Below is the table resulting from data cleaning and filtering:

This table/dictionary is also available on our GitHub repository as a .txt file.

There were 55 patients total, and an average of 6.236363636363 diagnoses per patient.

For this project, I used Python methods and data structures in a Jupyter Notebook via Visual Studio Code to clean and work with our data. First, I created a list that represented our directory of files (a folder in our GitHub repository that contained all the text files). Then I created a function to read through every file and create a dictionary with key value pairs. The keys were the file names, and the values were lists of strings (each string was a line from a file). I also wrote a function that filtered the data, so it got rid of all the newlines in the strings.

Then, I wrote a function that would find the diagnoses by finding them between two given strings. This gave me a dictionary with the keys being the file names and the values being the diagnoses in each file. After doing some more filtering and converting the dictionary to a list, I ended up with a dictionary with the keys being the unique diagnoses and the values being the number of times those diagnoses appeared.

Throughout the code file I called these functions on my data structures.

I did not use generative AI to write any code- I only used it to explain certain list and dictionary methods.

The work was split up this way:

Sheamin: all the coding and data/file processing

Adora: Both written patient descriptions

{'seizure': 1,

'metastatic melanoma': 1,

'bacteremia': 2,

'hypertension': 3,

'tracheobronchomalacia': 1,

'copd': 2,

'diabetes type ii': 2,

'obstructive [\*\*doctor last name \*\*] apnea': 1,

'acute myeloid leukemia (in remission)': 1,

'end state cirrhosis, cryptogenic': 1,

'renal failure': 2,

'upper gi bleed': 1,

'type 2 diabetes mellitus': 1,

'ovarian cancer': 1,

'large bowel obstruction': 1,

'postoperative fever': 1,

'coronary artery disease': 2,

'peripheral vascular disease with non healing left toe ulcer and ': 1,

'osteomyelitis': 1,

'aortic stenosis ': 1,

'dementia ': 1,

'diabetes mellitus ': 1,

'paroxysmal atrial fibrillation': 1,

'1. liver abscesses': 1,

'2. hilar mass-pathology pending': 1,

'3. copd flare': 1,

'cholangitis': 1,

'sepsis': 2,

'severe mitral regurgitation with congestive heart failure': 1,

'pneumonia': 4,

'chronic pulmomary obstructive disease': 1,

'pt expired on [\*\*3199-4-19\*\*] of cardiac arrest after being made cmo.': 1,

'meningioma': 1,

'epidural hematoma': 1,

'cerebral edema': 1,

'aspiration pneumonia ': 1,

"parkinson's disease ": 1,

'c-diff colitis ': 1,

'sepsis ': 1,

'uti ': 1,

'malnutrition, severe': 1,

'deceased': 2,

'1. acute myocardial infarction': 1,

'2. hypertension': 1,

'3. altered mental status': 1,

'anemia': 2,

'upper gi bleed.': 1,

'expired due to hypotension and shock in spite of aggressive ': 1,

'care. thought due to acute anteroseptal myocardial infarction ': 1,

'after orthopedic surgery at an outside hosptial.': 1,

'secondary: ': 4,

'coronary artery disease with past angioplasty x2, congestive ': 1,

'heart failure, paroxysmal atrial fibrillation, s/p aortic valve ': 1,

'replacement and mitral valve repair for rheumatic disease, ': 1,

'chronic renal insufficiency and acute renal failure, paroxysmal ': 1,

'atrial fibrillation (retroperitoneal bleed on coumadin [\*\*10-21\*\*]), ': 1,

'chronic anemia, copd, rheumatoid arthritis, lacunar infarct, ': 1,

'cortical atrophy, facial droop, peptic ulcer disease [\*\*2-20\*\*] nsaids, ': 1,

