Research log

Dennis Scheper

11-9-2020

1 Week 1 - original dataset

Date: 11 September, 2020 (week 1)

1.1 Introduction

The article, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records" states that hyperglycemia management has a significant impact on the outcome and readmission rates of hospitalized patients. The authors based this conclusion on the comprehensive assessment of 70.000 diabetes patient records retrieved from 140 US hospitals. The results that depict the relationship between readmission rates and the measurement of HbA1c levels can further enlargement of already existing diabetes tactics to reduce readmission rates.

1.2 Dataset and Attributes Information

All information used in this article comes from a database, consisting of 41 tables and totals 117 features, such as demographics (like gender, race, and age), inpatient or outpatient, and (in-hospital) mortality. Data came from 130 hospitals in the USA for over ten years (1998-2008) and contained around 74 million unique visits by 18 million unique patients. This research used information that needs to accede to the following specifications:

1. Is a hospital admission;
2. The encounter is a diagnosed with 'diabetes', any kind will satisfy;
3. The length of admission was at least one day up to eighteen days;
4. Laboratory test results are available; and
5. Medications were administered.

Of these five criteria points, 101.000 encounters fulfilled all specifications. After some considerations with removing encounters based on incomplete (weight and medical specialty) or biased data (discharge to a hospice or death), 69.984 encounters remained in the final dataset.

The initial dataset consists of 55 attributes, with the class attribute being an encounter of one patient. As there are way too many attributes to describe, please refer to the codebook for all descriptions; we only look at some essential attributes and their type and possible valuations. The age of a patient is nominal and is grouped into ten-year intervals. The admission type or for what specific reason a patient was hospitalized and comes in 9 distinct values while the type is nominal. Some attributes are numeric and count, such as the number of lab procedures, the number of medications, and the number of emergency visits. The database consists of three diagnosis attributes, which can have 848, 923, and 954 distinct values, respectively. The values are based on ICD9 three-letter codes and are of the nominal type. Other vital attributes are whether a patient changed medications (with the values ‘no change’ and ‘change’; nominal type) or had diabetes medication (‘yes’ or ‘no’ values and nominal typing). Twenty-four other attributes depict whether medicine is prescribed or not, if prescribed, then if the dosage was increased (‘up’), decreased (‘down’), or stayed the same (‘steady’) during the encounter. Readmission rates were calculated by looking at a nominal type (‘Readmitted’) with the possible valuations of ‘<30’ for a patient that was readmitted within 30 days, ‘>30’ for a patient that was readmitted after 30 days, and ‘No’ for patients that were not readmitted. The authors’ goal was to determine whether a relationship between readmission rates and HbA1c measurement exists; therefore, they introduced a new attribute ‘HbA1c’ with four different valuations, based on the information from the database: 1) no HbA1c test performed; 2) HbA1c performed and in the normal range; 3) HbA1c performed and the result is greater than 8% with no change in diabetic medication; and 4) Hb1Ac performed, the result is more significant than 8%, and diabetic medication was changed.

1.3 Research Question

Is it possible, using machine learning techniques, to predict whether a patient’s time in hospital is linked to the results of an HbA1c measurement?

2 Week 2 - EDA (Exploratory Data Analysis)

2.1 Date: 14 September, 2020

This section will perform an exploratory data analysis (EDA) on the dataset described above. With an EDA, we can explore our dataset and determine if any correlations exist between attributes and undermine any missing values. If any exist, we deal with them accordingly, looking to repair these values or remove them entirely. Additionally, we will look at variations between and in attributes for determining which ones are of most importance and interest to our research goals.

First, we load in our used packages, mostly used for visualization of data, and the data itself. Besides that, we also load in a codebook, which contains, i.e., a description for every attribute for the initial dataset. The codebook was not retrieved from any outside source but was constructed on the interpretation of the data.

To get a better picture of our dataset's distribution, we called upon the function glimpse(). We notice that the data contains 50 columns/ attributes. Additionally, a summary() function gives an outline of every attribute, showing each level with a count of its valuation.

```
## Rows: 101,766
## Columns: 50
## $ encounter_id             <int> 2278392, 149190, 64410, 500364, 16680, 357...
## $ patient_nbr              <int> 8222157, 55629189, 86047875, 82442376, 425...
## $ race                     <fct> Caucasian, Caucasian, AfricanAmerican, Cau...
## $ gender                   <fct> Female, Female, Female, Male, Male, Male, ...
## $ age                      <fct> [0-10), [10-20), [20-30), [30-40), [40-50)...
## $ weight                   <fct> ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ...
## $ admission_type_id        <int> 6, 1, 1, 1, 1, 2, 3, 1, 2, 3, 1, 2, 1, 1, ...
## $ discharge_disposition_id <int> 25, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 3, 6,...
## $ admission_source_id      <int> 1, 7, 7, 7, 7, 2, 2, 7, 4, 4, 7, 4, 7, 7, ...
## $ time_in_hospital         <int> 1, 3, 2, 2, 1, 3, 4, 5, 13, 12, 9, 7, 7, 1...
## $ payer_code               <fct> ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ...
## $ medical_specialty        <fct> Pediatrics-Endocrinology, ?, ?, ?, ?, ?, ?...
## $ num_lab_procedures       <int> 41, 59, 11, 44, 51, 31, 70, 73, 68, 33, 47...
## $ num_procedures           <int> 0, 0, 5, 1, 0, 6, 1, 0, 2, 3, 2, 0, 0, 1, ...
## $ num_medications          <int> 1, 18, 13, 16, 8, 16, 21, 12, 28, 18, 17, ...
## $ number_outpatient        <int> 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ number_emergency         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, ...
## $ number_inpatient         <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ diag_1                   <fct> 250.83, 276, 648, 8, 197, 414, 414, 428, 3...
## $ diag_2                   <fct> ?, 250.01, 250, 250.43, 157, 411, 411, 492...
## $ diag_3                   <fct> ?, 255, V27, 403, 250, 250, V45, 250, 38, ...
## $ number_diagnoses         <int> 1, 9, 6, 7, 5, 9, 7, 8, 8, 8, 9, 7, 8, 8, ...
## $ max_glu_serum            <fct> None, None, None, None, None, None, None, ...
```

```
## $ A1Cresult                <fct> None, None, None, None, None, None, None, ...
## $ metformin                <fct> No, No, No, No, No, No, Steady, No, No, No...
## $ repaglinide              <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ nateglinide              <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ chlorpropamide           <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ glimepiride              <fct> No, No, No, No, No, No, Steady, No, No, No...
## $ acetohexamide            <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ glipizide                <fct> No, No, Steady, No, Steady, No, No, No, St...
## $ glyburide                <fct> No, No, No, No, No, No, No, Steady, No, No...
## $ tolbutamide              <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ pioglitazone             <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ rosiglitazone            <fct> No, No, No, No, No, No, No, No, No, Steady...
## $ acarbose                 <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ miglitol                 <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ troglitazone             <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ tolazamide               <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ examide                  <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ citoglipton              <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ insulin                  <fct> No, Up, No, Up, Steady, Steady, Steady, No...
## $ glyburide.metformin      <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ glipizide.metformin      <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ glimepiride.pioglitazone <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ metformin.rosiglitazone  <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ metformin.pioglitazone   <fct> No, No, No, No, No, No, No, No, No, No, No...
## $ change                   <fct> No, Ch, No, Ch, Ch, No, Ch, No, Ch, Ch, No...
## $ diabetesMed              <fct> No, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes...
## $ readmitted               <fct> NO, >30, NO, NO, NO, >30, NO, >30, NO, NO,...

##   encounter_id        patient_nbr                      race
##  Min.   :    12522   Min.   :      135   ?              : 2273
##  1st Qu.: 84961194   1st Qu.: 23413221   AfricanAmerican:19210
##  Median :152388987   Median : 45505143   Asian          :  641
##  Mean   :165201646   Mean   : 54330401   Caucasian      :76099
##  3rd Qu.:230270888   3rd Qu.: 87545950   Hispanic       : 2037
##  Max.   :443867222   Max.   :189502619   Other          : 1506
##
##             gender            age                weight       admission_type_id
##  Female         :54708   [70-80):26068   ?        :98569   Min.   :1.000
##  Male           :47055   [60-70):22483   [75-100) : 1336   1st Qu.:1.000
##  Unknown/Invalid:    3   [50-60):17256   [50-75)  :  897   Median :1.000
##                          [80-90):17197   [100-125):  625   Mean   :2.024
##                          [40-50): 9685   [125-150):  145   3rd Qu.:3.000
##                          [30-40): 3775   [25-50)  :   97   Max.   :8.000
##                          (Other): 5302   (Other)  :   97
```

```
##  discharge_disposition_id admission_source_id time_in_hospital   payer_code
##  Min.   : 1.000           Min.   : 1.000      Min.   : 1.000   ?      :40256
##  1st Qu.: 1.000           1st Qu.: 1.000      1st Qu.: 2.000   MC     :32439
##  Median : 1.000           Median : 7.000      Median : 4.000   HM     : 6274
##  Mean   : 3.716           Mean   : 5.754      Mean   : 4.396   SP     : 5007
##  3rd Qu.: 4.000           3rd Qu.: 7.000      3rd Qu.: 6.000   BC     : 4655
##  Max.   :28.000           Max.   :25.000      Max.   :14.000   MD     : 3532
##                                                                (Other): 9603
##              medical_specialty num_lab_procedures num_procedures
##  ?                     :49949   Min.   :  1.0      Min.   :0.00
##  InternalMedicine      :14635   1st Qu.: 31.0      1st Qu.:0.00
##  Emergency/Trauma      : 7565   Median : 44.0      Median :1.00
##  Family/GeneralPractice: 7440   Mean   : 43.1      Mean   :1.34
##  Cardiology            : 5352   3rd Qu.: 57.0      3rd Qu.:2.00
##  Surgery-General       : 3099   Max.   :132.0      Max.   :6.00
##  (Other)               :13726
##  num_medications number_outpatient number_emergency  number_inpatient
##  Min.   : 1.00   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.:10.00   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median :15.00   Median : 0.0000   Median : 0.0000   Median : 0.0000
##  Mean   :16.02   Mean   : 0.3694   Mean   : 0.1978   Mean   : 0.6356
##  3rd Qu.:20.00   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 1.0000
##  Max.   :81.00   Max.   :42.0000   Max.   :76.0000   Max.   :21.0000
##
##      diag_1          diag_2          diag_3       number_diagnoses max_glu_serum
##  428    : 6862   276    : 6752   250    :11555   Min.   : 1.000   >200: 1485
##  414    : 6581   428    : 6662   401    : 8289   1st Qu.: 6.000   >300: 1264
##  786    : 4016   250    : 6071   276    : 5175   Median : 8.000   None:96420
##  410    : 3614   427    : 5036   428    : 4577   Mean   : 7.423   Norm: 2597
##  486    : 3508   401    : 3736   427    : 3955   3rd Qu.: 9.000
##  427    : 2766   496    : 3305   414    : 3664   Max.   :16.000
##  (Other):74419   (Other):70204   (Other):64551
##  A1Cresult     metformin      repaglinide     nateglinide     chlorpropamide
##  >7  : 3812   Down  :  575   Down  :     45   Down  :     11   Down  :      1
##  >8  : 8216   No    :81778   No    :100227   No    :101063   No    :101680
##  None:84748   Steady:18346   Steady:   1384   Steady:    668   Steady:     79
##  Norm: 4990   Up    : 1067   Up    :    110   Up    :     24   Up    :      6
##
##
##
##  glimepiride    acetohexamide    glipizide      glyburide      tolbutamide
##  Down  :  194   No    :101765   Down  :  560   Down  :  564   No    :101743
##  No    :96575   Steady:      1   No    :89080   No    :91116   Steady:     23
```

```
##  Steady: 4670                    Steady:11356   Steady: 9274
##  Up    :  327                    Up    :  770   Up    :  812
##
##
##
##  pioglitazone   rosiglitazone    acarbose        miglitol      troglitazone
##  Down  :  118   Down  :   87   Down  :      3   Down  :      5   No    :101763
##  No    :94438   No    :95401   No    :101458   No    :101728   Steady:      3
##  Steady: 6976   Steady: 6100   Steady:    295   Steady:     31
##  Up    :  234   Up    :  178   Up    :     10   Up    :      2
##
##
##
##   tolazamide     examide     citoglipton   insulin      glyburide.metformin
##  No    :101727   No:101766   No:101766   Down  :12218   Down  :      6
##  Steady:     38                          No    :47383   No    :101060
##  Up    :      1                          Steady:30849   Steady:    692
##                                          Up    :11316   Up    :      8
##
##
##
##  glipizide.metformin glimepiride.pioglitazone metformin.rosiglitazone
##  No    :101753       No    :101765            No    :101764
##  Steady:     13       Steady:      1           Steady:      2
##
##
##
##
##
##  metformin.pioglitazone change     diabetesMed readmitted
##  No    :101765          Ch:47011   No :23403   <30:11357
##  Steady:      1         No:54755   Yes:78363   >30:35545
##                                                NO :54864
##
##
##
##
```

