Group 1

Homework #4: Insurance Claims - Part 1

HS 256F Healthcare Data Analytics and Data Mining

Chujun Chen

Leah Peretsky

Hongxiao Luo

Nicole Levesque

Jianying Zhu

03/10/2021

INTRODUCTION:

The database for the market of health, the Administrative Claim data, explores data pertaining to inpatient discharge, outpatient procedures and services, and emergency departments. The data set covers a variety of information regarding case-specific diagnostic discharge, socio-demographic characteristics of patients, medical issues resulting in admission, treatment and services, duration of patients stay in the health facility, and lastly, total service-specific charges billed by the hospital. The patients observed suffer from a wide range of accidents and conditions, such as fractured tibias from car accidents, depression disorders, accidental poisoning from heroin, etc. This assignment uses Administrative Claim Data from Vermont in 2016 to illustrate narratives for patients, analyze trends in inpatient service portfolios by Major Diagnostic Category (MDC), and to present a thorough overview of the drug abuse crisis in the United States.

QUESTION 1:

859382, a male aged 30-34, was transferred from a non-healthcare facility, and was admitted to the emergency department at Rutland Regional Medical Center for what appeared to be heroin poisoning. Unfortunately, he has passed away after being treated in the medical facility for only one day. Though, reports have identified this act as being accidental, as he had no intention of developing such negative effects from the drug. The unfortunate reality for 859382 was acute respiratory failure, which led to a severe compression of his brain, leading to anoxic brain damage. He needed respiratory ventilation, in less than 24 consecutive hours. His principal payment, or insurance type is self-pay, and the total charges of his visit were equal to $13128.19.

Patient 1585831, a female, aged 40-44, was also admitted to the Rutland Regional Medical Center to the emergency department, after having been at a non-health facility. Her primary health insurance is through Medicare, and she incurred an emergency visit of $17093.79. Similar to patient 859382, this individual was also poisoned by heroin, in an instance which is believed to be accidental. As a result, the female has died in just a single day of being admitted to the hospital. She has suffered from acute respiratory failure with hypoxia, acute pulmonary edema, and acute and subacute infective endocarditis.

An 18-24 year old female, 200760, was transferred from a non-health facility, and admitted to University of Vermont Medical Center, and was believed to have been injured in an unspecified motor-vehicle accident, as a subsequent encounter. As a result, she displaced and fractured her medial malleolus of her left tibia. Additionally, she is reported to have gastro issues, specifically gastro-esophageal reflux disease without esophagitis, and was also diagnosed as having a major single episode of depressive disorder. After spending four days in the emergency room, she is able to return home to recover with home health-care tending to her urgent needs. She receives commercial insurance, and the cost of her emergency visit equates to approximately $49533.15. However, the revenue charge of the services needed equals $40422.72.

A male patient between the ages of 18-24, labeled as 3692, was originally transferred from a non-health facility, to the University of Vermont Medical Center to seek urgent attention and treatment. He suffers from bipolar disorder with severe psychotic features such as suicidal ideations. He is believed to be cannabis and nicotine dependent. After spending a total of 58 days in the hospital, patient 3692 was finally able to recover at home. His insurance is Blue Cross, and the total adjusted emergency visit costs equal $117895.3. Currently, he is at home either on his own, and receiving mental and emotional care from his family.

507033, a female between the age of 25-29 was transferred from a non-medical institution and arrived at the Northwest Medical Center for an elective admission, as she has been expecting a baby. She is expecting a full-term, uncomplicated delivery and is at 41 weeks of gestation, as her body is getting ready for labor. The patient's insurance is Blue Cross, and the total cost of her visit to the hospital is about $3233.29. After a successful delivery and spending only one day in the hospital, she is now at home resting with her newborn.

After transferring from a different hospital, patient 40436, who is an elderly female aged 70-74, was admitted to the University of Vermont Medical Center to receive urgent care. Having chronic illness such as type 2 diabetes with nuropathy and atherosceltic health disease, it was urgent that she received care promptly. Though, she was ultimately admitted for Non-ST elevation myocardial infarction and acute posthemorrhagic anemia. Fortunately, she only needed to stay a single day in the hospital and was able to go home on her own. Her primary insurance is Medicare, and the total cost of her urgent visit was relatively expensive at $7,0275.41. However, this patient was charged a total of $70300.41 for all of the services received at the hospital.

A female, 690326, aged 40-44 had cosmetic surgery at the University of Vermont Medical Center. Due to the voluntary nature of the procedure, she was not transferred from a health facility prior to her visit. Thus, this was an elective type appointment that had been planned for several months. She underwent a bariatric surgery as other weight loss methods have failed. Following the procedure, she had an uncomfortable itching sensation known as pruritus. It is unknown how many days she spent in the hospital, but with just minor muscle spasm, she was eventually able to go home without any further medical care assistance. She has a self-pay insurance type, and the total adjusted charges of the procedure equaled a grand total of about $43,425.53, while the unadjusted was $49,533.15.

QUESTION 2:

This second part of the report examines the service and cost profile of major insurance payers in the American healthcare industry. The three main insurance players are considered, for this analysis, to be Medicare, Medicaid, and commercial payers. Table 1 below presents a breakdown of cost in million $US by Major Diagnosis Category (MDC) for each of the three major insurance payers, along with total costs for each category and insurer. Of the $1,290,000,000 accounted for in the table, $683 million (53%) was covered by Medicare, $217 million (17%) was covered by Medicaid, and $390 (30%) million was covered by commercial payers.

