**Research Assistant work under professor Olena**

From 28th September 2020 to 29th January, I worked about 20 hours a week and from 29th January to 12 February 2021(30hrs/week) on tasks for Prof. Olena Maurenko. I worked on a Data Analytics project in health care on data set named ‘Encounter’. The task was to understand the data and perform data preprocessing (data cleaning, find the missing values, correlation between the columns, and describing each Columns in dataset.

The next step was to extract important information from the data e.g., maximum number of visits by a patient, minimum number of visits, median number of visits, overall distribution of visits of patients,

Plot graphs between two columns (e.g., Distribution of Discharge Disposition vs Care Setting Category)

The detailed explanation of each steps can be found below, and the python code can be found on GitHub (Detailed Analysis is mentioned in <https://github.com/mraunak/Research-Assistant-Work-/blob/main/Data%20Analytics%20in%20Health%20Care_%20RA.ipynb>

2nd Task

Task: Resolved 3 issues

1. Admit date and Discharge date being same for two encounters.
2. Admit date is continuous, count it as 1 encounter.
3. If there are duplicates with inpatient and emergency patients keep in patient record

The detailed python code can be found on GitHub https://github.com/mraunak/Research-Assistant-Work-/blob/main/Data%20processing%20using%20Pandas.ipynb

**Getting basic information about the data**

Number of rows: thirty million four hundred nineteen thousand ninety

Number of columns: 27

Name of Columns in Data Set

'PATIENT\_ID', 'PROVIDER\_IDS', 'PROVIDER\_TYPES', 'CARE\_SETTING\_CATEGORY',

'ADMIT\_DATE', 'DISCHARGE\_DATE', 'DISCHARGE\_DISPOSITION',

'OVERDOSE\_FLAG', 'OD\_DX\_DATE', 'OD\_NOTE\_DATE', 'OD\_TERMS',

'OD\_DRUG\_TERMS', 'OD\_AR\_FLAG', 'ADVERSE\_OPIOID\_EVENT\_FLAG',

'ADVERSE\_OPIOID\_EVENT\_DRUGS', 'ADVERSE\_OPIOID\_EVENT\_PROBLEMS',

'COMMERCIAL\_FLAG', 'OTHERGOV\_FLAG', 'SELFPAY\_FLAG', 'WORKERSCOMP\_FLAG',

'INSTITUTIONALIZED\_FLAG', 'CHARITY\_FLAG', 'MEDICARE\_FLAG',

'MEDICAID\_FLAG', 'LIFE\_FLAG', 'MDWISE\_FLAG', 'NODATA\_FLAG']

Task 1:

Find the following details:

Maximum number of visits by any patient: 1681

Minimum Number of Visits by any patient: 1

Patient id having maximum number of visits: 480347.

Average number of visits by patients overall: 24.919820248084704

Median number of visits by patients overall: 16

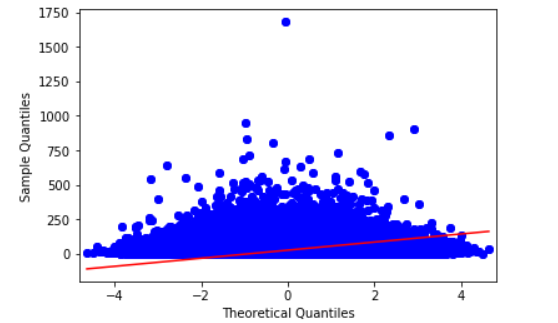
Find the total counts of patient id’s who visited more than the mean visits: 210253.

Find the patient id’s who visited more than the mean visits.