2.2 Missing values

As we see in the result determined by the glimpse() function, the columns race, weight, payer_code, gender, diag_3, and medical specialty have some or significant amounts of missing values. To get a better picture, we will zoom in on these attributes and determine the amount and the percentage of missing values.

	Missing.Values	Missing.Values.Percentage
race	2273.00	2.23
weight	98569.00	96.86
payer_code	40256.00	39.56
medical_specialty	49949.00	49.08
gender	3.00	0.00
diag_3	1423.00	1.40

As we notice in table 1, attribute weight is almost entirely made-up of missing values. For that reason, it is a candidate for removal. The same can be said for payer_code and medical_specialty, where 39.56% and 49.08% of values are missing, respectively. Payer_code has no significant value to our research goals and questions; therefore, it can be removed. Medical_specialty can be of value as it contains a range of useful information. Missing values can be set to 'Missing' or reevaluated based on other valuations; doing this is not without risk, as it is almost half of the attributes' values. Deleting all missing values is not doable as it removes half of all records! Setting it to 'Missing' seems to be the most logical option. Attributes race and diag_3 seem to have little amounts of missing data, but removing these values can alter our outcomes; therefore, we will use the same approach as to medical_specialty. Gender only has three missing values. Considering the amount does not account to even one percent, removal does not seem to harm, and gender is, in most cases, considered binary data.

In the next section, we will introduce a new attribute (HbA1c measurement) based on an A1C test and the response to that result, which is defined as a change in diabetic medication. This test was performed at the time of hospital admission. We consider four groups of encounters:

1. no HbA1c test performed;
2. HbA1c performed and in normal range;
3. HbA1c performed and the result is greater than 8
4. HbA1c performed, result is greater than 8

(Date: 18-20 September, 2020)

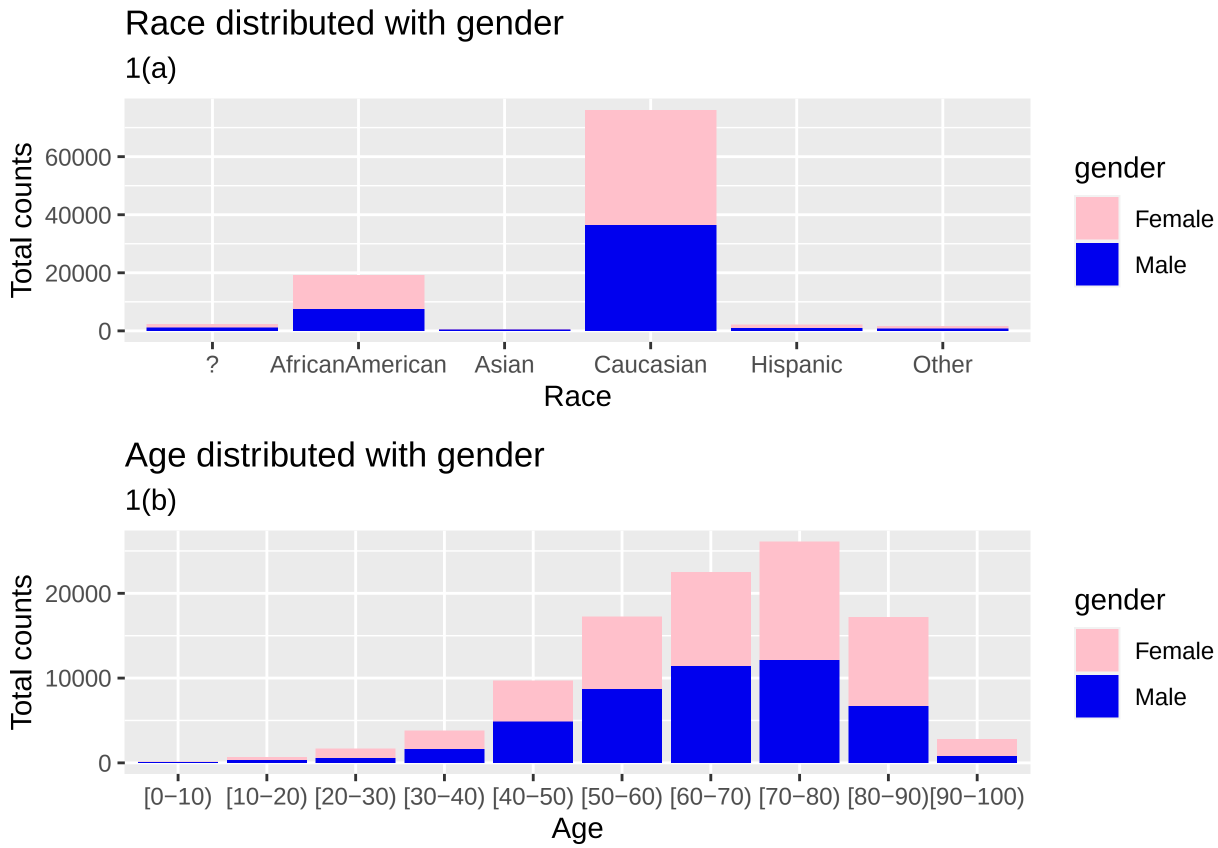
Before we start analyzing our dataset, it is crucial to check on duplicates as our dataset contains multiple inpatient visits for some patients. These observations cannot be statistically independent and would create noise. We thus declare that only one encounter per patient is optimal.

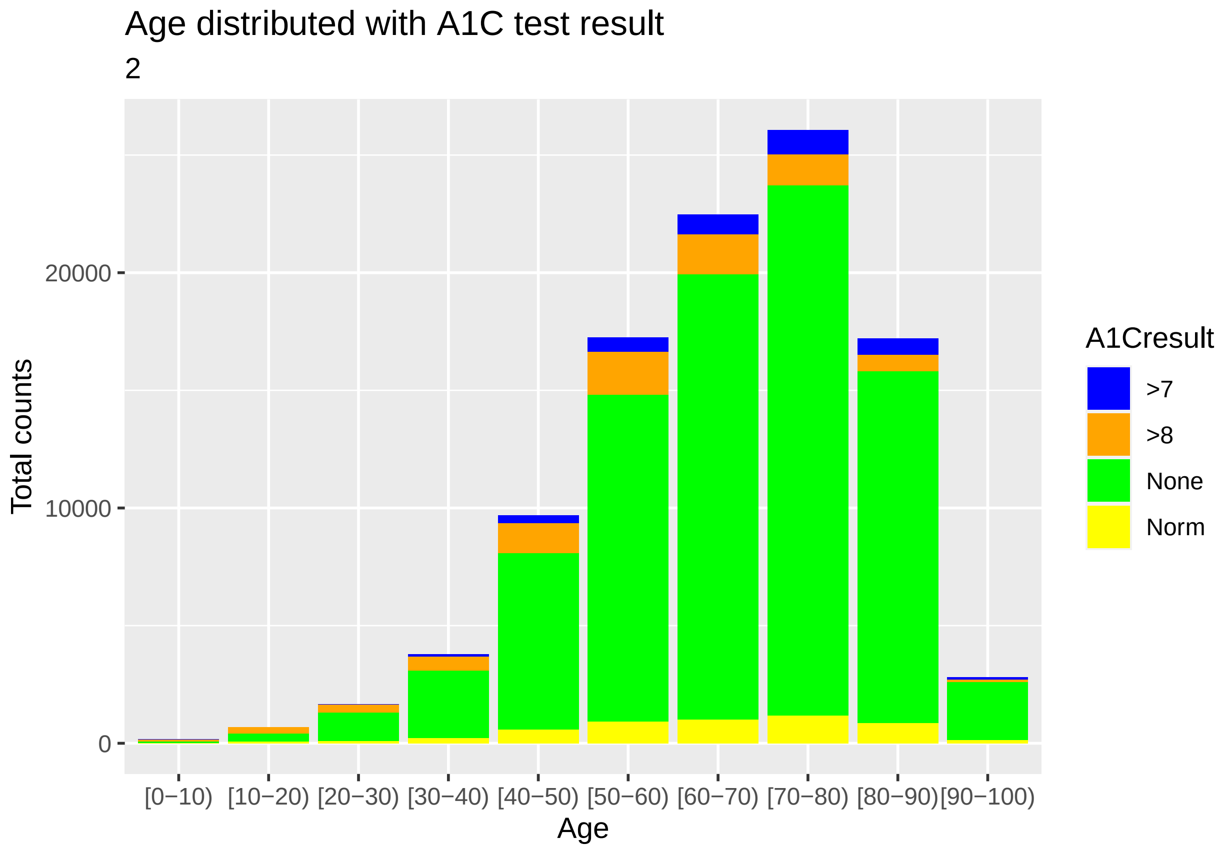
	Initial.number.of.records	Number.of.Duplicates	Difference.in.percent
1	101766	16773	-16.48