'diverticulosis, short bowel syndrome, neuropathy, recurrent ': 1,

'uti/pyelonephritis caused by serratia and klebsiella, s/p ': 1,

'colectomy 2003 for bowel ischemia, s/p ileostomy for recurrent ': 1,

'lower gi bleeding, s/p bilateral total knee replacement c/b ': 1,

'recurrent infection of the right knee (s/p incision and drainage ': 1,

'[\*\*5-21\*\*] for infection with klebsiella, proteus, e.coli), ': 1,

'degenerative disc diasease, s/p appendectomy, s/p ': 1,

'cholecystectomy, s/p hysterectomy, s/p tracheostomy ': 1,

'primary:': 2,

'1. coronary artery disease.': 1,

'2. nstemi': 1,

'3. congestive heart failure': 1,

'.': 5, 'secondary: ': 1,

'1. acute on chronic renal insufficiency - ': 1,

'2. hypertension ': 1,

'hypothermia': 1,

'urinary tract infection': 1,

'nstemi': 1,

'bladder cancer': 1,

'pneumonia, tia': 1,

'delayed gastric emptying': 1,

'1. thrombosed left aca aneurysm': 1,

'2. left aca/mca infarcts': 1,

'3. enterococcus urinary tract infection': 1,

'4. klebsiella urinary tract infection': 1,

'5. possible coagulase negative s. aureus bacteremia': 1,

'6. status post tracheostomy': 1,

'7. rheumatoid arthritis': 2,

'8. hypertension': 1,

'9. gastroesophageal reflux': 1,

'10. postherpetic neuralgia with l3-l5 dermatomes, followed at ': 1,

'[\*\*first name4 (namepattern1) \*\*] [\*\*last name (namepattern1) 1957\*\*] center': 1,

'11. right foot drop': 1,

'12. glaucoma': 1,

'13. history of right lower extremity cellulitis': 1,

'14. lumbar radiculopathy l4-l5': 1,

'15. bilateral femur replacements': 1,

'16. multiple hand surgeries': 1,

'17. status post appendectomy': 1,

'metastatic adenocarcinoma of colon': 1,

'primary: ': 2,

'j- tube displacement ': 1,

'1. dm type 2 ': 1,

'2. esrd ': 1,

'3. cad ': 1,

'4. chf- ef 15% [\*\*first name8 (namepattern2) \*\*] [\*\*hospital1 83\*\*] notes, s/p aicd placement ': 1,

'5. s/p multiple cvas ': 1,

'6. paroxysmal atrial fibrillation ': 1,

'7. history of multiple gi bleeds ': 1,

'8. sma syndrome ': 1,

'9. gerd ': 1,

'10. s/p cardiac arrest [\*\*3107-2-9\*\*] ': 1,

"11. h/o boerhaave's syndrome--> esophageal stricture ": 1,

'12. vre colonization ': 1,

'13. seizure disorder ': 1,

'14. hypertension ': 1,

'15. gastroparesis s/p g-tube/j-tube placement ': 1,

'16. h/o etoh abuse ': 1,

'17. aaa- 7cm ': 1,

'18. pseudomonal pna ': 1,

'19. c. diff colitis ': 1,

'aortic stenosis status post avr.': 1,

'none (pt passed away)': 1,

'congestive heart failure': 3,

'hypernatremia': 1,

'cholecystitis': 1,

'aids-related pancytopenia': 1,

'anuric renal failure': 1,

'mac': 1, 'sinusitis': 1,

'respiratory failure': 4,

'mental status changes': 1,

'carotid stenosis. tia': 1,

'history of htn': 1,

'history of hyperlipdemia': 1,

'primary diagnosis: ': 1,

'subacute right temporoparietal lobe stroke ': 1,

'urinary tract infection ': 1,

'atrial fibrillation with rapid ventricular response ': 1,

'sacral skin breakdown ': 1,

'secondary diagnosis: ': 1,

'cerebral amyloid angiopathy ': 1,

'left temporal hemorrhage [\*\*2906-11-29\*\*] ': 1,

'hypertension ': 1,

"stroke '[\*\*03\*\*] with resultant left sided deficits ": 1,

"stroke '[\*\*99\*\*] with left eye blindness ": 1,

'principal:': 1,

'traumatic comminuted c7 verterbral fracture [\*\*2825-2-12\*\*]': 1,

'orif/laminectomy/instrumentation/posterior fusion [\*\*2825-3-5\*\*]': 1,

'csf leak - wound infection s/p drainage and dural repair ': 1,

'[\*\*2825-3-22\*\*]': 1, 'incision and drainage and hardware exchange [\*\*2824-3-25\*\*]': 1,

