique patient ids.

In [7]:
```
# read in the patient csv
df3 = pd.read_csv('patient.csv')
```

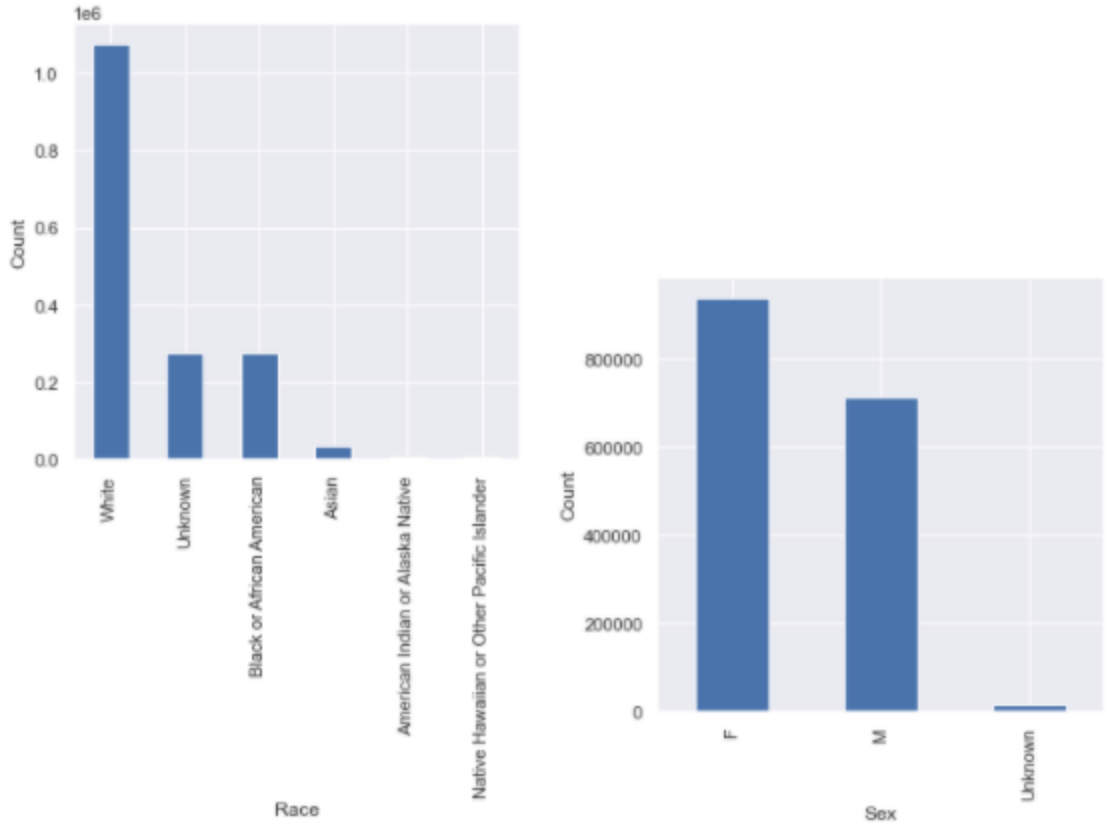
In [8]:
```
# basic stats
df3.describe()
```

Out[8]:

	year_of_birth	age_at_death	postal_code
count	1.655079e+06	21412.000000	0.0
mean	1.974726e+03	66.624977	NaN
std	2.149369e+01	17.767804	NaN
min	1.929000e+03	0.000000	NaN
25%	1.957000e+03	58.000000	NaN
50%	1.974000e+03	69.000000	NaN
75%	1.991000e+03	80.000000	NaN
max	2.020000e+03	90.000000	NaN

We see that we have over 1.6 million patients!
The mean age is (2021 - 1974 =) 47, with a standard deviation of 21 years.
The people who died, usually did at 66, but we have patients who have died as early as 58 or as late as 90.
Unfortunately, the postal code column is empty, so we wont be able to do any regional analysis :(

After dropping postal code, we had six rows of demographic information: id, sex, race, ethnicity, year of birth and age at death. Of these columns, only id is complete (no "unknown").