The initial dataset started with NA records, of which NA were duplicates. A potential removal of these records would leave us with 84993 records, a total loss of NA%. A small price to pay for a statistically independent dataset.

2.3 Categorial attributes

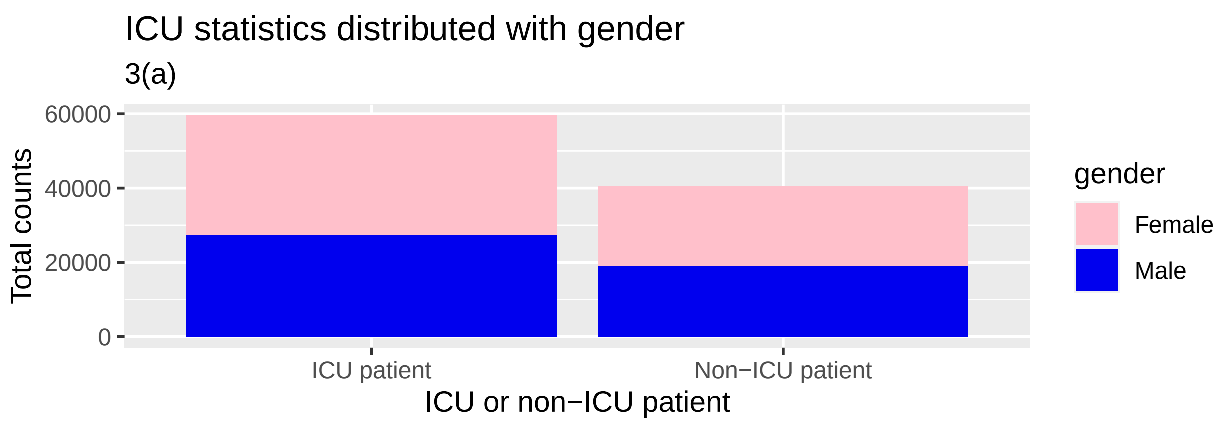
In this section, we take a look at variations in and between attributes. The attribute gender consists of three values ('female,' 'male,' and 'missing/unknown'). Since observe and use this attribute, it is essential to remove the abundant value.



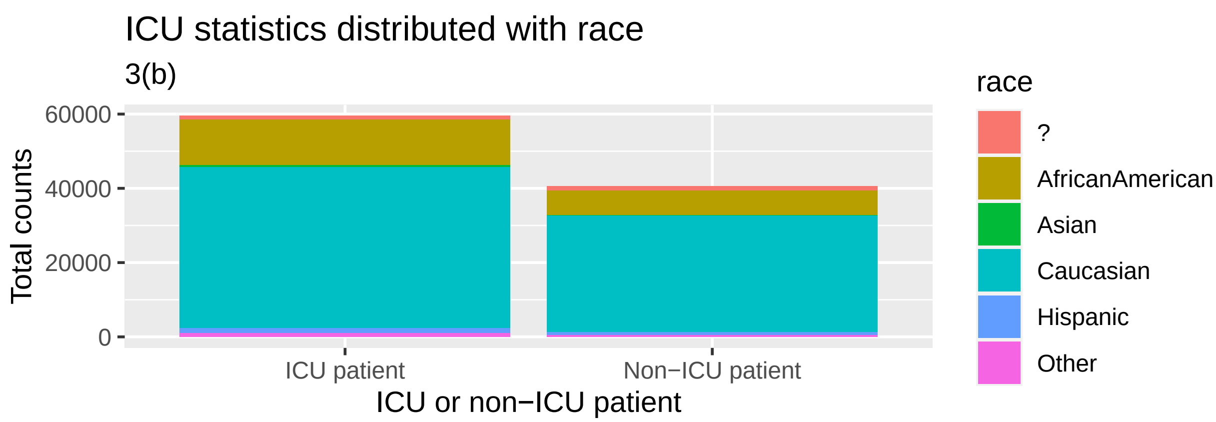


Looking at figure 1a, we notice that most patients are from the Caucasian race with almost equivalent male and female ratios. We expect a rise in the number of diabetes patients when looking at older population groups. Figure 1b shows this exact prediction; the data is negatively skewed and increases in numbers per older age group. We observe the same results in figure 2, where an increase in A1C test is shown when comparing older population groups. Additionally, the figure gives an essential insight for many patients - in most cases, the test was not conducted and shows that strategies surrounding testing diabetes are not normalized in hospital protocols.

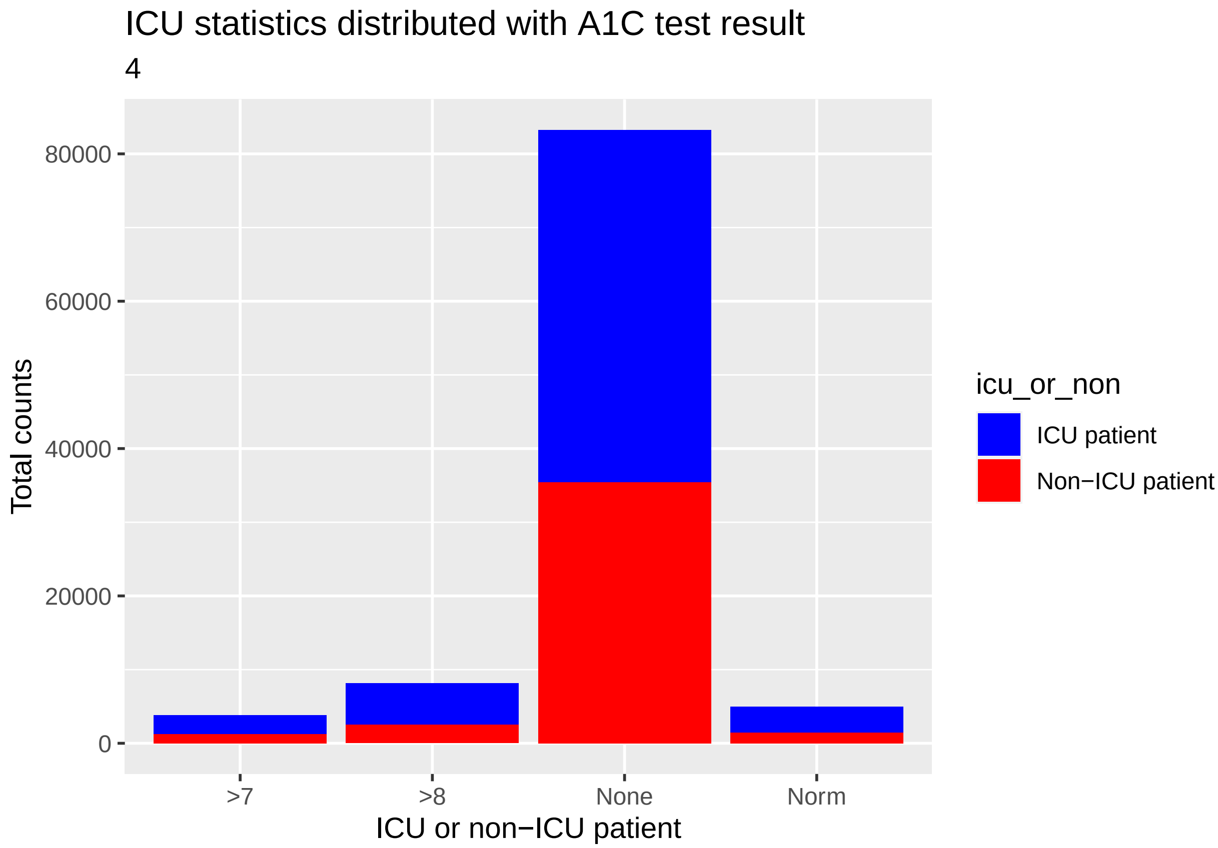
This graph does not, however, make a distinction between non-ICU and ICU patients. The authors of the original dataset analyses stated that ICU departments' protocols have a stricter policy surrounding testing for diabetes. When comparing non-ICU and ICU patient records, we need to construct a new attribute called 'icu_or_non,' comprising data from attributes admission_type_id, admission_source_id, and discharge_disposition_id. We distinguish between non-ICU and ICU patients.



