APPLYING DEEP LEARNING TO HOSPITALIZATION DATA

Springboard: Capstone 2

Ivette M Tapia

TABLE OF CONTENTS

INTRODUCTION

This report summarizes the overall process taken to build a deep learning network to predict a patient's procedure based on the hospitalization authorization data. Deep learning networks is a class of machine learning algorithms. However, deep learning differs from other machine learning algorithms in that:

1. Deep learning uses multiple layers for feature extraction and transformation. Each layer uses the output of the other layer as input. As such, it can handle fairly well non-linear processing and patterns.
2. It can learn in supervised or unsupervised manner.
3. Can learn multiple levels of representation, concepts, and abstraction.
4. Given all the layers and abstraction, it is often difficult to interpret and can be 'black-box'.

The data used to train the neural network is from the Authorization for Hospital Admission. This dataset is part of Brazil's SIHSUS Hospital Information System. This system manages the coordination and payment by Brazil's public healthcare system. The data is publically available on the web. In this application, I will be using data from 2015 – 2018. This represents 3.5 years' of data. A record in the AIH database is created when a hospital or healthcare unit generates a request for hospitalization. This dataset is large and highly dimensional.

The report starts by outlining the motivation for the project. After the motivation and background section, file conversation, extraction and initial data wrangling are discussed. In between, a summary of all the features available and action taken on the feature. After this, exploratory data analysis and further data wrangling on the features are discussed. Afterwards, feature engineering performed to enhance usefulness of the features and further reduce dimensionality is discussed. Finally, the deep neural network is described and predictive performance results are discussed.
The coding language used throughout this project is R, and python. The coding interface is Jupyter notebooks. The deliverables for the project is python code, a

detailed report, and presentation slides. Throughout the report references to specific notebooks will be provided.

MOTIVATION & BACKGROUND

Healthcare records have become increasingly more digitized. This is an open an opportunity to analyze and obtain patient data at an unprecedented detail and scale. While there is potential to gain greater insights, cost reduction and efficiencies in the healthcare space exist, great challenges remain. Some of these include data availability and complexity of services. Healthcare is particularly complex due to overlapping systems, diversity of providers, services and health issues.

This project would uses hospitalization authorization data that is publically available through the informatics department of Brazilian Ministry of Health. A deep neural network will be used to make predictions regarding procedures performed on a patient given the information given on a patient's authorization request. The input data is both categorical and numerical, and the output is categorical (procedures group performed). Originally, the project was going to use procedure performed as the output (y) variable. However, due to lack of computational resources the output variable was shifted to procedure group performed to save memory and simplify the deep neural network model.

Ability to accurately predict these three features of hospitalization can yield significant benefits. For example, knowing how many days a patient can be expected to be in the hospital will help hospital managers manage their capacity (especially in areas where beds are scarce). Length of stay and likely procedures can inform service charges and help all parties involved navigate the healthcare charge system better so likely costs are known in the front end.

Moreover, predicting healthcare expenditures can be tricky for insurers, providers and particularly consumers. One of the main factors that have been cited as a cause of rising healthcare expenditures is the inability of consumers to know in advance the cost of the healthcare services they consume.

The data that will be used is from the Authorization for Hospital Admission. This dataset is part of Brazil's SIHSUS Hospital Information System. This system manages the coordination and payment by Brazil's public healthcare system (covers around 34% of Brazil's population and pays for 80% of all hospitalizations). The data is publically available as .dbc files on the web. In this application, I will be using data from 2015 – 2018. This represents 3.5 years' of data and 41,537,081 unique hospitalizations.

A record in the AIH database is created when a hospital or healthcare unit generates a request for hospitalization. Providers submit demographic and health information about the patient. This request is approved, reduced, rejected, or rejected due to an error. While the patient is in the hospital, the record is updated to also contain information about procedures performed and discharge. Each row of information represents an hospitalization. If a patient is hospitalized more than 30 days, a new authorization is needed and a new record (i.e. row is created).

DATASET EXTRACTION, CONVERSION & SAMPLING

The dataset extraction was a fairly complex process with multi-steps. This was due to the fact that the data originally in .dbc format, was distributed in hundreds of files, spans multiple years, has millions of observations and is high-dimensional.

This characteristics presents challenges and opportunities. The challenges mainly stem from the file format and large size of the dataset. The .dbc format is propertary to the Brazilian Department of Health. It is basically a compressed version of a dbf files. The opportunities is that there are many features and lots of data to work with in this dataset. I will describe the extraction, conversion and sampling process below.

Step 1: Extraction

The hospitalization data was extracted from the DataSUS servers through the web. The website is as follows:

http://www2.datasus.gov.br/DATASUS/index.php?area=0901&item=1&acao=25.

There are several type of datasets choose from:

- Rejected hospitalization requests
- Professional Services
- Reduced

For this project, I extracted the reduced option. This dataset contains the hospitalization authorization request data. I extracted all the hospitalization data available for the years of 2015, 2016, 2017, and up to July 2018. As highlighted above this data roughly represents 80% of all hospitalizations in Brazil.

The 2015 year has 324 files, the 2016 has 324 files, the 2017 year has 317 files and the 2018 year has 176 files. This is for a total of 1,141 files that need to be converted to make them usable.

Step 2: Conversion from .dbc files to R dataframes to CSV

While it is possible to convert these files using python, R already has a package call **read.dbc** developed by Brazilian researchers specifically written to convert these files. Since a solution already existed in R, I implemented the file conversion in R.
Once all the files were converted to R data frames, I concatenated the data frames by their respective year. Four large data frames were created, one for each year represented. These four data frames were outputted as CSV files for further use. For more details refer to R code here.

Step 3: Create random sample

The entire dataset extracted and converted takes 35GB of memory in pandas and has 113 columns. Given, that the project involves creating a neural network, the computational resources are not available to use the entire dataset.
To solve this problem, I drew a random sample for each year. The goal was to extract 40% of the entire dataset randomly. For reproducibility the random seed throughout was 42.
Each year has different number of observations. To account for this, I forced the sampling process to take the same proportion that year represents of the total observations. The calculation and results were as follows:

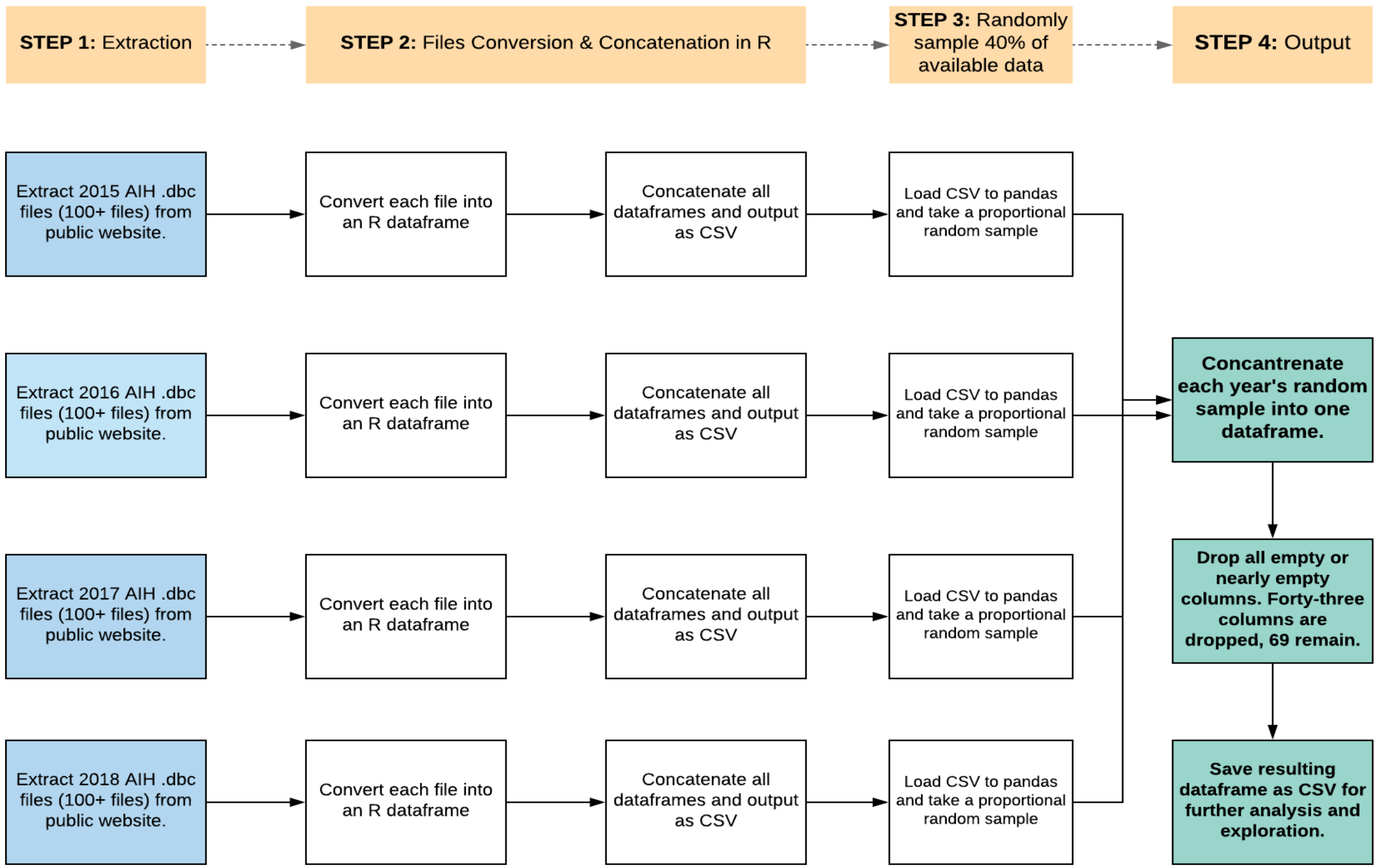
- Total Observations 2016 – 2018: 41,537,081
- 40% of total observations: 16,614,832
    - 2015 (28%): 4,655,541
    - 2016 (28%): 4,611,084
    - 2017 (26%): 4,624,383
    - 2018 (16%): 2,723,822
    - **Total:** 16,614,830

Once the samples has been extracted, I saved the results to as CSV files for further use. For more details refer to the python code here.

Step 4: Data Wrangling – *First Pass*

After the random sample was created for each year, I concatenated the four resulting yearly random sample data frames into one larger data frame. From this data frame, I dropped columns that were either completely empty of had more than 20% missing values. As a result 44 columns dropped, 69 columns / features remained. Please see section *'Dataset Features & Actions'* for a complete listing of features and actions taken on the features. For more details refer to the python code here.

Conversion, Sampling & Wrangling



EXPLORATORY DATA ANALYSIS & WRANGLING

The remaining columns can be grouped into four themes: (1) patient demographics, (2) patient diagnosis, (3) hospitalization services and, (4) financial features, and (5) auditor metadata. Given the large number of features remaining and keeping with the main interests of this project, I decided to not use the financial features and auditing metadata and focus on the patient demographics, diagnosis and hospitalization services.

PATIENT DEMOGRAPHIC FEATURES

The code for all the analysis and wrangling that will be described below can be found here.

Demographic Features Wrangling Summary

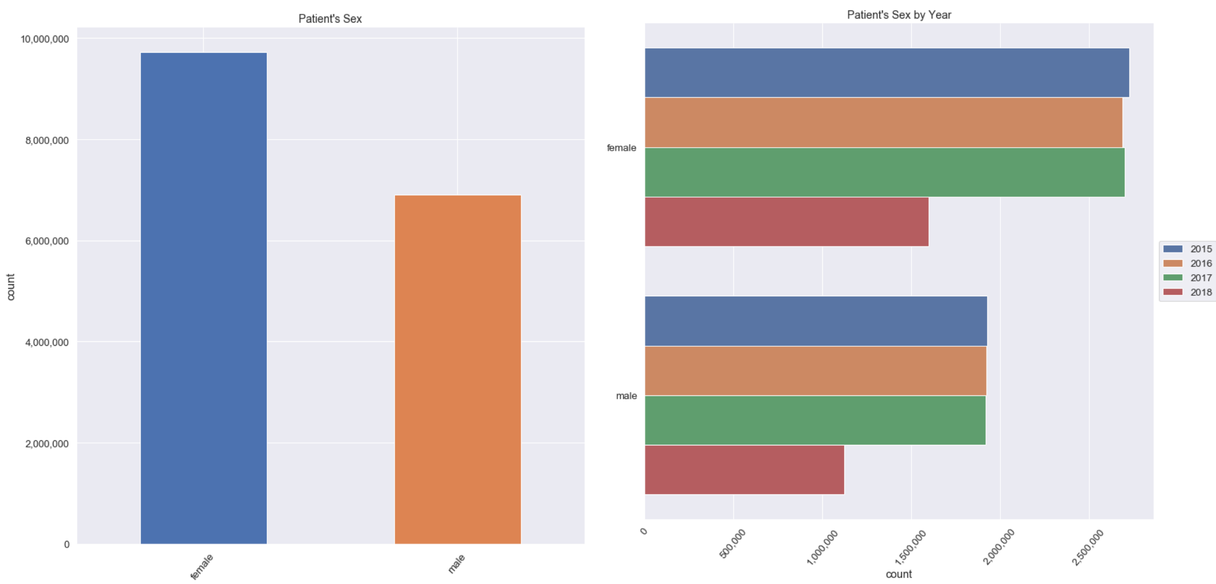
Demographic Feature	Description	Action
MUNIC_RES	Municipality of Residence	Declared categorical. Recoded using pandas cat.codes accessor.
SEXO	Sex	Declared categorical. Recoded using pandas cat.codes accessor.
IDADE	Age	None
MORTE	Death indicator	Declared categorical. Recoded using pandas cat.codes accessor.
NACIONAL	Nationality	Dropped due to extremely low variability.
NUM_FILHOS	Number of Children	Dropped due to concerns of data quality. Suspicion of serious errors in this feature.
INSTRU	Level of education	Dropped because it was more than 20% empty.
GESTRISCO	Pregnant at risk indicator	Dropped due to concerns of data quality. Suspicion of serious errors in this feature.
CBOR	Occupation	Dropped because it was more than 20% empty.
RACA_COR	Race	Declared categorical. Recoded using pandas cat.codes accessor.
ETNIA	Ethnicity	Declared categorical. Recoded using pandas cat.codes

		accessor.

The cleaned dataset was exported as CSV after exploratory analysis. Exploratory analysis will be described below.

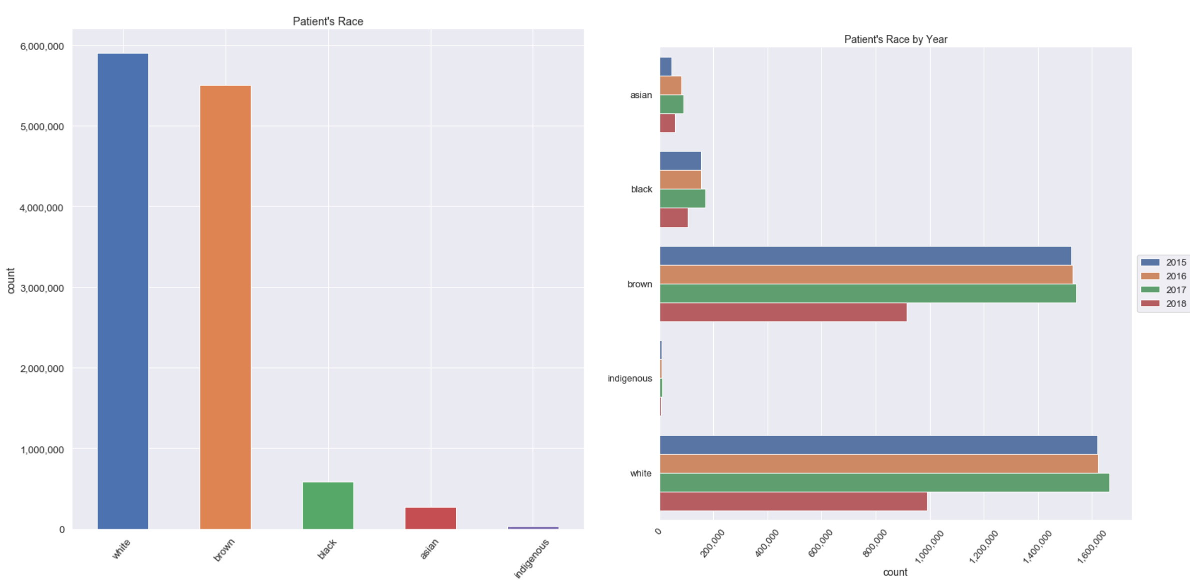
Demographic Features Exploratory Analysis

A. Patient's Sex



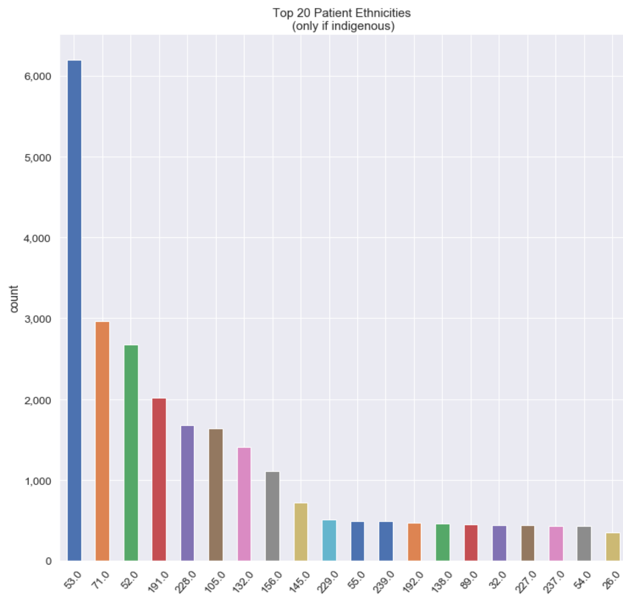
Somewhat more female patients than male. This holds for all the years under consideration.

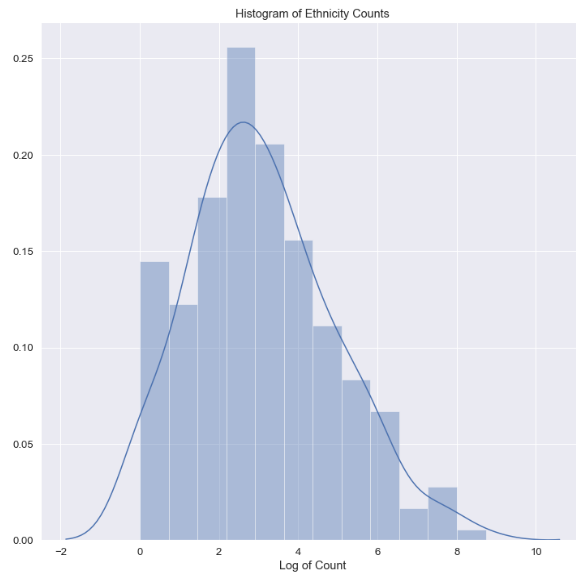
B. Patient's Race

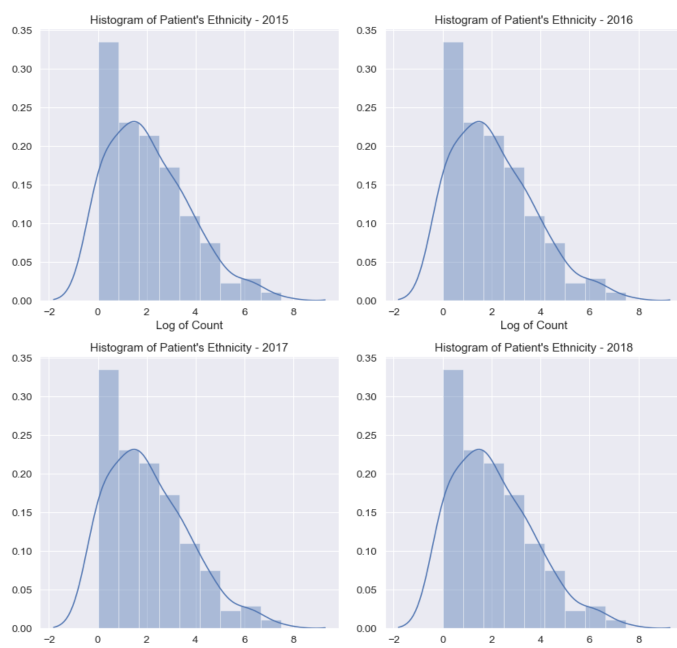


Most patients are identified as white or brown, with black, asian and indigenous being a minority of patients. The number of brown and white patients is close, with white being slightly higher. As highlighted before, this variable has 26% missing values. The pattern holds when broken down by year.

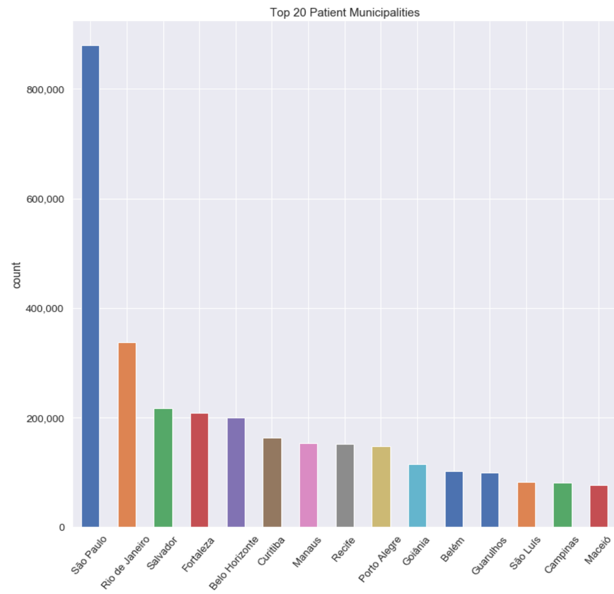
C. Patient’s Ethnicity

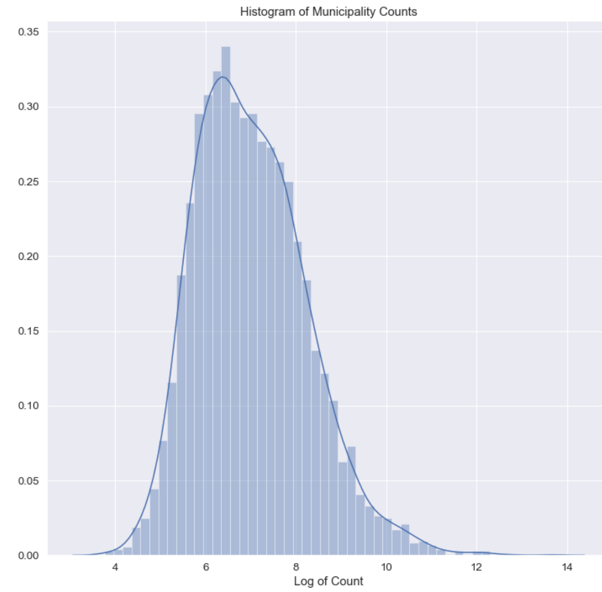


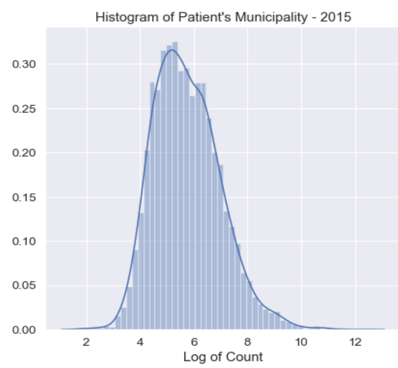


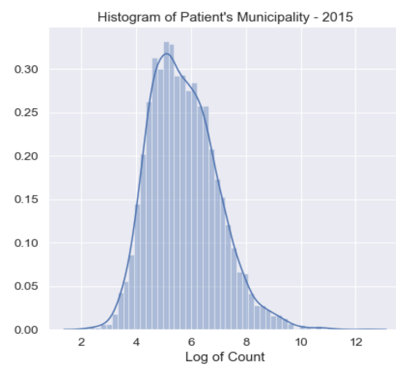


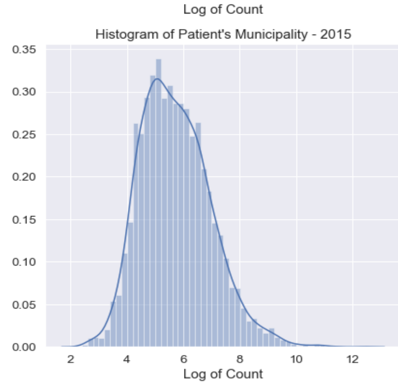
D. Patient’s Municipality

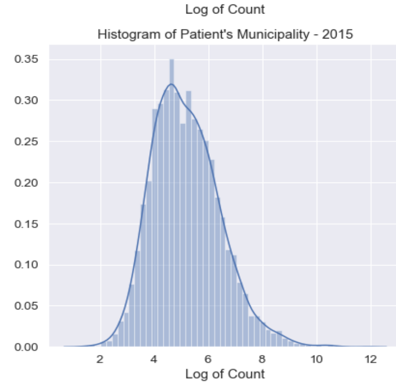






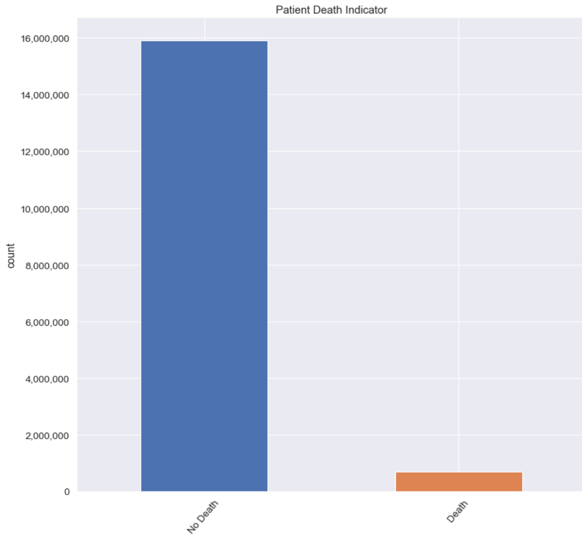


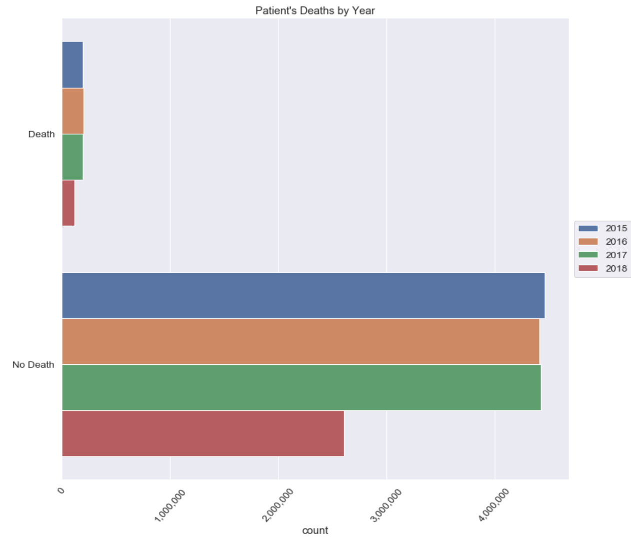




Sao Paulo, Rio de Janeiro and Salvador have the most hospitalization cases in the dataset. These are large cities. Nonetheless, these cities represent a small fraction the overall dataset. The counts follow a somewhat left skewed distribution, with a elongated normal shape.

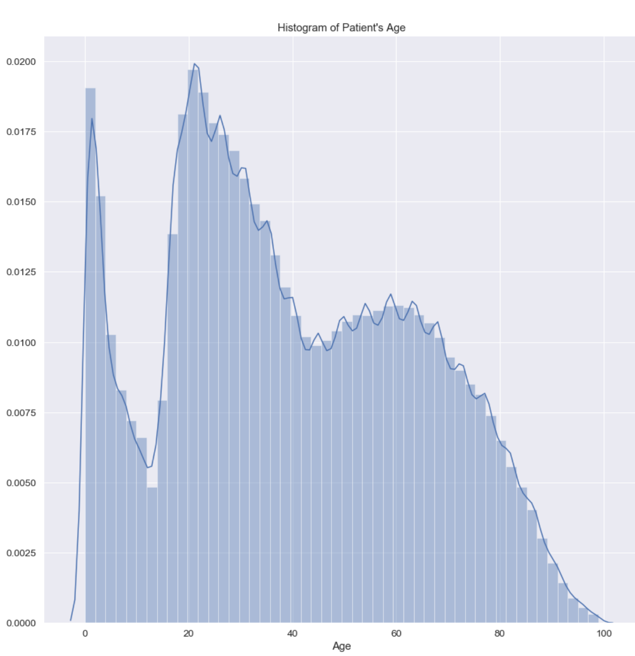
E. Patient Death Indicator

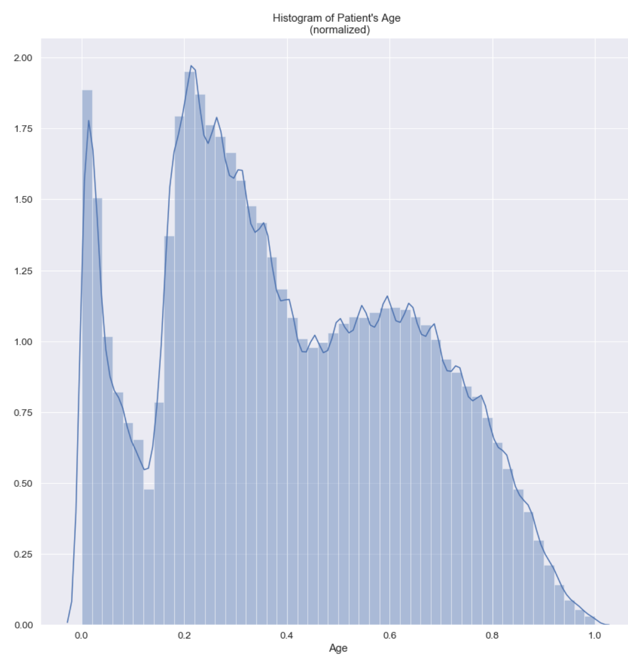




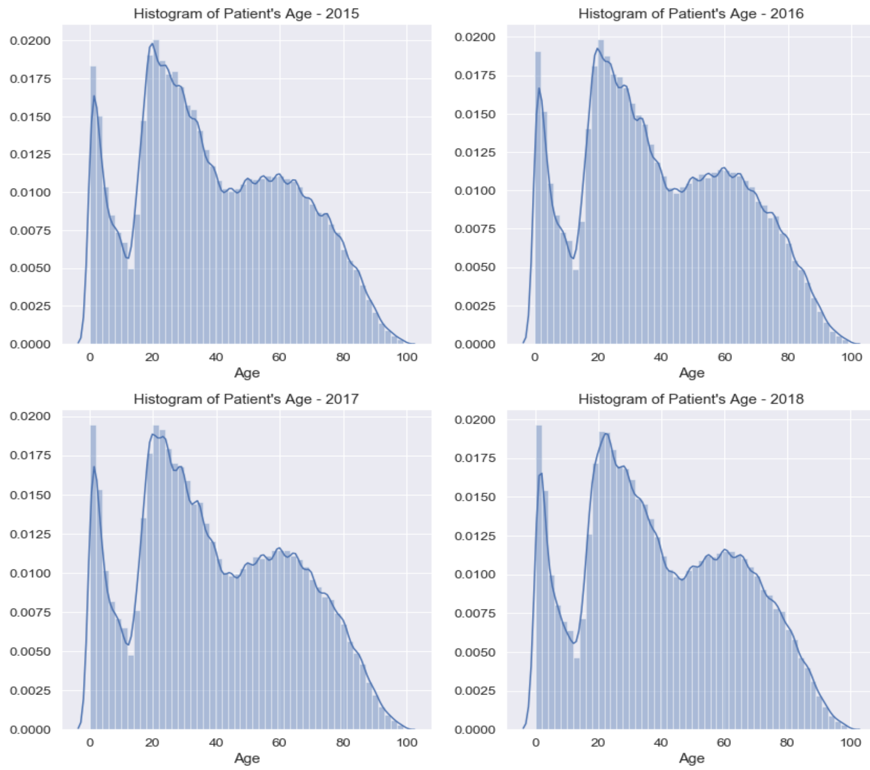
Most patients survive their hospitalization. This variable is very skewed, with a small fraction of patients with the death indicator. The same holds at the year by year breakdowns.