The diagnosis csv had 324,179,748 rows and six columns of data. Two of those columns, 'encounter_id and 'derived_by_TriNetX', did not provide any relevant information for the aims of the project and were dropped from the dataset. 'Patient_id', 'code_system', 'code' and 'date' are what remained. To access the data quality a few elements were explored including the agreement of between datasets common data elements like patient ids, and codes. These were in perfect agreement between one another. We also made sure the range of 'date' was logical for the dataset. 'Code_system' had only two unique values, 'ICD-10-CM' and 'ICD-9-CM'; which could be found in the 'terminology.csv'. Next the presence of null values in the data set was performed with only two rows being found. These were dropped because the column 'code' was the missing value representing the most critical information in the dataset. Finally duplicate rows were dropped from this dataset using the pandas duplicate function where nearly 10% of the rows were dropped. Because this was a significant number a manual check of about ten of these rows was conducted which found they were duplicates across all columns. With these steps taken we found that the diagnosis per patient had a mean of 163.78, median of 41, minimum of 1, and a maximum of 69,238. We were also able to produce a list of the most common codes:

```
Out[54]: {'E78.5': ['Hyperlipidemia, unspecified'],
          'E11.9': ['Type 2 diabetes mellitus without complications'],
          'K21.9': ['Gastro-esophageal reflux disease without esophagitis'],
          'F32.9': ['Major depressive disorder, single episode, unspecified'],
          'Z79.899': ['Other long term (current) drug therapy'],
          'I25.10': ['Atherosclerotic heart disease of native coronary artery without angina pectoris'],
          'E03.9': ['Hypothyroidism, unspecified'],
          'F41.9': ['Anxiety disorder, unspecified'],
          '250.00': ['Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrol
         led'],
          'G89.29': ['Other chronic pain'],
          'Z00.00': ['Encounter for general adult medical examination without abnormal findings'],
          'R05': ['Cough'],
          'M54.5': ['Low back pain'],
          'G47.33': ['Obstructive sleep apnea (adult) (pediatric)'],
          'J45.909': ['Unspecified asthma, uncomplicated'],
          'E66.9': ['Obesity, unspecified'],
          'D64.9': ['Anemia, unspecified'],
          'I48.91': ['Unspecified atrial fibrillation'],
          'R07.9': ['Chest pain, unspecified'],
```

Steps taken to wrangle (and integrate) the data.

The lab results csv was very large (over 630 million rows) and difficult to wrangle. The file was over 80 GB and contained eight rows; id, encounter id, code_system, code, date, the number value of a lab result (if applicable) and a text value of a lab result (if applicable). It was not feasible to load it in as a pandas dataframe on my computer to run a complete data assessment, so it was chunked into a smaller dataframe containing only "positive" text results in the text value row. After that dataframe was made, it was filtered for 17 COVID specific LOINC codes to make the final dataset with 181,950 positive COVID test results.

A similar strategy was used to wrangle the diagnosis csv. This csv, containing over 300 million rows was chunked into a smaller dataframe by first filtering for five COVID specific diagnoses. After filtering, this dataframe had 340,563 COVID diagnoses. This dataframe was then appended to the filtered lab results dataframe and sorted by date (descending). The tail end of the dataset had some older dates, so we decided to drop all the entries with a COVID diagnosis or positive lab result before 2019.

	patient_id	encounter_id	code_system	code	date
211377	b3990d354fb8caf63d269d33978152cd6a456991	526ccc470b713bb98558beb1303724137e36edc0	ICD-10-CM	U07.1	20040331
211392	b3990d354fb8caf63d269d33978152cd6a456991	526ccc470b713bb98558beb1303724137e36edc0	ICD-10-CM	U07.1	20040331
211375	b3990d354fb8caf63d269d33978152cd6a456991	526ccc470b713bb98558beb1303724137e36edc0	ICD-10-CM	U07.1	20040331
211395	b3990d354fb8caf63d269d33978152cd6a456991	526ccc470b713bb98558beb1303724137e36edc0	ICD-10-CM	U07.1	20040331
211391	b3990d354fb8caf63d269d33978152cd6a456991	526ccc470b713bb98558beb1303724137e36edc0	ICD-10-CM	U07.1	20040331