ICU statistics distributed with gender
3(a)
Total counts
60000
40000
20000
0
ICU patient
Non−ICU patient
ICU or non−ICU patient
gender
Female
Male

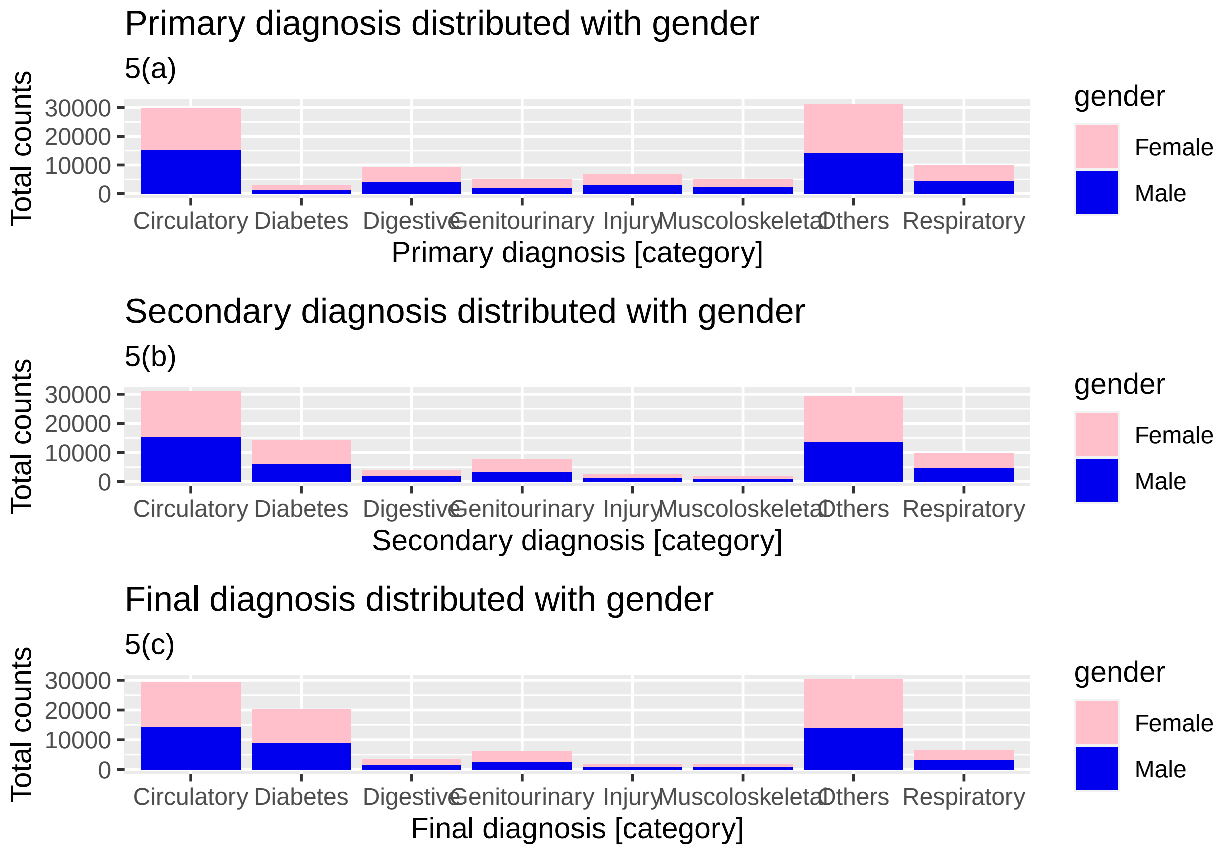


ICU statistics distributed with race
3(b)
Total counts
60000
40000
20000
0
ICU patient
Non−ICU patient
ICU or non−ICU patient
race
?
AfricanAmerican
Asian
Caucasian
Hispanic
Other



We observe the result from the distinction between ICU and non-ICU in figures 3a, 3b, and 4. We notice a smaller population of non-ICU patients in all mentioned figures, whereas the ratio of not only gender but also seen in 3b, where race is depicted, ratios stay consequently the same. Figure 4 shows that no matter the patient type (ICU or non-ICU), protocols surrounding diabetes testing is not firmly conducted. Keeping the distinction between non-ICU and ICU might not be necessary. However, it can be of complimentary use when starting with machine learning; the data is binary, and it allows the removal of three attributes.

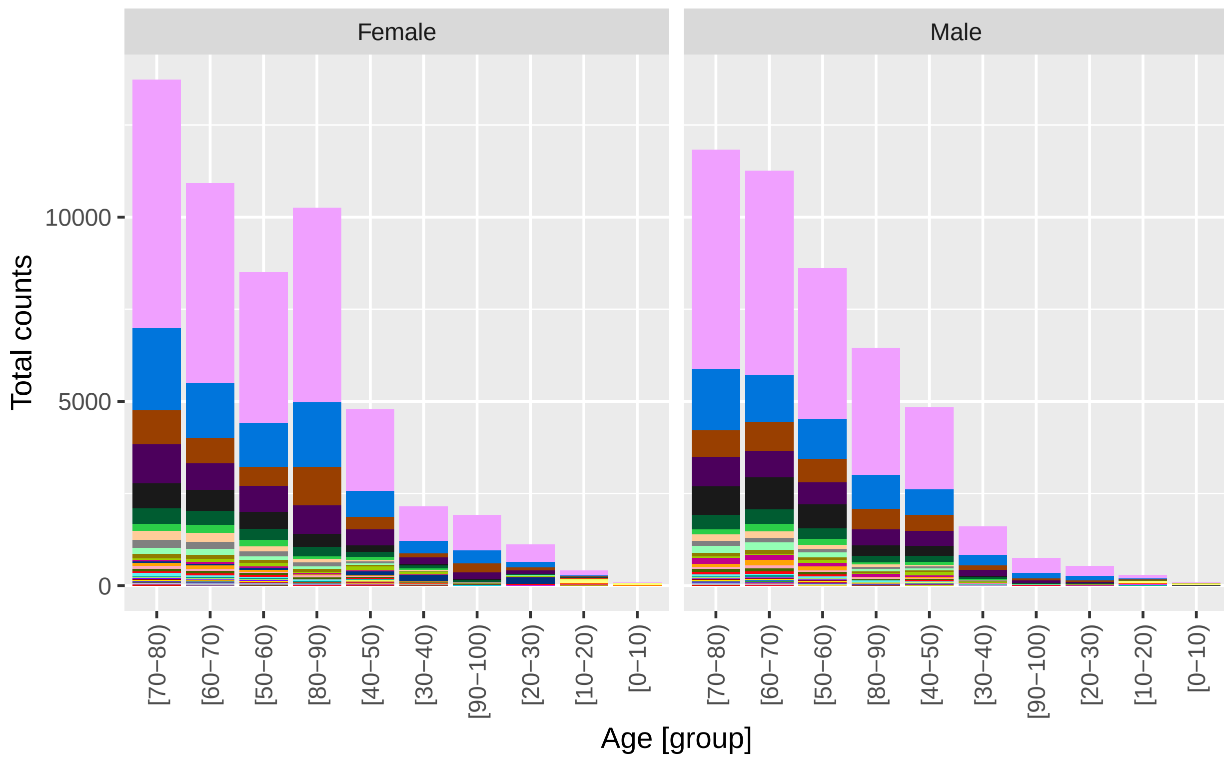
The attributes diag_1, diag_2, and diag_3 consist of many three-digit ICD codes. Many of these codes belong together in a subgroup. In the following section, we will construct a better way of describing the exact diagnosis. ICD codes descriptions are retrieved from [2].

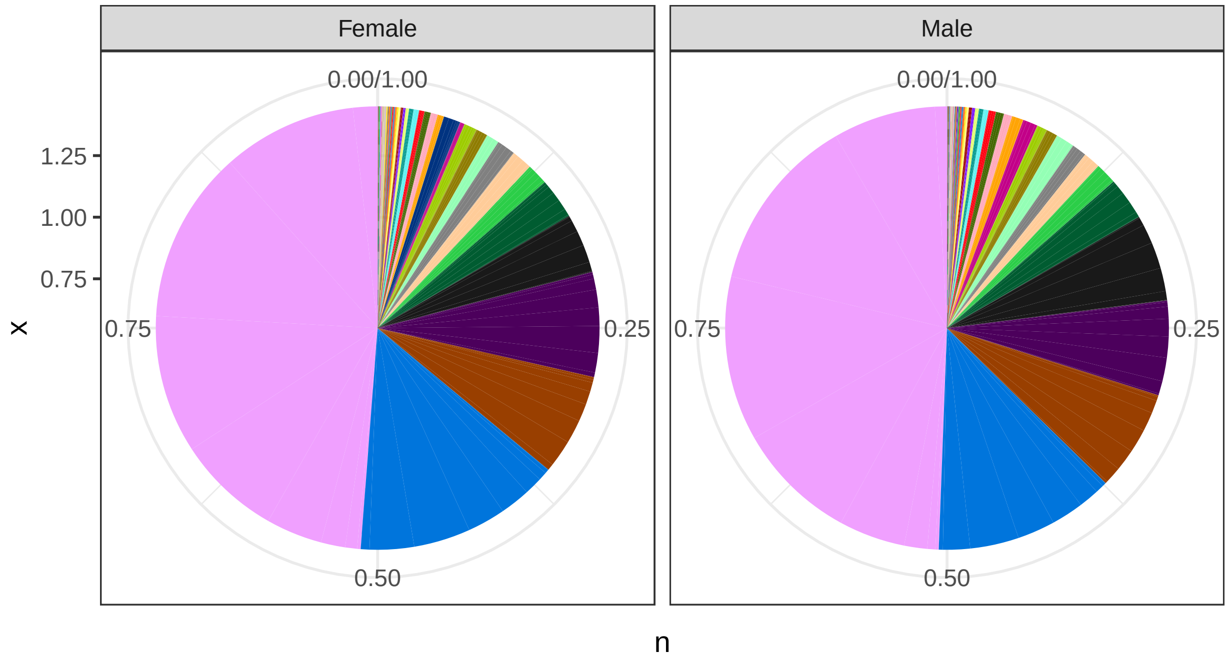


Looking at figure 5, we can see the results of the revaluation of attributes diag_1, diag_2, and diag_3. These attributes depict the primary, secondary, and final diagnosis of a patient, respectively. Interestingly, the difference between the figures is the increase in the number of diabetes diagnoses. Due to not testing diabetes on the initial hospitalization, readmission rates increase as a patient's primary diagnosis is not sustainable. The data classification now shows a clearer picture and would undoubtedly be of fair use in the final dataset. The original authors did not keep diag_2 and diag_3, as it would make records too complex to achieve their goals. Removal of these two attributes is still up for discussion, but the analysis does not give - for now - a clear indication for potential removal.