Count of outliers: 10610

Patient\_id which are outliers: line101

Q-Q plot on number of visits



Task 2

1. Explain the column Provider Types
2. Find the unique number of provider types: 3144
3. Clean the data in order to make it better for understanding

e.g. convert 'NA|NA|NA|NA', 'NA|NA|NA|NA|NA|NA|NA' to nan

Convert 'NA|NA|MEDICINE|INFECTIOUS D|INFECTIOUS D|INFECTIOUS DISE|NA'

To 'MEDICINE|INFECTIOUS D|INFECTIOUS D|INFECTIOUS DISE'

Task3: Perform Analyze the patients encounter with the hospitals based on 'CARE\_SETTING\_CATEGORY'

1. In Patient: Patient who are admitted to hospital normally.
2. Emergency: Patient who are admitted to hospital in emergency ward
3. Outgoing: Patient discharged from the hospital
4. Count of visits in each category

* O 24328691
* E 4474888
* I 1615511

1. How many unique patients each category has—465924
2. Unique Patients for I category ---- 357034

### Unique Patients for O category ---- 604026

### How many visit each patient had in each category---dummy e.g. how many visits olena has in O,E,I Categories?

### 

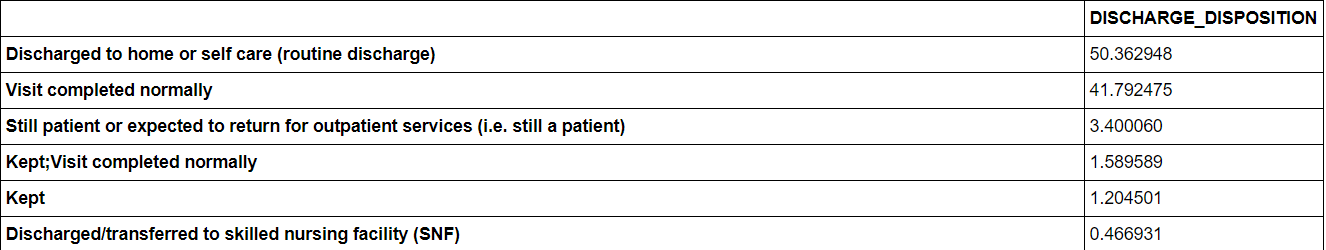
### BoxPlot: Number of Outliers in each category

### 

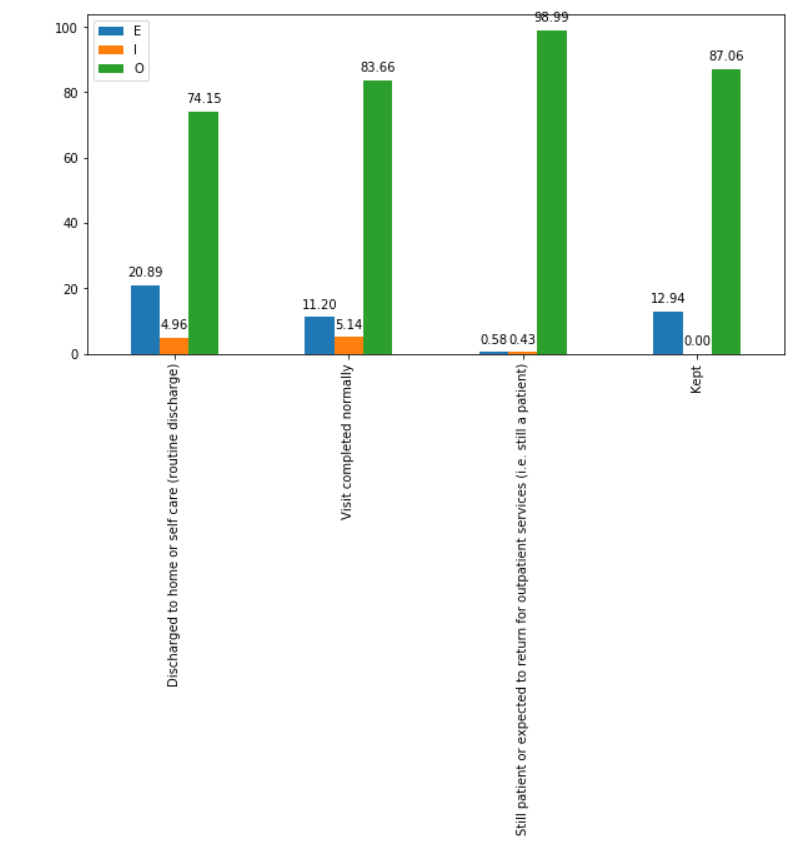
## Explanation of DISCHARGE\_DISPOSITION