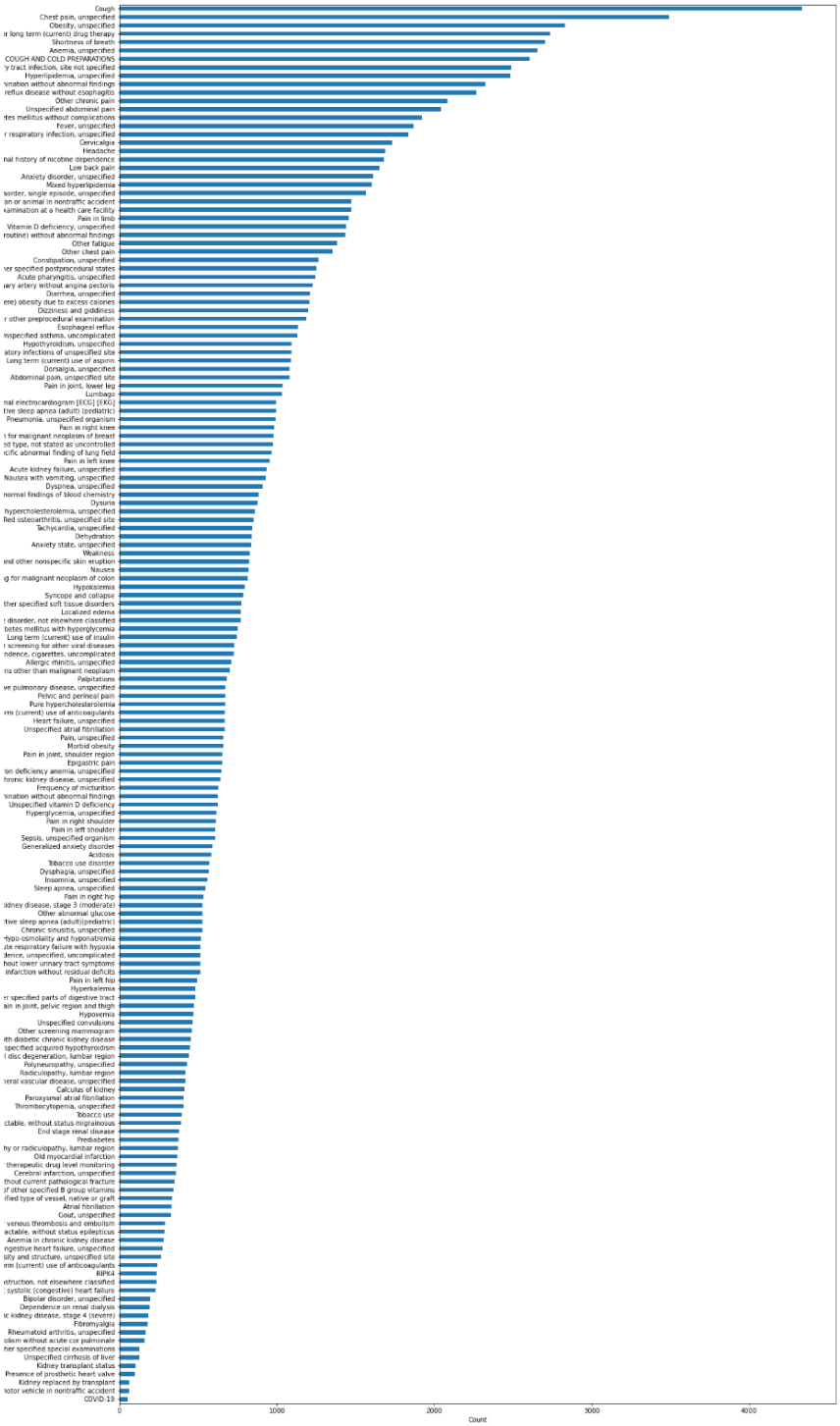
After these entries were dropped, we dropped duplicate patient ids keeping the only the first entry. This gave us a dataset with 189,205 patients with COVID which was merged to the patient csv to make the master file.