Table 1: Cost by MDG for major insurance payers, in million $US

	MEDICARE	MEDICAID	Commercial Payers	TOTAL
BRAIN AND CNS	57	13	38	**108**
EYE	0	0	0	**0**
EAR, NOSE & THROAT	3	2	2	**7**
RESPIRATORY	78	22	24	**124**
HEART CIRCULATORY	135	14	53	**202**
DIGESTIVE	63	14	31	**108**
LIVER & PANCREAS	17	10	12	**39**
MUSCULOSKELETAL	145	30	87	**262**
SKIN AND BREAST	11	3	5	**19**
ENDOCRINE	13	5	8	**26**
KIDNEY & URINARY	27	4	7	**38**
MALE REPRODUCTIVE	2	0	2	**4**
FEMALE REPRODUCTIVE	1	1	4	**6**
PREGNANCY, CHILDBIRTH AND THE PUERPERIUM	0	26	34	**60**
NEONATAL	0	30	28	**58**
SPLEEN & BLOOD	7	2	5	**14**
LYMPHATIC	8	2	5	**15**
INFECTION	61	13	17	**91**
MENTAL ILLNESS	24	14	9	**47**
SUBSTANCE ABUSE	2	4	1	**7**
INJURY, TOXIC EFFECTS	9	3	4	**16**
BURNS	1	0	0	**1**

ALL OTHER	15	3	6	**24**
TRAUMA	4	2	8	**14**
HIV	0	0	0	**0**
TOTAL	**683**	**217**	**390**	**1290**

Figures 1-3 below present a graphical representation of the inpatient service portfolio for each major insurance payer. Figure 1 shows the inpatient service portfolio for Medicare. The top five MDGS are musculoskeletal care (21.71%), heart and circulatory care (20.21%), respiratory care (11.68%), digestive care (9.42%), and infection care (9.13%). These MDCs represent areas in which older individuals covered by Medicare are more likely to require medical care. The musculoskeletal system in particular is known for including the expensive joint replacement surgeries that older Americans are prone to requiring. Mental illness notably makes up a small percentage of Medicare's inpatient service portfolio, perhaps due to stigma against mental illness for the older American generation. It is also not surprising that respiratory care constitutes such a large portion of the inpatient service, as disabled individuals insured by Medicare, similar to patient 1585831, who has died of acute respiratory failure induced by heroin, may be more prone to abuse drugs due to their disabled status.

An examination of Medicare beneficiaries by gender shows that females can attribute their spending to increased musculoskeletal care, heart and circulatory care makes up a larger portion of healthcare services for male Medicare beneficiaries for women. This could be due to increased prevalence of cardiac issues in older American males, while older women are more likely to suffer from broken bones and require expensive surgeries or joint replacements.

Figure 1: Medicare's Inpatient Services Portfolio

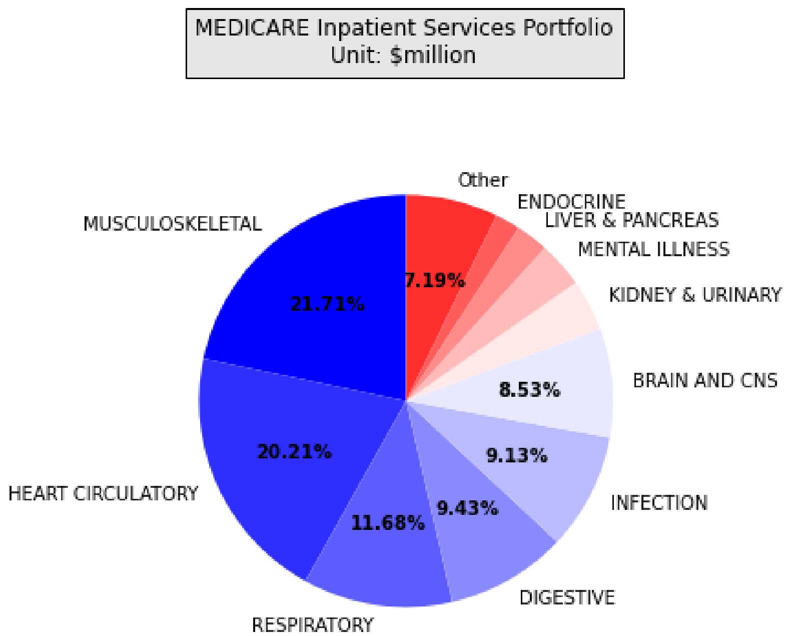
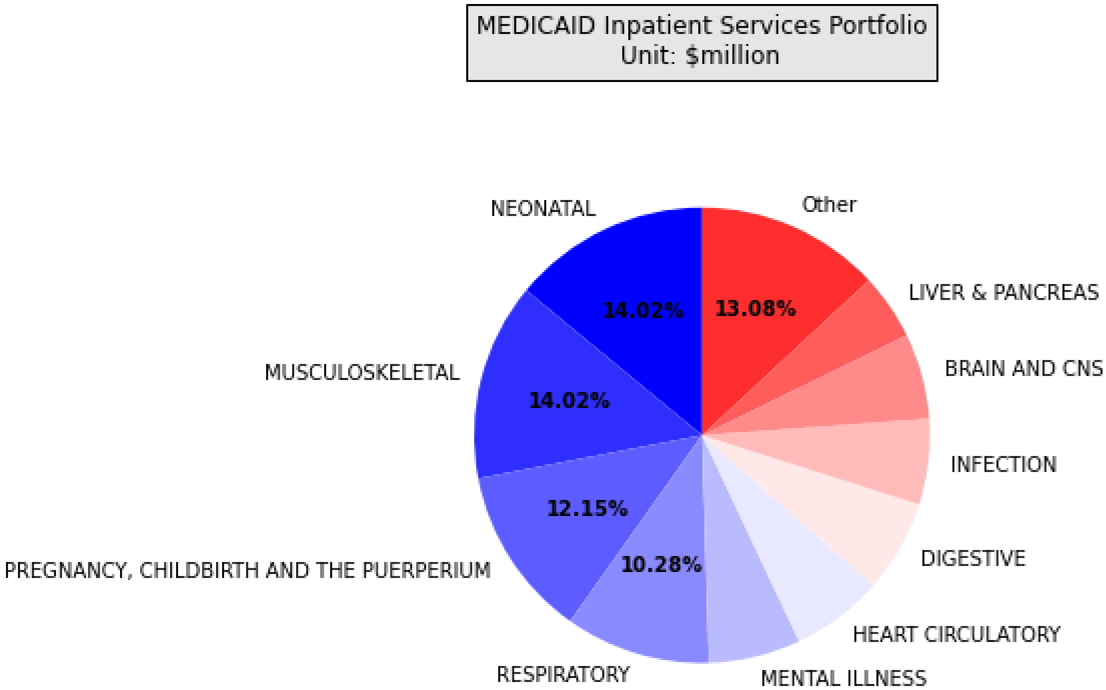