In the next section, we look at attribute medical_specialty and decide whether it is useful enough - considering the number of missing values - to be a candidate for the final dataset. We construct multiple plots that zoom in on the categories and valuations of this attribute.

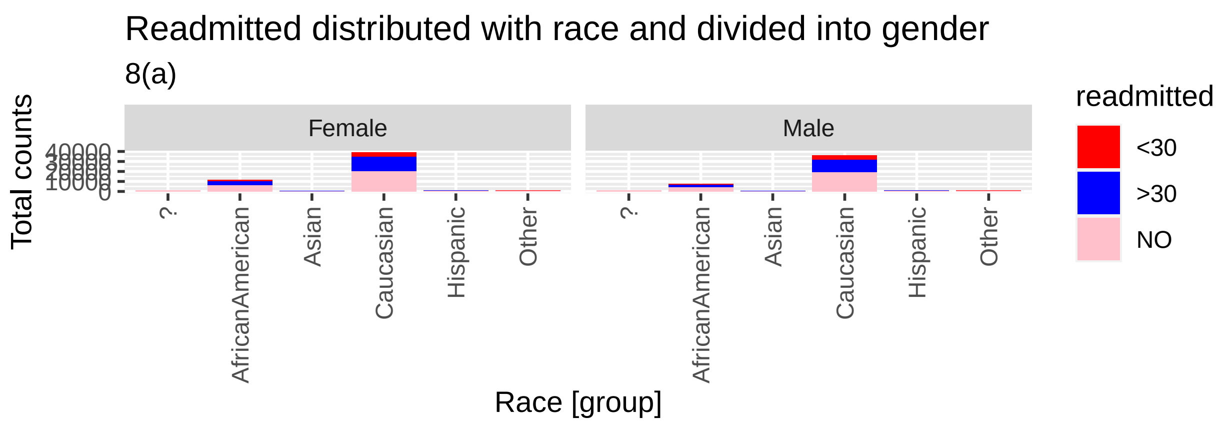
Medical specialty distributed with age groups and divided into gender
6

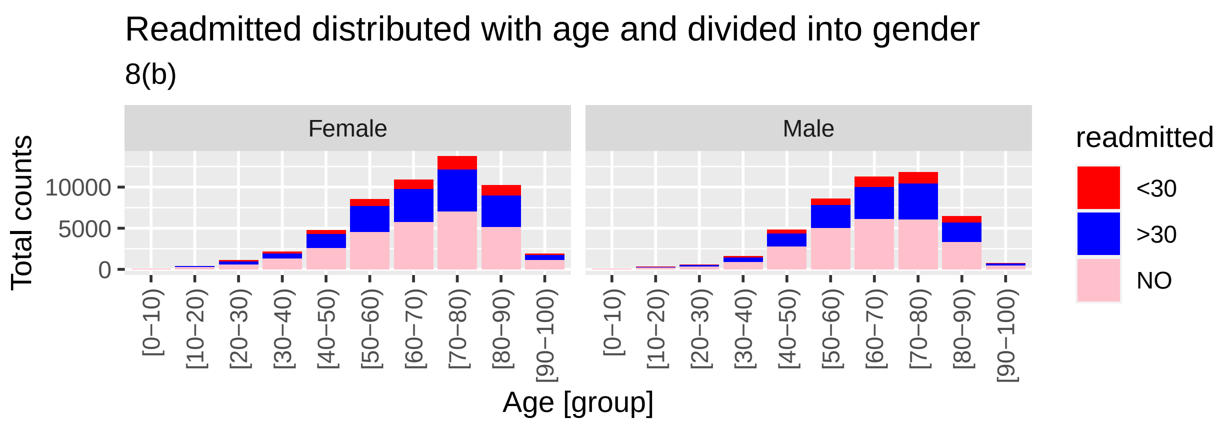


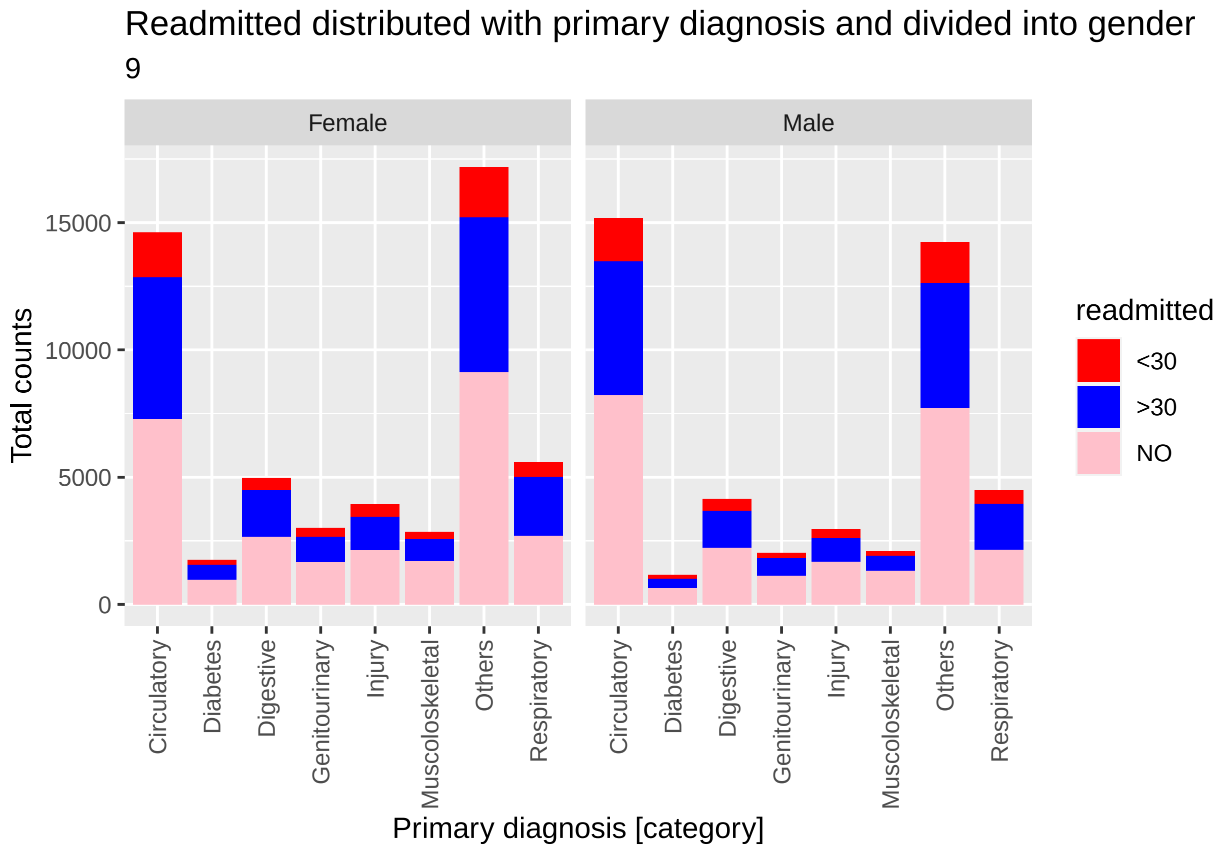


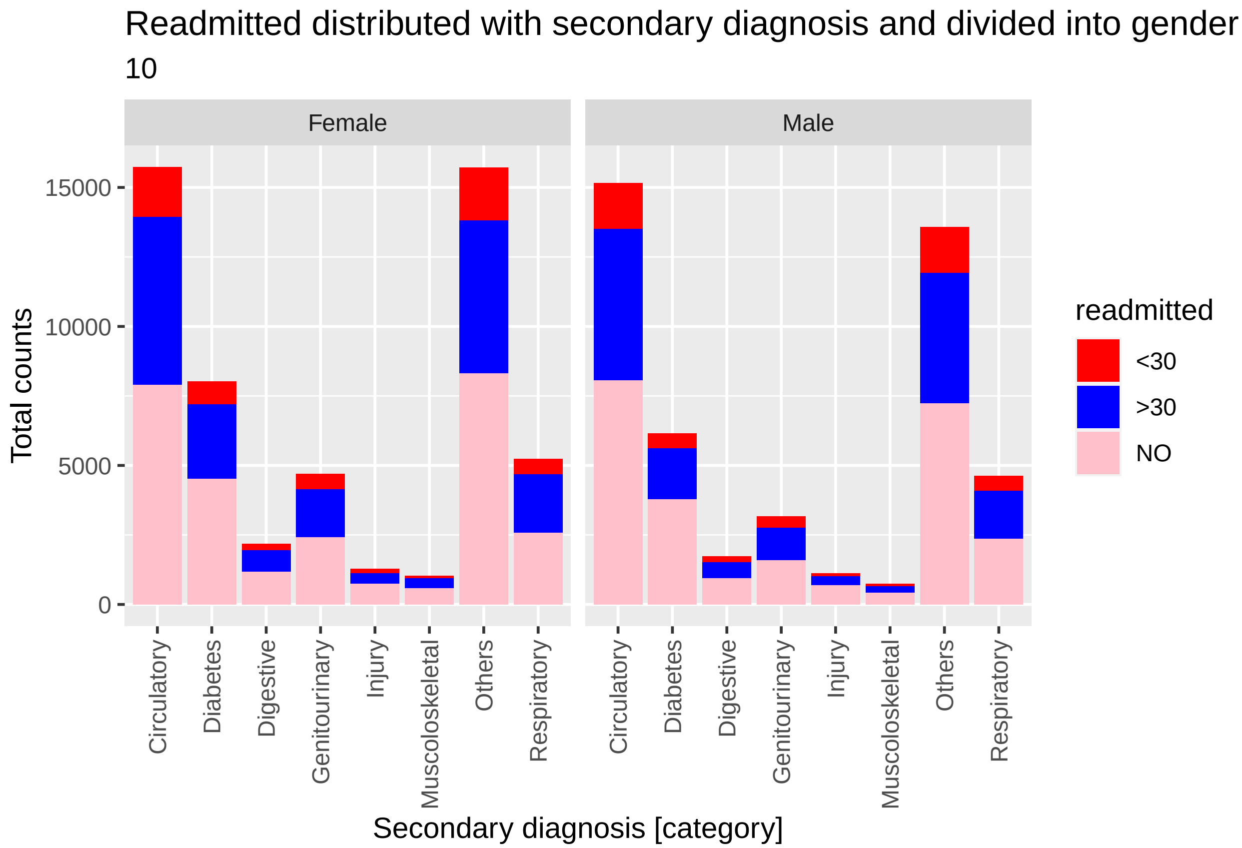
Looking at both figures 6 and 7, we observe a large amount of valuation between all demographic attributes used. Depicted in purple are the missing values, as already established, make-up almost 50 percent of total records. Medical specialty is different from, for example, admission_id, where we could shrink the attribute to a better format. This is an alternative to this attribute. We are considering this with the fact that medical specialty does not give better information than any diagnosis attribute, which also gives, maybe even more useful, guidance about an encounter's medical history. At this stage, removal of this attribute could not be of lethal harm.