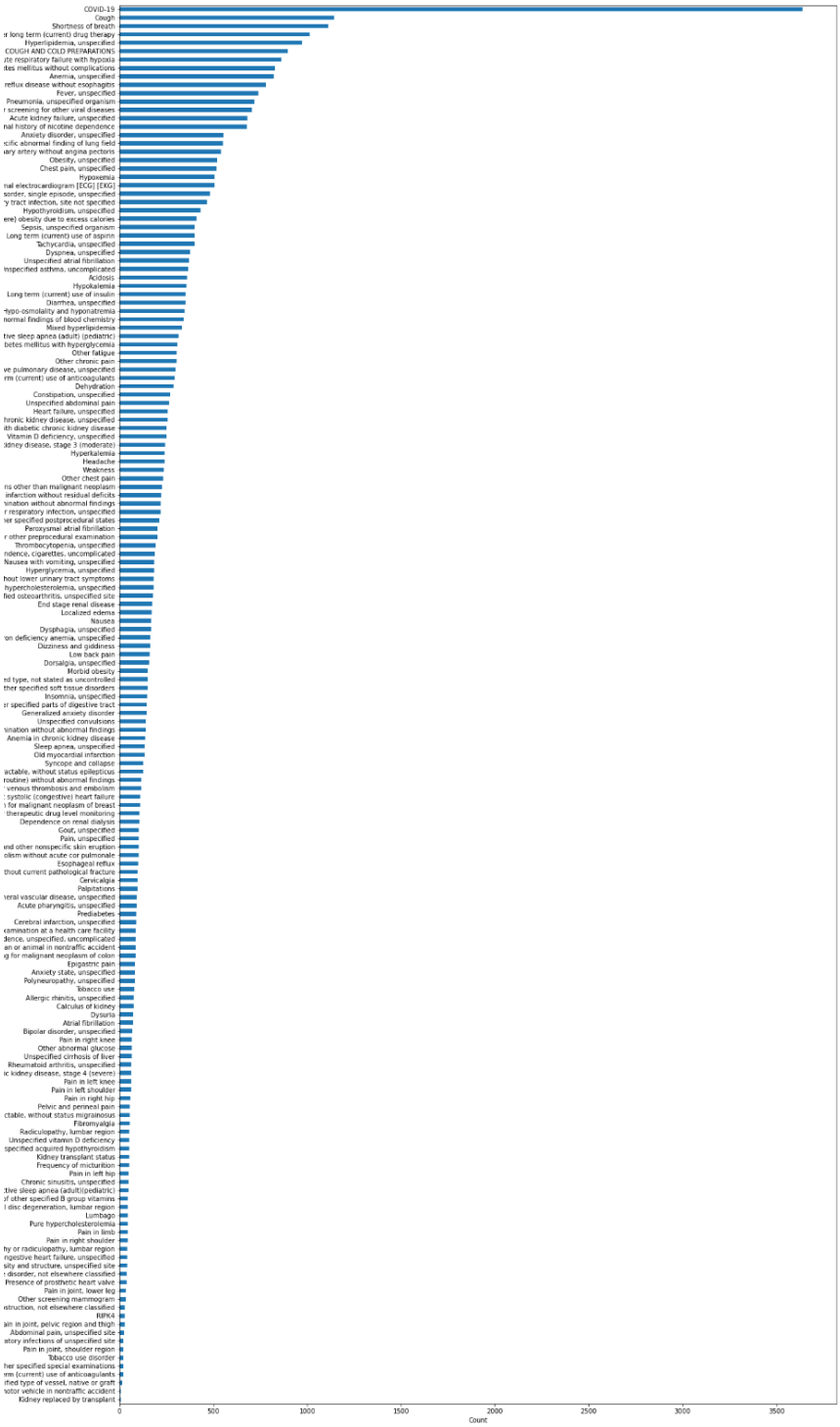
With the master file in hand, we were able to split the diagnosis and medication datasets by pre-COVID and (14 day) post-COVID datasets. For pre-COVID datasets we filtered on "date" according to a patient_id, date tuple and then the top 200 diagnoses

codes. For the post-COVID datasets, we converted the integer date field to a date type using pandas, added 14 days, then reconverted to integer in concordance with the other data files. Again we filtered based on a patient_id, 14_date tuple and then the top 200 diagnoses. These are the dataframes that we turned into a boolean array.