Figure 2 below shows a breakdown of Medicaid's inpatient services spending by MDC. The top five categories are neonatal care (14.02%), musculoskeletal care (14.02%), pregnancy, childbirth and the puerperium (12.15%), and respiratory care (10.28%), and mental illness (6.45%). The heavy presence of maternal and child care could be due to the eased enrollment requirements for pregnant women, their babies, and children living in poverty. Children in the foster care system and the adoption process are also covered by Medicaid, which could explain a large portion of the cost. Liver and pancreas care make up the largest share of Medicaid's inpatient services portfolio, which is not unique for Medicaid. This could indicate that in general, liver and pancreas spending is low due to a low incidence rate or inexpensive procedures within the MDC.

Neonatal care for Medicaid beneficiaries under age 1 does not appear to vary by gender, but for women, pregnancy and related care makes up a significant portion of healthcare expenses. Women of childbearing age can attribute a substantial portion of their inpatient services to such medical care. This finding is unsurprising, as pregnancy, childbirth, and the puerperium is unique to women. Similar to the results for Medicare beneficiaries, the costs associated with heart and circulatory care are greater for older male Medicaid beneficiaries than for women.

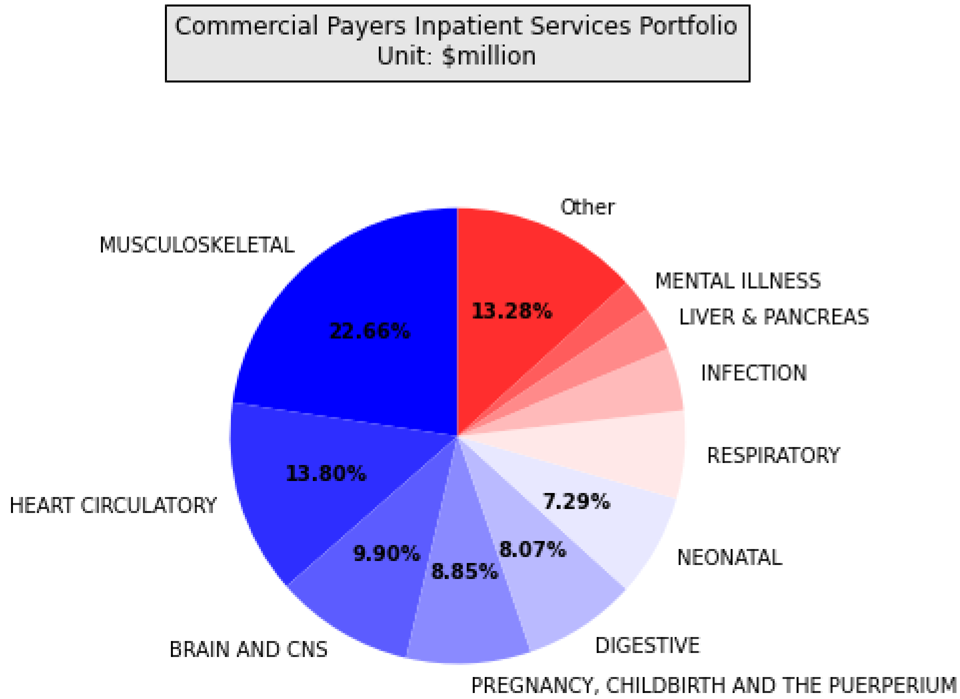
Figure 2: Medicaid's Inpatient Services Portfolio



Finally, Figure 3 below presents the inpatient services portfolio for commercial payers in the American healthcare system. The top five MDCs are musculoskeletal care (22.66%), heart and circulatory care (13.8%), brain and CNS care (9.9%), pregnancy, childbirth and the puerperium (8.85%), and digestive care (8.07). An interesting distinction with commercial payers vs. state-funded insurers is that brain and CNS care makes up a substantial portion of the services portfolio, perhaps because its clients can afford to pay for insurance premiums and out-of-pocket payments associated with unnecessary testing such as MRIs for common headaches. Similar to Medicare and Medicaid, liver and pancreas care makes up a small percentage of commercial payers' inpatient services portfolio, which could show its overall small contribution to inpatient spending. Interestingly, mental illness spending is very low for commercial payers, perhaps because high-income payers are less likely to be in the position to require inpatient mental illness care. In fact, 18-24 year old patient 200760, who uses commercial insurance was diagnosed with a major single depression episode disorder, yet she was not admitted to the hospital for this reason. Thus, there may be a relationship with those that are on commercial insurance, and individuals not willing to receive treatment for mental illness, possibly due to the stigma surrounding this condition.

Similar to Medicare, female beneficiaries of commercial insurance receive a substantial amount of care related to pregnancy, childbirth, and the puerperium, while male beneficiaries do not. Men covered by commercial insurance can attribute more of their spending to respiratory care than women can, perhaps due to increased rates of smoking among men than women. Additionally, men with commercial healthcare insurance coverage spend more on heart and circulatory care than women, indicating that the increased prevalence of cardiac issues among American men applies to individuals across insurance providers, income levels, and lifestyles.