For our final categorical attribute, and one of the most influential -according to the original authors- we will discuss readmitted. Readmitted is an attribute with three different valuations: '<30' for readmission within 30 days after release, '>30' for readmission after 30 days release, and 'NO' for no readmission.


Readmitted distributed with race and divided into gender
8(a)
Female
Male
Total counts
?
AfricanAmerican
Asian
Caucasian
Hispanic
Other
Race [group]
readmitted
<30
>30
NO


Readmitted distributed with age and divided into gender
8(b)
Female
Male
Total counts
10000
5000
0
[0−10)
[10−20)
[20−30)
[30−40)
[40−50)
[50−60)
[60−70)
[70−80)
[80−90)
[90−100)
Age [group]
readmitted
<30
>30
NO


Readmitted distributed with primary diagnosis and divided into gender
9
Female
Male
Total counts
15000
10000
5000
0
Circulatory
Diabetes
Digestive
Genitourinary
Injury
Muscoloskeletal
Others
Respiratory
Primary diagnosis [category]
readmitted
<30
>30
NO



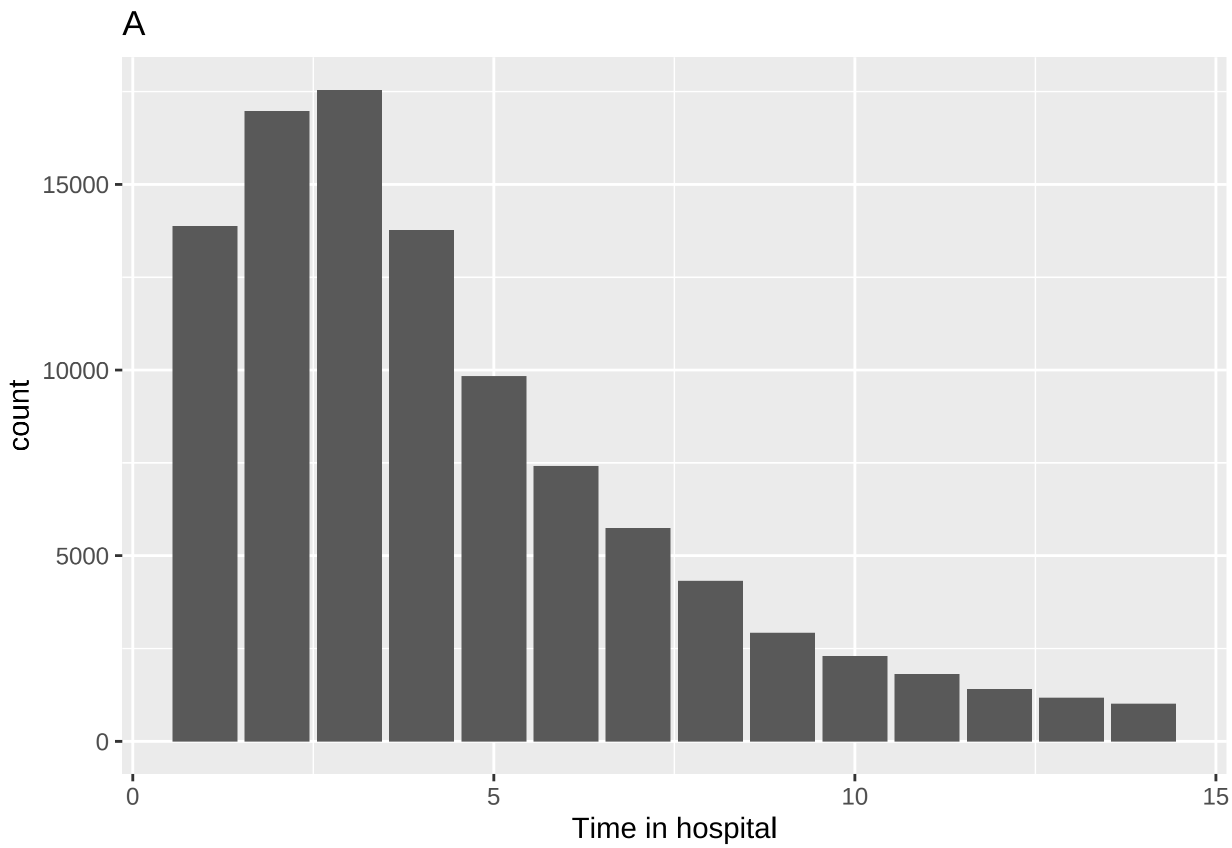
We observe the results of analyzing the readmitted attribute. We looked at readmitted with different demographics (gender, age, and race) and two of the three diagnosis attributes. Figure 8a shows that the Caucasian race has the most readmission rates in both genders. Observations made in figure 8b depict that as age increases, readmission rates also increase for females and males. This makes sense as immunity decreases with age, and the chance of recovery diminishes. Readmission rates decrease after the age of 80. The reasoning behind this might be death or transferring to a hospice or other facility.

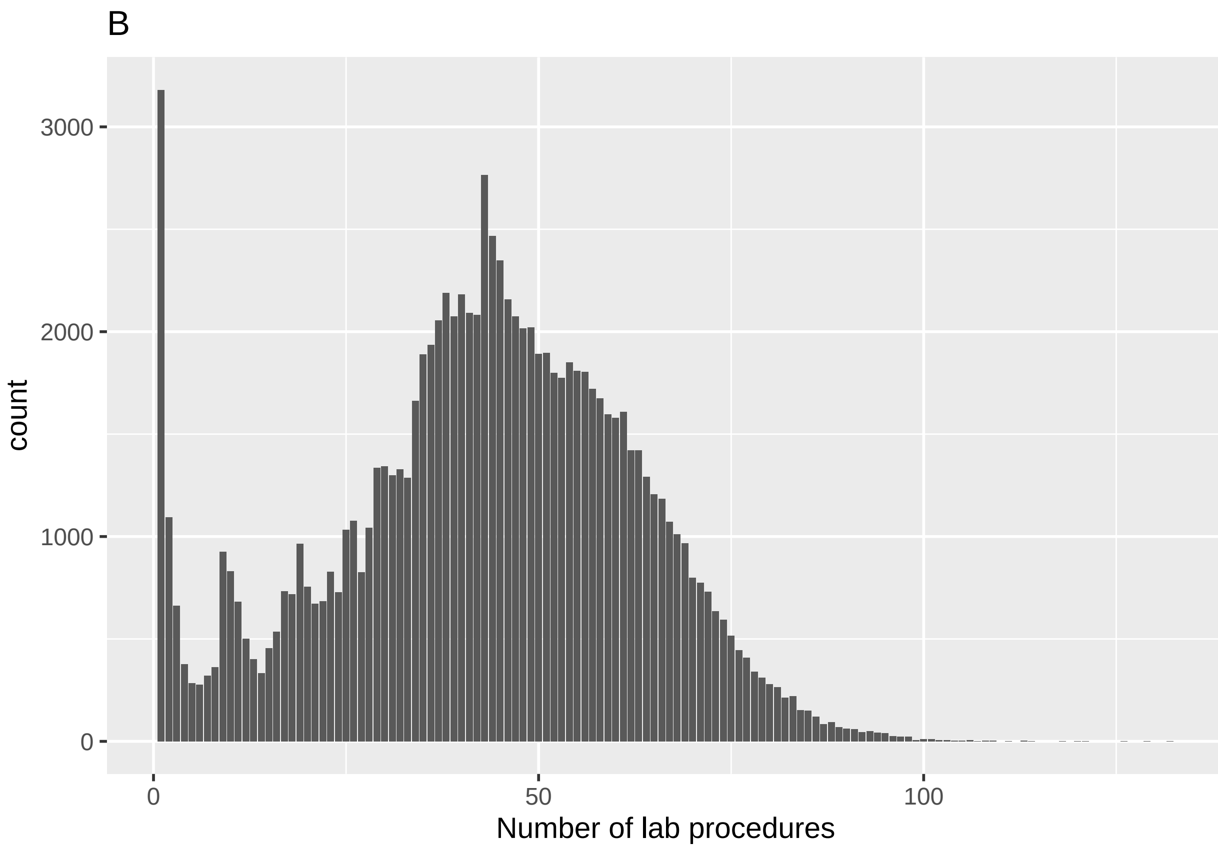
Comparing the results in figures 9 and 10, we see a familiar trend: the increase of diabetes diagnoses between the primary and secondary diagnoses. This typically means that patients admitted were diagnosed differently, and over time, are getting a new diabetes diagnosis much quicker.