### Find the number of unique types of Discharge disposition: 91

##### Percentage of Missing value

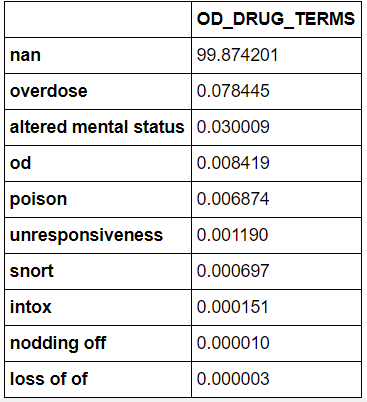
1. DISCHARGE\_DISPOSITION 56.242162
2. Distribution of Discharge Disposition
3. 

Distribution of Discharge Disposition vs Care Setting Category



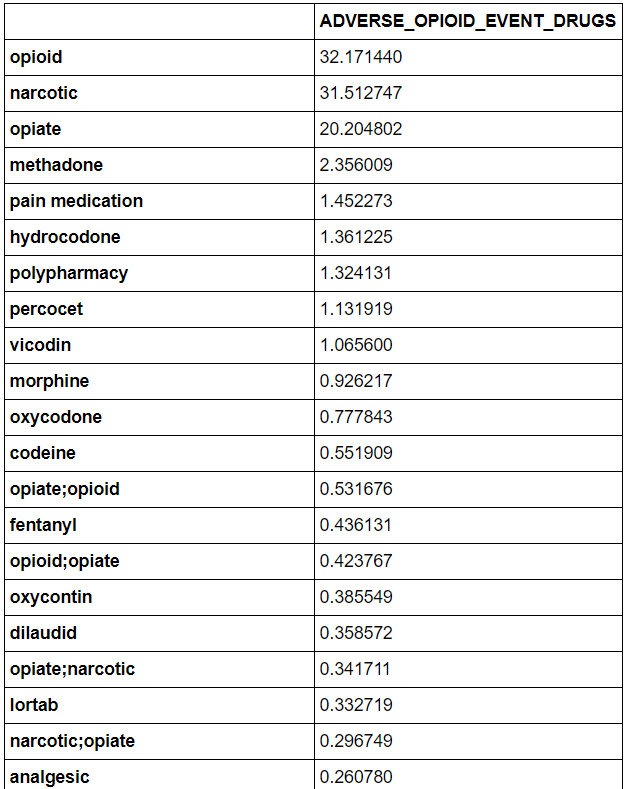
## Explaination of OD\_DRUG\_TERMS

1. Percentage of missing values 99.869365



## Explanation of ADVERSE\_OPIOID\_EVENT\_DRUGS with their percentage for all the patients who visited the hospital.

There are 209 unique ADVERSE\_OPIOID\_EVENT\_DRUGS



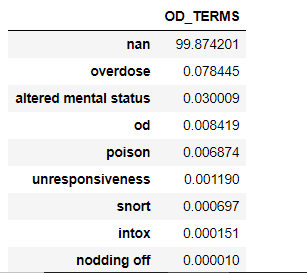
## Explaination of OD\_TERMS

1. Data Cleaning

Convert 'altered\nmental status' or 'altered mental\nstatus','altered mental \nstatus','altered mental \ or \nstatus','altered\n\nmental status' to 'altered mental status'