Figure 3: Commercial Payers' Inpatient Services Portfolio



QUESTION 3:

To further examine the opioid epidemic in the United States, this section of the report examines trends among hospital visits for drug use and abuse in the state of Vermont in 2016. During the year, 2151 emergency department visits have been identified as visits for drug users or abusers. A common myth is that drug use and abuse is a male problem, and that women avoid drug abuse and overdoses worthy of emergency admission. The results of a Fisher's Exact Test to test this relationship prove that there is only a marginal difference between the drug use emergency admission rates for men and women, as this relationship is not statistically significant.

Table 3 Gender and drug use

Gender	Drug User	Non-Drug User
Male	1009	123149
Female	1141	140553
odds ratio	1.00928	
p-value	0.42395	

In addition to the health effects, the opioid crisis has a substantial financial impact on the communities in which it is prevalent. In total, the full dollar amount spent on drug use/abuse cases for the patients identified with drug use and abuse admissions in Vermont in 2016 is $30,741,219.53. Of this spending, Medicare covered 56%, Medicaid covered 22%, and commercial payers covered 19%. The remaining 3% could be out-of-pocket payments or uncompensated costs that are picked up by hospitals.

The use and abuse of synthetic amphetamines is also a substantial problem in the United States. In total, there were 467 patients who are identified as drug use or abuse admissions due to synthetic narcotics or amphetamines. 467 patients experienced poisoning or another adverse health effect due to synthetic narcotics or amphetamines, including intentional, unintentional, initial encounters, and sequelas. The use of synthetic narcotics or amphetamines accounted for 22% of all drug use and abuse admissions in Vermont in 2016.

The three regions in Vermont with the highest number of drug use and abuse cases are the 05400-05499 zip code range (excluding Burlington to Saint Albans), the 05700-05799 zip code range (excluding 05701), and Rutland.

Table 4: Top 3 zip codes and drug use

TXTZIP	Highest Numbers of Drug Users
054	379
057	249
05701	226

Drug use and abuse cases include a variety of diagnoses. The ten most common in the sample are important to note so policymakers in Vermont can work to address this problem. Accidental, unintentional initial encounter poisonings by heroin is the most common, with 256 (12%) cases attributed to it. 256 (12%) of cases are due to adverse effects of other opioids. 123 (0.6%) of cases are due to intentional, self-harm poisoning by benzodiazepines. 114 (0.5%) cases were caused by the adverse effect of benzodiazepines at the initial encounter. 112 (0.5%) cases were caused by the adverse effect of unspecified narcotics at the initial encounter. 82 (4%) cases were due to intentional, self-harm poisoning by selective serotonin reuptake inhibitors at the initial encounter. 81 (4%) cases were due to accidental, unintentional poisoning by other opioids at the initial encounter. 80 (4%) cases were caused by accidental, unintentional poisoning by benzodiazepines at the subsequent encounter. 75 (3%) cases were caused by the adverse effect of other antiepileptic and sedative-hypnotic drugs at the subsequent encounter. Finally, 74 (3%) cases were due to intentional self-harm poisoning by other antiepileptic and sedative-hypnotic drugs at the initial encounter.

Table 5: Codes and Patients

ICD-10 Codes	Numbers of Patients
T404xxx and T4362xxx	156

T404xxx and T4362xxx, ATYPE: Emergency and Urgent	155
T404xxx	113
T4362xxx	43

CONCLUSION:

Although difficult to obtain access to, insurance claims data is unparalleled in the insight it can provide for the state of the American healthcare industry. This report has examined such data for patients in Vermont in the year 2016 to examine trends within the insurance industry and the opioid crisis. First, a presentation of narratives for several patients has provided an individual view into the lives of seven Americans who relied on emergency and inpatient hospital care in the state of Vermont in 2016. A description of their demographics, diagnoses, services, and billing makes the insurance claims more personal and highlights the importance of the provision of quality care. Second, this report presented general data for insurance spending for the MDCs by payer and in total, showing which services are currently of critical importance in the state of Vermont. It then provided an overview of the inpatient services portfolios for the three major payers in the American healthcare insurance system: Medicare, Medicaid, and commercial payers. An in-depth view of spending by payer, analyzed by the patients' age and sex allowed for a stronger understanding of the critical health issues for each demographic. Finally, an analysis of administrative claim data relevant to the use and abuse of drugs resulted in a discussion of pressing questions for the major issues relating to the drug crisis, and could assist policymakers in finding effective solutions.