F. Patient's Age

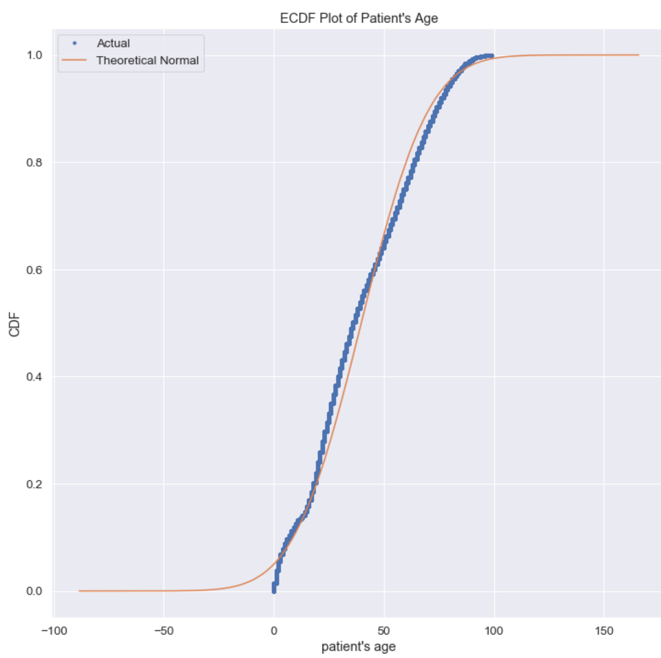




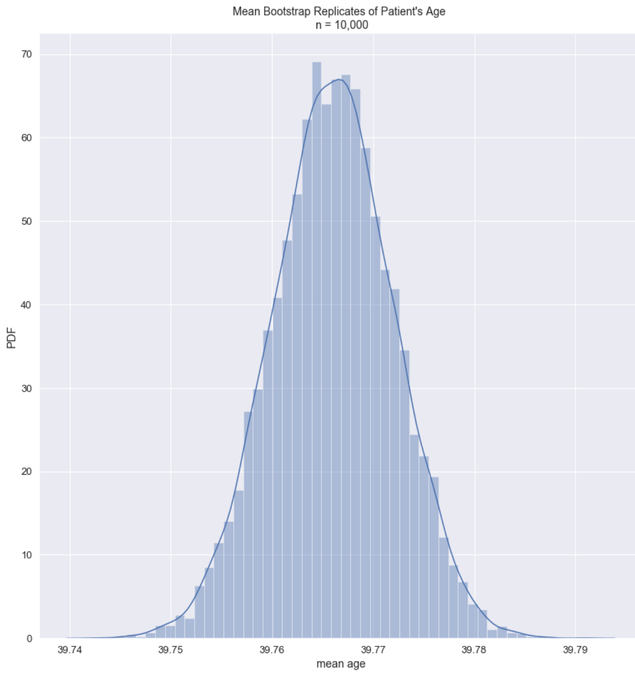
Age has three peaks: (1) around young age, (2) around ~30s, and (3) in late middle age.



The pattern of the three peaks repeat throughout all the four years under consideration.



The ECDF plot shows that the patient's age distribution follows certain parts of the theoretical normal and slightly diverges at other points. The age variable cannot be negative and is unlikely to pass 100 (given human life expectancies), so it cannot completely follow a normal theoretical distribution. Normality tests[1] suggests that the distribution is not normally distributed.

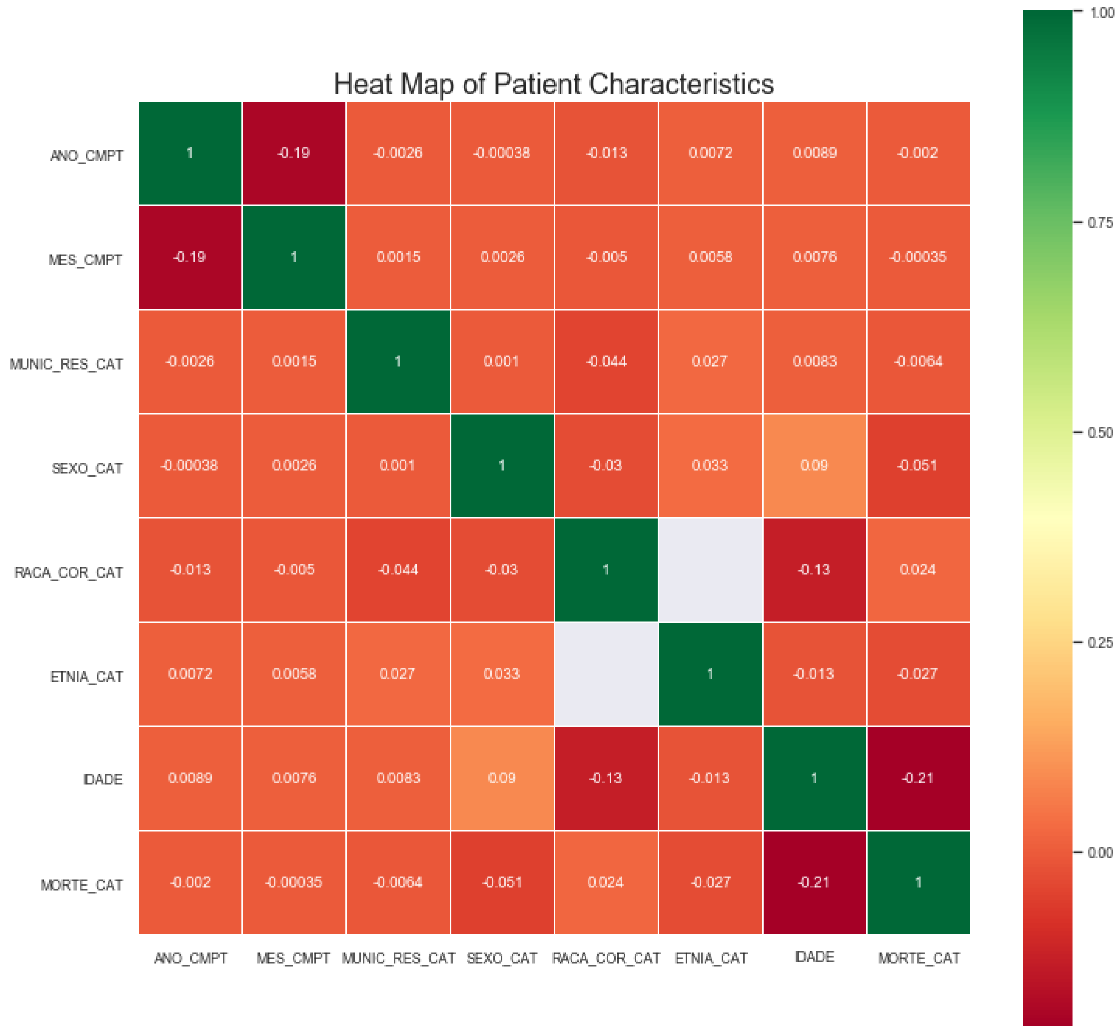


Bootstrap 95% Confidence Interval: [39.75 – 39.77], p-value = 0.4983

The bootstrap mean replicates show a 95% confidence interval for the population mean is between 47.34 and 47.57. This confidence interval is very close together. This range contains the sample mean of 39.7 years. The p-value is 0.49 which is above the set alpha level of 0.05, this means we cannot reject the hypothesis that the patient's mean age is 39.7 years. A one sample t-test yielded a similar conclusion.

[1] D' Agostino and Pearson's Normality Test, Distribution Statistics and Anderson-Darling Test. Every time "normality tests" is used in this report it refers to these three tests.

G. Correlation Heat map of Patient's Demographic Features



- None of the features have a strong correlation with each other.
- Most correlations are extremely weak.
- Age and mortality has a weak correlation.

PATIENT DIAGNOSIS FEATURES

The code for all the analysis and wrangling that will be described below can be found here.

Diagnosis Features Wrangling Summary

Diagnosis Feature	Description	Action
DIAG_PRINC	Principal diagnosis (ICD – 10 coding)	Declared categorical. Recoded using pandas cat.codes accessor.
DIAG_SECUN	Secondary diagnosis	Dropped because it was more than 20% empty.
CAP	Diagnosis chapter	Diagnosis chapter. ICD – 10 reference, the AIH dataset does not contain this information. Recoded using pandas cat.codes accessor.
DES_CAP	Diagnosis chapter description	Diagnosis chapter. ICD – 10 reference, the AIH dataset does not contain this information.
DES_GRP	Diagnosis group description	Diagnosis chapter. ICD – 10 reference, the AIH dataset does not contain this information. Recoded using pandas cat.codes accessor.
CAT	Diagnosis category	Diagnosis chapter. ICD – 10 reference, the AIH dataset does not contain this information. Recoded using pandas cat.codes accessor.
DES_CAT	Diagnosis category description	Diagnosis chapter. ICD – 10 reference, the AIH dataset does not contain this information.
SUB_CAT	Diagnosis category sub-category	Diagnosis chapter. ICD – 10 reference, the AIH dataset does not contain this information. Dropped because principal diagnosis and subcategory are the same.
SUBCAT_DES	Diagnosis sub-category description	Diagnosis chapter. ICD – 10 reference, the AIH dataset does not contain this information.

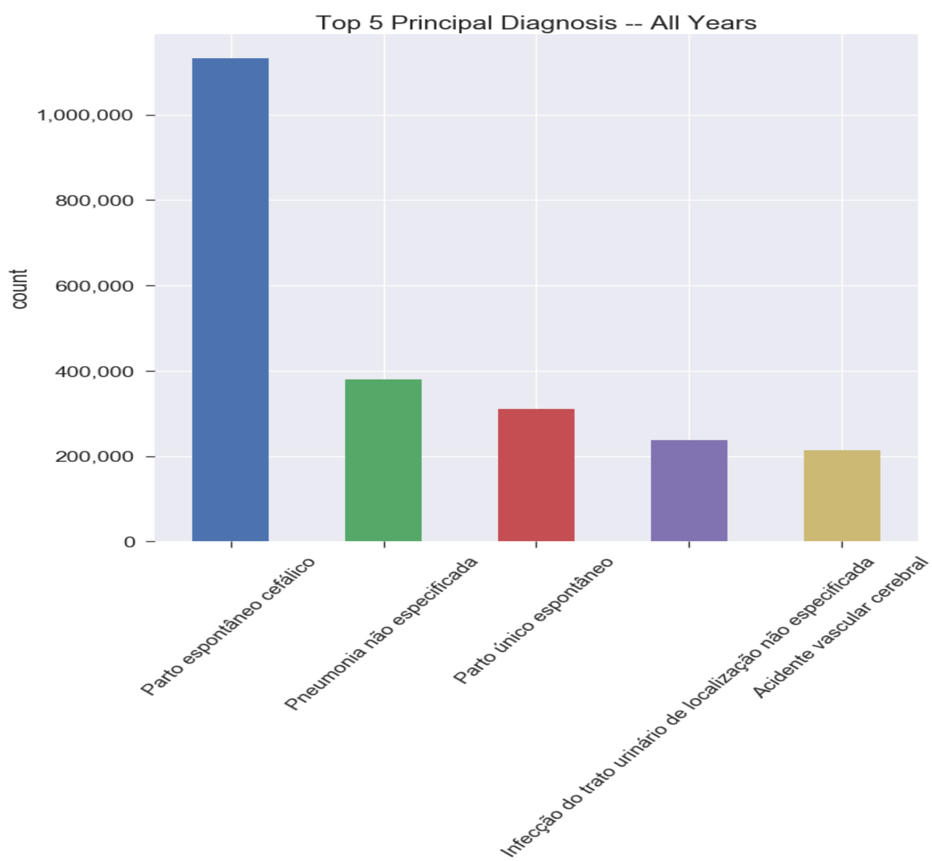
The cleaned dataset was exported as CSV after exploratory analysis. Exploratory analysis will be described below.

Diagnosis Features Exploratory Analysis

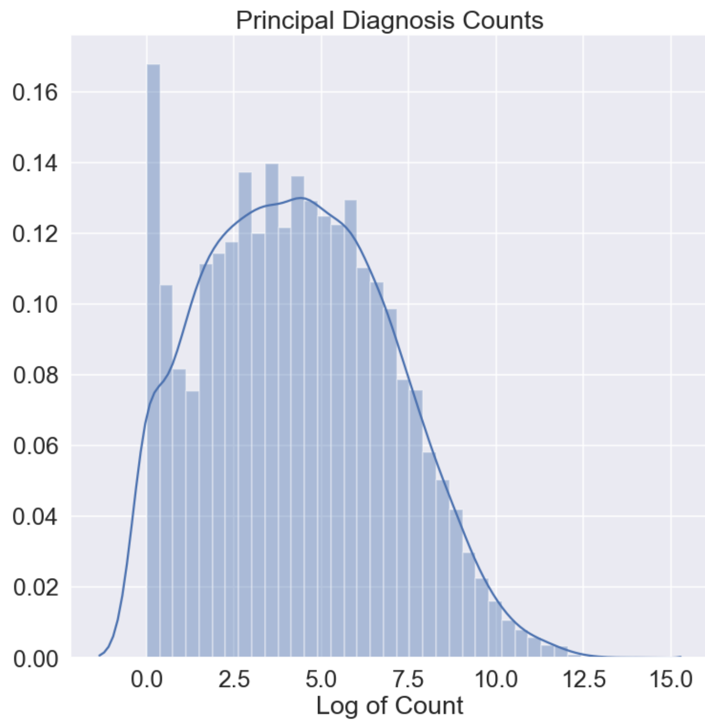
<u>Main Findings:</u>

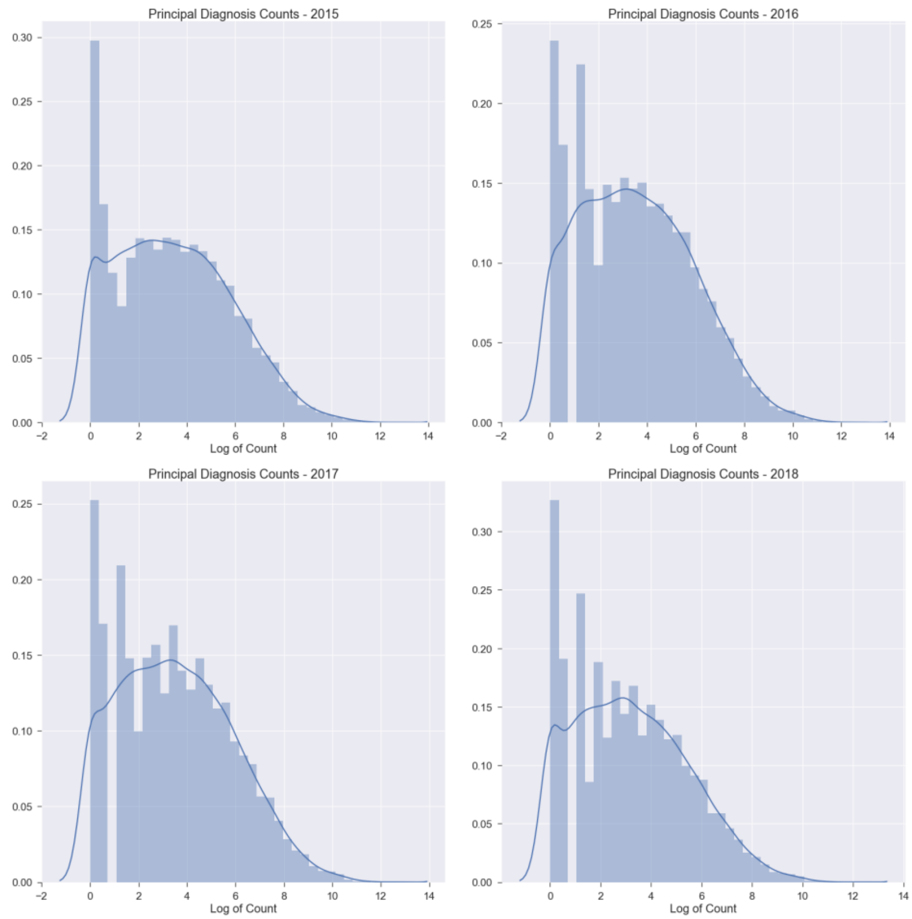
- The diagnosis features are coded using ICD-10 codes. Specifically, providers use the subcategory codes of this codebook to record a patient's diagnosis.
- To aid interpretation I added, the diagnosis codes description, groups, categories and chapters.
- Variables were declared as categorical and re-coded for consistency.
- The distribution of diagnosis is highly unbalanced. While there is diagnosis that are more common than others, the top diagnosis are still a fraction of the total. As such, there is a lot of heterogeneity in this variable.
- The point above is evidenced by the large confidence intervals for the mean diagnosis counts.

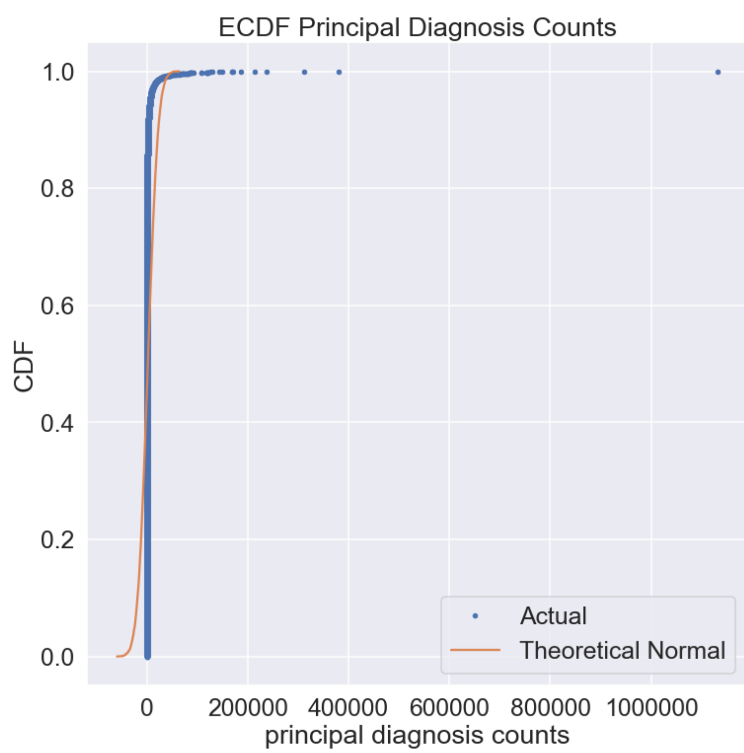
A. Patient's Principal Diagnosis



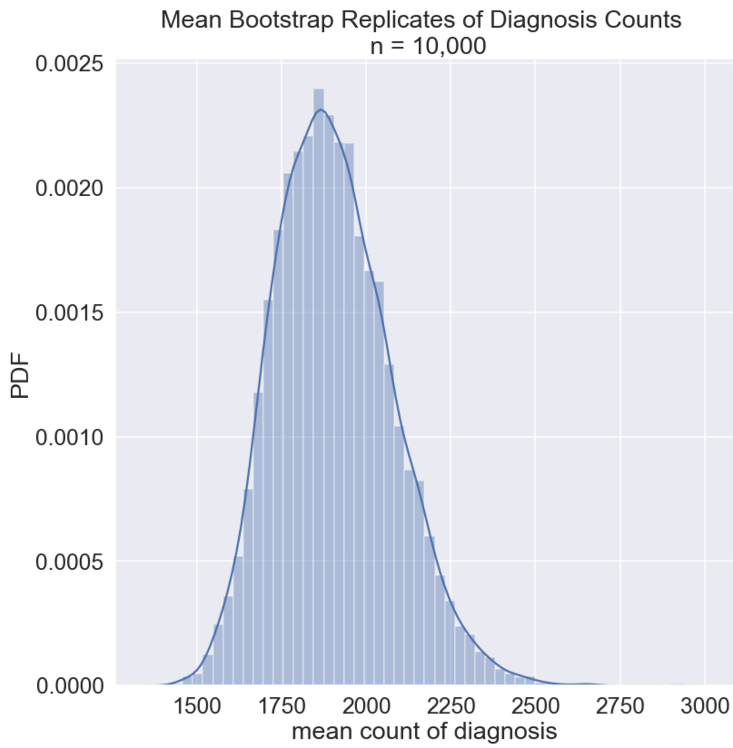
The most common diagnosis is spontaneous cefalic births, pneumonia, unique spontaneous birth. The most common case is birth. However, it is worth noting that the list of diagnoses is long. There are 8,721 unique diagnosis spread over 16+M hospitalizations.




Principal Diagnosis Counts - 2015
Log of Count
Principal Diagnosis Counts - 2016
Log of Count
Principal Diagnosis Counts - 2017
Log of Count
Principal Diagnosis Counts - 2018
Log of Count


ECDF Principal Diagnosis Counts
CDF
principal diagnosis counts
Actual
Theoretical Normal

The ECDF plot shows that the patient's principal distribution follows certain parts of the theoretical normal and slightly diverges at other points. Normality tests[2] suggests that the distribution is not normally distributed.

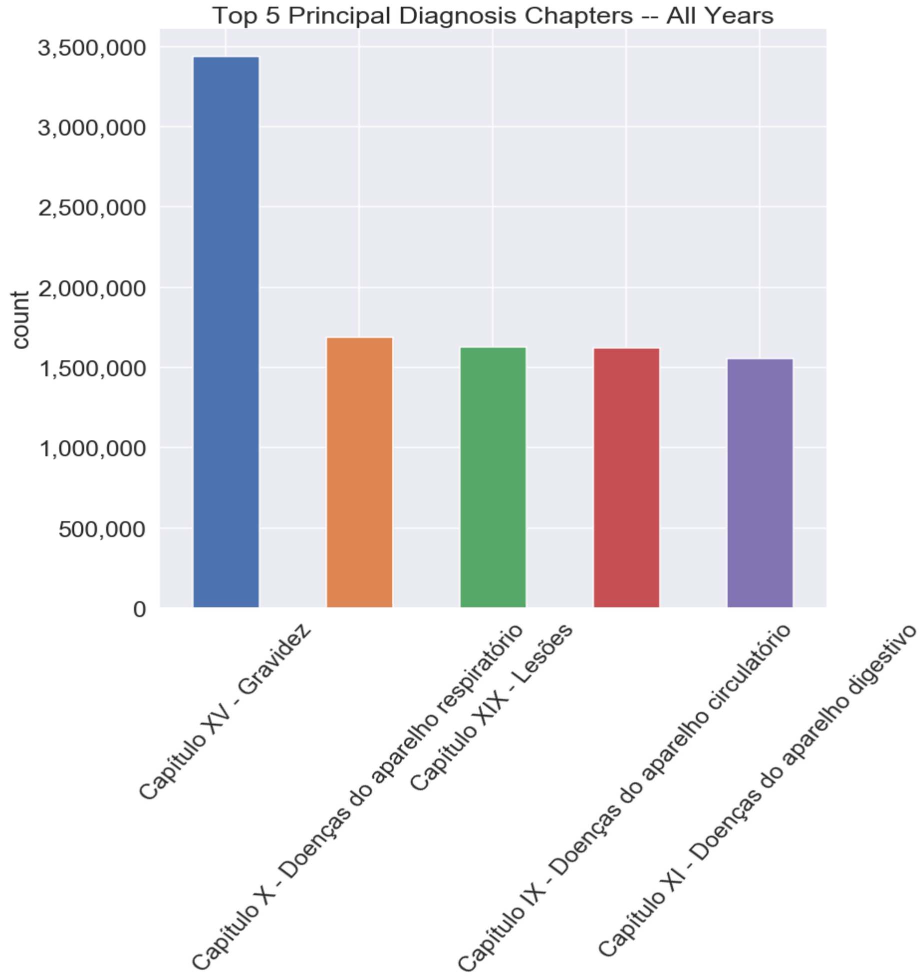


Bootstrap 95% Confidence Interval: [1,607.79 – 2,273.45], p-value = 0.5309

The bootstrap mean replicates show a 95% confidence interval for principal diagnosis counts is between 1,607 and 2,273. This is a wide interval. This range contains the sample mean of 1,905. The p-value is 0.53 which is above the alpha level of 0.05, this means we cannot reject the hypothesis that the mean age is 1,905 cases per principal diagnosis. A one sample t-test yielded a similar conclusion.

[2] D' Agostino and Pearson's Normality Test, Distribution Statistics and Anderson-Darling Test. Every time "normality tests" is used in this report it refers to these three tests.

B. Patient's Principal Diagnosis Chapter



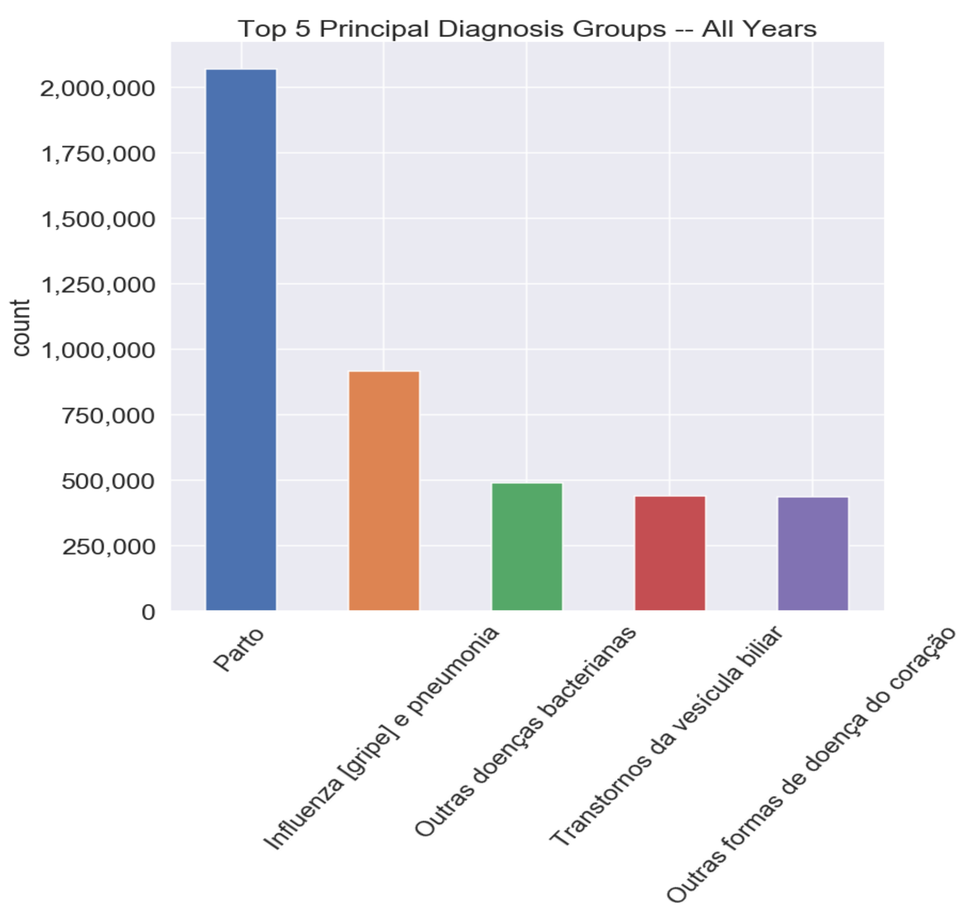
The most common chapters are gravity, respiratory illnesses, lesions, circulatory problems and digestive problems. This pattern repeats for all the years in the sample.