Repeat data quality assessment of the wrangled data

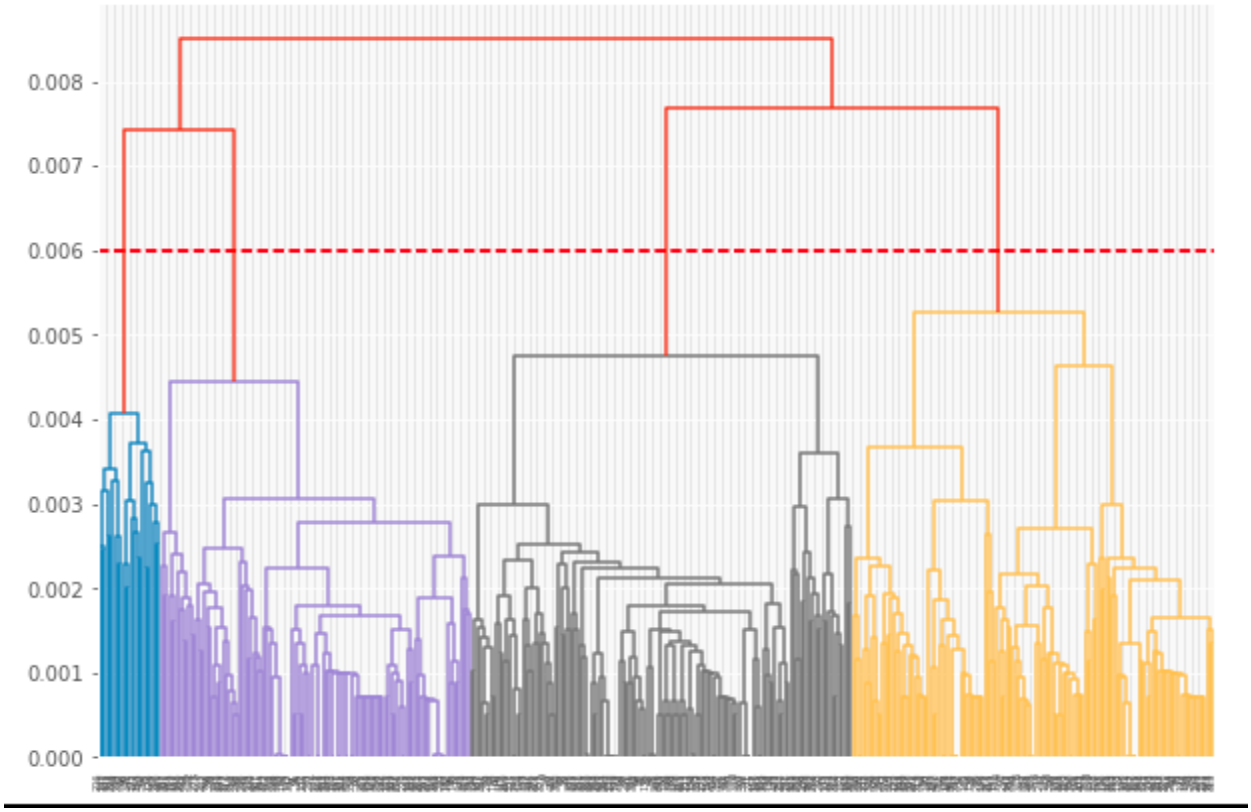
To assess the quality of the wrangled data, before we converted the the diagnosis dataset into a boolean array we needed to discover if the post-COVID dataframe was different from the pre-COVID dataframe and whether or not it contained “long-haul COVID” diagnoses. To this end we were supplied a paper by a group of researchers in Sweden who conducted a study of long-haul COVID in groups of health care workers with mild cases of COVID-19. In their study, they reported; anosmia, ageusia, fatigue and dyspnea as the most common effects[3]. In our dataset we discovered fatigue and dyspnea in our post-COVID data set as well as many other interesting long-haul candidates.