RELEVANT CODING

Question 2

```
import pandas as pd
import numpy as np
df_in= pd.read_csv("VTINP16_upd.txt")
a= [1,2,6,7]

df_in_4insurance =df_in[df_in['PPAY'].isin([1,2,6,7])]

#df_in_4insurance.apply(df_in_4insurance.CHRGS.to_numeric, errors = ‘coerce’)
df_in_4insurance=df_in_4insurance.apply(pd.to_numeric,errors='coerce')
df_in_4insurance["CHRGS"] = pd.to_numeric(df_in_4insurance["CHRGS"])

df_in_4insurance = df_in_4insurance.dropna(subset=["CHRGS"])

pd.options.display.max_columns = None

MDC_Ins=df_in_4insurance.groupby([ 'MDC','PPAY'])["CHRGS"].apply(lambda x :
x.astype(int).sum()) \
        .unstack(fill_value=0) \
        .reset_index() \
        .rename_axis(None, axis=1)

cats_insurance = ['MEDICARE', 'MEDICAID', 'BLUE CROSS', 'COMMERCIAL INSURANCE']
cats_mdc = ['BRAIN AND CNS','EYE','EAR, NOSE & THROAT','RESPIRATORY','HEART
CIRCULATORY','DIGESTIVE',
'LIVER & PANCREAS','MUSCULOSKELETAL','SKIN AND BREAST','ENDOCRINE','KIDNEY &
URINARY','MALE REPRODUCTIVE',
'FEMALE REPRODUCTIVE','PREGNANCY, CHILDBIRTH AND THE
PUERPERIUM','NEONATAL','SPLEEN & BLOOD','LYMPHATIC',
'INFECTION','MENTAL ILLNESS','SUBSTANCE ABUSE','INJURY, TOXIC
EFFECTS','BURNS','ALL OTHER','TRAUMA','HIV']

MDC_Ins=MDC_Ins.set_index('MDC')
MDC_Ins_table = pd.DataFrame(data = MDC_Ins.values,
index=cats_mdc,columns=cats_insurance)

MDC_Ins_table['Commercial Payers'] =  MDC_Ins_table.iloc[:,2]+MDC_Ins_table.iloc[:,3]
MDC_Ins_table_3ins = MDC_Ins_table.drop(['BLUE CROSS' ,'COMMERCIAL
INSURANCE'],axis=1)
MDC_Ins_table_3ins = MDC_Ins_table_3ins[:]/1000000
MDC_Ins_table_3ins =MDC_Ins_table_3ins.astype(int)
```

```
display(MDC_Ins_table_3ins)

MDC_Ins_table_3ins_wo_other = MDC_Ins_table_3ins.drop('ALL OTHER')

MDC_Ins_table_MEDICARE= pd.DataFrame(MDC_Ins_table_3ins_wo_other['MEDICARE'])
MDC_Ins_table_MEDICAID= pd.DataFrame(MDC_Ins_table_3ins_wo_other['MEDICAID'])
MDC_Ins_table_ComPayers= pd.DataFrame(MDC_Ins_table_3ins_wo_other['Commercial
Payers'])

MEDICARE_order = MDC_Ins_table_MEDICARE.sort_values('MEDICARE', ascending = False)

import matplotlib.pyplot as plt
temp = MEDICARE_order
temp2 = temp.head(10)
if len(temp) > 10:
   temp2.loc['Other'] = temp[10:].sum()

name = list(temp2.index)
sizes = temp2.iloc[:,0]
labels = name

fig1, ax1 = plt.subplots(figsize=(5, 5))
fig1.subplots_adjust(0.3, 0, 1, 1)

theme = plt.get_cmap('bwr')
ax1.set_prop_cycle("color", [theme(1. * i / len(sizes))
                             for i in range(len(sizes))])

def autopct_generator(limit):
   """Remove percent on small slices."""
   def inner_autopct(pct):
      return ('%.2f%%' % pct) if pct > limit else ''
   return inner_autopct

_, _, autotexts = ax1.pie(
   sizes, autopct=autopct_generator(7), startangle=90, radius=1.8 * 1000,labels=name)
for autotext in autotexts:
   autotext.set_weight('bold')
ax1.axis('equal')

plt.title("MEDICARE Inpatient Services Portfolio\n"+"Unit: $million", bbox={'facecolor':'0.9',
'pad':5})
plt.show()
```

```
MEDICAID_order = MDC_Ins_table_MEDICAID.sort_values('MEDICAID', ascending = False)

import matplotlib.pyplot as plt
temp = MEDICAID_order
temp2 = temp.head(10)
if len(temp) > 10:
    temp2.loc['Other'] = temp[10:].sum()


name = list(temp2.index)

import matplotlib.pyplot as plt

sizes = temp2.iloc[:,0]
labels = name

fig1, ax1 = plt.subplots(figsize=(5, 5))
fig1.subplots_adjust(0.3, 0, 1, 1)

theme = plt.get_cmap('bwr')
ax1.set_prop_cycle("color", [theme(1. * i / len(sizes))
                             for i in range(len(sizes))])

def autopct_generator(limit):
    """Remove percent on small slices."""
    def inner_autopct(pct):
        return ('%.2f%%' % pct) if pct > limit else ''
    return inner_autopct

_, _, autotexts = ax1.pie(
    sizes, autopct=autopct_generator(7), startangle=90, radius=1.8 * 1000,labels=name)
for autotext in autotexts:
    autotext.set_weight('bold')
ax1.axis('equal')

plt.title("MEDICAID Inpatient Services Portfolio\n"+"Unit: $million", bbox={'facecolor':'0.9',
'pad':5})
plt.show()


Commercial_Payers_order = MDC_Ins_table_ComPayers.sort_values('Commercial Payers',
ascending = False)

import matplotlib.pyplot as plt
```

```
temp = Commercial_Payers_order
temp2 = temp.head(10)
if len(temp) > 10:
    temp2.loc['Other'] = temp[10:].sum()


name = list(temp2.index)

import matplotlib.pyplot as plt

sizes = temp2.iloc[:,0]
labels = name

fig1, ax1 = plt.subplots(figsize=(5, 5))
fig1.subplots_adjust(0.3, 0, 1, 1)

theme = plt.get_cmap('bwr')
ax1.set_prop_cycle("color", [theme(1. * i / len(sizes))
                             for i in range(len(sizes))])


def autopct_generator(limit):
    """Remove percent on small slices."""
    def inner_autopct(pct):
        return ('%.2f%%' % pct) if pct > limit else ''
    return inner_autopct

_, _, autotexts = ax1.pie(
    sizes, autopct=autopct_generator(7), startangle=90, radius=1.8 * 1000,labels=name)
for autotext in autotexts:
    autotext.set_weight('bold')
ax1.axis('equal')

plt.title("Commercial Payers Inpatient Services Portfolio\n"+"Unit: $million",
bbox={'facecolor':'0.9', 'pad':5})
plt.show()

df_in_all = df_in.apply(pd.to_numeric,errors='coerce')
df_in_all["CHRGS"] = pd.to_numeric(df_in_all["CHRGS"])
df_in_all= df_in_all.dropna(subset=["CHRGS"])

df_in_all['intage'].replace({1:'Under 1', 2: '1-17',3:'18-24',4:'25-29',5:'30-34',6:'35-39',7:'40-44',
                             8:'45-49',9:'50-54',10:'55-59',11:'60-64',12:'65-69',13:'70-74',14:'75 and
over',15:'Unknown'}, inplace=True)
```

```
df_in_all['sex'].replace({1:'Male', 2: 'Female'}, inplace=True)
df_in_all['MDC'].replace({1:'BRAIN AND CNS',2:'EYE',3:'EAR, NOSE &
THROAT',4:'RESPIRATORY',5:'HEART CIRCULATORY',6:'DIGESTIVE',
7:'LIVER & PANCREAS',8:'MUSCULOSKELETAL',9:'SKIN AND
BREAST',10:'ENDOCRINE',11:'KIDNEY & URINARY',12:'MALE REPRODUCTIVE',
13:'FEMALE REPRODUCTIVE',14:'PREGNANCY, CHILDBIRTH AND THE
PUERPERIUM',15:'NEONATAL',16:'SPLEEN & BLOOD',17:'LYMPHATIC',
18:'INFECTION',19:'MENTAL ILLNESS',20:'SUBSTANCE ABUSE',21:'INJURY, TOXIC
EFFECTS',22:'BURNS',23:'ALL OTHER',24:'TRAUMA',25:'HIV'}, inplace=True)
df_in_all['PPAY'].replace({1:'MEDICARE', 2: 'MEDICAID',6: 'Commercial Payers',7: 'Commercial
Payers'}, inplace=True)

df_in_3co_wo_cat =df_in_all[df_in_all['PPAY'].isin(['MEDICARE','MEDICAID','Commercial
Payers'])]
df_in_3co_wo_cat_new = df_in_3co_wo_cat[['PPAY','MDC','sex','intage','CHRGS']]
```