No. Cases per Diagnosis Chapter

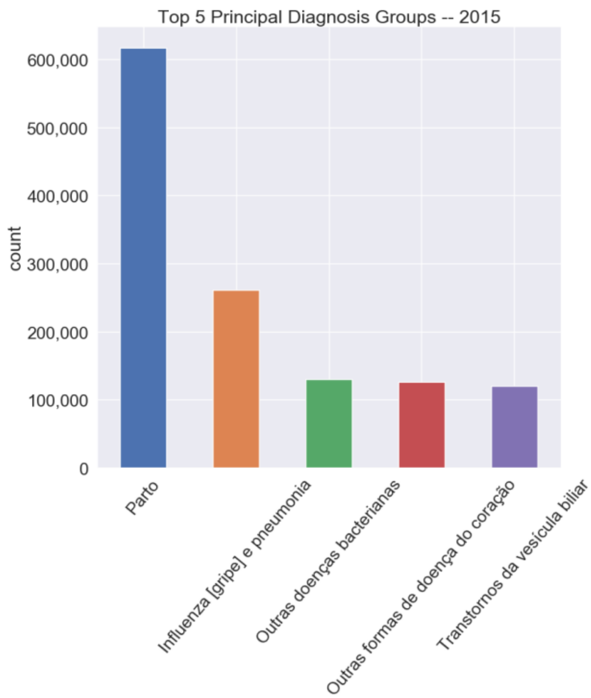
Diagnosis Group	No. Cases
Capítulo XV - Gravidez	3438854
Capítulo X - Doenças do aparelho respiratório	1687929
Capítulo XIX - Lesões	1628304
Capítulo IX - Doenças do aparelho circulatório	1622716
Capítulo XI - Doenças do aparelho digestivo	1558606
Capítulo I - Algumas doenças infecciosas e par...	1135365
Capítulo XIV - Doenças do aparelho geniturinário	1132084
Capítulo II - Neoplasias [tumores]	1113542
Capítulo V - Transtornos mentais e comportamen...	547177

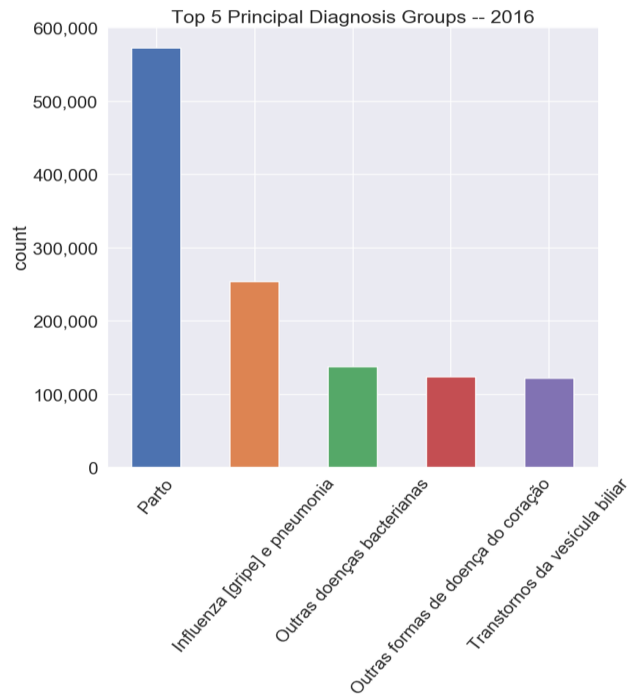
Capítulo XVI - Algumas afecções originadas no ...	385634
Capítulo XII - Doenças da pele e do tecido sub...	359660
Capítulo IV - Doenças endócrinas	352362
Capítulo XXI - Fatores que influenciam o estad...	328253
Capítulo VI - Doenças do sistema nervoso	324632
Capítulo XIII - Doenças do sistema osteomuscul...	299629
Capítulo XVIII - Sintomas	259193
Capítulo VII - Doenças do olho e anexos	155166
Capítulo III - Doenças do sangue e dos órgãos...	141372
Capítulo XVII - Malformações congênitas	116735
Capítulo VIII - Doenças do ouvido e da apófise...	27606
Capítulo XXII - Códigos para propósitos especiais	11

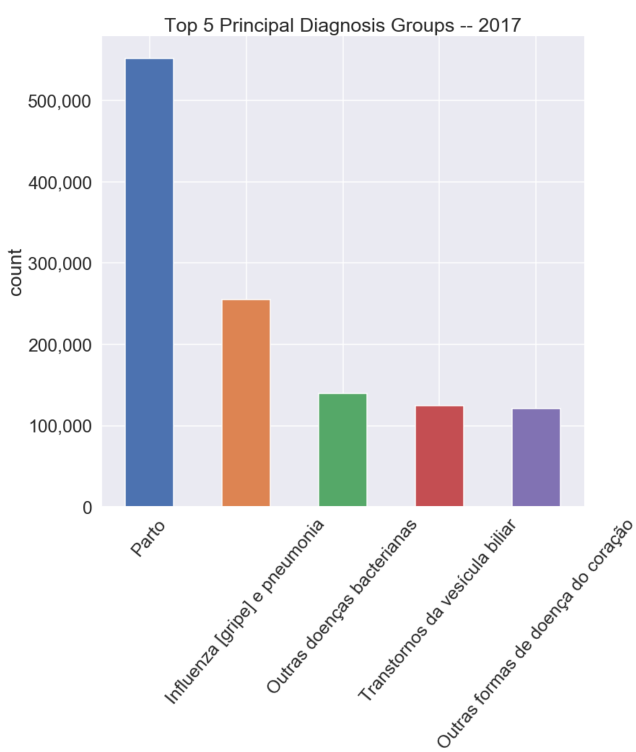
C. Patient's Principal Diagnosis Group

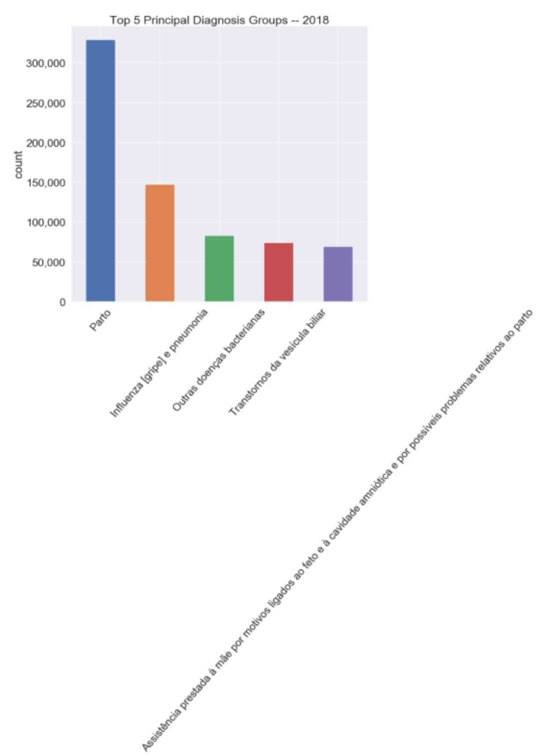


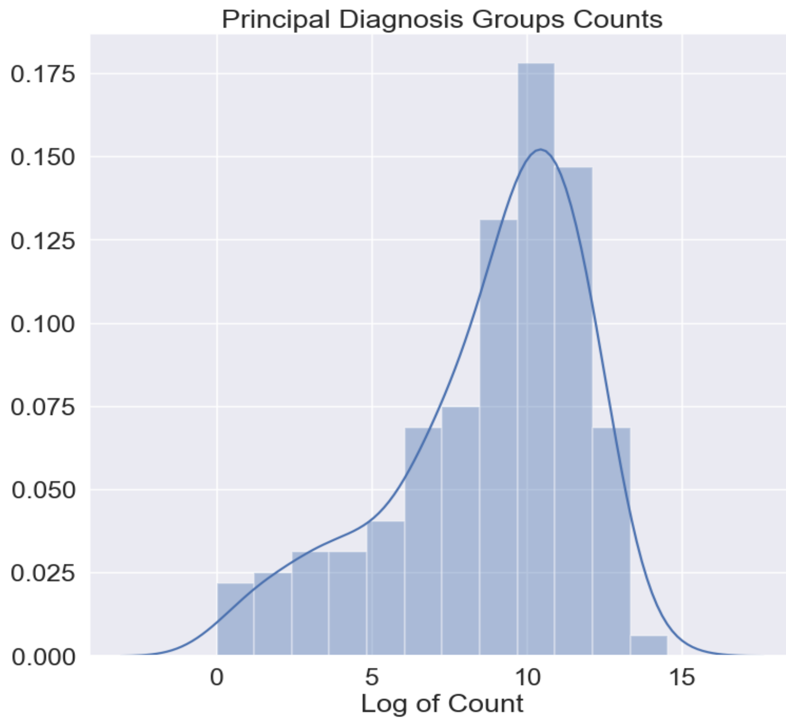
The most common diagnosis is group are birth, pneumonia, bacterial diseases, heart disease, and gallbladder issues. There are 264 unique diagnosis groups in the sample. It is worth noting that there are groups in which cases are very rare.

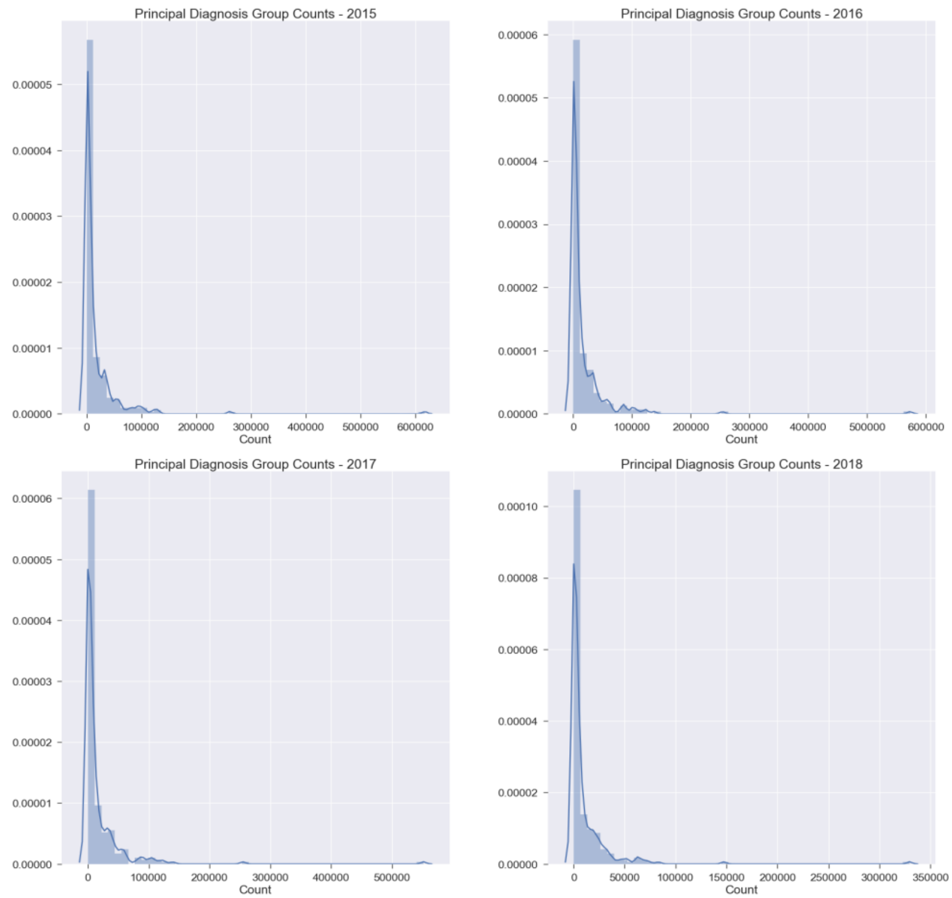

Top 5 Principal Diagnosis Groups -- 2015
count
0
100,000
200,000
300,000
400,000
500,000
600,000
Parto
Influenza [gripe] e pneumonia
Outras doenças bacterianas
Outras formas de doença do coração
Transtornos da vesícula biliar

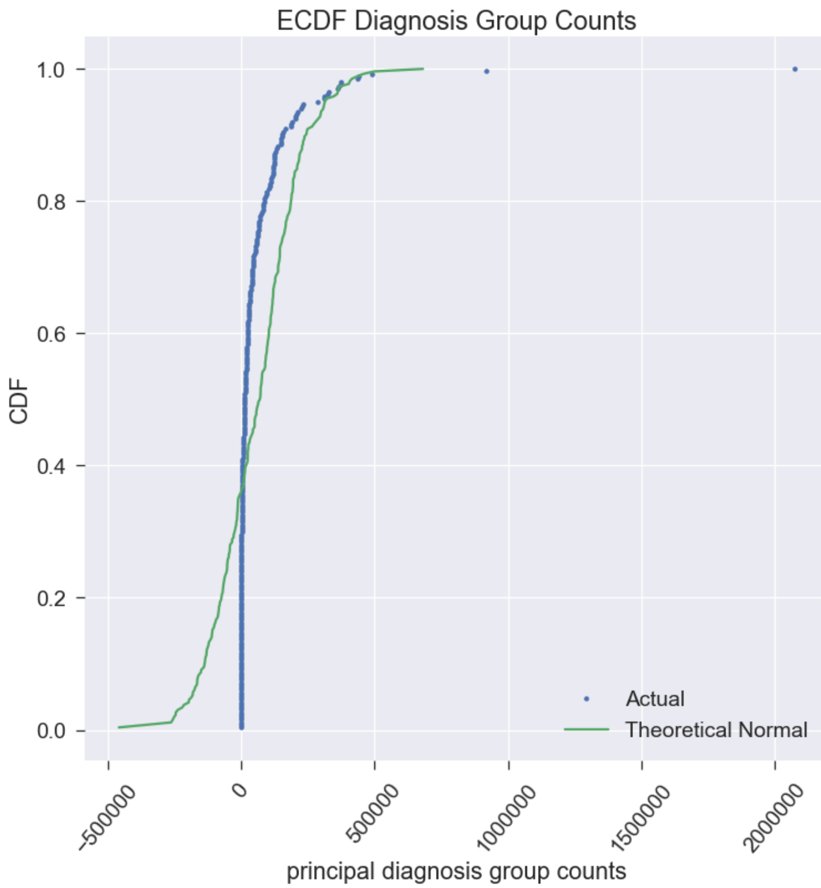

Top 5 Principal Diagnosis Groups -- 2016
count
0
100,000
200,000
300,000
400,000
500,000
600,000
Parto
Influenza [gripe] e pneumonia
Outras doenças bacterianas
Outras formas de doença do coração
Transtornos da vesícula biliar


Top 5 Principal Diagnosis Groups -- 2017
count
0
100,000
200,000
300,000
400,000
500,000
Parto
Influenza [gripe] e pneumonia
Outras doenças bacterianas
Transtornos da vesícula biliar
Outras formas de doença do coração


Top 5 Principal Diagnosis Groups -- 2018
count
0
50,000
100,000
150,000
200,000
250,000
300,000
Parto
Influenza [gripe] e pneumonia
Outras doenças bacterianas
Transtornos da vesícula biliar
Assistência prestada à mãe por motivos ligados ao feto e à cavidade amniótica e por possíveis problemas relativos ao parto

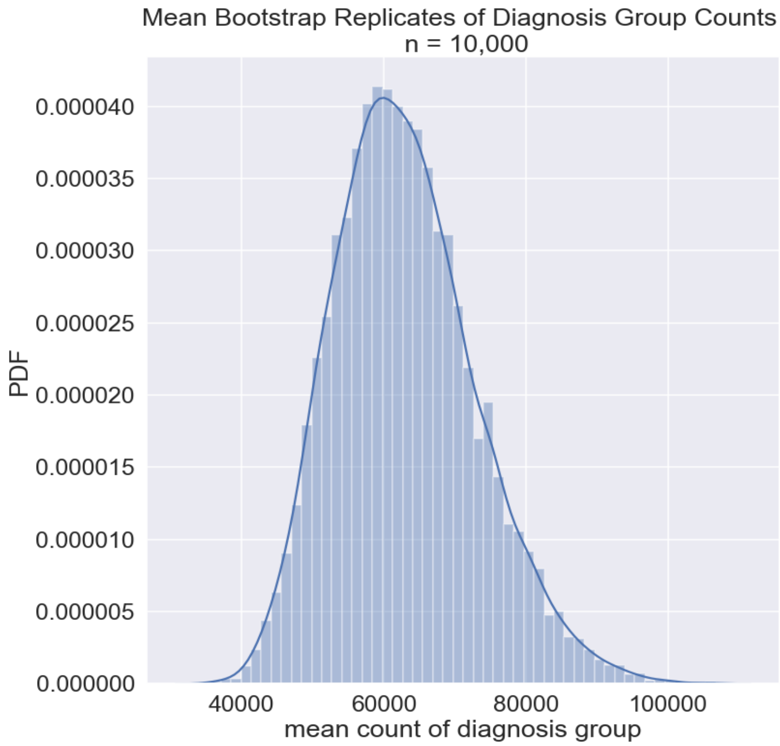

Principal Diagnosis Groups Counts
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
5
10
15
Log of Count


Principal Diagnosis Group Counts - 2015
Count
Principal Diagnosis Group Counts - 2016
Count
Principal Diagnosis Group Counts - 2017
Count
Principal Diagnosis Group Counts - 2018
Count



The ECDF plot shows that the patient's principal diagnosis group does not follow a normal theoretical at all. Normality tests[3] further suggests that the distribution is not normally distributed.

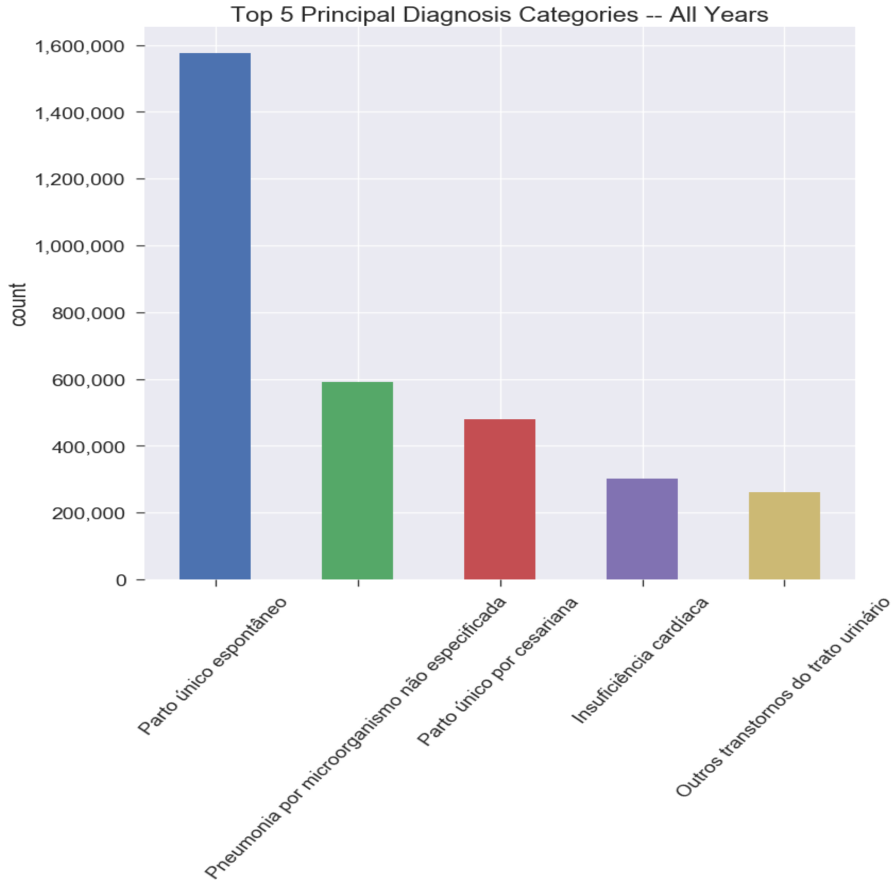
[3] D' Agostino and Pearson's Normality Test, Distribution Statistics and Anderson-Darling Test. Every time "normality tests" is used in this report it refers to these three tests.



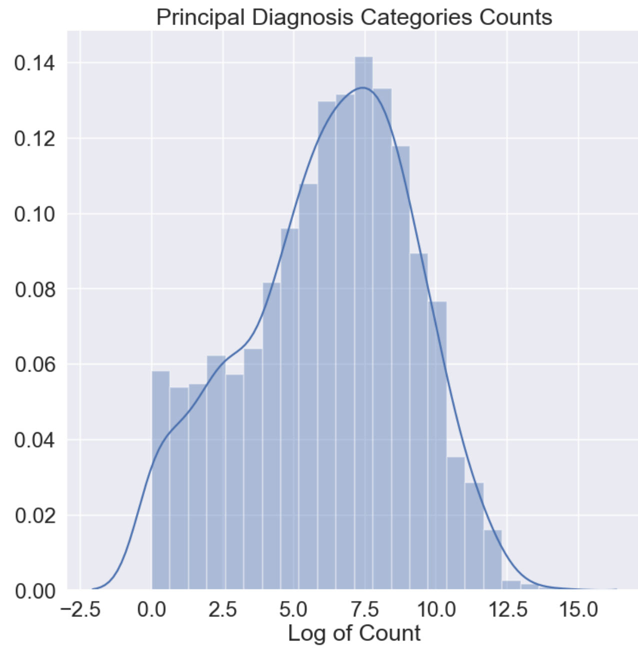
Bootstrap 95% Confidence Interval: [46,009.52– 84,638.17], p-value = 0.5302

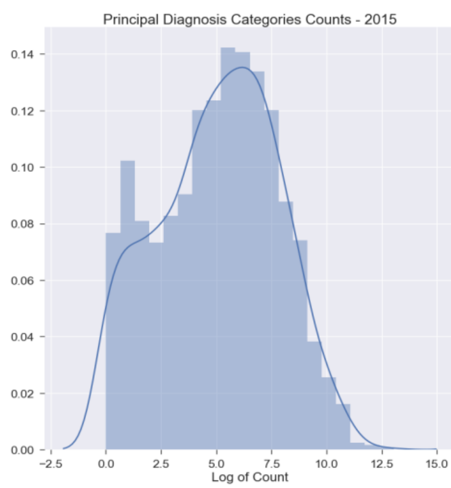
The bootstrap mean replicates shows a 95% confidence interval for diagnosis group counts is between 46,009 and 84,638. This is a very wide interval that suggest a lot of uncertainty about the mean value for each group. This range contains the sample mean of 62,912. The p-value is 0.53 which is above the alpha level of 0.05, this means we cannot reject the hypothesis that the mean age is 1,905 cases per principal diagnosis. A one sample t-test yielded a similar conclusion.

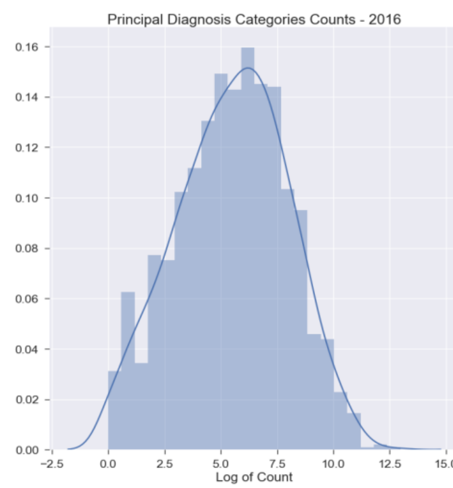
D. Patient’s Principal Diagnosis Category

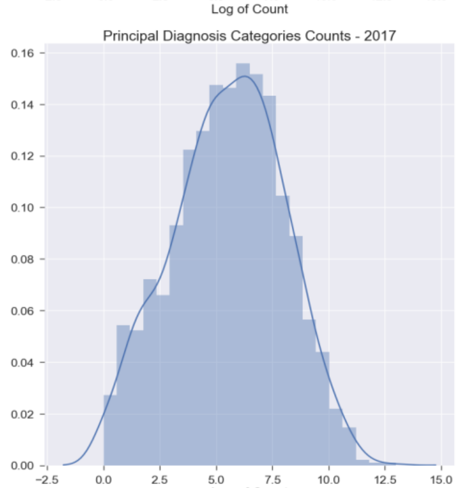


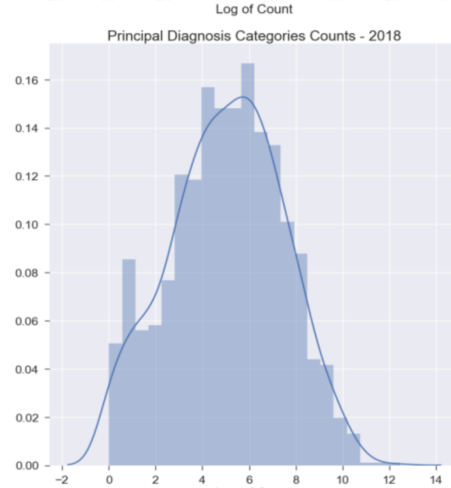
There are 1,829 unique diagnosis categories in the dataset. The most common diagnosis categories are: spontaneous birth, pneumonia by microorganism, birth by cesarean surgery, cardiac insufiency, and urinary tract disorders. This pattern holds for all the years under consideration. While these represent a large number of cases they are still a fraction of the total 16+M cases. It is worth noting that there are categories in which cases are very rare. As such this is somewhat of an unbalanced distribution.

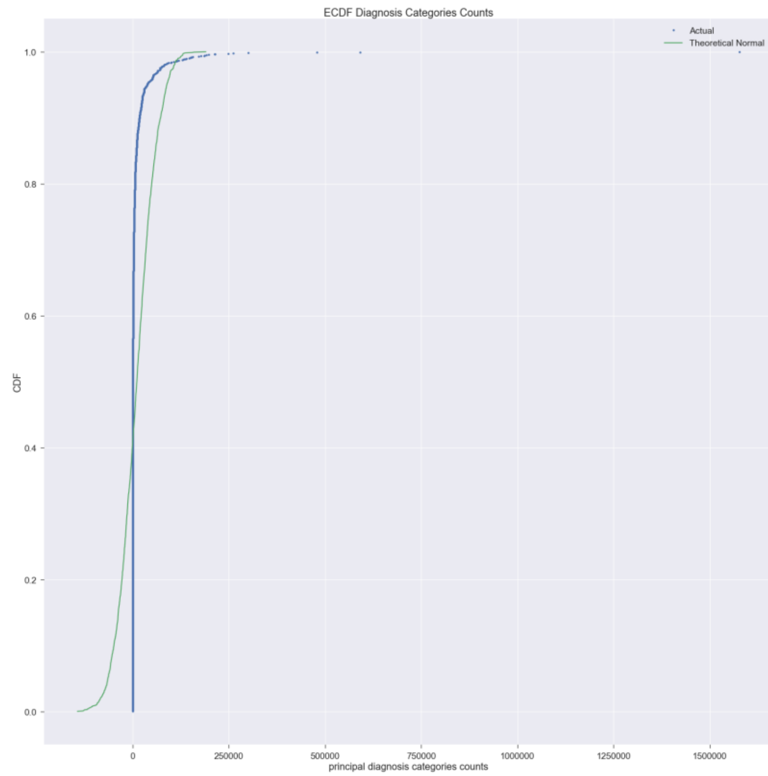

Principal Diagnosis Categories Counts
Log of Count


Principal Diagnosis Categories Counts - 2015
Log of Count


Principal Diagnosis Categories Counts - 2016
Log of Count

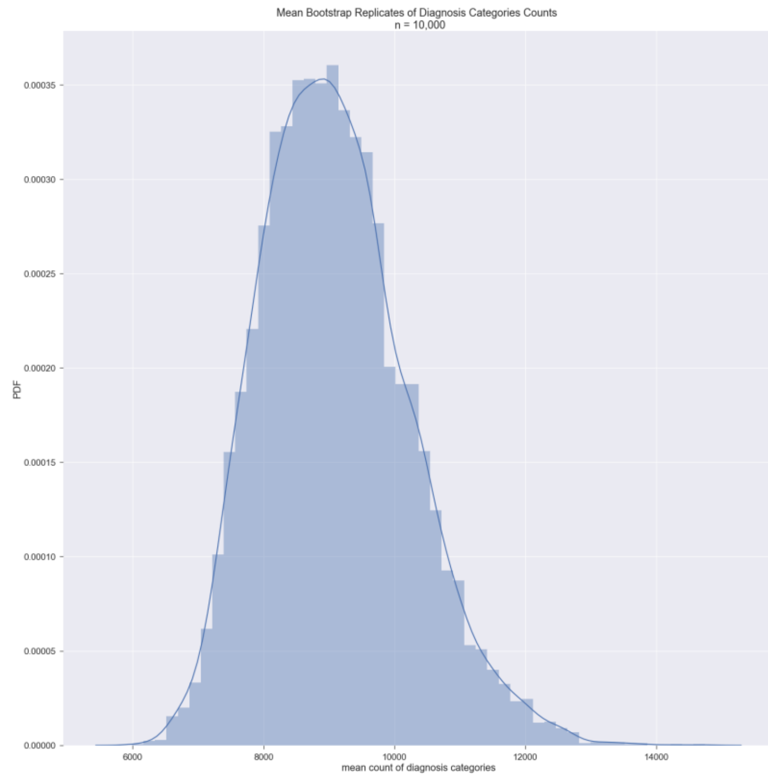

Principal Diagnosis Categories Counts - 2017
Log of Count


Principal Diagnosis Categories Counts - 2018
Log of Count



The ECDF plot shows that the patient's principal diagnosis group does not follow a normal theoretical at all. Normality tests[4] further suggests that the distribution is not normally distributed.

[4] D' Agostino and Pearson's Normality Test, Distribution Statistics and Anderson-Darling Test. Every time "normality tests" is used in this report it refers to these three tests.



Bootstrap 95% Confidence Interval: [7,232.28 – 11,541.97], p-value = 0.5317

The bootstrap mean replicates show a 95% confidence interval for diagnosis categories counts is between 7,232 and 11,541. This is a very wide interval that suggests a lot of uncertainty about the mean value for each group. This range contains the sample mean of 9,084. The p-value is 0.53 which is above the alpha level of 0.05, this means we cannot reject the hypothesis that the mean age is 9,084 cases per principal diagnosis. A one sample t-test yielded a similar conclusion.

HOSPITALIZATION FEATURES

The code for all the analysis and wrangling that will be described below can be found here.

Hospitalization Features Wrangling Summary

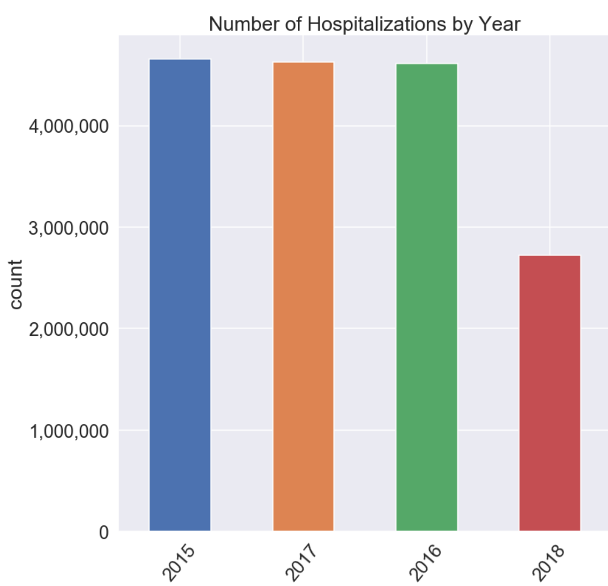
Hospitalization Feature	Description	Action
ESPEC	Type of bed	Declared categorical. Recoded using pandas cat.codes accessor.
CGC_HOSP	Hospital ID	Declared categorical. Recoded using pandas cat.codes accessor.
UTI_MES_TO	Intensive care nights	None
UTI_INT_TO	Num nights in intermediate intensive care	None
DIAR_ACOM	Number of companion nights	None
PROC_REA	Procedure performed code	None
PROC_SOLIC	Procedure requested code	Dropped due to redundancy with high collinearity with procedure realized.
COBRANCA	Reason for stay/exit	Dropped due to concerns of data quality. Suspicion of serious errors in this feature.
IND_VDRL	Venereal exam indicator	Dropped because it was more than 20% empty.
DIAS_PER	Total days of hospitalization	None
QT_DIARIAS	Total number of nights	Dropped due to redundancy with nights of stay.
CAR_INT	Character of hospitalization code	Declared categorical. Recoded using pandas cat.codes accessor.
CONTRACEP1	Contraception used 1	Declared categorical. Recoded using pandas cat.codes accessor.
CONTRACEP2	Contraception used 2	Declared categorical. Recoded using pandas cat.codes accessor.