Cough
Chest pain, unspecified
Obesity, unspecified
r long term (current) drug therapy
Shortness of breath
Anemia, unspecified
COUGH AND COLD PREPARATIONS
y tract infection, site not specified
Hyperlipidemia, unspecified
ination without abnormal findings
reflux disease without esophagitis
Other chronic pain
Unspecified abdominal pain
tes mellitus without complications
Fever, unspecified
r respiratory infection, unspecified
Cervicalgia
Headache
nal history of nicotine dependence
Low back pain
Anxiety disorder, unspecified
Mixed hyperlipidemia
sorder, single episode, unspecified
an or animal in nontraffic accident
amination at a health care facility
Pain in limb
Vitamin D deficiency, unspecified
routine) without abnormal findings
Other fatigue
Other chest pain
Constipation, unspecified
er specified postprocedural states
Acute pharyngitis, unspecified
ary artery without angina pectoris
Diarrhea, unspecified
ere) obesity due to excess calories
Dizziness and giddiness
r other preprocedural examination
Esophageal reflux
nspecified asthma, uncomplicated
Hypothyroidism, unspecified
atory infections of unspecified site
Long term (current) use of aspirin
Dorsalgia, unspecified
Abdominal pain, unspecified site
Pain in joint, lower leg
Lumbago
nal electrocardiogram [ECG] [EKG]
tive sleep apnea (adult) (pediatric)
Pneumonia, unspecified organism
Pain in right knee
for malignant neoplasm of breast
d type, not stated as uncontrolled
cific abnormal finding of lung field
Pain in left knee
Acute kidney failure, unspecified
Nausea with vomiting, unspecified
Dyspnea, unspecified
normal findings of blood chemistry
Dysuria
hypercholesterolemia, unspecified
fied osteoarthritis, unspecified site
Tachycardia, unspecified
Dehydration
Anxiety state, unspecified
Weakness
nd other nonspecific skin eruption
Nausea
g for malignant neoplasm of colon
Hypokalemia
Syncope and collapse
ther specified soft tissue disorders
Localized edema
disorder, not elsewhere classified
betes mellitus with hyperglycemia
Long term (current) use of insulin
r screening for other viral diseases
ndence, cigarettes, uncomplicated
Allergic rhinitis, unspecified
ns other than malignant neoplasm
Palpitations
ve pulmonary disease, unspecified
Pelvic and perineal pain
Pure hypercholesterolemia
rm (current) use of anticoagulants
Heart failure, unspecified
Unspecified atrial fibrillation
Pain, unspecified
Morbid obesity
Pain in joint, shoulder region
Epigastric pain
on deficiency anemia, unspecified
hronic kidney disease, unspecified
Frequency of micturition
ination without abnormal findings
Unspecified vitamin D deficiency
Hyperglycemia, unspecified
Pain in right shoulder
Pain in left shoulder
Sepsis, unspecified organism
Generalized anxiety disorder
Acidosis
Tobacco use disorder
Dysphagia, unspecified
Insomnia, unspecified
Sleep apnea, unspecified
Pain in right hip
idney disease, stage 3 (moderate)
Other abnormal glucose
tive sleep apnea (adult)(pediatric)
Chronic sinusitis, unspecified
hypo-osmolality and hyponatremia
ite respiratory failure with hypoxia
dence, unspecified, uncomplicated
hout lower urinary tract symptoms
infarction without residual deficits
Pain in left hip
Hyperkalemia
r specified parts of digestive tract
sin in joint, pelvic region and thigh
Hypoxemia
Unspecified convulsions
Other screening mammogram
th diabetic chronic kidney disease
specified acquired hypothyroidism
l disc degeneration, lumbar region
Polyneuropathy, unspecified
Radiculopathy, lumbar region
eral vascular disease, unspecified
Calculus of kidney
Paroxysmal atrial fibrillation
Thrombocytopenia, unspecified
Tobacco use
clable, without status migrainosus
End stage renal disease
Prediabetes
hy or radiculopathy, lumbar region
Old myocardial infarction
therapeutic drug level monitoring
Cerebral infarction, unspecified
thout current pathological fracture
of other specified B group vitamins
ified type of vessel, native or graft
Atrial fibrillation
Gout, unspecified
venous thrombosis and embolism
ectable, without status epilepticus
Anemia in chronic kidney disease
ngestive heart failure, unspecified
sity and structure, unspecified site
rm (current) use of anticoagulants
RIPK4
struction, not elsewhere classified
systolic (congestive) heart failure
Bipolar disorder, unspecified
Dependence on renal dialysis
ic kidney disease, stage 4 (severe)
Fibromyalgia
Rheumatoid arthritis, unspecified
olism without acute cor pulmonale
her specified special examinations
Unspecified cirrhosis of liver
Kidney transplant status
Presence of prosthetic heart valve
Kidney replaced by transplant
otor vehicle in nontraffic accident
COVID-19
0
1000
2000
3000
4000
Count