Medicare data analytics 1

```
#Medicare data
df_in_3co_wo_cat_new_medicare = df_in_3co_wo_cat_new[df_in_3co_wo_cat_new['PPAY']
=='MEDICARE']

mdc_intage_table_medicare = df_in_3co_wo_cat_new_medicare.groupby([
'MDC','intage','sex'])["CHRGS"].apply(lambda x : x.astype(int).sum()) \
                .unstack(fill_value=0) \
                .reset_index() \
                .rename_axis(None, axis=1)

#Filter top 10 MDC
MEDICARE_order_top10 = pd.DataFrame(MEDICARE_order.index[:10])
MEDICARE_order_top10.columns=['MDC']
MEDICARE_order_top10_table = pd.merge(MEDICARE_order_top10
,mdc_intage_table_medicare,how='left',on=['MDC'])
display(MEDICARE_order_top10_table)

MEDICARE_order_top10_table["Markersize_Male"] =
MEDICARE_order_top10_table.iloc[:,3]/40000
fig = plt.figure(figsize=(8, 6))
#plot categorical scatter plot
plt.scatter(MEDICARE_order_top10_table.MDC, MEDICARE_order_top10_table.intage, s =
MEDICARE_order_top10_table.Markersize_Male, edgecolors = "black", c = "white", zorder = 1)
#plot grid behind markers
#plt.grid(ls = "--", zorder = 1)
#take care of long labels
fig.autofmt_xdate()
```

```
plt.tight_layout()
plt.title("MEDICARE Top 10 Inpatient Services Portfolio\n"+"Male", bbox={'facecolor':'0.9',
'pad':1})
plt.show()

MEDICARE_order_top10_table["Markersize_Female"] =
MEDICARE_order_top10_table.iloc[:,2]/40000
fig = plt.figure(figsize=(8, 6))
#plot categorical scatter plot
plt.scatter(MEDICARE_order_top10_table.MDC, MEDICARE_order_top10_table.intage, s =
MEDICARE_order_top10_table.Markersize_Female, edgecolors = "black", c = "white", zorder =
1)
#plot grid behind markers
#plt.grid(ls = "--", zorder = 1)
#take care of long labels
fig.autofmt_xdate()
plt.tight_layout()
plt.title("MEDICARE Top 10 Inpatient Services Portfolio\n"+"Female", bbox={'facecolor':'0.9',
'pad':1})
plt.show()
```

Medicare data analytics 2

```
df_in_3co_wo_cat_new_medicaid = df_in_3co_wo_cat_new[df_in_3co_wo_cat_new['PPAY']
=='MEDICAID']

mdc_intage_table_medicaid = df_in_3co_wo_cat_new_medicaid.groupby([
'MDC','intage','sex'])["CHRGS"].apply(lambda x : x.astype(int).sum()) \
            .unstack(fill_value=0) \
            .reset_index() \
            .rename_axis(None, axis=1)

MEDICAID_order_top10 = pd.DataFrame(MEDICAID_order.index[:10])
MEDICAID_order_top10.columns=['MDC']
MEDICAID_order_top10_table = pd.merge(MEDICAID_order_top10
,mdc_intage_table_medicaid,how='left',on=['MDC'])
display(MEDICAID_order_top10_table)

MEDICAID_order_top10_table["Markersize_Male"] =
MEDICAID_order_top10_table.iloc[:,3]/40000
fig = plt.figure(figsize=(8, 6))
#plot categorical scatter plot
plt.scatter(MEDICAID_order_top10_table.MDC, MEDICAID_order_top10_table.intage, s =
MEDICAID_order_top10_table.Markersize_Male, edgecolors = "blue", c = "white", zorder = 1)
#plot grid behind markers
```

```
#plt.grid(ls = "--", zorder = 1)
#take care of long labels
fig.autofmt_xdate()
plt.tight_layout()
plt.title("MEDICAID Top 10 Inpatient Services Portfolio\n"+"Male", bbox={'facecolor':'0.9',
'pad':1})
plt.show()

MEDICAID_order_top10_table["Markersize_Female"] =
MEDICAID_order_top10_table.iloc[:,2]/40000
fig = plt.figure(figsize=(8, 6))
#plot categorical scatter plot
plt.scatter(MEDICAID_order_top10_table.MDC, MEDICAID_order_top10_table.intage, s =
MEDICAID_order_top10_table.Markersize_Female, edgecolors = "blue", c = "white", zorder = 1)
#plot grid behind markers
#plt.grid(ls = "--", zorder = 1)
#take care of long labels
fig.autofmt_xdate()
plt.tight_layout()
plt.title("MEDICAID Top 10 Inpatient Services Portfolio\n"+"Female", bbox={'facecolor':'0.9',
'pad':1})
plt.show()
```