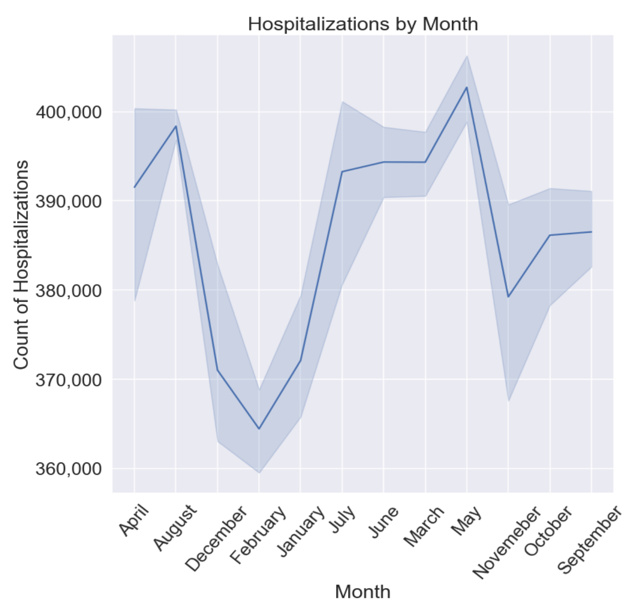
INSC_PN	Number of pregnant women in pre-natal care	Dropped because it was more than 20% empty.
CID_ASSO	ICD-10 code of cause	Dropped because it was more than 20% empty.
CID_MORTE	ICD-10 code of death	Dropped because it was more than 20% empty.
COMPLEX	Complexity code	Declared categorical. Recoded using pandas cat.codes accessor.
MARCA_UCI	Type of intensive care unit used	Dropped due to data quality issues.
DT_SAIDA	Exit date	Dropped due to redundancy with days of stay.
DT_INTER	Entry date	Dropped due to redundancy with days of stay.

The cleaned dataset was exported as CSV after exploratory analysis. Exploratory analysis will be described below.

Hospitalization Features Exploratory Analysis

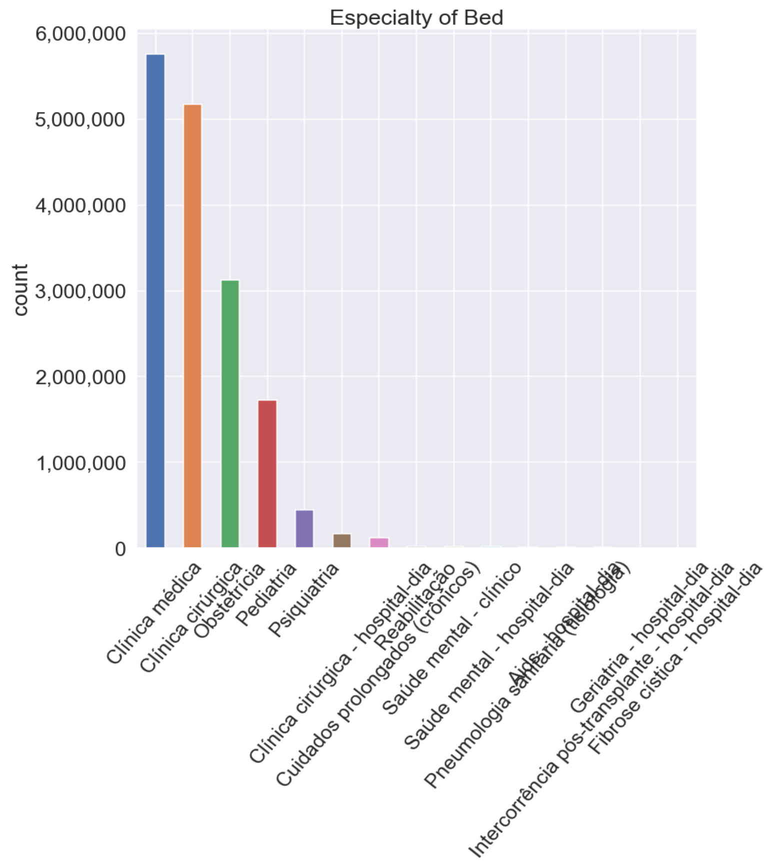
A. Trends in Hospitalization





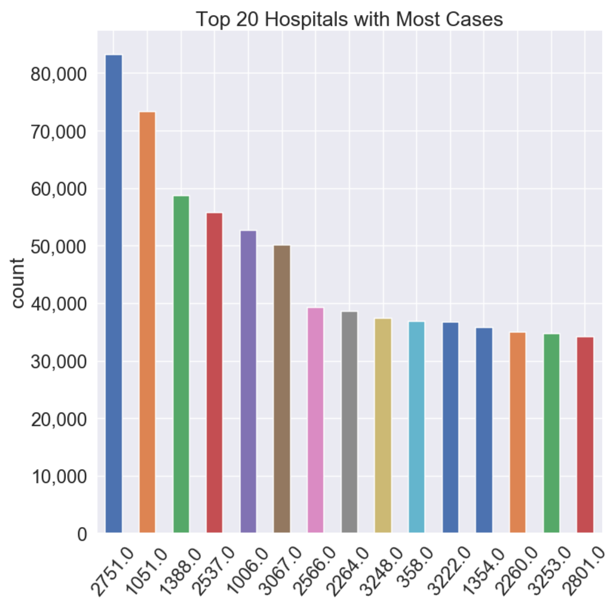
The years of 2015, 2016 and 2017 have around the same number of hospitalization; 2018 has less. This is due to the proportion sampling process described above. The most hospitalizations are 400K in a month and least 360K. In general, the peak tends to be March and May and lowest point on December, January, February.

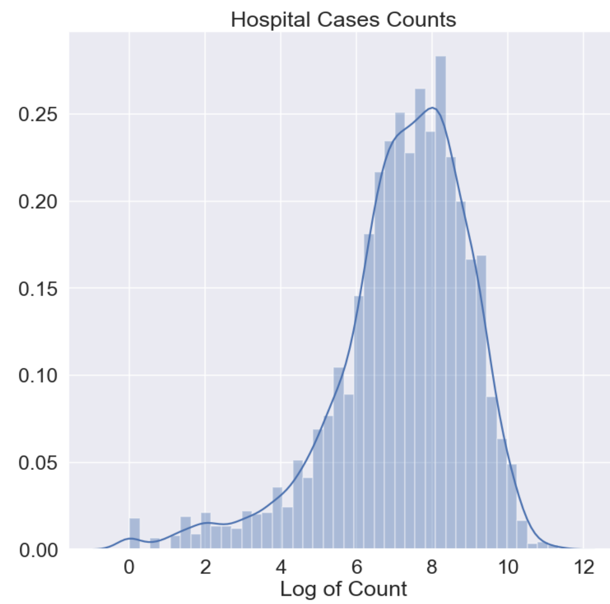
B. Specialty of Bed



The majority of types of bed are medical clinic, surgery and obstetrics. This aligns with the most common diagnoses discussed above.

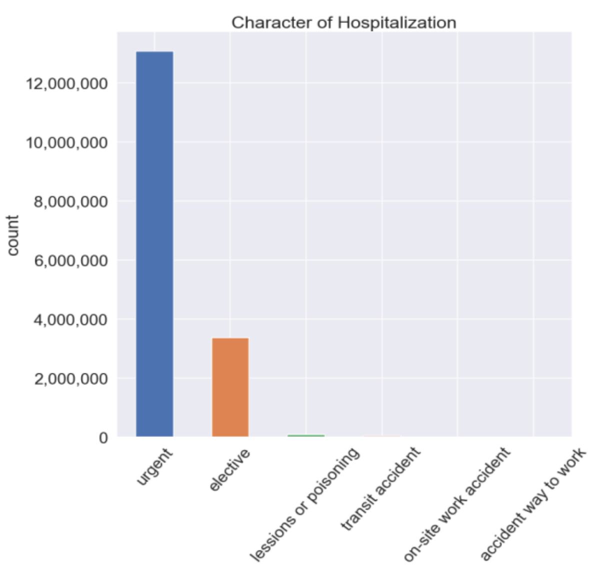
C. Hospitals and Hospitalization Cases

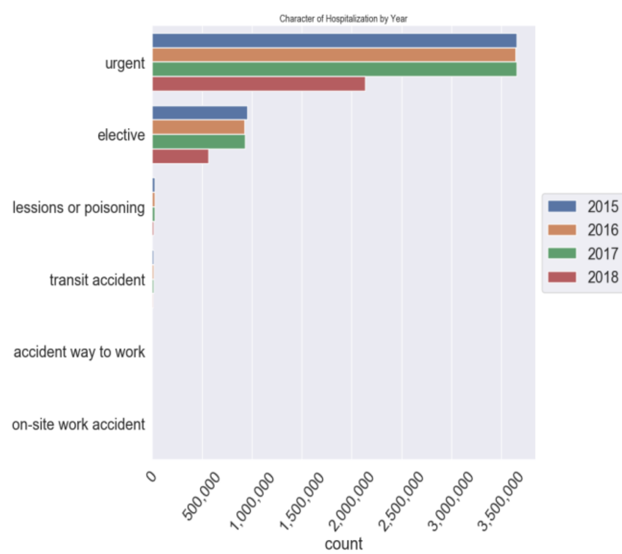




Cases are distributed across hospitals, no one hospitals dominate the sample. The hospital with most cases has 80K cases in the dataset. Distribution is somewhat right skewed. This suggests that there are a few hospitals with large amounts of cases when relative with other hospitals.

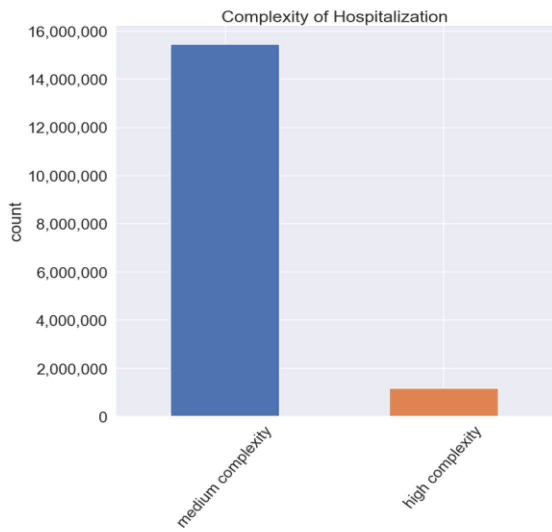
D. Character of Hospitalization

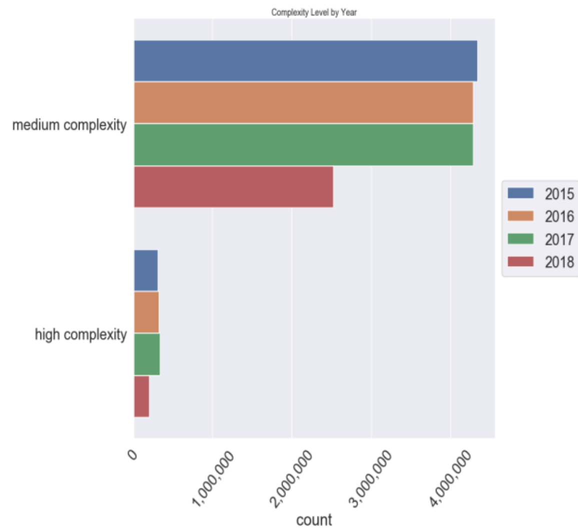




Most common type of hospitalization by far is urgent. This pattern holds when broken down by year.

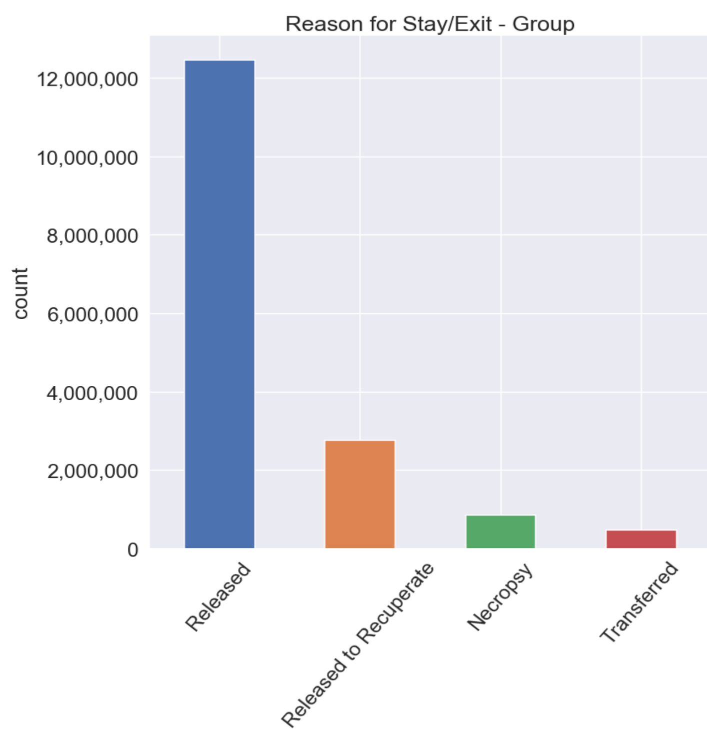
E. Complexity Level



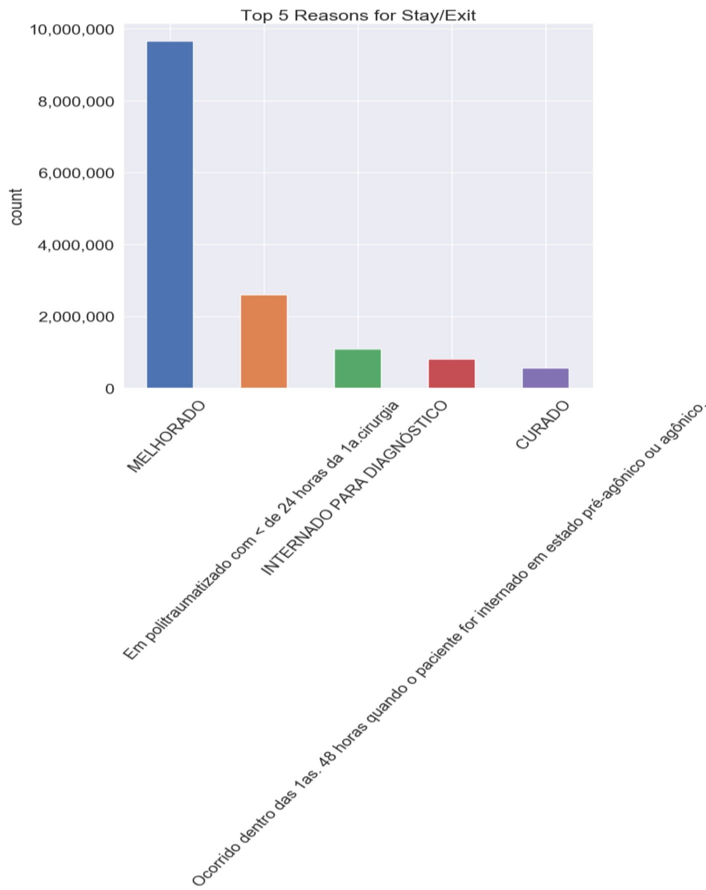


Complexity level is medium complexity by far. No hospitalization was marked as 'basic attention', which is an option for this feature.

F. Reason for Stay/Exit – Groups



G. Reasons for Stay/Exit

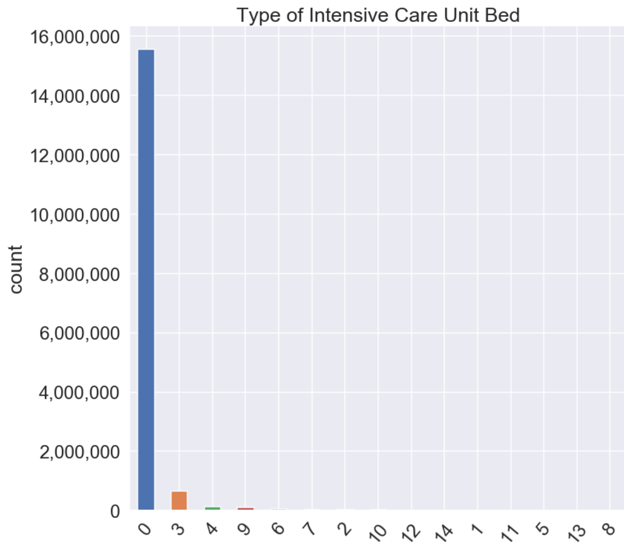


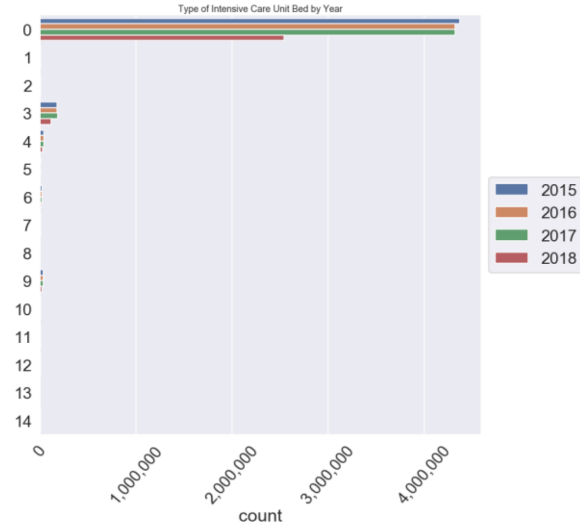
Improved is by far the most common reason for stay/exit.

Reason for stay exit	Number of Cases
MELHORADO	9665636
Em politraumatizado com < de 24 horas da 1a.cirurgia	2604471
INTERNADO PARA DIAGNÓSTICO	1095517
Ocorrido dentro das 1as. 48 horas quando o paciente for internado em estado pré-agônico ou agônico.	805235
CURADO	553550
TISIOLOGIA	487374
POR CARACTERÍSTICAS PRÓPRIAS DA DOENÇA	322336

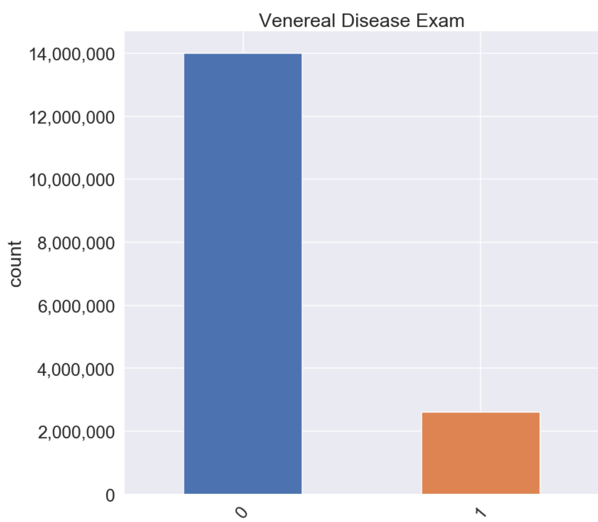
PARA OUTRA INTERNAÇÃO (OUTRO DIAGNÓSTICO)	149074
Em politraumatizado 24 a 48 horas após a 1a.cirurgia	134612
A PEDIDO	121665
POR INTERCORRÊNCIA DO PROCEDIMENTO	114573
ADMINISTRATIVA	79160
EVASÃO	62238
POR MOTIVO SOCIAL	54305
POR IMPOSSIBILIDADE DE VIVÊNCIA SÓCIO-FAMILIAR	31620
Ocorrido dentro das 1 as 48 horas quando o paciente não for internado em estado pré-agônico ou agônico.	29156
Ocorrido a partir de 48 horas após a internação.	25317
Em politraumatizado > de 72 hs. Após a 1a. cirurgia	20556
Em politraumatizado 48 a 72 horas após a 1a.cirurgia	9398
PARA COMPLEMENTAÇÃO	5889
POR DOENÇA CRÔNICA	544
PSIQUIATRIA	447
Em cirurgia de emergência 24 a 48 horas após a primeira cirurgia.	198
Em cirurgia de emergência com menos de 24 da primeira cirurgia	144
Em cirurgia de emergência 48 a 72 horas após a primeira cirurgia.	84

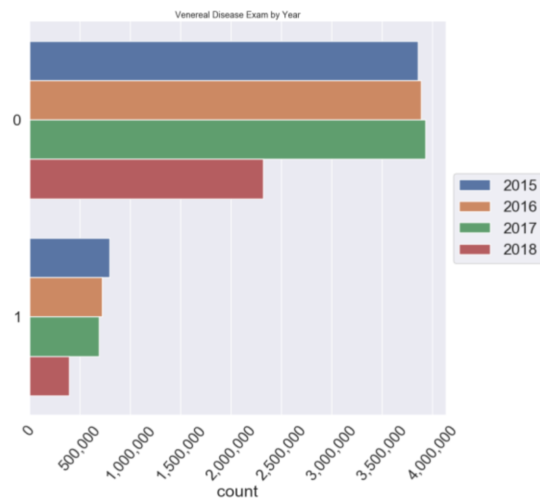
H. Type of Intensive Care Unit Bed





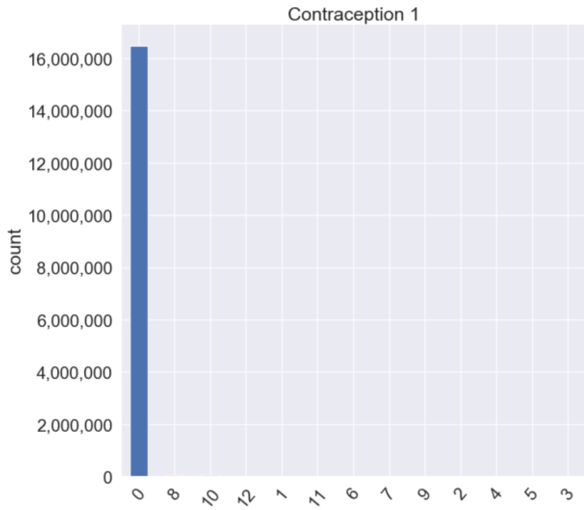
I. Venereal Exam

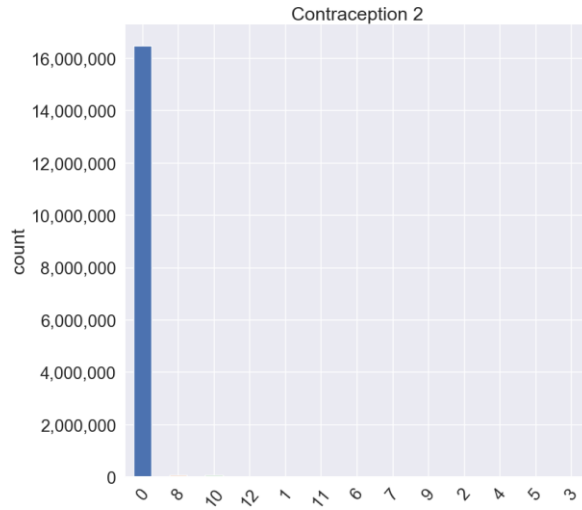




The largest proportion of of patients do not get a venereal exam performed.

J. Contraception Used 1 & 2



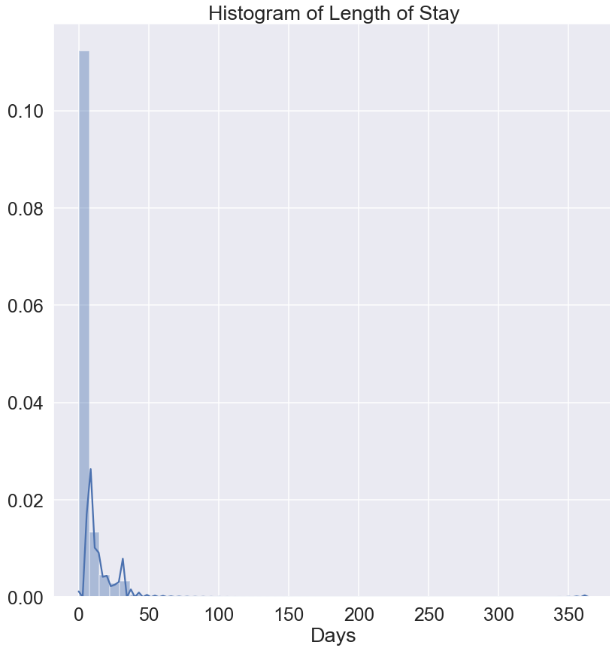


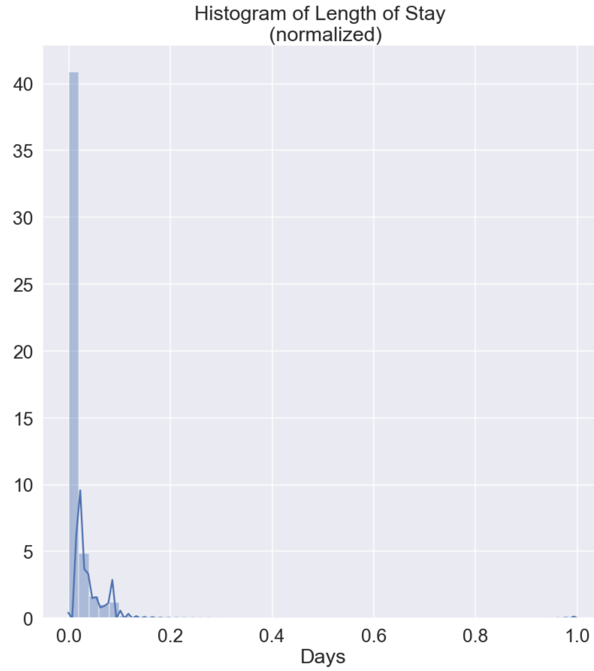
Vast majority of cases did not use contraception in their hospitalization. There are values in each of the categories of this feature, however the values are relatively very small when compare with 0 or 'no contraception'.

K. Days of Stay: *Length of Hospitalization, ICU days, Companion Days, Intermediary Unit*

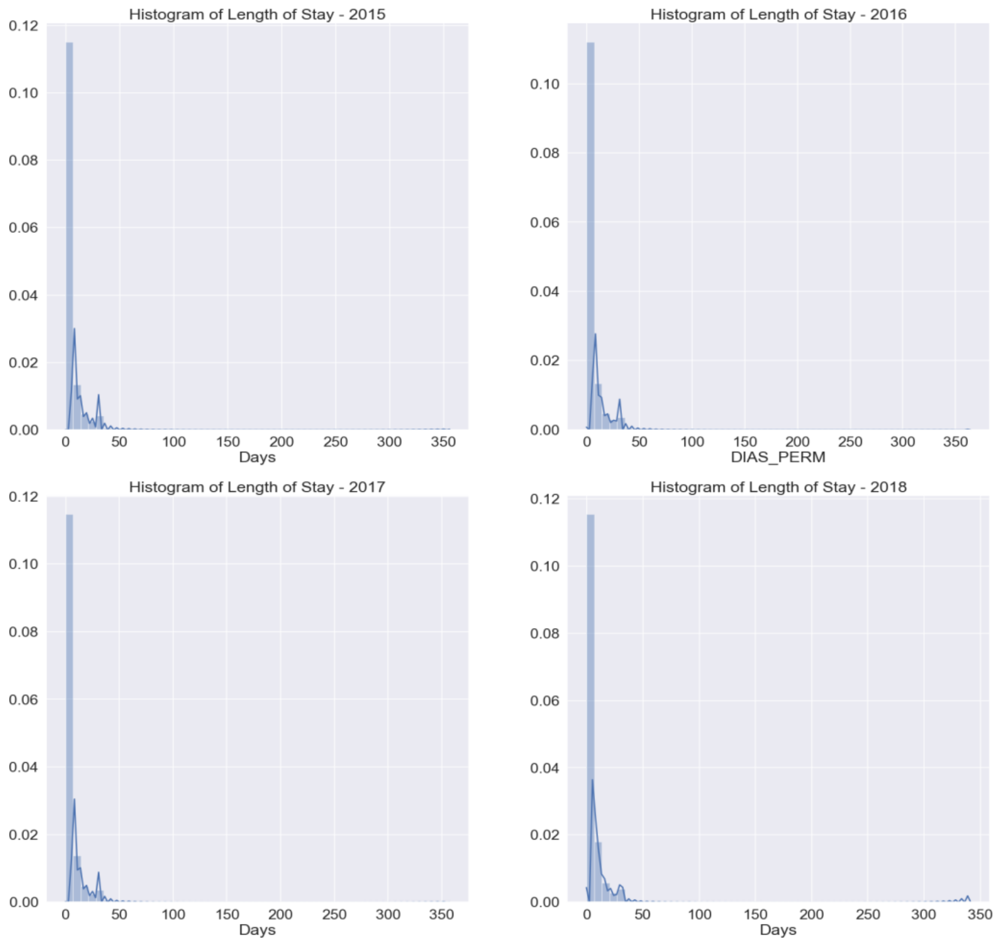
	UTI_MES_TO	UTI_INT_TO	DIAR_ACOM	DIAS_PERM
count	16,614,830	16,614,830	16,614,830	16,614,830
mean	0.46	0.05	1.95	5.39
std	2.89	0.93	4.38	8.02
min	0.00	0.00	0.00	0.00
25%	0.00	0.00	0.00	2.00
50%	0.00	0.00	0.00	3.00
75%	0.00	0.00	2.00	6.00
max	302.00	228.00	340.00	364.00

- Total Intensive Care Unit: mean is ~46 days, with max 302 days. Heavy skew with 75% being 0.
- Intermediate Intensive Care Unit: mean 49 days, with max 228 days.
- Companion Nights: mean 49.4 days, with max 340 days.
- Total Length of Stay mean hospitalization stay is 5.39 days, 75th 6 days and max 364 days.