COVID-19
Cough
Shortness of breath
r long term (current) drug therapy
Hyperlipidemia, unspecified
COUGH AND COLD PREPARATIONS
ite respiratory failure with hypoxia
des mellitus without complications
Anemia, unspecified
reflux disease without esophagitis
Fever, unspecified
Pneumonia, unspecified organism
r screening for other viral diseases
Acute kidney failure, unspecified
nal history of nicotine dependence
Anxiety disorder, unspecified
cific abnormal finding of lung field
ary artery without angina pectoris
Obesity, unspecified
Chest pain, unspecified
Hypoxemia
nal electrocardiogram [ECG] [EKG]
sorder, single episode, unspecified
y tract infection, site not specified
Hypothyroidism, unspecified
ene) obesity due to excess calories
Sepsis, unspecified organism
Long term (current) use of aspirin
Tachycardia, unspecified
Dyspnea, unspecified
Unspecified atrial fibrillation
nspecified asthma, uncomplicated
Acidosis
Hypokalemia
Long term (current) use of insulin
Diarrhea, unspecified
-hypo-osmolality and hyponatremia
normal findings of blood chemistry
Mixed hyperlipidemia
tive sleep apnea (adult) (pediatric)
betes mellitus with hyperglycemia
Other fatigue
Other chronic pain
ve pulmonary disease, unspecified
rm (current) use of anticoagulants
Dehydration
Constipation, unspecified
Unspecified abdominal pain
Heart failure, unspecified
hronic kidney disease, unspecified
th diabetic chronic kidney disease
Vitamin D deficiency, unspecified
idney disease, stage 3 (moderate)
Hyperkalemia
Headache
Weakness
Other chest pain
ns other than malignant neoplasm
infarction without residual deficits
nination without abnormal findings
r respiratory infection, unspecified
er specified postprocedural states
Paroxysmal atrial fibrillation
r other preprocedural examination
Thrombocytopenia, unspecified
ndence, cigarettes, uncomplicated
Nausea with vomiting, unspecified
Hyperglycemia, unspecified
hout lower urinary tract symptoms
hypercholesterolemia, unspecified
fied osteoarthritis, unspecified site
End stage renal disease
Localized edema
Nausea
Dysphagia, unspecified
on deficiency anemia, unspecified
Dizziness and giddiness
Low back pain
Dorsalgia, unspecified
Morbid obesity
d type, not stated as uncontrolled
ther specified soft tissue disorders
Insomnia, unspecified
r specified parts of digestive tract
Generalized anxiety disorder
Unspecified convulsions
nination without abnormal findings
Anemia in chronic kidney disease
Sleep apnea, unspecified
Old myocardial infarction
Syncope and collapse
actable, without status epilepticus
routine) without abnormal findings
venous thrombosis and embolism
systolic (congestive) heart failure
for malignant neoplasm of breast
therapeutic drug level monitoring
Dependence on renal dialysis
Gout, unspecified
Pain, unspecified
nd other nonspecific skin eruption
olism without acute cor pulmonale
Esophageal reflux
thout current pathological fracture
Cervicalgia
Palpitations
veral vascular disease, unspecified
Acute pharyngitis, unspecified
Prediabetes
Cerebral infarction, unspecified
amination at a health care facility
dence, unspecified, uncomplicated
an or animal in nontraffic accident
g for malignant neoplasm of colon
Epigastric pain
Anxiety state, unspecified
Polyneuropathy, unspecified
Tobacco use
Allergic rhinitis, unspecified
Calculus of kidney
Dysuria
Atrial fibrillation
Bipolar disorder, unspecified
Pain in right knee
Other abnormal glucose
Unspecified cirrhosis of liver
Rheumatoid arthritis, unspecified
ic kidney disease, stage 4 (severe)
Pain in left knee
Pain in left shoulder
Pain in right hip
Pelvic and perineal pain
ctable, without status migrainosus
Fibromyalgia
Radiculopathy, lumbar region
Unspecified vitamin D deficiency
specified acquired hypothyroidism
Kidney transplant status
Frequency of micturition
Pain in left hip
Chronic sinusitis, unspecified
tive sleep apnea (adult)(pediatric)
of other specified B group vitamins
l disc degeneration, lumbar region
Lumbago
Pure hypercholesterolemia
Pain in limb
Pain in right shoulder
hy or radiculopathy, lumbar region
ngestive heart failure, unspecified
sity and structure, unspecified site
disorder, not elsewhere classified
Presence of prosthetic heart valve
Pain in joint, lower leg
Other screening mammogram
struction, not elsewhere classified
BMI 40
ain in joint, pelvic region and thigh
Abdominal pain, unspecified site
atory infections of unspecified site
Pain in joint, shoulder region
Tobacco use disorder
her specified special examinations
rm (current) use of anticoagulants
ified type of vessel, native or graft
otor vehicle in nontraffic accident
Kidney replaced by transplant
0
500
1000
1500
2000
2500
3000
3500
Count