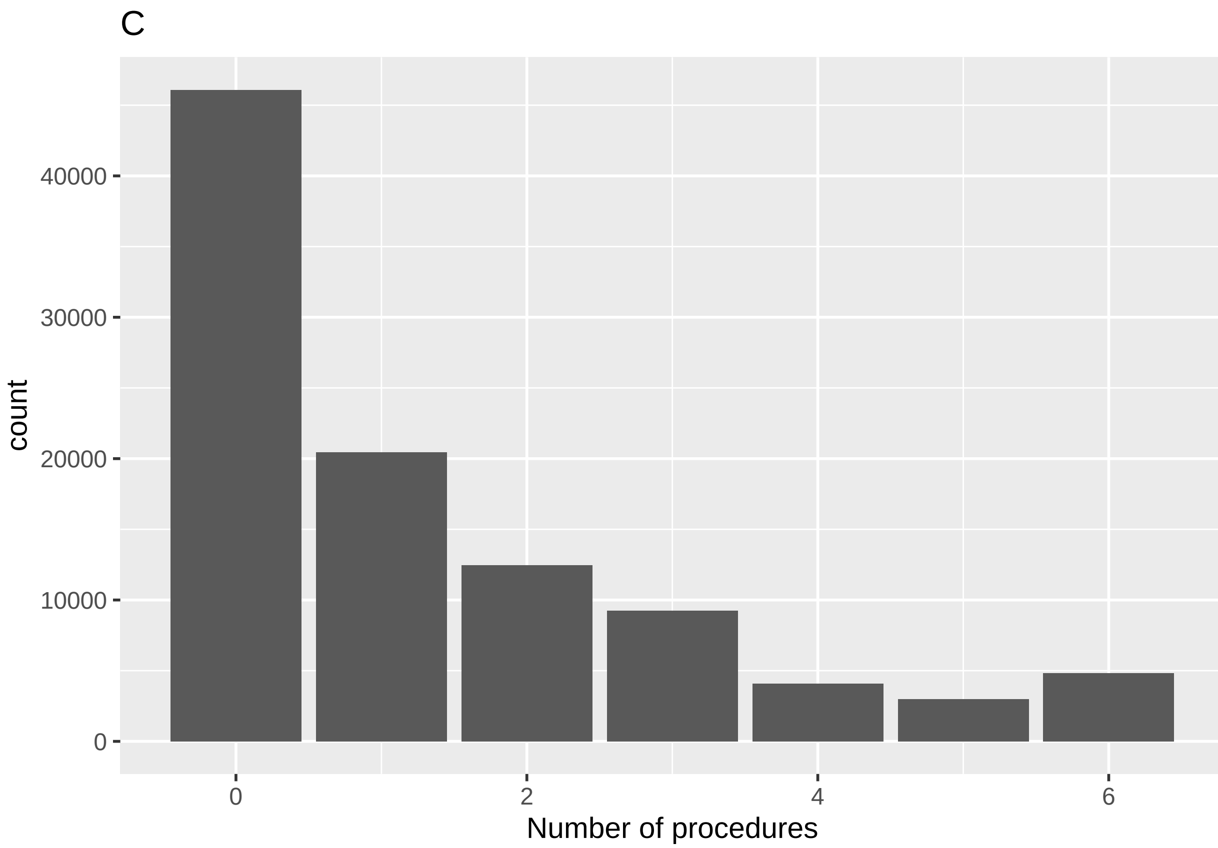
Seeing as we have multiple diagnosis categories but are somewhat interested in diabetes only, we can alter these attributes to have only two valuations ('diabetes' and 'other'). However, doing this can affect later analysis as we remove potentially valuable information.

2.4 Distribution - numeric attributes

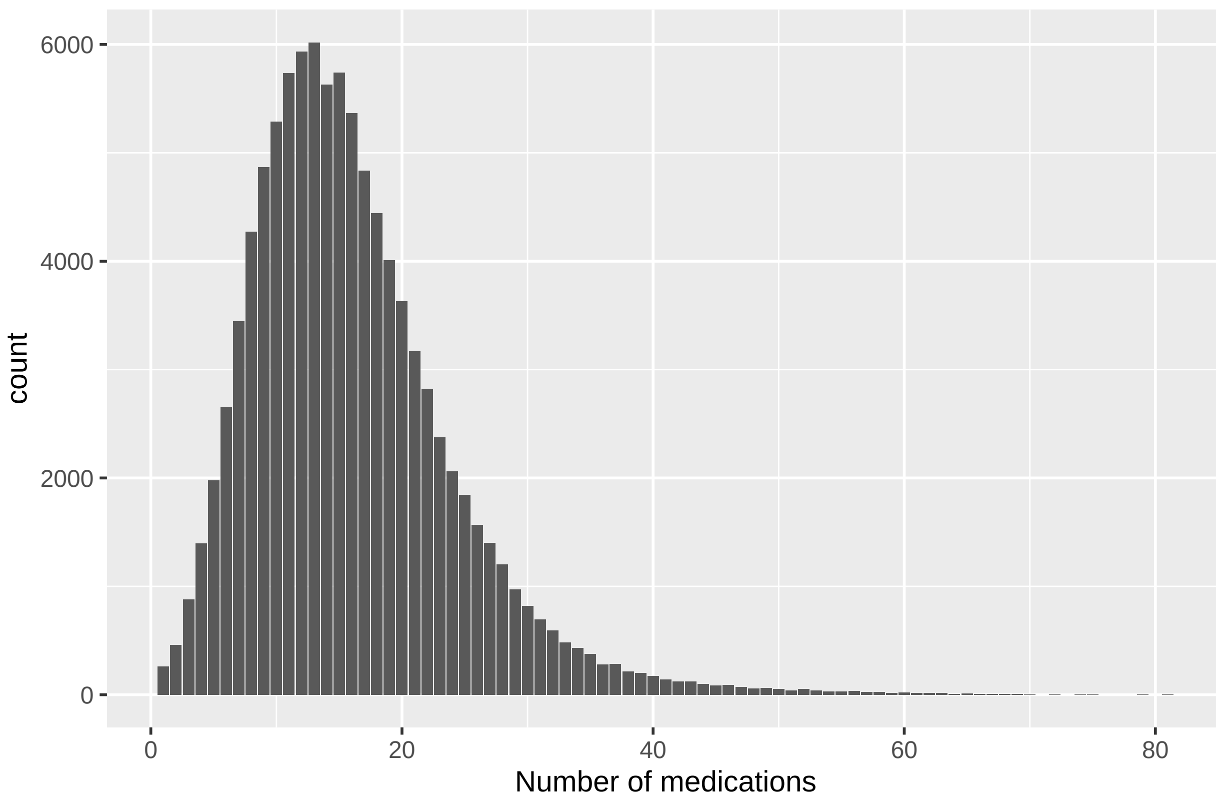
We discussed many categorical data attributes, now turn to our few numeric data. It is important to have a good distribution between numeric data. If the range of distribution is too large, then normalization is necessary. We construct multiple histograms, which we get through the histogram.plotter() function.