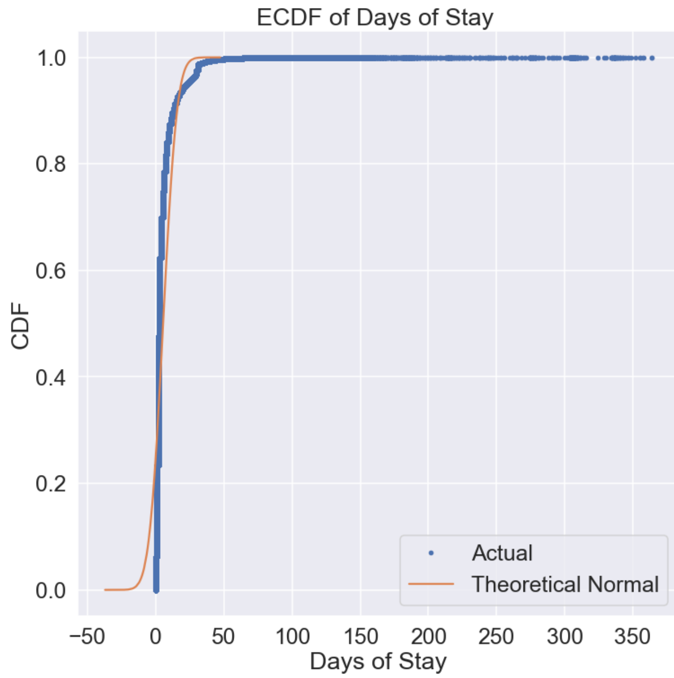




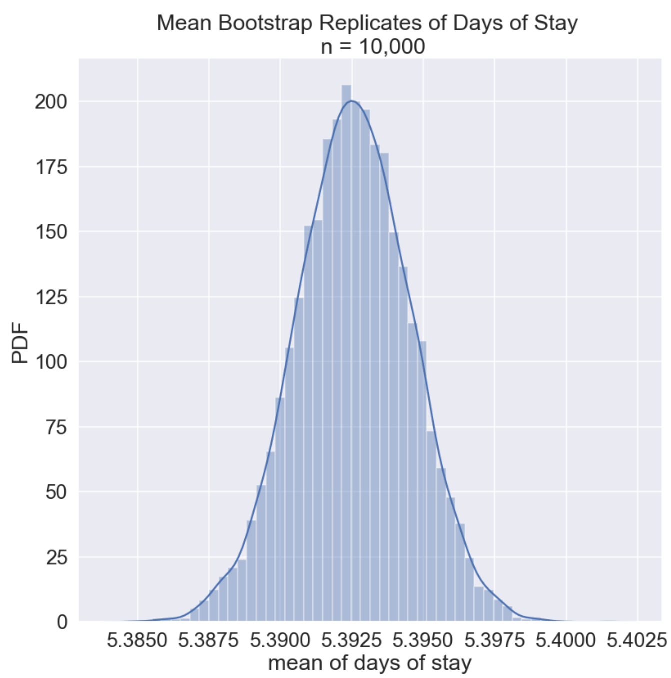
Length of stay has a heavy skew to the left, with most hospitalization being somewhat short and some outliers with long hospitalizations. This pattern repeats when broken down by year (see below).



Histogram of Length of Stay - 2015
Days
Histogram of Length of Stay - 2016
DIAS_PERM
Histogram of Length of Stay - 2017
Days
Histogram of Length of Stay - 2018
Days



The actual distribution follow the theoretical normal distribution up to CDF < 0.8. After 0.8 it diverges, with the actual being longer at the right tail. Normality tests[5] further suggests that the distribution is not normally distributed.

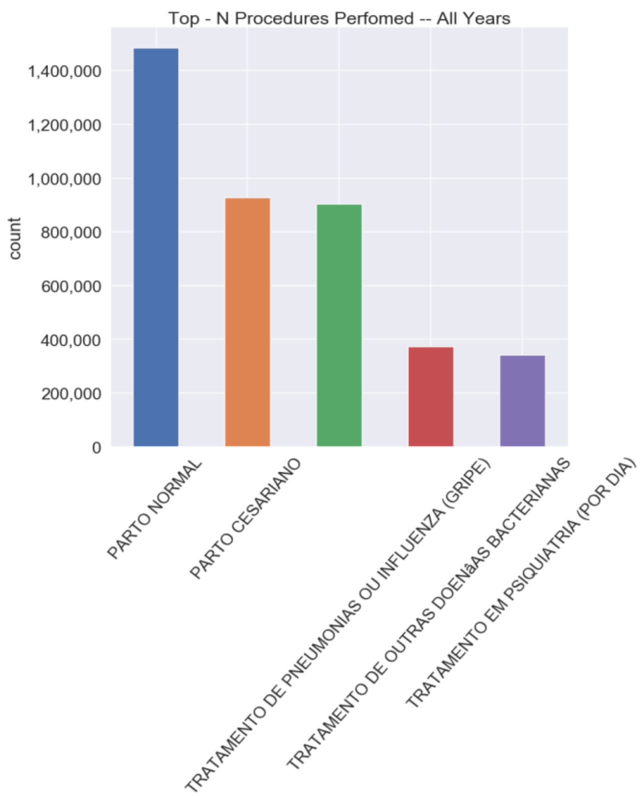


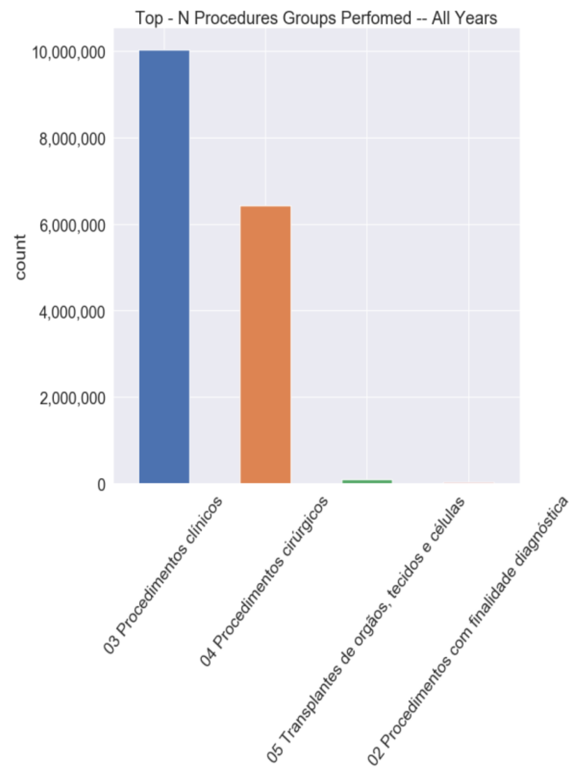
Bootstrap 95% Confidence Interval: [5.38 – 5.39], p-value = 0.5061

[5] D' Agostino and Pearson's Normality Test, Distribution Statistics and Anderson-Darling Test. Every time "normality tests" is used in this report it refers to these three tests.

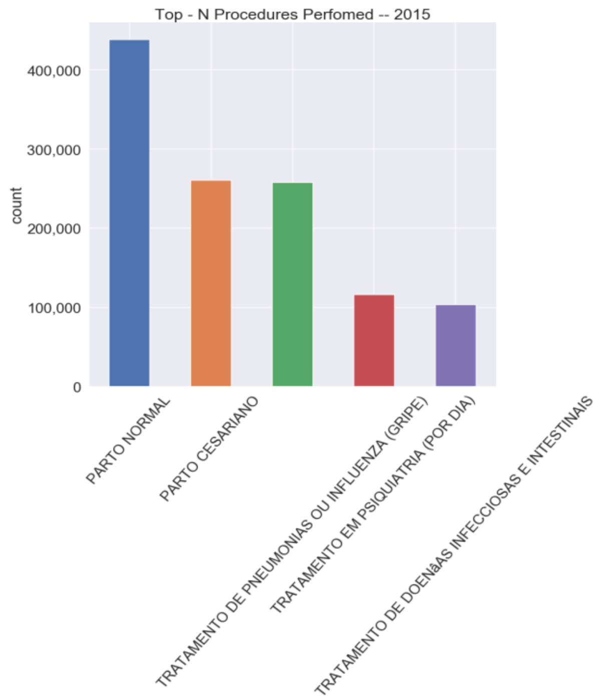
The bootstrap mean replicates show a 95% confidence interval for diagnosis counts is between 5.39 and 5.39. This is very tight interval. This range contains our sample mean of 5.39. The p-value is 0.51 which is above the alpha level of 0.05, this means we cannot reject the hypothesis that the mean age is 5.39 cases per diagnosis. A one sample t-test yielded a similar conclusion.

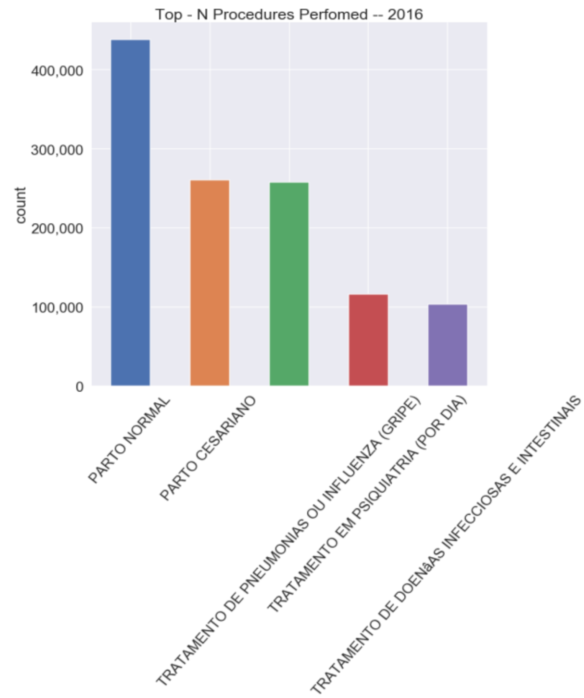
L. Procedure Performed

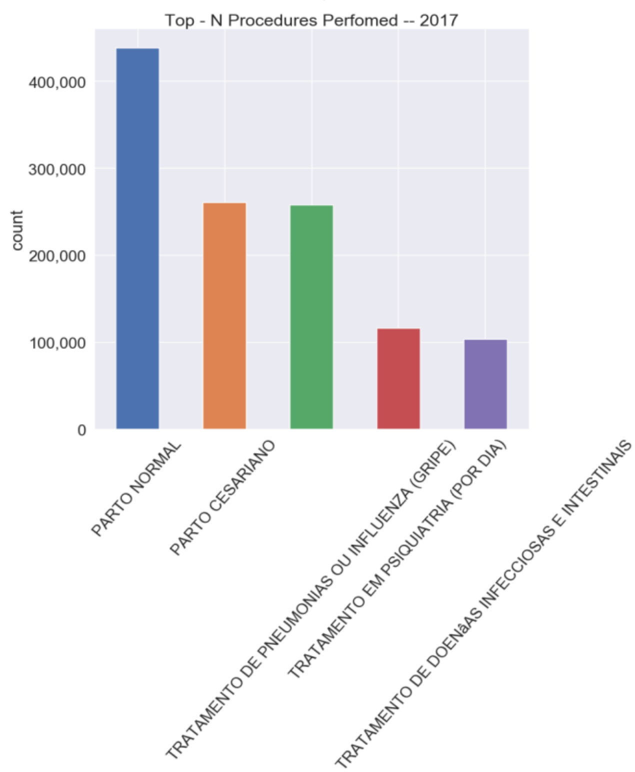


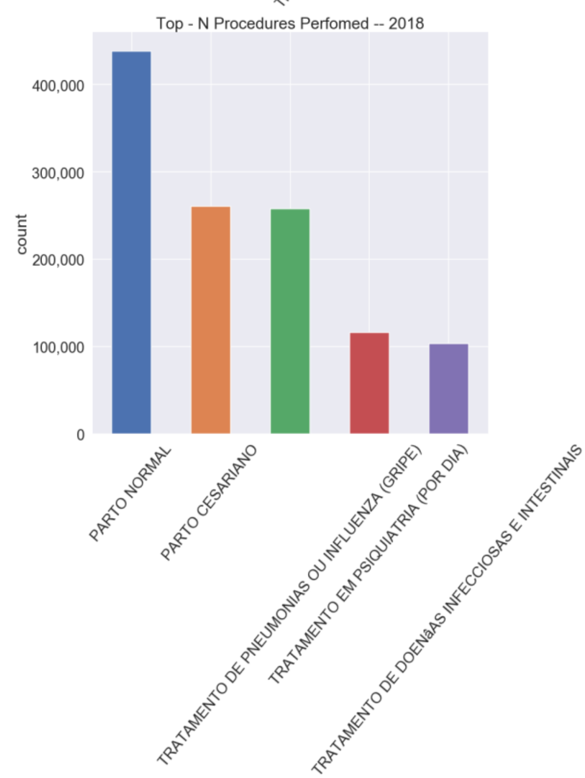


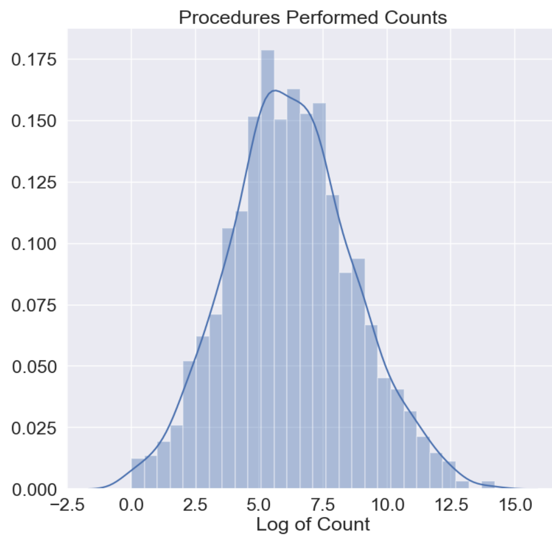
There are 1,829 unique procedure categories in the dataset. Top procedures performed are normal birth, cesarean birth, treatment for pneumonia, treatment for bacterial diseases, treatment for psychiatric disorders. While these are the most common procedures, these are a portion of the total 16+M procedures and much diversity exists of procedures performed. These patterns hold when broken down by year (see below).

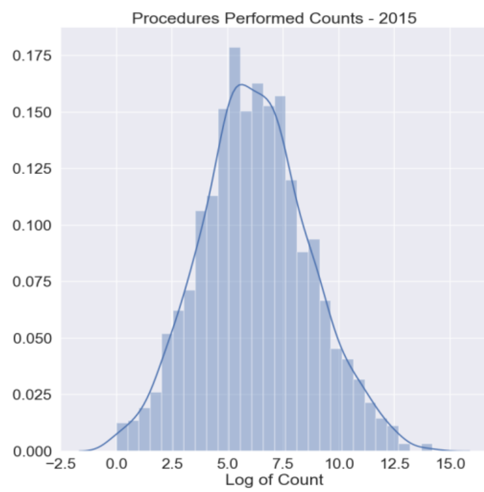

Top - N Procedures Perfomed -- 2015
count
400,000
300,000
200,000
100,000
0
PARTO NORMAL
PARTO CESARIANO
TRATAMENTO DE PNEUMONIAS OU INFLUENZA (GRIPE)
TRATAMENTO EM PSIQUIATRIA (POR DIA)
TRATAMENTO DE DOENãAS INFECCIOSAS E INTESTINAIS

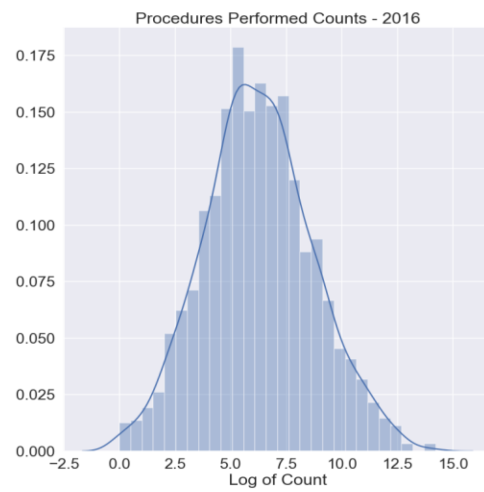

Top - N Procedures Perfomed -- 2016
count
400,000
300,000
200,000
100,000
0
PARTO NORMAL
PARTO CESARIANO
TRATAMENTO DE PNEUMONIAS OU INFLUENZA (GRIPE)
TRATAMENTO EM PSIQUIATRIA (POR DIA)
TRATAMENTO DE DOENãAS INFECCIOSAS E INTESTINAIS

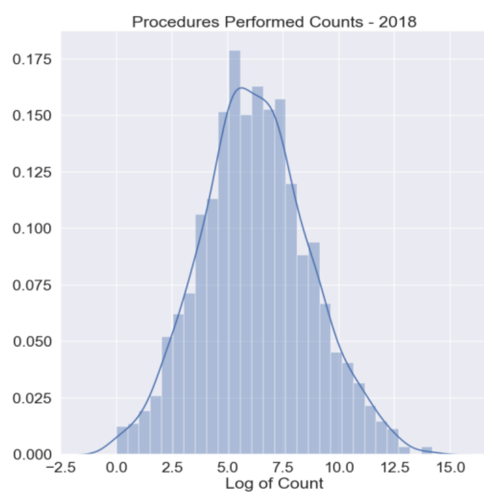

Top - N Procedures Perfomed -- 2017
count
400,000
300,000
200,000
100,000
0
PARTO NORMAL
PARTO CESARIANO
TRATAMENTO DE PNEUMONIAS OU INFLUENZA (GRIPE)
TRATAMENTO EM PSIQUIATRIA (POR DIA)
TRATAMENTO DE DOENãAS INFECCIOSAS E INTESTINAIS

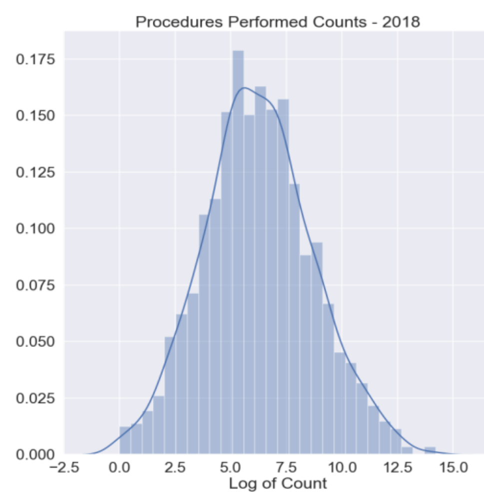

Top - N Procedures Perfomed -- 2018
count
400,000
300,000
200,000
100,000
0
PARTO NORMAL
PARTO CESARIANO
TRATAMENTO DE PNEUMONIAS OU INFLUENZA (GRIPE)
TRATAMENTO EM PSIQUIATRIA (POR DIA)
TRATAMENTO DE DOENãAS INFECCIOSAS E INTESTINAIS

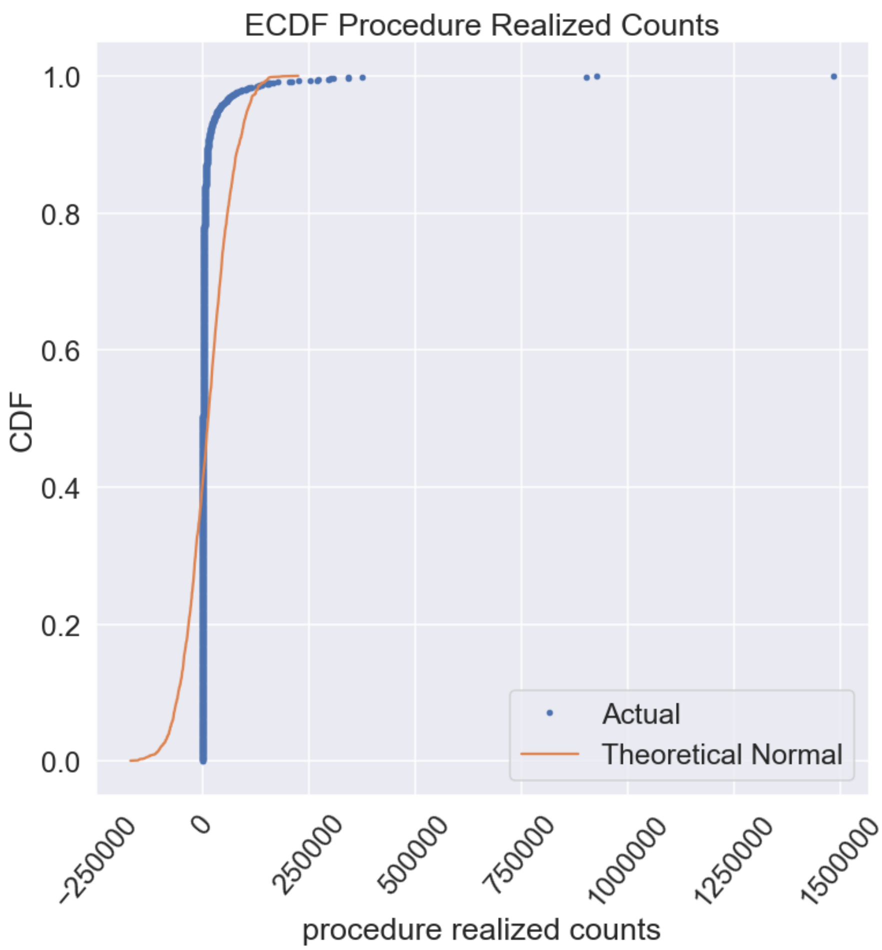

Procedures Performed Counts
Log of Count


Procedures Performed Counts - 2015
Log of Count


Procedures Performed Counts - 2016
Log of Count

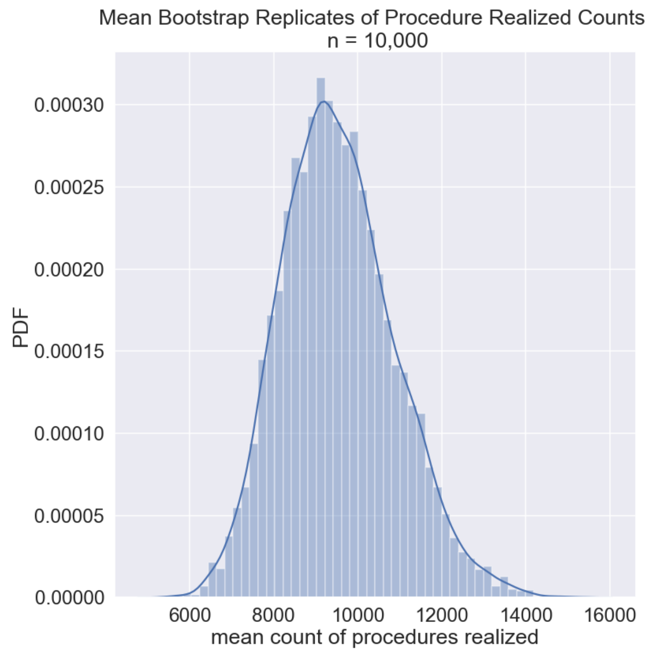

Procedures Performed Counts - 2018
Log of Count


Procedures Performed Counts - 2018
Log of Count



The distribution of procedure realized counts does not follow a theoretical normal distribution at all. Normality tests[6] further suggests that the distribution is not normally distributed.

[6] D' Agostino and Pearson's Normality Test, Distribution Statistics and Anderson-Darling Test. Every time "normality tests" is used in this report it refers to these three tests.



Bootstrap 95% Confidence Interval: [7,183.50 – 12,393.45], p-value = 0.5293

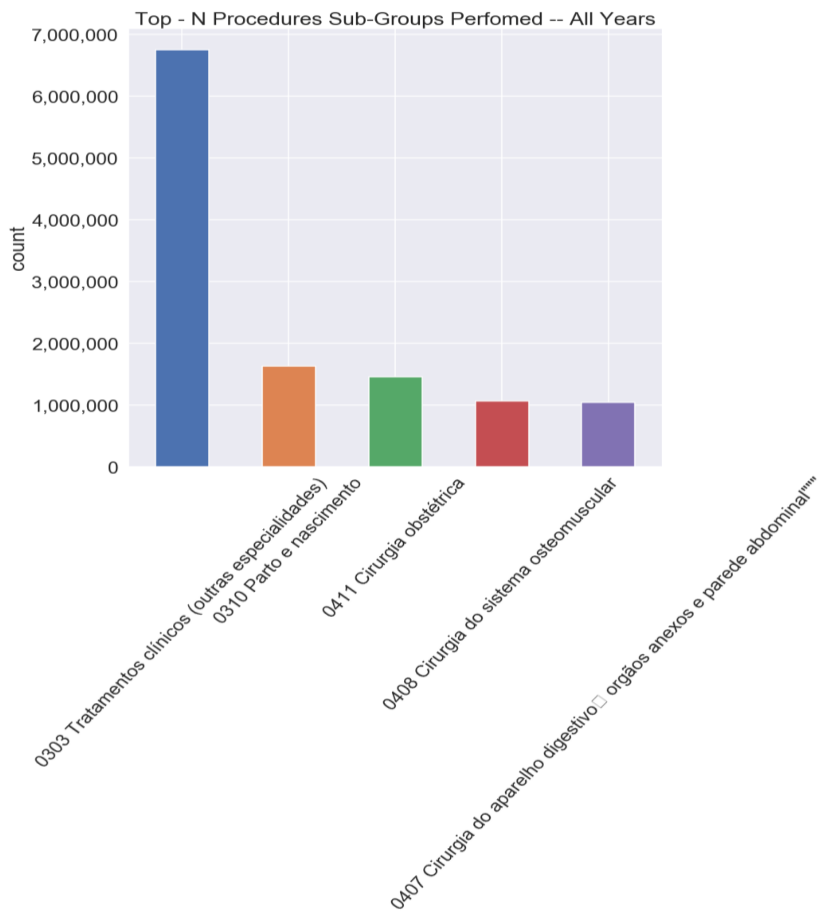
The bootstrap mean replicates show a 95% confidence interval for diagnosis counts is between 7,183.50 and 12,393. This is a wide interval. This range contains the sample mean of 9,084. The p-value is 0.43 which is above the alpha level of 0.05, this means we cannot reject the hypothesis that the mean cases per category is 9,537. A one sample t-test yielded a similar conclusion.

M. Procedure Performed – *Procedure Sub-Group*

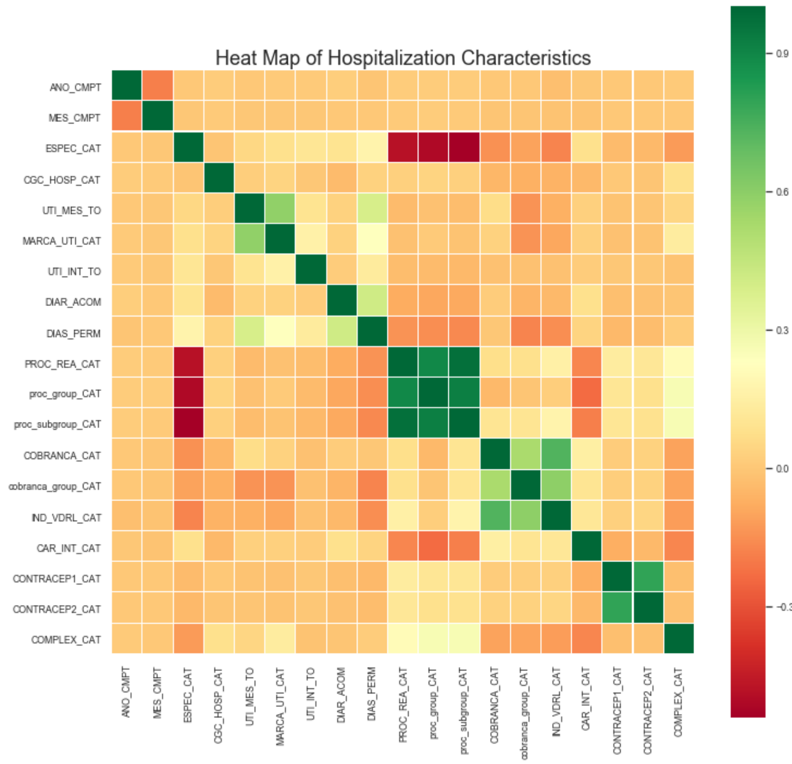
Procedure Sub-Group Name	Cases
Tratamentos clínicos (outras especialidades)	6,752,990
Parto e nascimento	1,636,961
Cirurgia obstétrica	1,452,495
Cirurgia do sistema osteomuscular	1,068,710
Cirurgia do aparelho digestivo\t orgãos a...	1,045,207
Outras cirurgias	732,356
Cirurgia do aparelho geniturinário	676,902
Consultas / Atendimentos / Acompanhamentos	524,848
Tratamento em oncologia	442,079
Cirurgia do aparelho circulatório	403,030
Tratamento em nefrologia	342,469
Tratamento de lesões\t envenenamentos e o...	318,025
Cirurgia em oncologia	197,010
Cirurgia das vias aéreas superiores\t da ...	182,954
Pequenas cirurgias e cirurgias de pele\t ...	141,905
Cirurgia do aparelho da visão	139,815
Cirurgia do sistema nervoso central e per...	118,581
Cirurgia reparadora	82,275

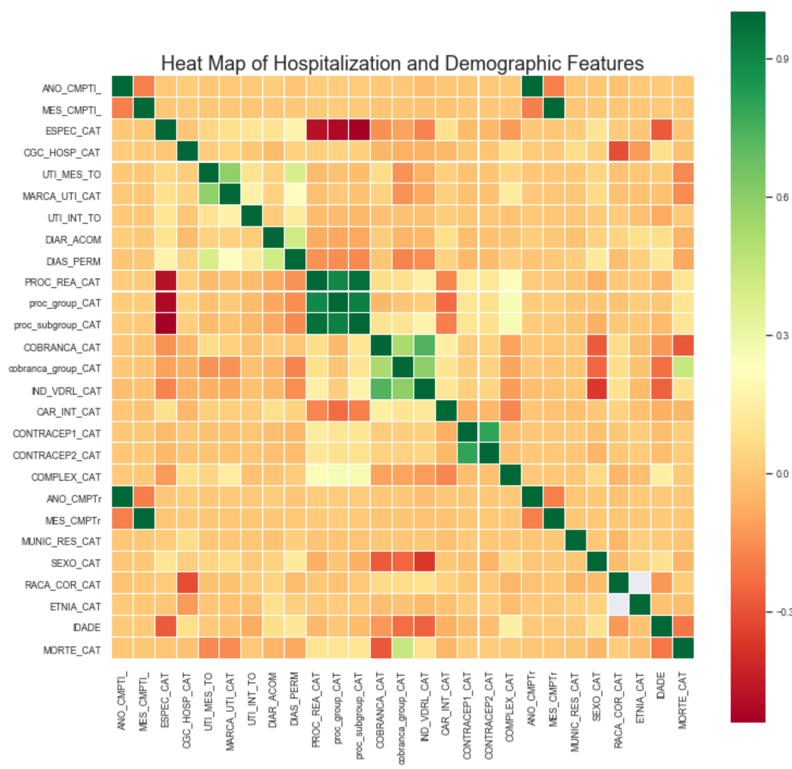
Cirurgia torácica	80,033
Cirurgia de mama	48,925
Acompanhamento e intercorrências no pré e...	45,899
Ações relacionadas à doação de orgãos e t...	29,975
Tratamentos clínicos (outras especialidad...	26,840
Transplante de orgãos	19,263
Pequenas cirurgias e cirurgias de pele	18,787
Cirurgia de glândulas endócrinas	17,345
Bucomaxilofacial	17,072
Coleta de material	15,080
Métodos diagnósticos em especialidades	9,903
Diagnóstico por endoscopia	9,155
Cirurgia das vias aéreas superiors	4,983
Cirurgia do sistema nervoso central e per...	4,527
Processamento de tecidos para transplante	2,444
Cirurgia do aparelho circulatório	2,350
Cirurgia torácica	2,349
Coleta e exames para fins de doação de or...	1,141
Cirurgia de mama	130
Coleta de material	1

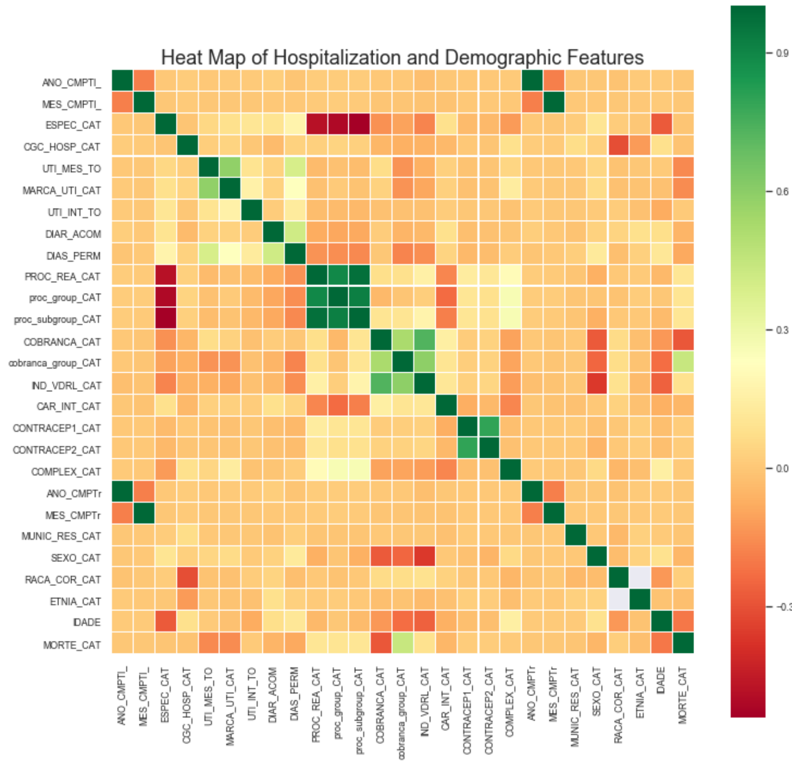
Top five procedure subgroups are: (1) clinical treatments, (2) birth, (3) obstetric surgery, (4) osteomuscular surgery and, (5) surgery of the digestive system.