Any evidence to show that the wrangled data is usable such early clustering results, values for surgical measures or descriptive statistics of pre and post covid diagnoses/medication

In order to see if a clustering analysis was possible on the data set, a dendrogram plot was created in order to visualize clusters within the dataset. To begin, the csv file containing the boolean array for diagnosis codes along with patient demographics was read into a pandas dataframe. Certain categorical data columns were dropped and dummy data (i.e. columns with 0 or 1) were created for the "sex" and "race" columns. The data was normalized using the sklearn preprocessing library function "normalize". The dendrogram was created using the "dendrogram" function from the scipy.cluster.hierarchy library.



The results of creating the dendrogram can be seen above. This shows that using the diagnosis codes and demographic data the patients can be separated into many groups. The largest vertical distance is often used in dendrograms to represent the largest amount of difference present between clusters. Using this method here, the patients would be separated into four clusters.

Lessons Learned

The major challenge with this project was learning how to handle large datasets, these were anywhere from 50-128gb in size! We handled this by chunking our datasets into smaller files. For example, the medications dataset was chunked into 88 smaller files with each having 10^7 rows at most. For this project, python in the jupyter notebook was the primary tool but with how long it would take to filter a pandas dataframe, we tried to use a select statement in SQLite. However it ended it up being roughly 10% slower. A major lesson we all learned was time and task management. This is the first time all of us had to manage large datasets before so even simpler tasks would take a while to run. And if we were to get too greedy, it would take hours to see the wrong results, so with task management, the code should be tested on the smallest amount possible (chunk of chunks).

References

1. Rando, Halie M et al. “Challenges in defining Long COVID: Striking differences across literature, Electronic Health Records, and patient-reported information.” *medRxiv : the preprint server for health sciences* 2021.03.20.21253896. 26 Mar. 2021, doi:10.1101/2021.03.20.21253896. Preprint.
2. Huang, Chaolin, et al."6-month Consequences of COVID-19 in Patients Discharged from Hospital: A Cohort Study." *The Lancet (British Edition)* 397.10270 (2021): 220-32. Web
3. Havervall, Sebastian, Rosell, Axel, Phillipson, Mia, Mangsbo, Sara M, Nilsson, Peter, Hober, Sophia, and Thålin, Charlotte. "Symptoms and Functional Impairment Assessed 8 Months After Mild COVID-19 Among Health Care Workers." *JAMA : The Journal of the American Medical Association* (2021): JAMA : the Journal of the American Medical Association, 2021-04-07. Web.
4. Haodi, Grigorios, Robert, et al. “Clustering datasets with demographics and diagnosis codes” Journal of Biomedical Informatics Volume 102, February 2020, 103360
5. Sudre, C.H., Murray, B., Varsavsky, T. et al. Attributes and predictors of long COVID. Nat Med 27, 626–631 (2021). https://doi.org/10.1038/s41591-021-01292-y