'mrsa meningitis': 1,

'mrsa pneumonia': 1,

'left heart failure': 1,

'non-st elevation myocardial infarction': 1,

'left occipital stroke vs mrsa cerebritis': 1,

'rle deep venous thrombosis': 1,

'pulmonary embolism': 1, 'non-sustained ventricular tachycardia': 1,

'hypersensitivity desquamative dermatitis (rifampin vs ': 1,

'vancomycin)eosinophilia': 1,

'hypoxic respiratory failure ': 1,

'septic vs. anaphylactic shock': 1,

'delirium': 2,

'cholestasis': 1,

'rue paresis': 1,

'bilateral lower extremity myopathy': 1,

'dysphagia': 1,

'gi bleed': 1,

'nosocomial lll pneumonia': 1,

'anemia - multifactorial: illness, blood loss, ckd.': 1,

'sacral and heel ulcers': 1,

'mrsa/vre colonization': 1,

'candidemia': 1,

'secondary:': 3,

'diabetes mellitus type ii uncontrolled w/ complications': 1,

'coronary artery disease s/p cabg x 3': 1,

'anxiety': 1,

'hypercholesterolemia': 1,

'l3-l4 surgery': 1,

'bph': 2,

'aspiration pneumonia. right hip fracture. gram negative rod ': 1,

"bacteremia. hypotension. alzheimer's dementia. anemia. ": 1,

'hypothroidism.': 1,

'change in mental status secondary to glioblastoma': 1,

'brain tumor (final pathology pending)': 1,

'narcotics overdose': 1,

'cocaine abuse/dependence': 1,

'chronic back pain': 1,

'intracranial mass': 1,

'diverticular bleed s/p clipping': 1,

'diverticulitis': 1,

'coagulase negative staph uti': 1,

'secondary diagnoses': 2,

'atrial fibrillation': 2,

'chronic kidney disease': 1,

'dementia': 1,

'depression': 1,

'htn': 3,

'h/o lv thrombus': 1,

'h/o cva': 1,

'1. change in mental status': 1,

'2. urinary tract infection': 1,

'3. chronic respiratory failure': 1,

'1. hypertension': 1,

'2. diabetes mellitus': 1,

'3. end stage renal disease on hd': 1,

'aortic stenosis': 1,

'chronic afib': 1,

'right lung nodule': 1,

'gout': 1,

'niddm': 1,

'nephrolithiasis': 1,

'elev. chol.': 1,

'chronic lle edema': 1,

'? cirrhosis': 1,

'hep b': 1,

'vre': 1,

"bell's palsy": 1,

'mental status change': 1,

'partial diabetes insipidus': 1,

'?nms': 1,

'dehydration': 1,

'cad and elevated troponin': 1,

'schizophrenia': 1,

'rigth frontal glioma ': 1,

'end stage renal disease': 1,

'1)urosepsis ': 1,

'2)pyelonephritis ': 1,

'3)atypical chest pain ': 1,

'1)peptic ulcer disease ': 1,

'2)depression ': 1,

'3)hypertension ': 1,

'4)hypothyroidism ': 1,

'5)cluster headache ': 1,

"6)raynaud's ": 1,

'7)asthma ': 1,

'thoracoabdominal pneumococcal aneurysm': 1,

'hypovolvemia': 1,

'anemia requiring multiple blood transfusions': 1,

'renal faioure': 1,

'liver failure': 1,

'ileus': 1,

'splenic ischemiahypertension': 1,

'osteoporosis': 1,

'multiple organ system failure': 1,

'systemic inflammatory response sysndrome': 1,

'hypotension requiring pressors': 1,

'hypokalemia': 1,

'hypomagnesemia': 1,

'hypocalemia': 1,

'adrenal insufficiency': 1,

'upper gastrointestinal bleed': 1,

'ischemic colitis': 1,

'liver ischemia': 1,

'cardiac arrest': 1,

'pleural effusions bilaterally': 1,

'acute respiratory distress syndrome': 1,

'nutritional compromise requring parenteral nutrition': 1,

'coagulopathy': 1,

'septic shock': 2,

'primary diagnosis:': 2,

'1. urosepsis': 1,

'2. cardiomyopathy': 1,

'3. pneumonia': 1,

'4. thrombocytopenia': 1,

'5. tia': 1,

'secondary diagnosis:': 2,

'6. hypertension': 1,

'prolonged gingival bleeding after dental procedure, probably due ': 1,

'to vitamin k deficiency vs inhibitors': 1,

's/p struck by train': 1,

'subdural hematoma and diffuse axonal injury': 1,

'left arm traumatic amputation': 1,

'primary diagnoses:': 1,

'gib (likely [\*\*12-24\*\*] diverticulosis)': 1,

'rue swelling': 1,

'crf': 1,

'hyperlipidemia': 1,

'lung cancer- cytology consistent with nsclc': 1,

'hypercalcemia causing nephrogenic diabetes insipidus': 1,

'right renal artery embolus': 1,

'intramural sigmoid abscess': 1,

'rheumatoid arthritis': 1,

'patient expired': 1,

'primary: copd exacerbation, diastolic chf, frontal sinus osteoma': 1,

'type ii dm, well-controlled': 1,

'cad': 1,

'morbid obesity': 1,

'decubitus ulcer': 1,

'hemorrhagic cystitis': 1,

'metabolic acidosis': 1,

'ipf': 1,

'primary': 1,

'1. multifactorial respiratory distress from etiologies including ': 1,

'idiopathic pulmonary fibrosis, pneumonia, and hypercarbic ': 1,

'respiratory failure requiring tracheostomy': 1,

'2. gastrointestinal bleed': 1,

'3. c. diff colitis': 1,

'4. malnutrition': 1,

'5. sepsis with multiple organ failure, improved': 1,

'secondary': 1,

'1. severe restrictive lung disease': 1,

'2. breast cancer s/p mastectomy': 1}