Commercial Payers data analytics

```
df_in_3co_wo_cat_new_compay = df_in_3co_wo_cat_new[df_in_3co_wo_cat_new['PPAY']
=='Commercial Payers']

mdc_intage_table_compay = df_in_3co_wo_cat_new_compay.groupby([
'MDC','intage','sex'])["CHRGS"].apply(lambda x : x.astype(int).sum()) \
                .unstack(fill_value=0) \
                .reset_index() \
                .rename_axis(None, axis=1)

Commercial_Payers_order_top10 = pd.DataFrame(Commercial_Payers_order.index[:10])
Commercial_Payers_order_top10.columns=['MDC']
Commercial_Payers_order_top10_table = pd.merge(Commercial_Payers_order_top10
,mdc_intage_table_compay,how='left',on=['MDC'])
display(Commercial_Payers_order_top10_table)

Commercial_Payers_order_top10_table["Markersize_Male"] =
Commercial_Payers_order_top10_table.iloc[:,3]/40000
fig = plt.figure(figsize=(8, 6))
#plot categorical scatter plot
```

```
plt.scatter(Commercial_Payers_order_top10_table.MDC,
Commercial_Payers_order_top10_table.intage, s =
Commercial_Payers_order_top10_table.Markersize_Male, edgecolors = "purple", c = "white",
zorder = 1)
#plot grid behind markers
#plt.grid(ls = "--", zorder = 1)
#take care of long labels
fig.autofmt_xdate()
plt.tight_layout()
plt.title("Commercial Payers Top 10 Inpatient Services Portfolio\n"+"Male",
bbox={'facecolor':'0.9', 'pad':1})
plt.show()

Commercial_Payers_order_top10_table["Markersize_Female"] =
Commercial_Payers_order_top10_table.iloc[:,2]/40000
fig = plt.figure(figsize=(8, 6))
#plot categorical scatter plot
plt.scatter(Commercial_Payers_order_top10_table.MDC,
Commercial_Payers_order_top10_table.intage, s =
Commercial_Payers_order_top10_table.Markersize_Female, edgecolors = "purple", c = "white",
zorder = 1)
#plot grid behind markers
#plt.grid(ls = "--", zorder = 1)
#take care of long labels
fig.autofmt_xdate()
plt.tight_layout()
plt.title("Commercial Payers Top 10 Inpatient Services Portfolio\n"+"Female",
bbox={'facecolor':'0.9', 'pad':1})
plt.show()
```