N. Heat maps of Hospitalization, Demographics and Diagnosis Features







- Weak correlation between diagnosis and: procedure performed, reasons for stay/exit and, character of hospitalization.
- There are some features with strong positive correlation but this is due to the fact that they are either closely related or derived from each other.

DATASET FEATURES & ACTIONS

Field_Name	Type of Field	Description	Action
UF_ZI	char(6)	Municipality Manager	Auditor metadata - Not Used
ANO_CMPT	char(4)	Year of AIH processing, in yyyy format.	**Part of all Groups**
MÊS_CMPT	char(2)	Month of AIH processing, in mm format.	**Part of all Groups**
ESPEC	char(2)	Specialty of Bed	**Hospitalization Group**
CGC_HOSP	char(14)	CNPJ of the Establishment	**Hospitalization Group**
N_AIH	char(13)	Number of AIH	Auditor metadata - Not Used
IDENT	char(1)	Identification of the type of AIH	Auditor metadata - Not Used
CEP	char(8)	CEP of the patient	Auditor metadata - Not Used
MUNIC_RES	char(6)	Municipality of Patient's Residence	**Demographic Group**
NASC	char(8)	Date of birth of the patient (yyyammdd)	Demographic Group – Not Used
SEXO	char(1)	Sex of patient	**Demographic Group**
UTI_MES_IN	numeric(2)	Reset	Dropped – First Pass
UTI_MES_AN	numeric(2)	Reset	Dropped – First Pass
UTI_MES_AL	numeric(2)	Reset	Dropped – First Pass
UTI_MES_TO	numeric(3)	Number of ICU days in the month	**Hospitalization Group**
MARCA_UTI	char(2)	Indicates the type of ICU used by the patient	**Hospitalization Group**
UTI_INT_IN	numeric(2)	Reset	Dropped – First Pass
UTI_INT_AN	numeric(2)	Reset	Dropped – First Pass
UTI_INT_AL	numeric(2)	Reset	Dropped – First Pass
UTI_INT_TO	numeric(3)	Number of nights in intermediate unit	**Hospitalization Group**
DIAR_ACOM	numeric(3)	Number of companion nights	**Hospitalization Group**
QT_DIARIAS	numeric(3)	Number of nights	Hospitalization Group – Not Used
PROC_SOLIC	char(10)	Procedure requested	Hospitalization Group – Not Used
PROC_REA	char(10)	Procedure performed	**Hospitalization Group**
VAL_SH	numeric(13,2)	Value of hospital services	Financial Group – Not Used
VAL_SP	numeric(13,2)	Value of professional services	Financial Group – Not Used
VAL_SADT	numeric(13,2)	Reset	Dropped – First Pass
VAL_RN	numeric(13,2)	Reset	Dropped – First Pass
VAL_ACOMP	numeric(13,2)	Reset	Dropped – First Pass
VAL_ORTP	numeric(13,2)	Reset	Dropped – First Pass
VAL_SANGUE	numeric(13,2)	Reset	Dropped – First Pass
VAL_SADTSR	numeric(11,2)	Reset	Dropped – First Pass
VAL_TRANSP	numeric(13,2)	Reset	Dropped – First Pass
VAL_OBSANG	numeric(11,2)	Reset	Dropped – First Pass
VAL_PED1AC	numeric(11,2)	Reset	Dropped – First Pass
VAL_TOT	numeric(14,2)	Total value of the AIH	Financial Group – Not Used
VAL_UTI	numeric(8,2)	Value of ICU	Financial Group – Not Used
US_TOT	numeric(10,2)	Total value, in US dollars	Financial Group – Not Used
DI_INTER	char(8)	Date of hospitalization in aaammdd format	Hospitalization Group – Not Used
DT_SAIDA	char(8)	Exit date in yyyammdd format	Hospitalization Group – Not Used

DIAG_PRINC	char(4)	Code of the main diagnosis (CID10)	**Diagnosis Group**
DIAG_SECUN	char(4)	Secondary diagnosis code (ICD10). Filled with zeros from 201501.	Diagnosis Group – Not Used
COBRANCA	char(2)	Reason for Exit / Stay	**Hospitalization Group**
NATUREZA	char(2)	Legal nature of the hospital (with content until May / 12). It was used the classification of Regime and Nature.	Auditor metadata - Not Used
NAT_JUR	char(4)	Legal nature of the establishment, as the Commission National classification - CONCLA	Auditor metadata - Not Used
DESTAO	char(1)	Type of hospital management	Auditor metadata - Not Used
RUBRICA	numeric(5)	Reset	Dropped – First Pass
IND_VDRL	char(1)	Indicates VDRL exam	**Hospitalization Group**
MUNIC_MOV	char(6)	Municipality of the Establishment	Auditor metadata - Not Used
COD_IDADE	char(1)	Unit of measure of age	Demographic Group – Not Used
IDADE	numeric(2)	Age	**Demographic Group**
DIAS_PERM	numeric(5)	Days of Stay	**Hospitalization Group**
MORTE	numeric(1)	Indicates Death	**Demographic Group**
NACIONAL	char(2)	Code of nationality of the patient	Demographic Group – Not Used
NUM_PROC	char(4)	Reset	Dropped – First Pass
CAR_INT	char(2)	Character of hospitalization	**Hospitalization Group**
TOT_PT_SP	numeric(6)	Reset	Dropped – First Pass
CPF_AUT	char(11)	Reset	Dropped – First Pass
HOMONIMO	char(1)	Indicator if the patient of the AIH is homonymous with the another AIH.	Auditor metadata - Not Used
NUM_FILHOS	numeric(2)	Number of children of the patient	Demographic Group – Not Used
INSTRU	char(1)	Degree of instruction of the patient	Demographic Group – Not Used
CID_NOTIF	char(4)	CID of Notification	Auditor metadata - Not Used
CONTRACEP1	char(2)	Type of contraceptive used	**Hospitalization Group**
CONTRACEP2	char(2)	Second type of contraceptive used	**Hospitalization Group**
GESTRISCO	char(1)	Indicator if pregnant at risk	Demographic Group – Not Used
INSC_PN	char(12)	Number of the pregnant woman in prenatal care	Hospitalization Group – Not Used
SEQ_AIH5	char(3)	Long-stay sequential (AIH type 5)	Auditor metadata - Not Used
CBOR	char(3)	Occupancy of the patient, according to the Brazilian Occupations - CBO.	Demographic Group – Not Used
CNAER	char(3)	Work accident code	Auditor metadata - Not Used
GESTOR_COD	char(3)	Reason for authorization of the AIH by the Manager	Auditor metadata - Not Used
GESTOR_TP	char(1)	Type of manager	Auditor metadata - Not Used
GESTOR_CPF	char(11)	Manager's CPF number	Auditor metadata - Not Used
GESTOR_DT	char(8)	Date of authorization given by the Manager (yyyammdd)	Dropped – First Pass
CNES	char(7)	CNES code of the hospital	Auditor metadata - Not Used
CNPJ_MANT	char(14)	CNPJ of the maintainer	Auditor metadata - Not Used
INFEHOSP	char(1)	Hospital infection status	Auditor metadata - Not Used
CID_ASSO	char(4)	CID causes	Hospitalization Group – Not Used
CID_MORTE	char(4)	CID of death	Hospitalization Group – Not Used
COMPLEX	char(2)	Complexity	**Hospitalization Group**

FINANC	char(2)	Type of financing	Financial Group – Not Used
FAEC_TP	char(6)	Financing subtype FAEC	Financial Group – Not Used
REGCT	char(4)	Contract rule	Auditor metadata - Not Used
RACA_COR	char(4)	Race / Color of the patient	**Demographics Group**
ETNIA	char(4)	Ethnicity of patient, if race color is indigenous	**Demographics Group**
SEQUENCIA	numeric(9)	Sequential of the AIH in the consignment	Auditor metadata - Not Used
REMESSA	char(21)	Shipping number	Auditor metadata - Not Used
AUD_JUST	char (50)	Auditor's justification for acceptance of the IAI without the National Health Card.	Auditor metadata - Not Used
SIS_JUST	char (50)	Rationale of the establishment for acceptance of the AIH without number of the National Health Card	Auditor metadata - Not Used
VAL_SH_FED	numeric (10, 2)	Value of the federal complement of hospital services. It is included in the total value of the AIH.	Financial Group – Not Used
VAL_SP_FED	numeric (10, 2)	Value of the federal complement of professional services. It is included in the total value of the AIH.	Financial Group – Not Used
VAL_SH_GES	numeric (10, 2)	Value of the complement of the manager (state or municipal) of hospital services. It is included in the total value of the AIH.	Financial Group – Not Used
VAL_SP_GES	numeric (10, 2)	Value of the complement of the manager (state or municipal) of profesional services.It is included in the total value of the AIH.	Financial Group – Not Used
VAL_UCI	numeric (10, 2)	Value of ICU.	Financial Group – Not Used
MARCA_UCI	char (2)	Type of ICU used by the patient.	Hospitalization Group – Not Used
DIAGSEC1	char (4)	Secondary diagnosis1	Dropped – First Pass
DIAGSEC2	char (4)	Secondary diagnosis2	Dropped – First Pass
DIAGSEC3	char (4)	Secondary diagnosis3	Dropped – First Pass
DIAGSEC4	char (4)	Secondary diagnosis4	Dropped – First Pass
DIAGSEC5	char (4)	Secondary diagnosis5	Dropped – First Pass
DIAGSEC6	char (4)	Secondary diagnosis6	Dropped – First Pass
DIAGSEC7	char (4)	Secondary diagnosis7	Dropped – First Pass
DIAGSEC8	char (4)	Secondary diagnosis8	Dropped – First Pass
DIAGSEC9	char (4)	Secondary diagnosis9	Dropped – First Pass
TPDISEC1	char(1)	Type of secondary diagnosis 1	Dropped – First Pass
TPDISEC2	char(1)	Type of secondary diagnosis 2	Dropped – First Pass
TPDISEC3	char(1)	Type of secondary diagnosis 3	Dropped – First Pass
TPDISEC4	char(1)	Type of secondary diagnosis 4	Dropped – First Pass
TPDISEC5	char(1)	Type of secondary diagnosis 5	Dropped – First Pass
TPDISEC6	char(1)	Type of secondary diagnosis 6	Dropped – First Pass
TPDISEC7	char(1)	Type of secondary diagnosis 7	Dropped – First Pass
TPDISEC8	char(1)	Type of secondary diagnosis 8	Dropped – First Pass
TPDISEC9	char(1)	Type of secondary diagnosis 9	Dropped – First Pass

FEATURE ENGENEERING

The data was randomly divided and shuffled into on training, testing and validation using an 85%, 5% and 10% split respectively. The training set has 14,621,050 hospitalizations, the testing set has 319,005 hospitalizations, and the validation set has 1,674,775 hospitalizations.

The dataset has a significant proportion of features that are categorical. As shown in the exploratory analysis section, these features are not only categorical, but they also have a large number of categories within them. The most common practice in handling categorical variables is one-hot encoding. However, one-hot encoding proved unfeasible given the computational resources available. When the variables were converted to one-hot encoded vectors, the memory size of the data grew substantially and computational resources available (AWS EC2 instance and Google Collaboratory) were not sufficient to both hold the data in memory and run the deep neural network models.

To handle the categorical data, an embedding layer is applied within the neural network Embedding is often used in natural language processing settings. In this case, the embedding layer assigns a 10-dimensional vector to each category and the layer learns the weights needed to find the position of the category in a lookup table. Finally, numerical features were normalized within a 0 to 1 range.

DEEP LEARNING NEURAL NETWORK

MODELING STRATEGY

The approach of this project was to conduct several experimental models first before tuning and building final model. Eight experimental models were built, tested and cross – validated. Each model is an attempt to test a specific modeling/architecture strategy and its impact on performance. Given time constraints and computational resources not the entire universe of deep network architecture could be tested. Neural network architecture choices were made based on the data, problem domain and results of modeling experiments.

Given the nature of the data, problem and deep neural networks, all the models tested are complex. However, within the group of experimental models there are some that are more complex than others. The aspects of the models that worked the best where used to build and parameter optimize a final model. Please see table below and model description for more details. Moreover, code for all models can be found [HERE].

MODEL EVALUATION STRATEGY

The primary metric is accuracy. While accuracy is not the only evaluation metric available and not appropriate for all problems, in this case the main interest is the extent the neural network can predict the correct procedure for a patient. Given that several models were tested, a secondary metric is the false positive rate and a third metric is average accuracy (i.e. accuracy by class). Beyond these three-metrics, the extent of model overfitting, complexity, and training time will be taken into account when evaluating models.

SUMMARY OF MODEL RESULTS

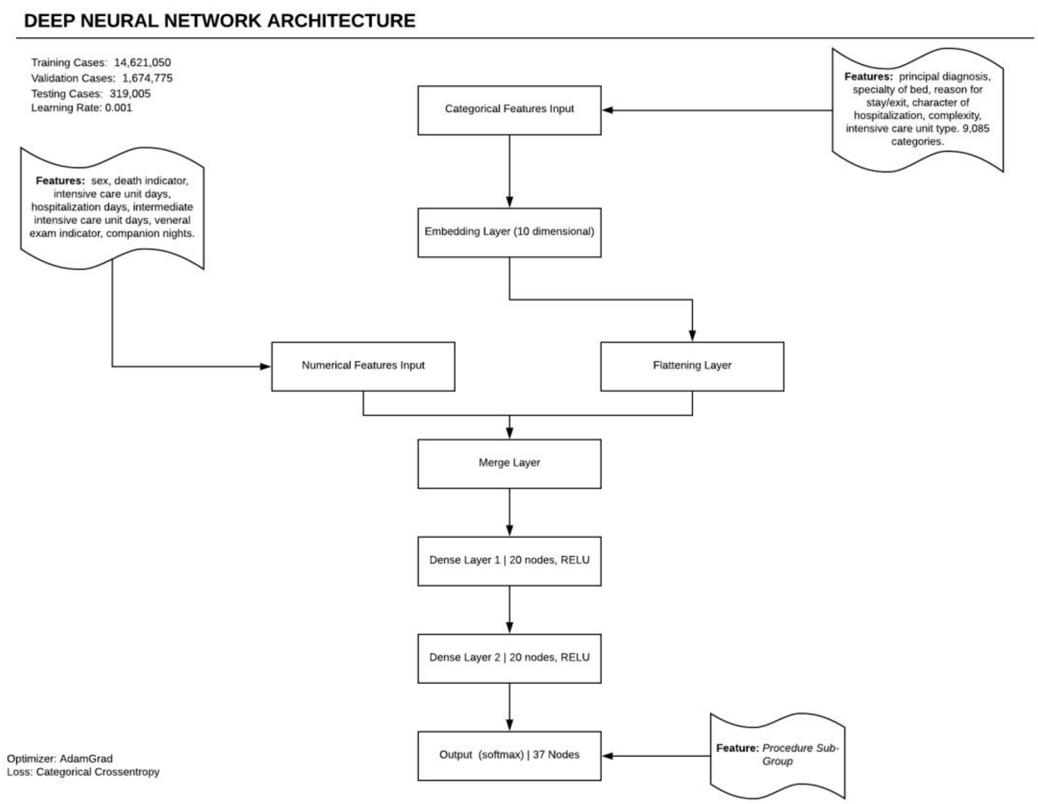
Model Number	Training & Testing Data				Validation Data		Total Training Time
	Training Accuracy	Testing Accuracy	Average Accuracy*	False Positive Rate	CV (k = 5)	Diff Cross Validation & Training	
Model 1	90.14%	90.12%	57.9%	0.003	85.86% (+/-0.21%)	4.28	3.5 hours
Model 2	89.80%	89.78%	55.6%	0.003	85.77% (+/- 0.67%)	4.03	4.25 hours
Model 3	83.68%	83.72%	31.7%	0.005	78.52% (+/- 1.37%)	5.16	3.5 hours
Model 4	89.08%	89.11%	49.1%	0.003	85.90% (+/- 0.41%)	3.18	5.9 hours
Model 5	82.81%	82.86%	26.7%	0.005	71.94% (+/- 2.21%)	10.87	3.5 hours
Model 6	85.63%	88.32%	43.2%	0.003	85.27% (+/- 0.44%)	0.36	4.25 hours
Model 7	89.38%	89.89%	82.0%	0.003	83.38% (+/- 0.92%)	6.00	4 hours
Model 8	89.99%	89.88%	80.1%	0.003	N/A	N/A	5 hours
Final Model	87.31%	90.31%	80.4%	0.003			19 hours

**Unweighted average of accuracy by class*

MODELS MAIN ARCHITECTURE FEATURES

Model Number	Embedding Layer	Layers	Nodes *(at each dense layer)*	Regularization	Learning Rate	Oversampling
Model 1	Yes	2	20	No	0.001	No
Model 2	Yes	4	20	No	0.001	No
Model 3	Yes	2	20	L2	0.001	No
Model 4	Yes	6	20	No	0.001	No
Model 5	Yes	2	20	L1	0.001	No
Model 6	Yes	2	20	Dropout	0.001	No
Model 7	Yes	2	20	No	0.001	Random
Model 8	Yes	2	20	No	0.001	SMOTE
Final Model	Yes	2	20	Dropout	0.001	Random

MODEL 1: BASELINE



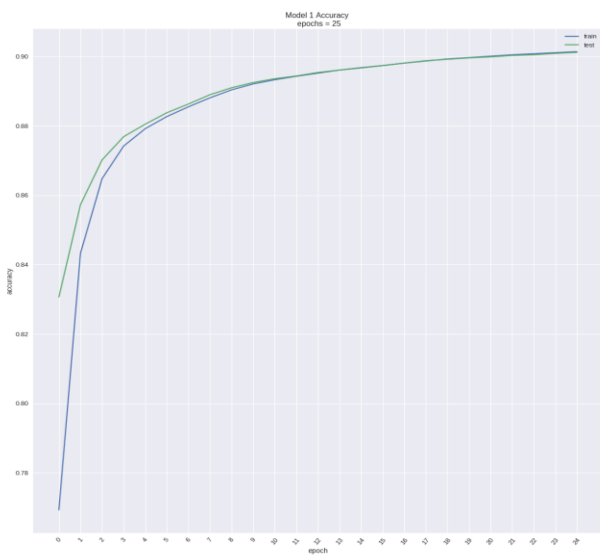
Model 1 starts with an embedding layer that converts each categorical feature's observation into a 10-dimensional array. The total number of categories being used is 9,085 across six categorical features. A rule of thumb is that the embedding vector dimension be the 4th root of the number of categories. As such, 9,085**0.25 = 9.76 which rounded leads to 10.

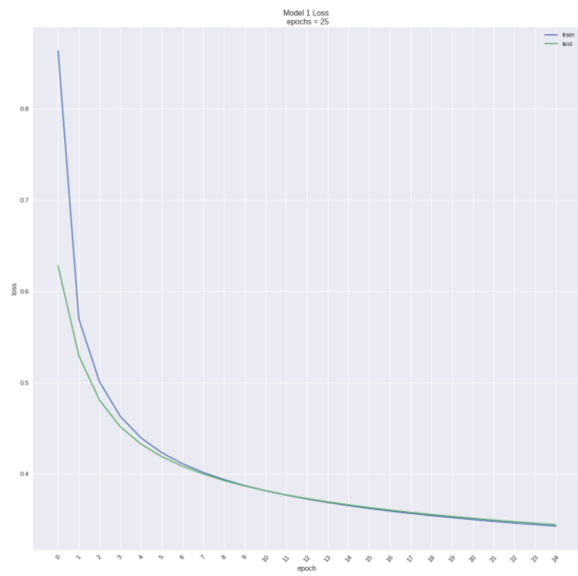
Once the embedding layer processes the categorical features and learns the weights, the numerical features are brought in using an input layer. The next step is to concatenate the results of the embedding layer and the numerical features. These inputs are then fed into two dense layers and the output layer is a softmax layer. The optimizer used is Adam Gradient Descent with a learning rate of 0.001 and the loss function is categorical crossentropy.

> "*Cross-entropy loss, or log loss, measures the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability diverges from the actual label. So predicting a probability of .012 when the actual observation label is 1 would*

be bad and result in a high loss value. A perfect model would have a log loss of 0."[7]

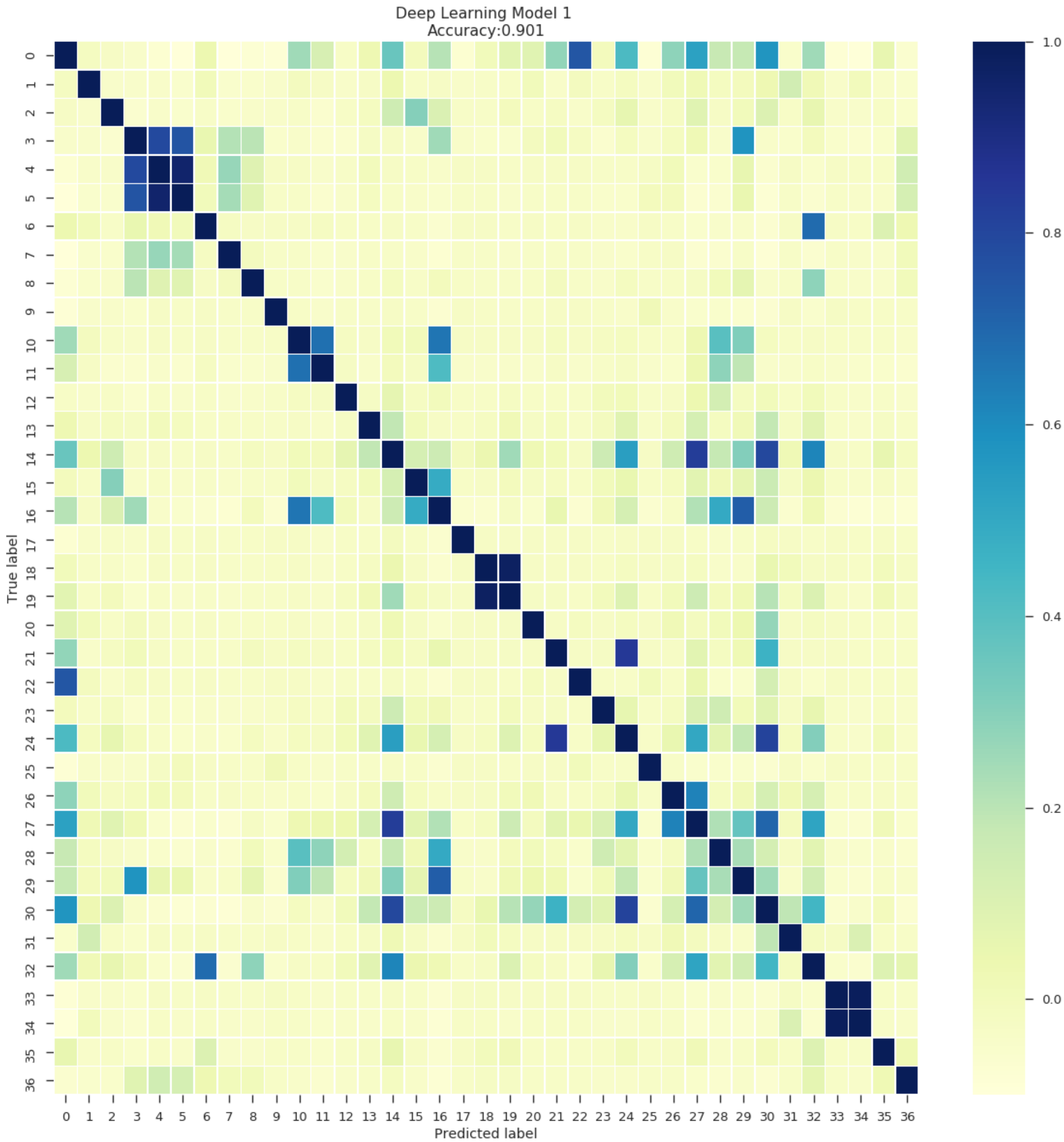
Results indicate that training accuracy is 90.14% and testing accuracy is 90.12%. Five-fold cross validation suggested a moderate level of overfitting, less than 5%. The graphs below show model accuracy and loss by epoch.





When examining the predictions using testing data it is clear that model performance by class is unbalanced. The average unweighted accuracy is 57.9%. The model performed remarkably well in certain categories such as breast surgery, consultation, treatment of lesions and poorly in others such as actions related to organ donation, small surgery, and re-constructive surgery. The heat map below shows the association between true labels and predicted labels. Labels that are being misclassified will have some association with its true label.