Convert ‘poisoning','poisoned' to poisoned

Percentage distribution of OD\_Terms



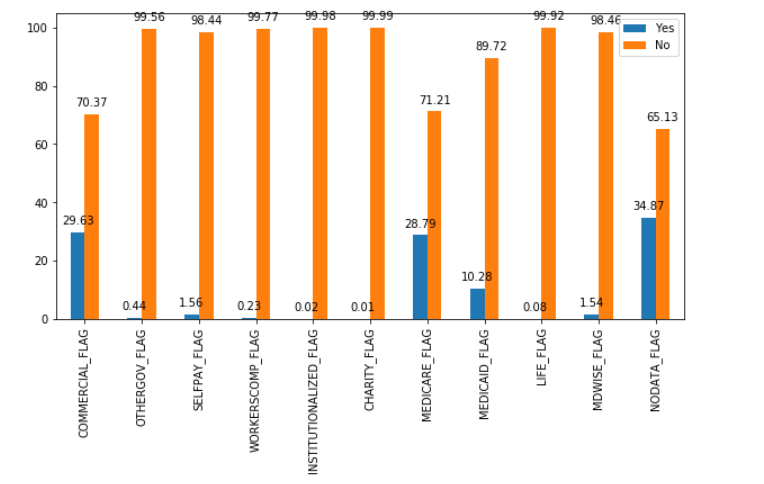
Detailed Analysis of Insurance Flags

Different Types of Insurance Flags are:

'COMMERCIAL\_FLAG', 'OTHERGOV\_FLAG', 'SELFPAY\_FLAG', 'WORKERSCOMP\_FLAG',

'INSTITUTIONALIZED\_FLAG', 'CHARITY\_FLAG', 'MEDICARE\_FLAG',

'MEDICAID\_FLAG', 'LIFE\_FLAG', 'MDWISE\_FLAG', 'NODATA\_FLAG'



Tasks to calculate the number of patients who visited the hospital with different types of insurance

flags

##### Calculate Number of Visits with COMMERCIAL\_FLAG and MEDICARE\_FLAG: 944019

##### Calculate Number of Visits with COMMERCIAL\_FLAG and MEDICAID\_FLAG: 708976

Calculate Number of Visits with COMMERCIAL\_FLAG OTHERGOV\_FLAG NODATA\_FLAG:

83

Detailed Analysis is mentioned at

https://github.com/mraunak/Research-Assistant-Work-/blob/main/Data%20Analytics%20in%20Health%20Care\_%20RA.ipynb

Task: Resolved 3 major issues

1. Admit date and Discharge date being same for two encounters.
2. Admit date is continuous, count it as 1 encounter.
3. If there are duplicates with inpatient and emergency patients keep in patient record

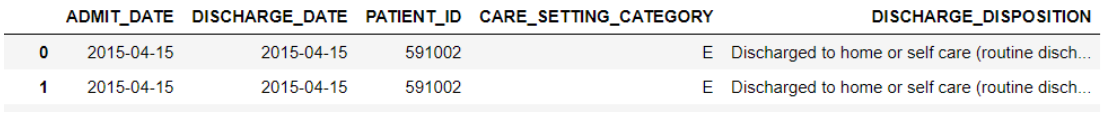
The detailed python code can be found on GitHub https://github.com/mraunak/Research-Assistant-Work-/blob/main/Data%20processing%20using%20Pandas.ipynb

Data Processing

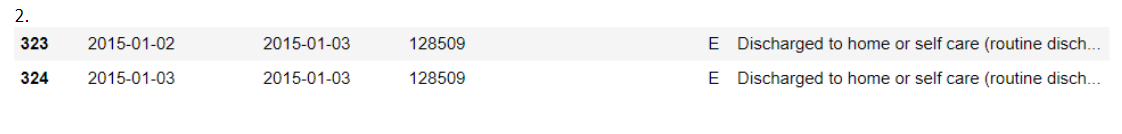
After the discussion conclusion was to use data after the date October 2014

Issues

There are two scenarios:

1. Admit date and Discharge date being same for two encounters: 

Solution: remove the duplicates



In this case, we were considering these both encounters as a single encounter as the admit date were continuous and were ignoring the discharge date, these two are unique encounters as the patient was admitted on 2nd and was discharged on 3 rd and again was admitted on 3rd.

Solution: merge the two rows and final output: 2015-01-02 – 2015-01-03 only one line

3. If there are duplicates with inpatient and emergency patients keep in patient record