Question 3

```
# -*- coding: utf-8 -*-
"""
Created on Wed Mar  3 17:25:40 2021

@author: 14830
"""
import csv
import numpy as np
import pandas as pd
import scipy.stats as stats

#read text files and transform to CSV format and load CSV file
```

```
df = pd.read_csv("VTINP16_upd.txt")
df.to_csv('VTINP16_upd.csv', index = None)
inpatient = pd.read_csv('VTINP16_upd.csv')

df2 = pd.read_csv("VTREVCODE16.txt")
df2.to_csv('VTREVCODE16.csv', index = None)
REVCODE = pd.read_csv('VTREVCODE16.csv')

df3 = pd.read_csv("VTED16.txt")
df3.to_csv('VTED16.csv', index = None)
Emergency = pd.read_csv('VTED16.csv', low_memory = False)

#%% 3.1
#filter columns with DX1-20
Emergency['DX1'].unique()
dx = Emergency.loc[:, Emergency.columns.str.startswith("DX")].astype(str)
#dx1 = Emergency.filter(regex='^DX',axis=1).astype(str)
print(dx)

#filter codes with T40|T41|T42|T43 and add new column in ED
#a = dx.DX1.str.startswith(('T40', 'T41', 'T42', 'T43'))
a = ['T40', 'T41', 'T42', 'T43']
flag = np.zeros(len(dx))
for i in range(len(dx)) :
    for j in range(dx.shape[1]):
        temp = dx.iloc[i,j]
        for k in a:
            if temp.startswith(k):
                flag[i] = 1

#Numbers of ED visits exactly have been diagnosed as drug user/abuser
flag = pd.DataFrame(flag)
sumflag = sum(flag[0])
print(sumflag)

#if dx.iloc[i,j].str.contains(('T40', 'T41', 'T42', 'T43')).any() :
#flag[i] = 1


#%% 3.2
#filter gender
Emergency['flag'] = flag
emg = Emergency[Emergency['sex'].isin(['1','2'])]
```

```
#2*2 table
table1=pd.crosstab(index = emg['sex'], columns = emg['flag'])
table1 = table1.reindex(columns=[1,0])
#table1 = table1.reindex([1,0], axis = 1)
table1.index = ['Male', 'Female']
table1.columns =['Drug User', 'Non-Drug User']
#table1 = table1.rename(columns={1:'Drug User', 0:'Non-Drug User'}, index={'1':'Male',
'2':'Female'})
print(table1)
#fisher test: 1009/123149 > 1141/140553 that means male are likely to be drug users than
female
oddsratio,pvalue = stats.fisher_exact(table1, alternative='greater')
print(pvalue)
print(oddsratio)


#%% 3.3
#Exact dollar amount for your identified patients
drugab = Emergency[Emergency['flag'] ==1]
sumcharge = drugab.CHRGS.astype(float).sum()
print(sumcharge)

#Of the three insurances in Question 2, what was share of each of the total payments?
medicare = drugab[drugab.PPAY.isin([1])]
medicare_sum = medicare.CHRGS.astype(float).sum()/sumcharge
medicaid = drugab[drugab.PPAY.isin([2])]
medicaid_sum = medicaid.CHRGS.astype(float).sum()/sumcharge
comm = drugab[drugab.PPAY.isin([6,7])]
comm_sum = comm.CHRGS.astype(float).sum()/sumcharge
print("Medicare, Medicaid, Comm's perc are:", medicare_sum,medicaid_sum,comm_sum)
#%% 3.4
#filter code start with T404|T4362 and add new column in ED
i = 0
j = 0
b = 'T404'
c = 'T4362'
#b = ['T404', 'T4362']
flag1 = np.zeros(len(dx))

for i in range(len(dx)):
    for j in range(dx.shape[1]):
        temp1 = dx.iloc[i,j]
        if temp1.startswith(b):
                flag1[i] = 1
```

```
        elif temp1.startswith(c):
            flag1[i] = 2

"""
#original
for i in range(len(dx)):
    for j in range(dx.shape[1]):
        temp1 = dx.iloc[i,j]
        for m in b:
            if temp1.startswith(m):
                flag1[i] = 1
flag1 = pd.DataFrame(flag1)
sumflag1 = int(sum(flag1[0]))
print(sumflag1)
"""

flag1 = pd.DataFrame(flag1)
sumflag1 = flag1[flag1[0] == 1]
sumflag2 = flag1[flag1[0] == 2]

Emergency['flag1'] = flag1
emg1 = Emergency[Emergency['flag1'].isin(['1','2'])]
T404 = emg1[emg1['flag1'] == 1]
T4362 = emg1[emg1['flag1'] == 2]

#emergency and urgent
emg_urgent = emg1[emg1['ATYPE'].isin(['1' , '2'])]

#Patient and their diagnosis code for ED
table2 = pd.DataFrame({'ICD-10 Codes':['T404xxx and T4362xxx', 'T404xxx and T4362xxx,
ATYPE: Emergency and Urgent', 'T404xxx', 'T4362xxx'],
                'Numbers of Patient':[len(sumflag1)+len(sumflag2), len(emg_urgent), len(T404),
len(T4362)]})
print(table2)


#%% 3.5
#extract columns with zipcode and DX1-20
drugab1 = drugab.iloc[:,3:29]
drugab1.columns
#drugab2 = drugab1.drop(['dstat', 'sex', 'PPAY', 'CHRGS'], axis = 1)
drugab2 = pd.concat(drugab1.iloc[:, l] for l in range(6, drugab1.shape[1]))

#filter codes with T40|T41|T42|T43
```

```
ziprank = drugab2.copy().to_frame().astype(str)
ziprank['TXTZIP'] = drugab['TXTZIP']
ziprank = ziprank[['TXTZIP', 0]]
ziprank = ziprank[ziprank[0].str.contains('T40|T41|T42|T43')]
ziprank.TXTZIP.unique()

#3 zip code regions with the highest numbers of drug use/abuse cases.
top3_zip = ziprank.groupby('TXTZIP').count()
top3_zip.columns = ['Highest Numbers of Drug Users']
top3_zip = top3_zip.sort_values(by=['Highest Numbers of Drug Users'], ascending = False)
top3_zip = top3_zip.reset_index(level='TXTZIP', drop = False)
top3_zip.index = np.arange(1, len(top3_zip)+1)
print(top3_zip.head(3))

#age distirbution among top 3 zipcodes
ziprank1 = ziprank.copy().astype(str)
ziprank1['Age'] = drugab1['intage']

age = ziprank1[ziprank1['TXTZIP'].isin(['054','057', '05701'])]
age = age.groupby(['TXTZIP', 'Age'])[0].count().to_frame().reset_index()
#age.TXTZIP.unique()
age.columns = ['TXTZIP', 'Age', 'Total Numbers']
#age = age.sort_values(by = ['TXTZIP', 'Age', 'Total Numbers'], ascending = [True, True, False])
age.index = np.arange(1, len(age)+1)
print(age)


#%%3.6
#10 most common diagnoses of drug use/abuse
#drugab1a = drugab.iloc[:,9:29]
drugab1.nunique()
#firstcolumnnum = list(set(drugab1.DX1))
drugab3 = pd.concat(drugab1.iloc[:,l] for l in range(6, drugab1.shape[1]))
coderank = drugab3.copy().to_frame().astype(str)
#series.str.startswith(tuple(a)).to_frame()
coderank['ICD-10 Codes'] = coderank[0]
coderank = coderank[coderank[0].str.contains('T40|T41|T42|T43')]
top10_code = coderank.groupby('ICD-10 Codes').count()
top10_code.columns = ['Numbers']
top10_code = top10_code.sort_values(by='Numbers', ascending = False)
top10_code = top10_code.reset_index(level='ICD-10 Codes', drop = False)
top10_code.index = np.arange(1, len(top10_code)+1)
print(top10_code.head(10))
```