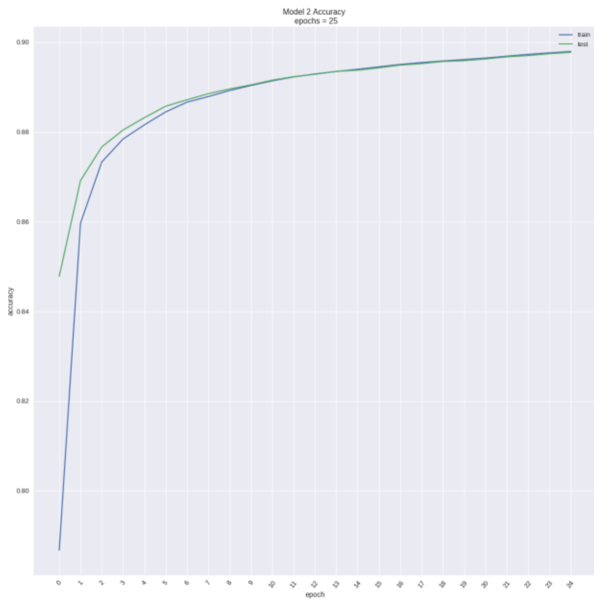
[7] https://ml-cheatsheet.readthedocs.io/en/latest/loss_functions.html

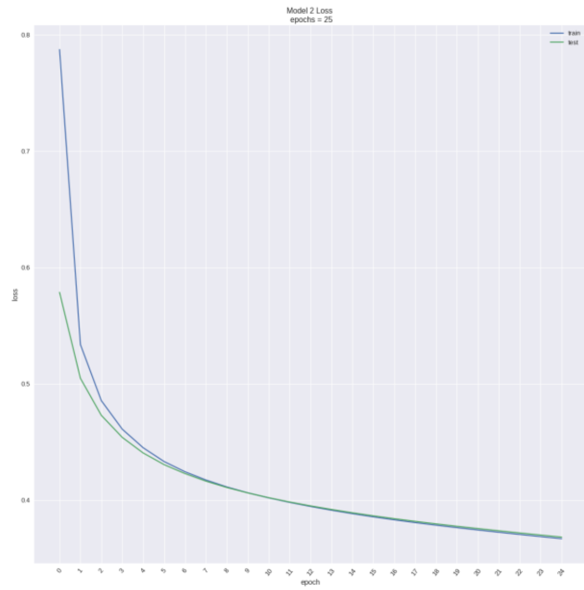

Deep Learning Model 1
Accuracy:0.901
True label
Predicted label
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
1.0
0.8
0.6
0.4
0.2
0.0

MODEL 2: BASELINE + 2 LAYERS

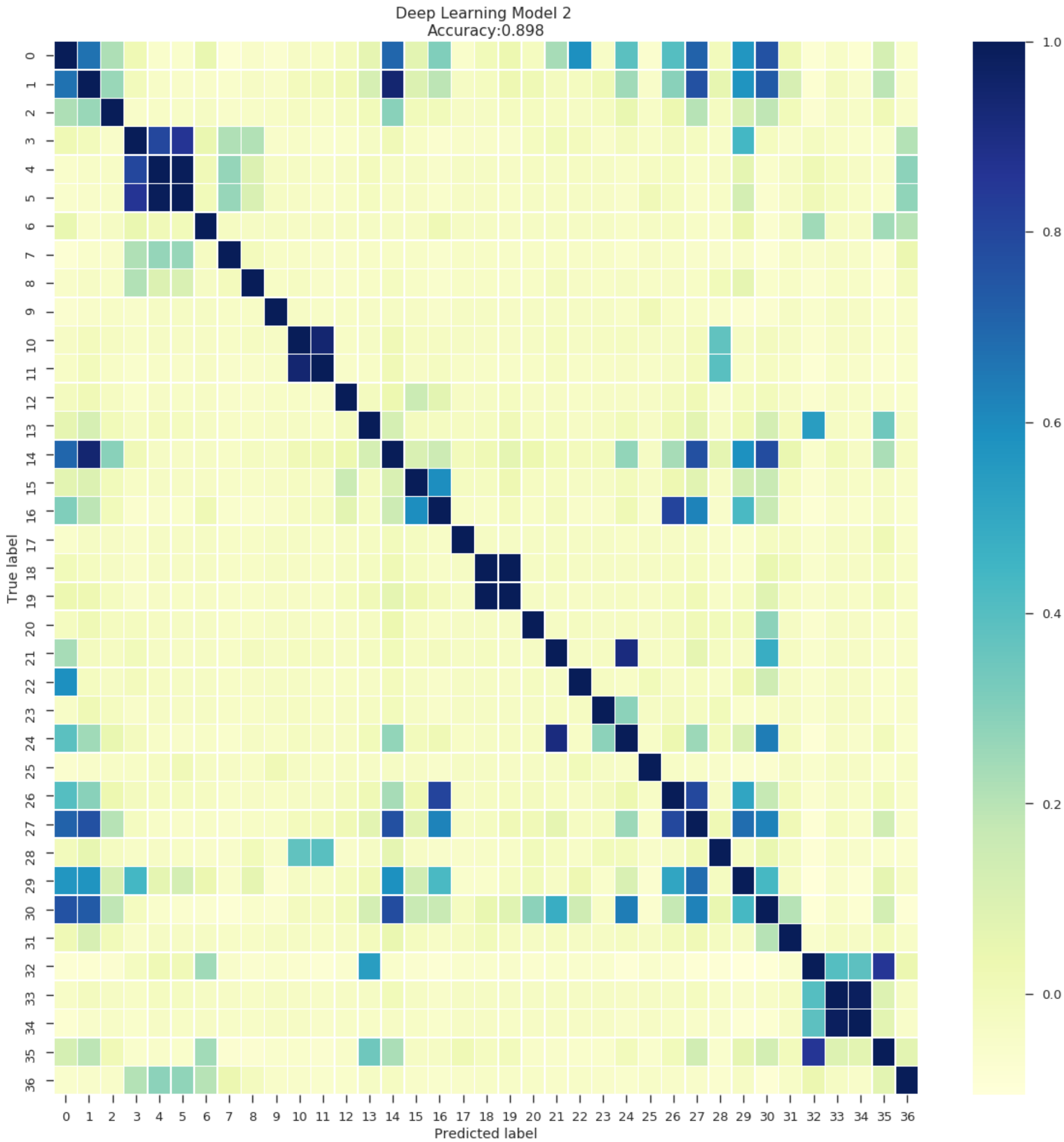
Model 2 is similar than model 1. The difference in this model is the addition of two more dense layers with 20 nodes and ReLU activation function. The purpose of this model is to test the extent the addition of layers improves or worsens model performance.

Results indicate that training accuracy is 89.80% and testing accuracy is 89.78%. Five-fold cross validation suggested a moderate level of overfitting, around 4%. The graphs below show model accuracy and loss by epoch. The addition of two layers did not improve or significantly worsen the performance of the model.





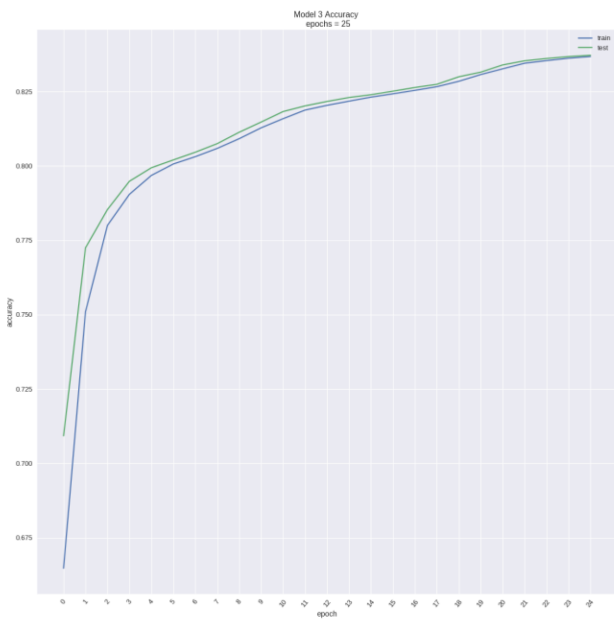
Again, when examining the predictions using testing data it is clear that model performance by class is unbalanced. The average unweighted accuracy is 55.6%. The model performed remarkably well in certain categories such as breast surgery, consultation, treatment of lesions and poorly in others such as actions related to organ donation, small surgery, and re-constructive surgery. The heat map below shows the association between true labels and predicted labels. Labels that are being misclassified will have some association with its true label.

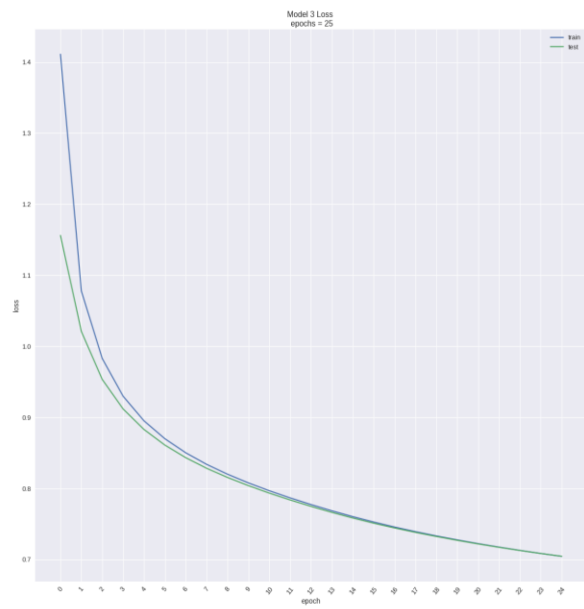

Deep Learning Model 2
Accuracy:0.898
True label
Predicted label
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
1.0
0.8
0.6
0.4
0.2
0.0

MODEL 3: BASELINE + L2 REGULARIZATION

Model 3 is similar than model 1. The difference is the addition of L2 regularization term in the two dense layers. The purpose of this model is to test whether L2 regularization decreases the level of model overfitting.

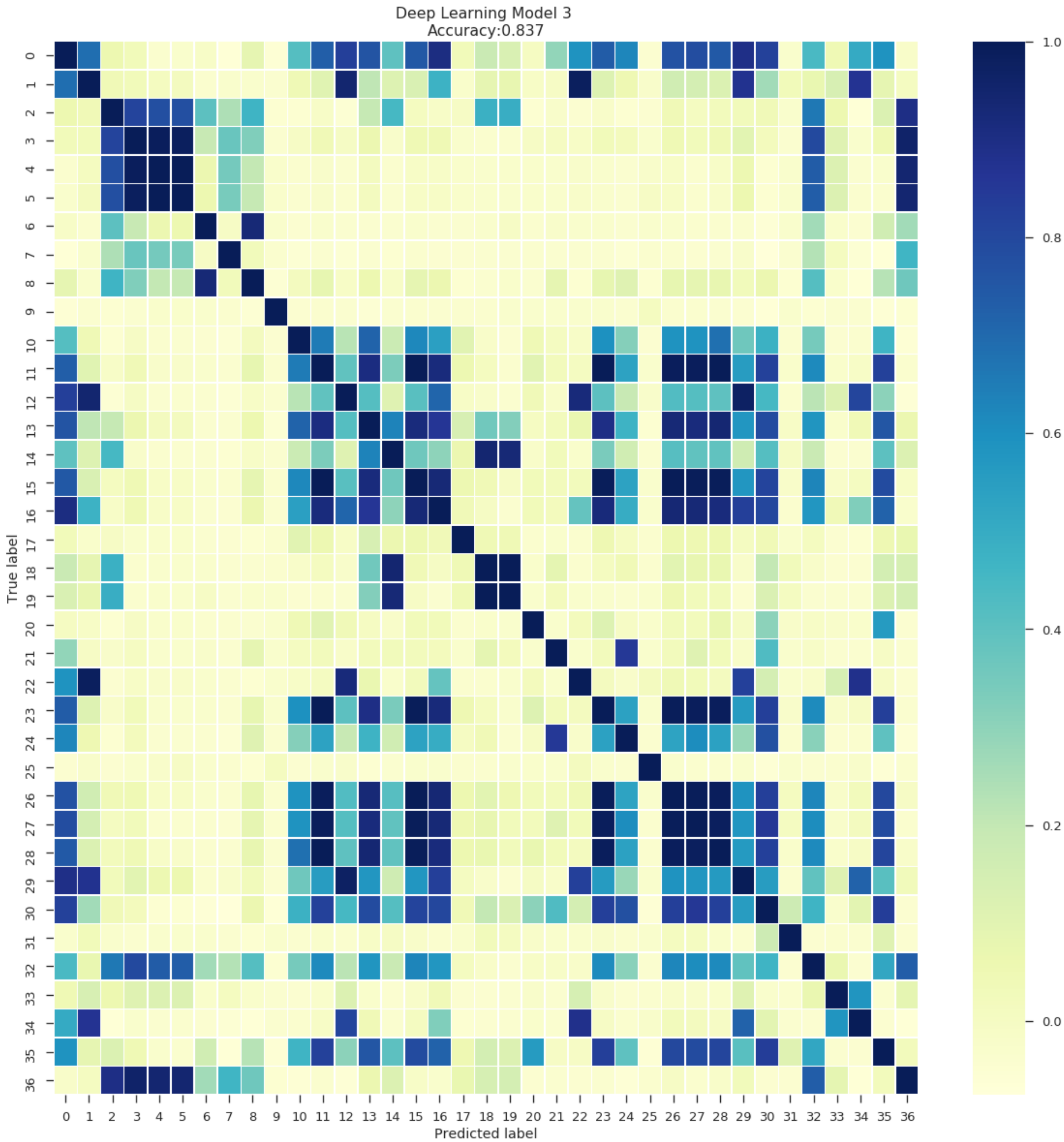
Results indicate that training accuracy is 83.16% and testing accuracy is 83.72%. Five-fold cross validation suggested a moderate level of overfitting, a bit over 5%. The L2 regularization term seem to have moderately worsen overfitting and accuracy performance. It is possible that the L2 term is penalizing the weights too much and the network is not sufficiently learning.





Again, when examining the predictions using testing data it is clear that model performance by class is unbalanced. The average unweighted accuracy is 31.7%. This is significantly worse than models 1 and 2. The model performed remarkably well in certain categories such as breast surgery, consultation, treatment of lesions and poorly in others such as actions related to organ donation, small surgery, and re-constructive surgery.

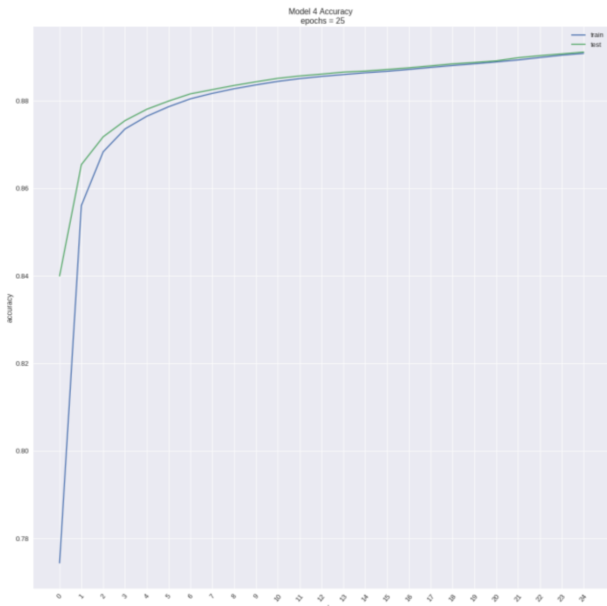
The heat map below shows the association between true labels and predicted labels. Labels that are being misclassified will have some association with its true label. When compared with the other two models above, this model is misclassifying at a higher rate.

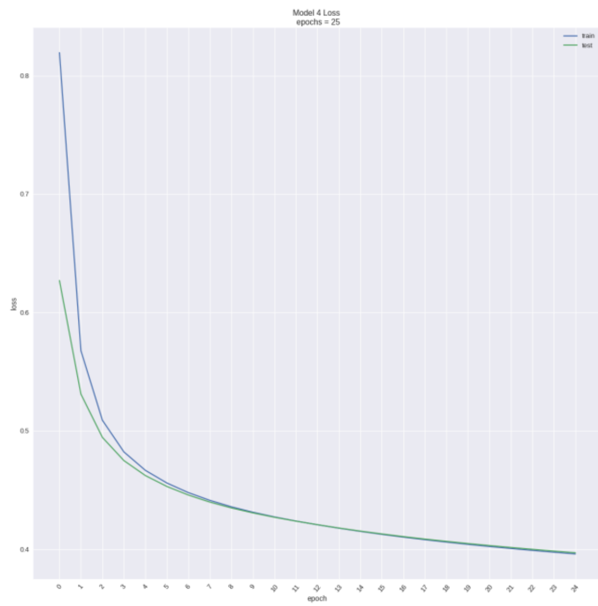

Deep Learning Model 3
Accuracy:0.837
True label
Predicted label
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
1.0
0.8
0.6
0.4
0.2
0.0

MODEL 4: BASELINE + 4 LAYERS

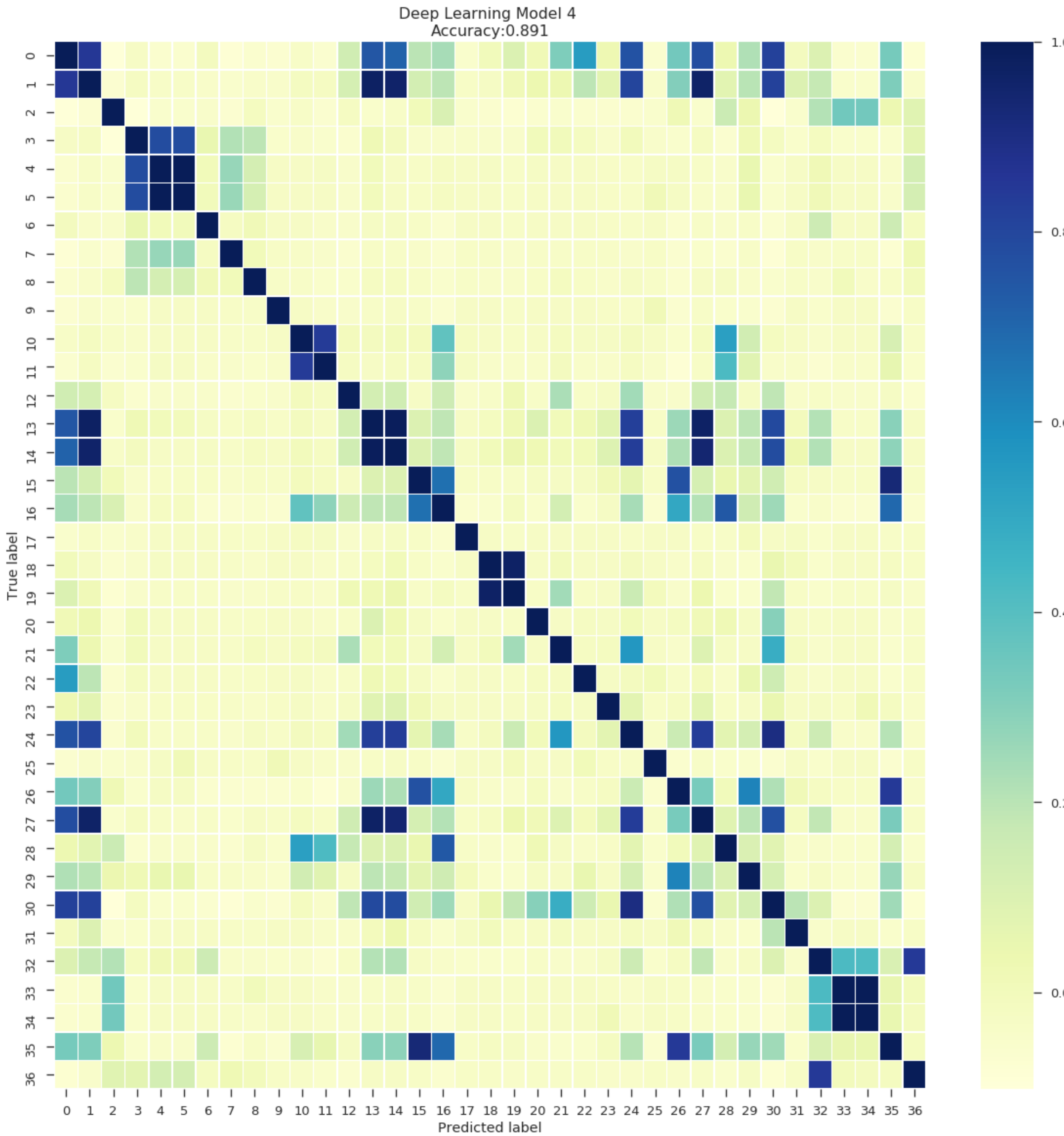
Model 4 is similar than model 2. The difference is the addition of two more dense layers for a total of six layers, no regularization used. The purpose of this model is to test further the extent additional layers increases or worsens model performance.

Results indicate that training accuracy is 89.08% and testing accuracy is 89.11%. Five-fold cross validation suggested a moderate level of overfitting, around 3%. The graphs below show model accuracy and loss by epoch. The fact that the model had six layers four more than model 1 and two more than model 2 did not improve or significantly worsen the performance of the model.





Again, when examining the predictions using testing data it is clear that model performance by class is unbalanced. The average unweighted accuracy is 49.1%. The model performed remarkably well in certain categories such as breast surgery, consultation, treatment of lesions and poorly in others such as actions related to organ donation, small surgery, and re-constructive surgery. The heat map below shows the association between true labels and predicted labels. Labels that are being misclassified will have some association with its true label.

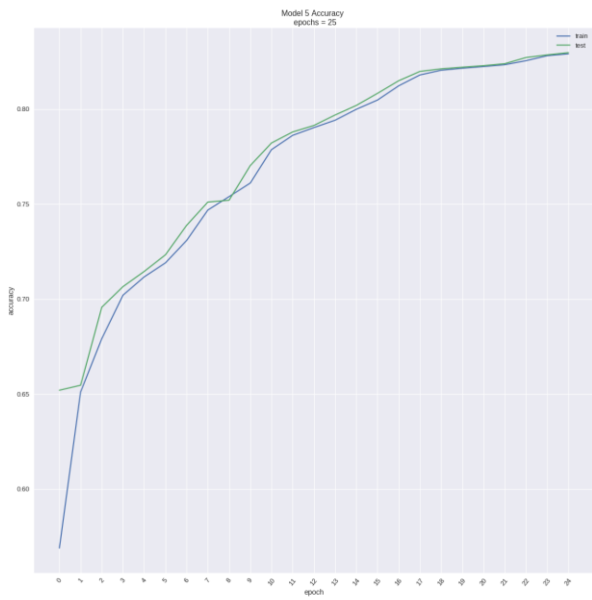


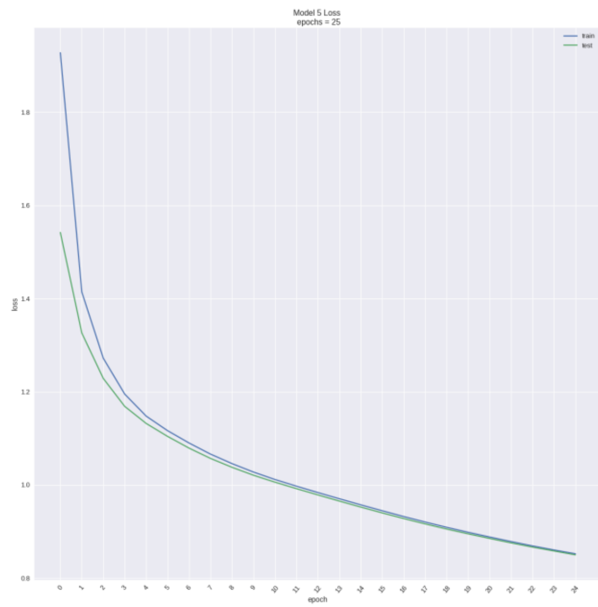
Deep Learning Model 4
Accuracy:0.891
True label
Predicted label
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
1.0
0.8
0.6
0.4
0.2
0.0

MODEL 5: BASELINE + L1 REGULARIZATION

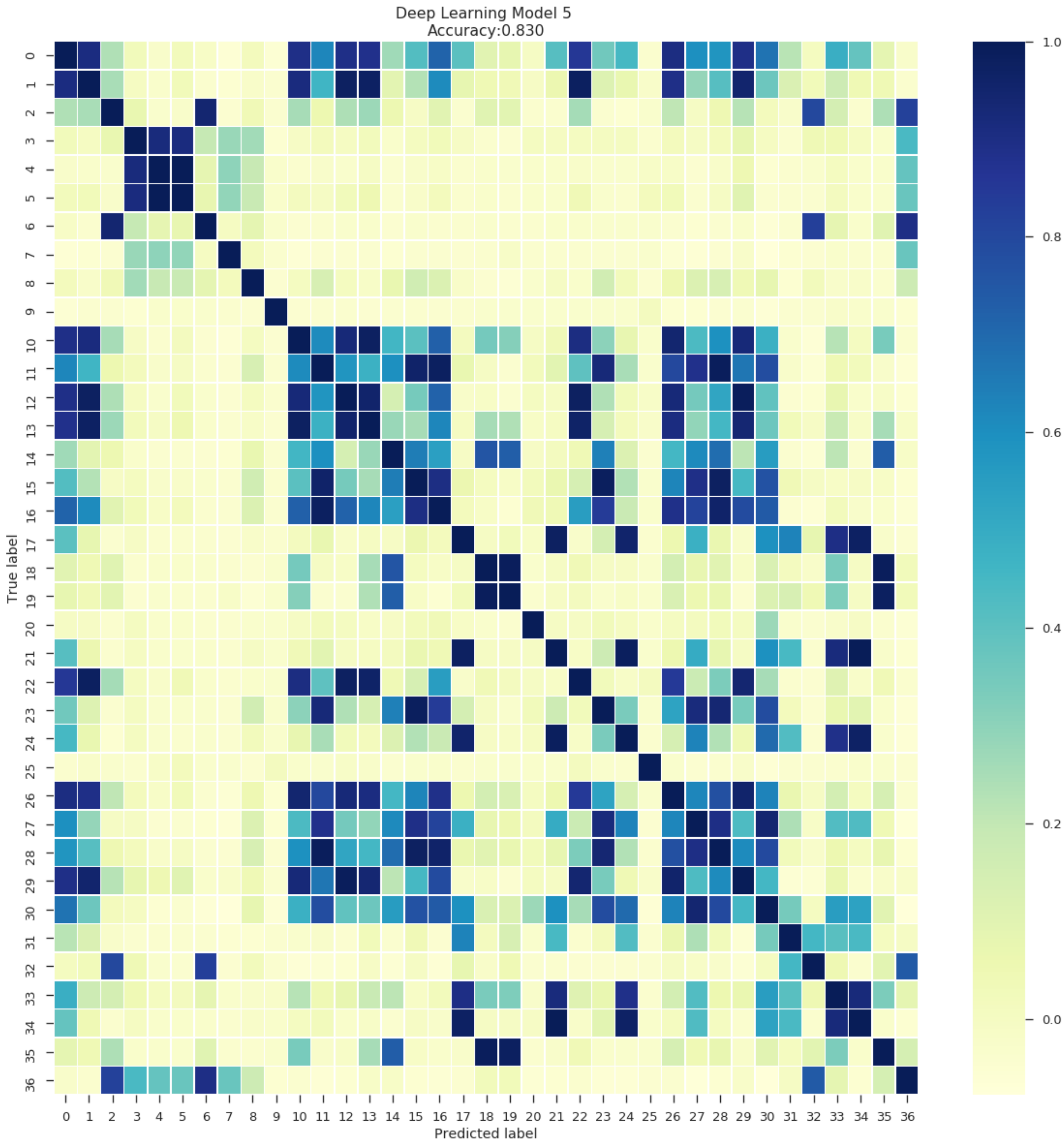
Model 5 is similar than model 1. The difference is the addition of L1 regularization term in the two dense layers. The purpose of this model is to test whether L1 regularization will decrease the level of model overfitting.

Results indicate that training accuracy is 82.81% and testing accuracy is 82.86%. Five-fold cross validation suggested a significant level of overfitting, around 10%. The graphs below show model accuracy and loss by epoch. Of all the experimental models, this was the worst performing in all metrics and the model with the most overfitting.





Again, when examining the predictions using testing data it is clear that model performance by class is unbalanced. The average unweighted accuracy is 26.7%. The heat map below shows the association between true labels and predicted labels. Labels that are being misclassified will have some association with its true label.

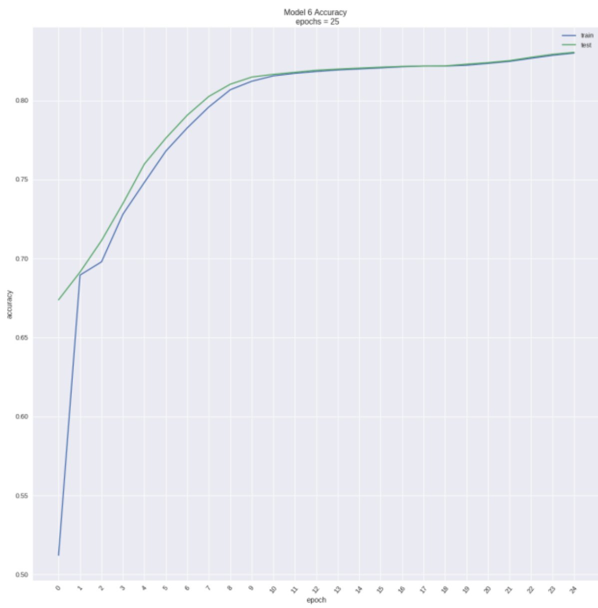


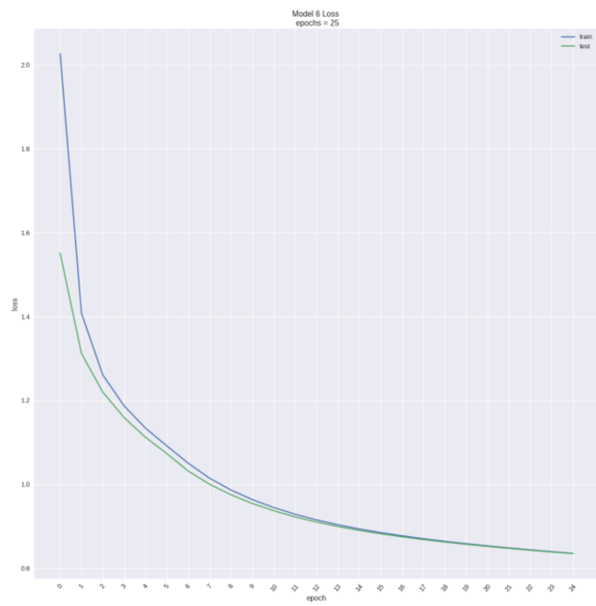
Deep Learning Model 5
Accuracy:0.830
True label
Predicted label
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
1.0
0.8
0.6
0.4
0.2
0.0

MODEL 6: BASELINE + DROPOUT

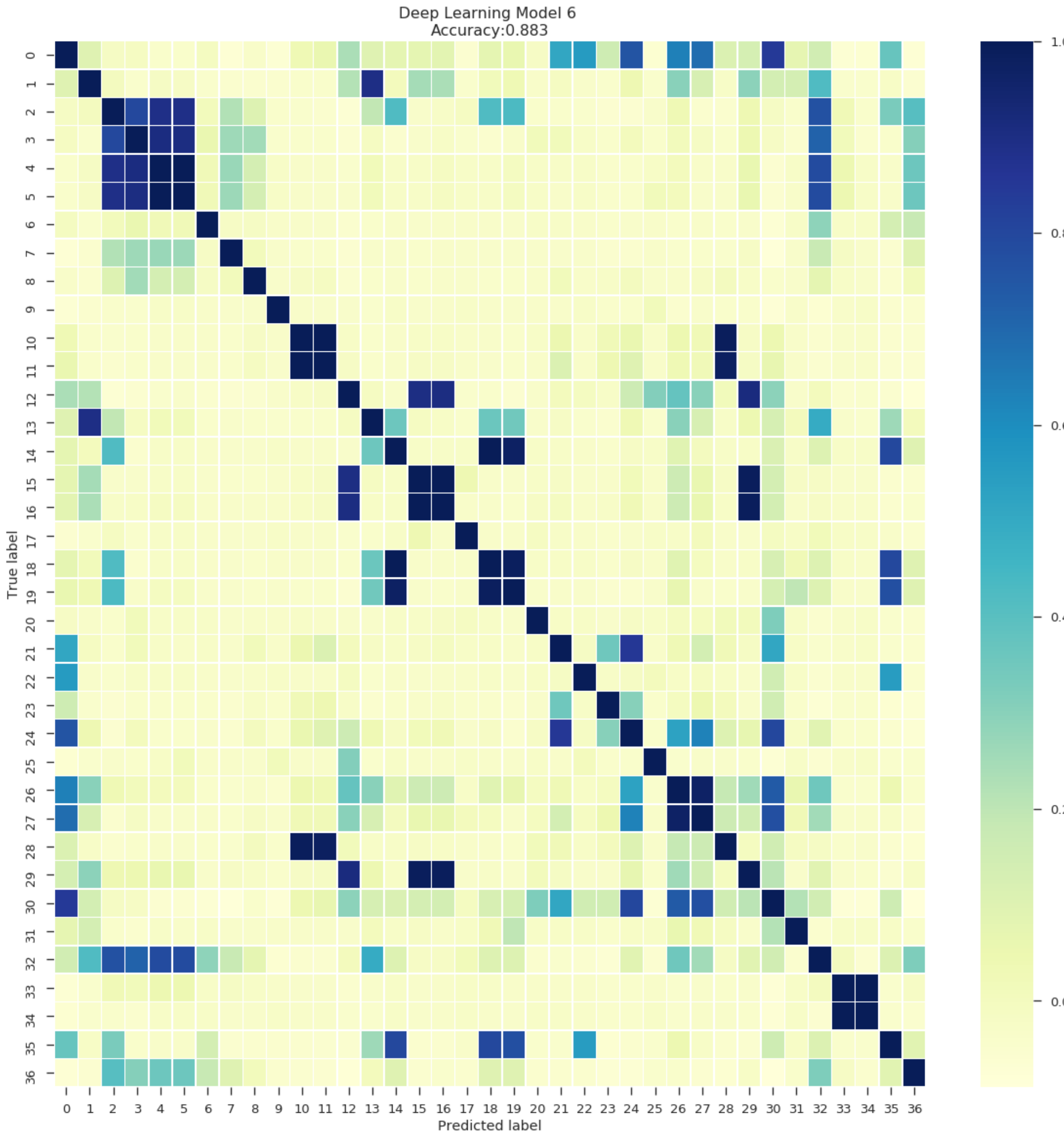
Model 6 is similar than model 1. The difference is the addition of dropout regularization layers before each of the two dense layers. The purpose of this model is to test whether dropout regularization will decrease the level of model overfitting.

Results indicate that training accuracy is 85.63% and testing accuracy is 88.32%. Results suggested an small amount of overfitting with the difference between the average cross-validation accuracy and training accuracy at 0.36%. The graphs below show model accuracy and loss by epoch. While dropout did decrease the accuracy of the model somewhat when compared with model 1, it significantly decreased overfitting as well.





Again, when examining the predictions using testing data it is clear that model performance by class is unbalanced. The average unweighted accuracy is 43.2%. The model performed remarkably well in certain categories such as breast surgery, consultation, treatment of lesions and poorly in others such as actions related to organ donation, small surgery, and re-constructive surgery. The heat map below shows the association between true labels and predicted labels. Labels that are being misclassified will have some association with its true label.

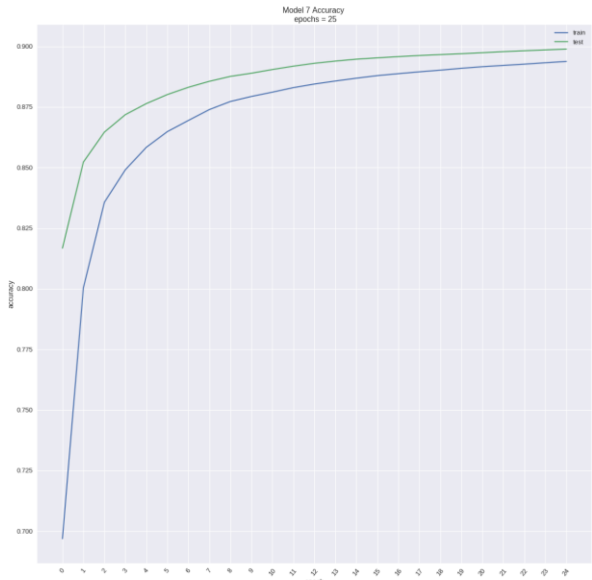

Deep Learning Model 6
Accuracy:0.883
True label
Predicted label
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
1.0
0.8
0.6
0.4
0.2
0.0

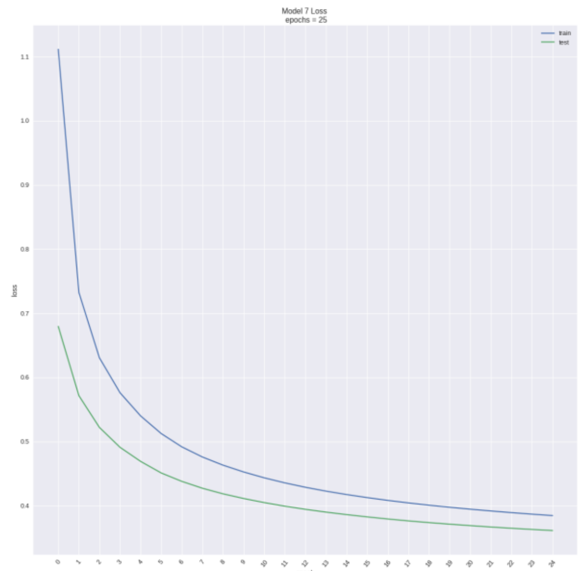
MODEL 7: BASELINE + RANDOM OVERSAMPLING

The performance of all the models discussed up to this point has been unbalanced. Meaning that the model has performed very well for certain classes and poorly in others. The purpose of model 7 is to test whether model performance could be improved by randomly oversampling the minority classes. Random oversampling works by duplicating some of the original samples of the minority class.

Model 7 is the same as model 1 in structure. However, the training data has been oversampled. Testing data was not oversampled. The sampling strategy was to randomly oversample all classes with less 100,000 observations. The random oversampling was done using the imbalanced-learn library[8].

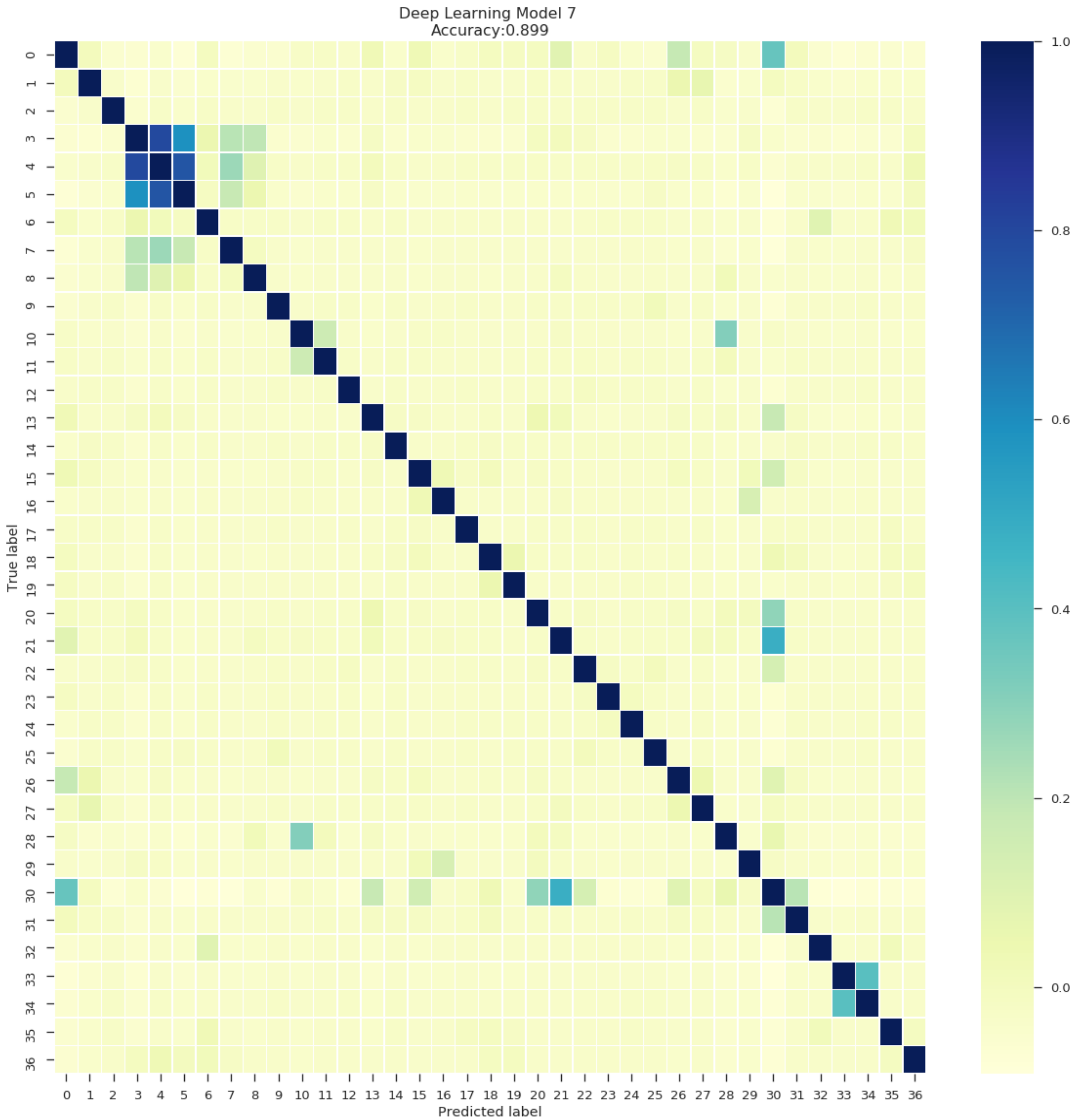
Results indicate that training accuracy is 89.38% and testing accuracy is 89.89%. Five-fold cross validation suggested a significant level of overfitting, around 6%. The graphs below show model accuracy and loss by epoch.





The average unweighted accuracy is 82%. This is a significant improvement in average unweighted accuracy from previous models. The heat map below shows the association between true labels and predicted labels. Labels that are being misclassified will have some association with its true label.

[8] https://imbalanced-learn.readthedocs.io/en/stable/

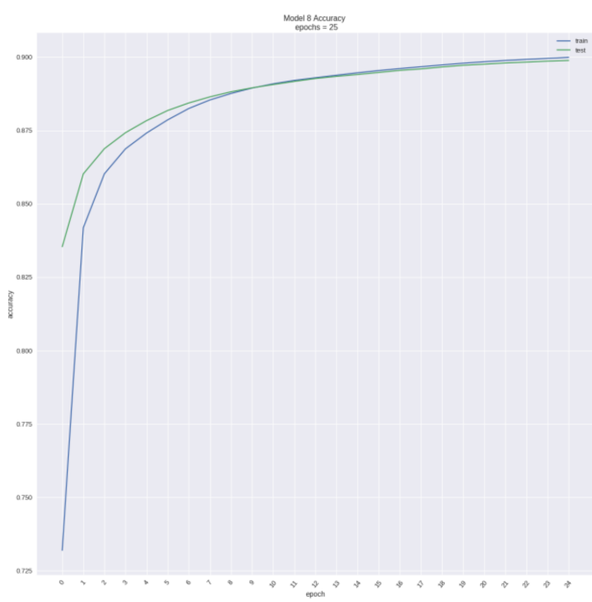


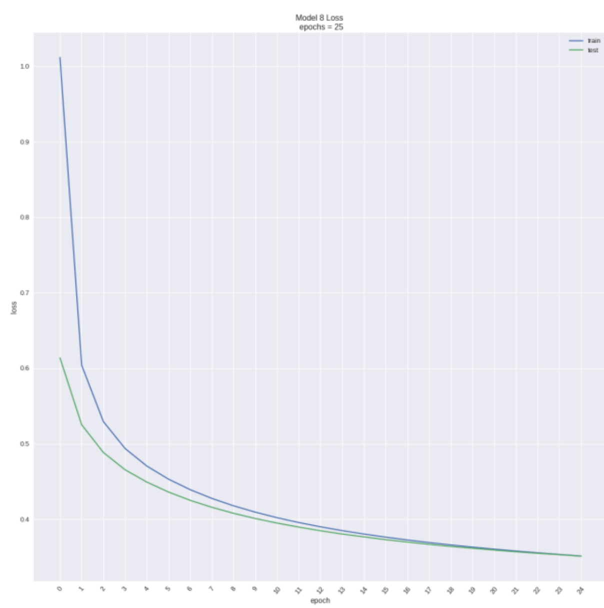
Deep Learning Model 7
Accuracy:0.899
True label
Predicted label
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
1.0
0.8
0.6
0.4
0.2
0.0

MODEL 8: BASELINE + SMOTE-NC OVERSAMPLING

The purpose of model 8 is to test whether model performance could be improved by oversampling the minority classes using SMOTE-NC. SMOTE is the acronyms of *'Synthetic Minority Over-Sampling Technique'.* This is a technique that generates new samples by interpolation. SMOTE uses k-Nearest Neighbors to find similar samples and generate new ones. Thus, SMOTE-NC is an extension of SMOTE that allows for using both numerical and categorical data. Again, only training data was oversampled, testing data remains as the original.

The sampling strategy was to oversample all classes with less 100,000 observations. The SMOTE-NC oversampling was done using the imbalanced-learn library[9]. Results indicate that training accuracy is 89.99% and testing accuracy is 89.88%. Five-fold cross validation was not possible for with model because in some cross-validation folds there was not enough data in some classes to perfom k-nearest neighbors to create new samples. In short, SMOTE-NC did not improve overall accuracy performance and when compared with the random oversampling strategy it performed similarly in terms of average unweighted accuracy across classes. The graphs below show model accuracy and loss by epoch.





The average unweighted accuracy is 80.1%, slightly below random oversampling. This is a significant improvement in average unweighted accuracy over models 1-6, but slighly below model 7. The heat map below shows the association between true labels and predicted labels. Labels that are being misclassified will have some association with its true label.

[9] https://imbalanced-learn.readthedocs.io/en/stable/

SUMMARY OF EXPERIMENTAL MODELS

The worst performing model was model 5, this model used L1 regularization. Model 3 which used L2 regularization, decreased model performance and over fitted to the same extent as model 1, which had not regularization at all.

The models with additional layers over model 1 (models 2 and 4) did not improve performance over the simpler model 1. Model 6, which used dropout regularization decreased testing accuracy when compared with model 1, however it performed similarly to model 1 in cross-validation and significantly decreased overfitting.

Models 7 and 8, used oversampled data to fit model 1. They performed slightly worse than model 1, however the average accuracy by class was significantly improved. This suggests that performance across classes was more balanced than previous models. During cross-validation model 7 (random oversampling) showed signs of a moderate overfitting (6%). The false positive rate was similar cross models for exception of models 3 and 5, these two models had higher levels of false positives.

PARAMETER TESTING

Five-fold cross validation was used to test model performance using different parameters for batch size, learning rate and embedding dimensions. These parameters are considered to be key in this model and therefore prioritized for further testing. Model 1 was used for all the tests.

The goal was to do some parameter tuning of key parameters such as learning rate, batch size and embedding layer dimensions. Given computational resources and time required, the parameters tested was kept small in both scope and range. With more resources and time, a wider a larger range of values could be tested and additional parameters such as nodes and regularization term value could be tested as well.

Results suggests that a smaller batch size of 32 and larger embedding dimensions yield better performance in terms of accuracy. A smaller learning rate needs a larger number of epochs to learn more, as such the larger learning rate of 0.001 tested yielded higher accuracy in these experiments.

Batch Size	Learning Rate	Embedding Dimension	**5 fold CV** *(epochs = 25)*	
			Training Accuracy	**5 Fold CV Evaluation**
128	0.001	10	85.86%	85.86% (+/- 0.21%)
128	0.0001	10	48.08%	48.10% (+/- 3.77%)
128	0.0001	30	58.46%	55.98% (+/- 4.01%)
32	0.0001	10	55.07%	55.45% (+/- 3.18%)
32	0.001	10	86.43%	86.39% (+/- 0.77%)
32	0.0001	30	66.99%	67.13% (+/- 1.80%)
32	0.001	30	88.05%	87.97% (+/- 0.31%)

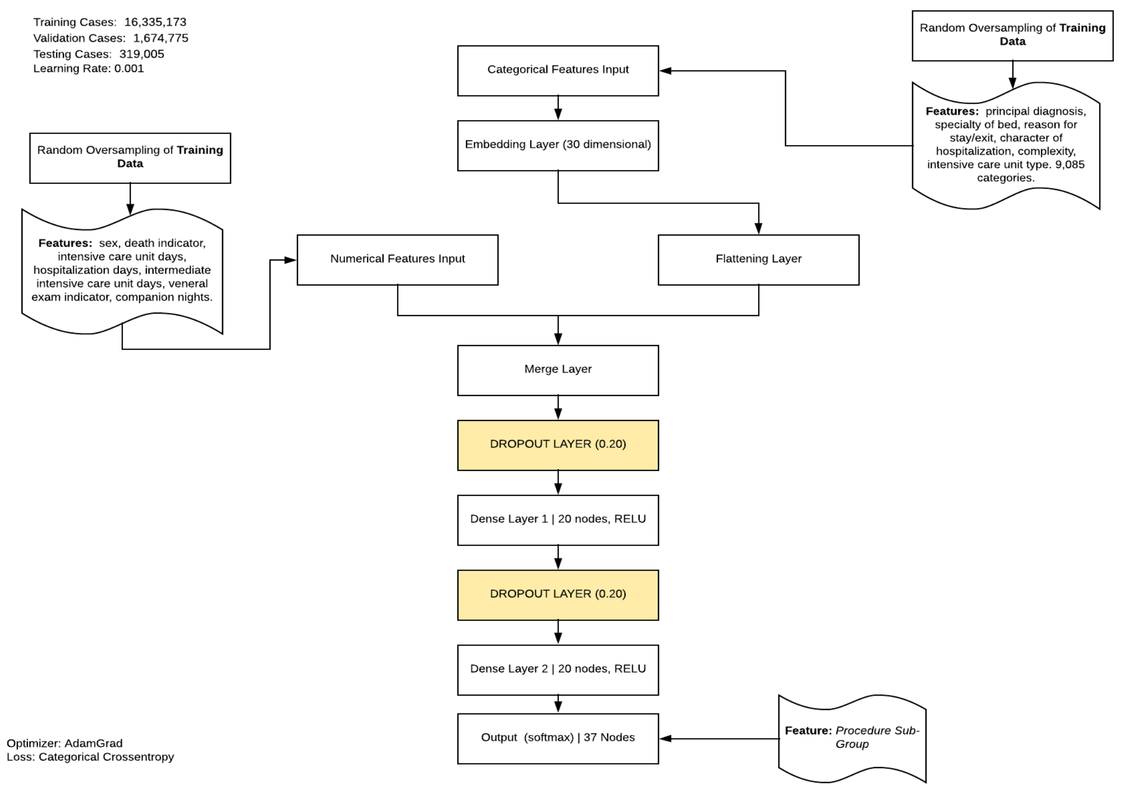
FINAL MODEL

Given the results of experimental models and parameter tuning a final model was constructed. This model attempts to maximize model accuracy both overall and by class, and minimize the false positive rate. To improve average accuracy by class a random oversampling strategy was used to help the model learn the minority classes better. Given that random oversampling and SMOTE-NC oversampling performed similarly, the simpler and faster random oversampling method was used.

The main parameters used were:

- Batch Size = 64
- Learning Rate = 0.001
- Embedding Dimensions = 30
- Dropout Rate = 0.20
- Epochs = 50

DEEP NEURAL NETWORK ARCHITECTURE - Final Model

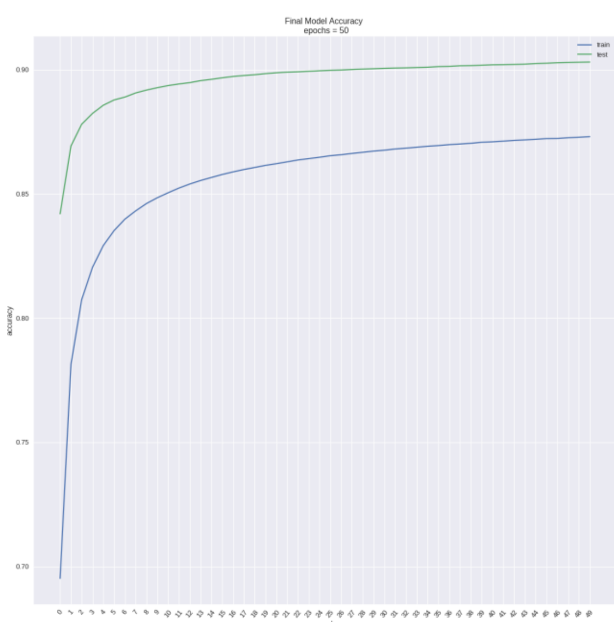


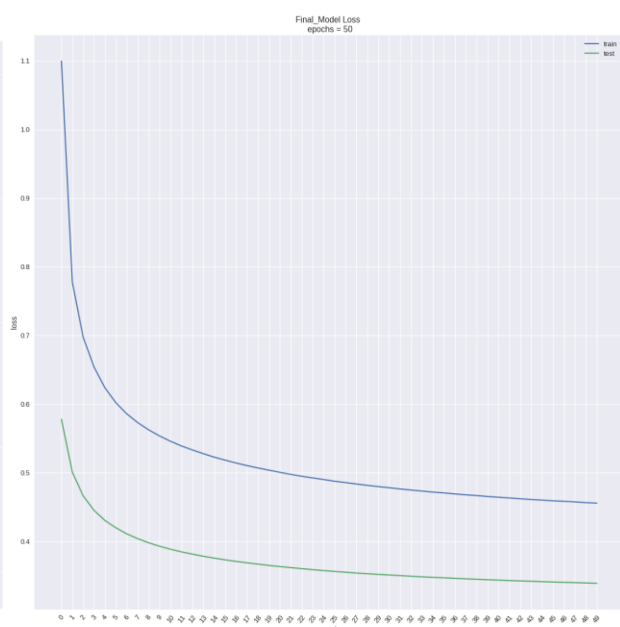
The main considerations for the batch size and the learning rate were computational resources and time. A model was run with a smaller batch size and smaller learning rate than the final model presented here. This model's training accuracy was 72.45%. testing accuracy was 82.63 and five-fold cross validation score was 62.22% (+/- 1.32%). Moreover, this model's average accuracy by class was fairly low at 33.6%. Which meant that the model was misclassifying a significant number of classes.

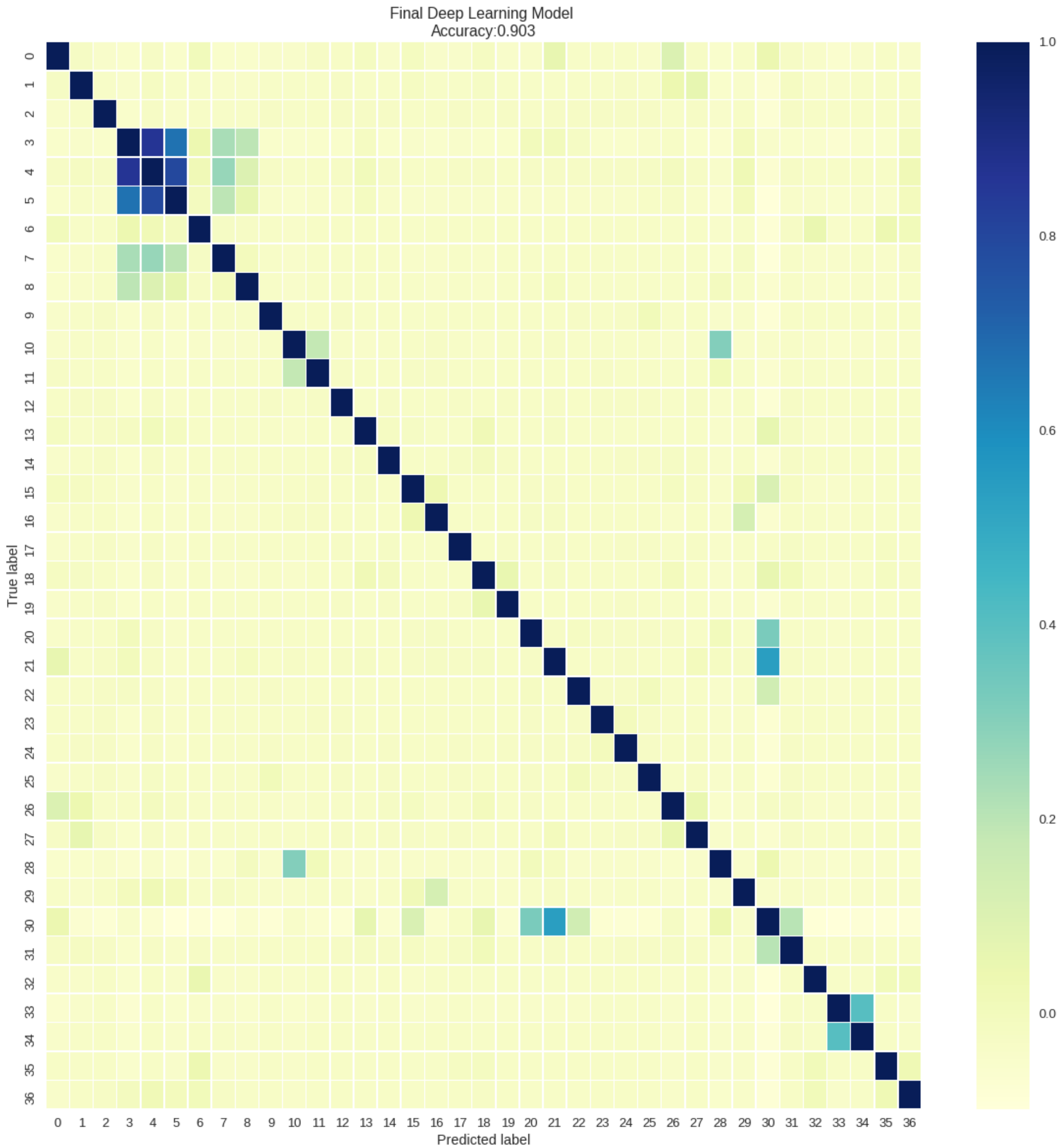
This lead to the conclusion that setting the batch size and learning rate smaller required a much larger number of epochs and time to train to achieve similar or slightly better results to the final model presented here. The model presented here takes an estimated 20 hours to train.

The final model results indicate a training accuracy is 87.31% and testing accuracy is 90.31%. Five-fold cross validation results suggests moderate overfitting, less than 3%. Finally, average accuracy by class is 84%. This is the best average accuracy by class achieved throughout all the models tested.

Model Number	**Training & Testing Data**				**Validation Data**		**Total Training Time**
	Training Accuracy	**Testing Accuracy**	**Average Accuracy***	**False Positive Rate**	**CV (k = 5)**	**Diff CV & Training**	
Final Model	87.31%	90.31%	80.4%	0.003	84.62 (+/- 0.35%)	2.69	19 hours







Final Deep Learning Model
Accuracy:0.903
True label
Predicted label
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
1.0
0.8
0.6
0.4
0.2
0.0

RECOMMENDATIONS AND NEXT STEPS

Opportunities to further optimize the neural network remain. Parameter optimization could be performed on key parameters, such as learning rate, batch size, embedding dimensions.

Further, parameter tuning of epochs, nodes and the regularization term was not performed due to resource and time constraints. Moreover, the architecture here is not the only option to use. Nonetheless, given the problem and evaluation results it showed to be a reasonable choice. Other choices were to use either hashing trick or grouping the categories instead of an embedding layer to process the large amounts of categorical data.

My recommended next steps are with level of priority:

1. Hyper Parameter Tuning and Optimization
    - Dropout Rate
    - Learning Rate
    - Batch Size
    - Embedding Dimensions
    - Nodes
    - Regularization Term
2. Further data engineering: from the graph, it is clear that there is a cluster of categories which the network is having difficulty with. It is likely that these maybe profiles of patients or procedures that are related.
3. Pruning the Network: removing nodes that do not contribute much to the predictions.
4. Identifying Most Important Nodes and Features for the Prediction

Given that the domain is healthcare, it is reasonable to expect some level of review of the predictions. Before embarking in further optimization and testing is important to consider the trade-offs and the cost-benefit ratio to the business. Building, training and tuning a neural network is time consuming and computationally expensive. Thus, a key question is how much the business is willing to invest to increase network performance by a few percentage points more. The final model performance is as follows: (1) training accuracy is 87.31%, (2) testing accuracy is 90.31%, (3) average testing accuracy (i.e. accuracy by class) is 80.4%, cross-validation accuracy score is 84.62 (+/- 0.35%), and false positive rate 0.003.

Moreover, given that in the field of healthcare having certain business processes for processing and using predictions, as well handling rarer procedures categories is advised. In summary, given the time allotted and computational resources available an adequately performing model in terms of accuracy and overfitting was achieved. Opportunities remain for the improving model.

SOURCES

1. Choi et al. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. 2016. https://arxiv.org/abs/1511.05942
2. de Araujo Lima et al. Successful Brazilian Experiences in the Field of Health Information – Final Report. 2016. https://www.measureevaluation.org/our-work/health-information-systems/health-information-system-strengthening-in-lac-region-2005-2010/his-brazil-english-august2007.pdf
3. Malhotra et al. Long Short Term Memory Networks for Anomaly Detection in Time Series. 2015. Long Short Term Memory Networks for Anomaly Detection in Time Series. https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2015-56.pdf
4. Ralf Carvalho & Alex Paino. Deep Learning for Fraud Detection. https://engineering.siftscience.com/deeplearning-fraud-detection/
5. Sapronova et al. Deep learning for wind power production forecast. 2017. http://ceur-ws.org/Vol-1818/paper3.pdf
6. Shipmon et al. Time Series Anomaly Detection. https://ai.google/research/pubs/pub46283
7. Tang and Mendis. Unsupervised fraud detection in Medicare Australia. 2011. https://pdfs.semanticscholar.org/e7b6/9c1ba648de68d30aacb84b47ab43fcdf75e9.pdf
8. Teodoro et al. ORBDA: An openEHR benchmark dataset for performance assessment of electronic health record servers. 2017. https://www.ncbi.nlm.nih.gov/pubmed/29293556
9. Tianwei Yue & Haohan Wang. Deep Learning for Genomics: A Concise Overview. 2018. https://arxiv.org/abs/1802.00810
10. Viscondi et al. Cost Description of Prevention and Treatment of Cervical Cancer. https://www.researchgate.net/profile/Juliana_Viscondi/publication/299483604_COST_DESCRIPTION_OF_PREVENTION_AND_TREATMENT_OF_CERVICAL_CANCER_IN_BRAZIL_IN_2012/links/59b6786ea6fdcc7415bd3940/COST-

DESCRIPTION-OF-PREVENTION-AND-TREATMENT-OF-CERVICAL-CANCER-IN-BRAZIL-IN-2012.pdf