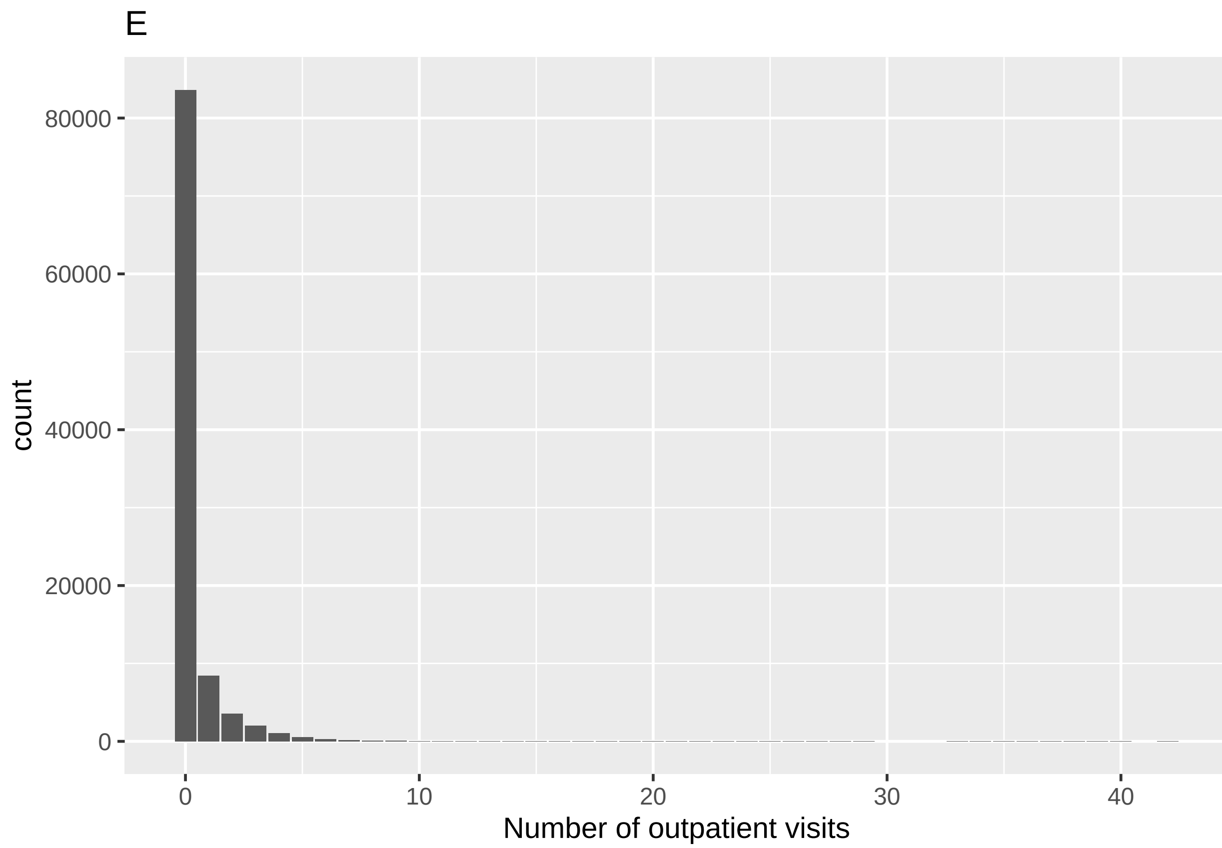


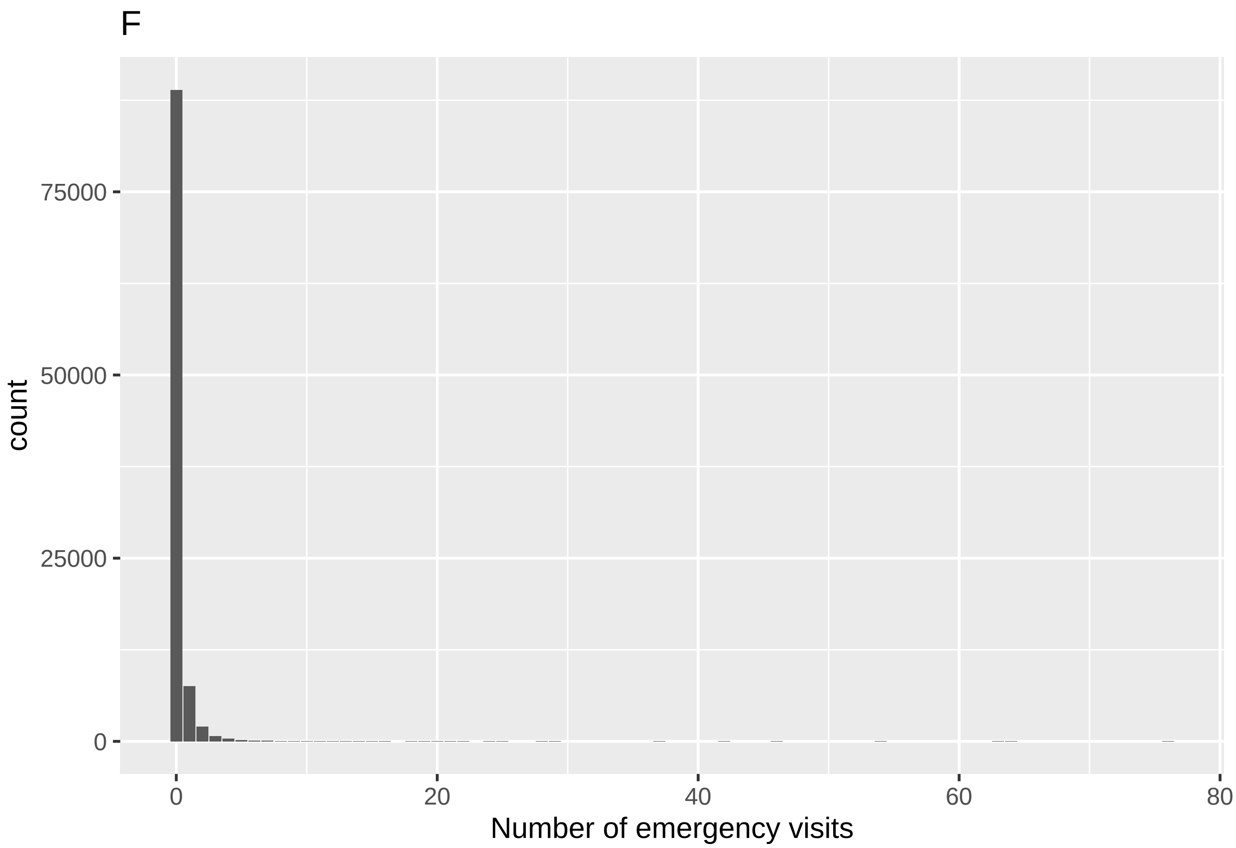

B
3000
2000
1000
0
count
0
50
100
Number of lab procedures


C
count
40000
30000
20000
10000
0
0
2
4
6
Number of procedures

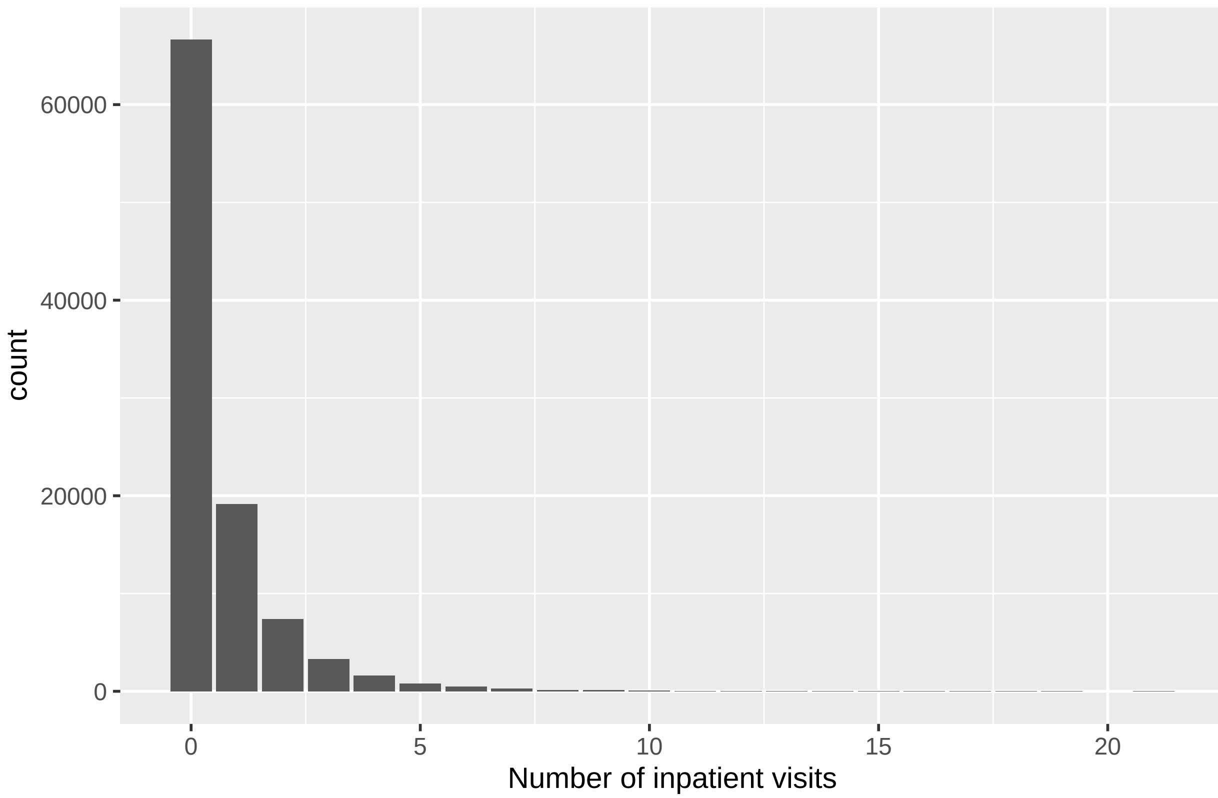
D



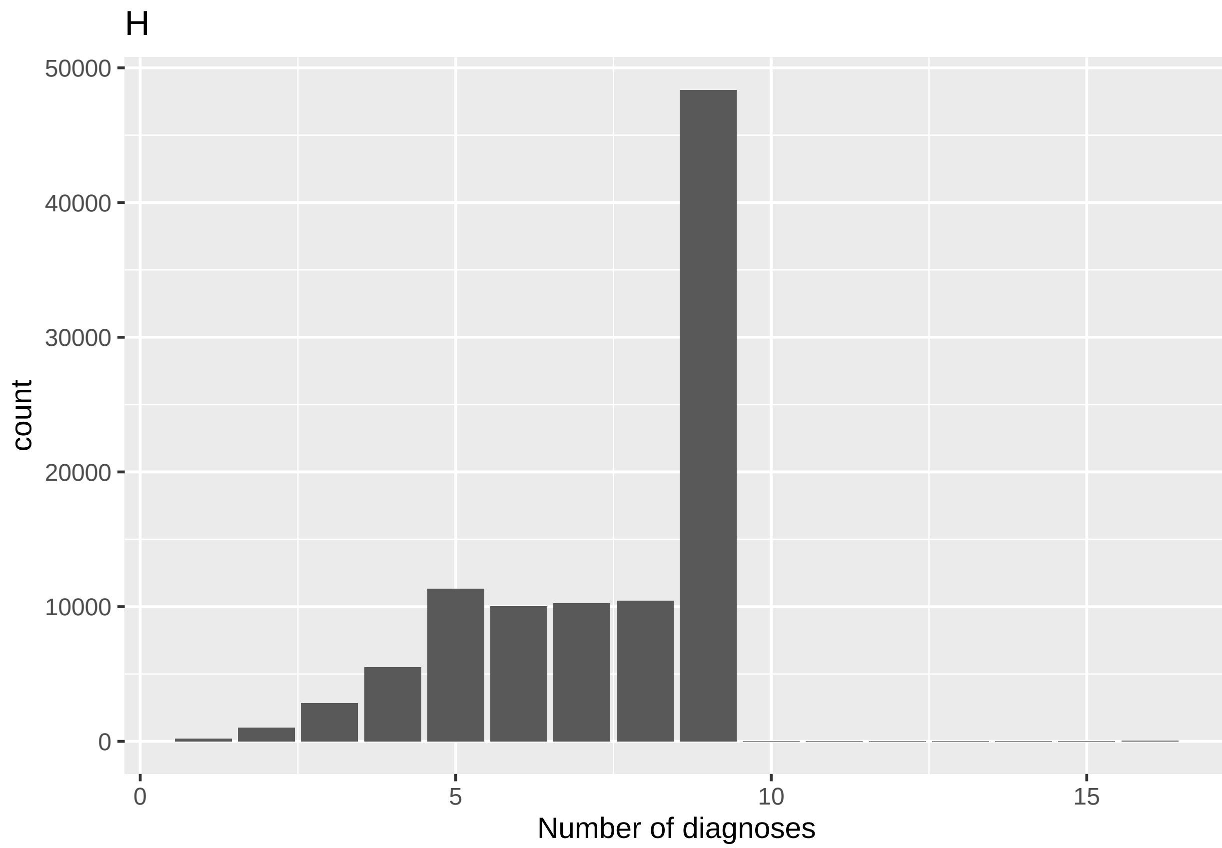

E
count
80000
60000
40000
20000
0
0
10
20
30
40
Number of outpatient visits

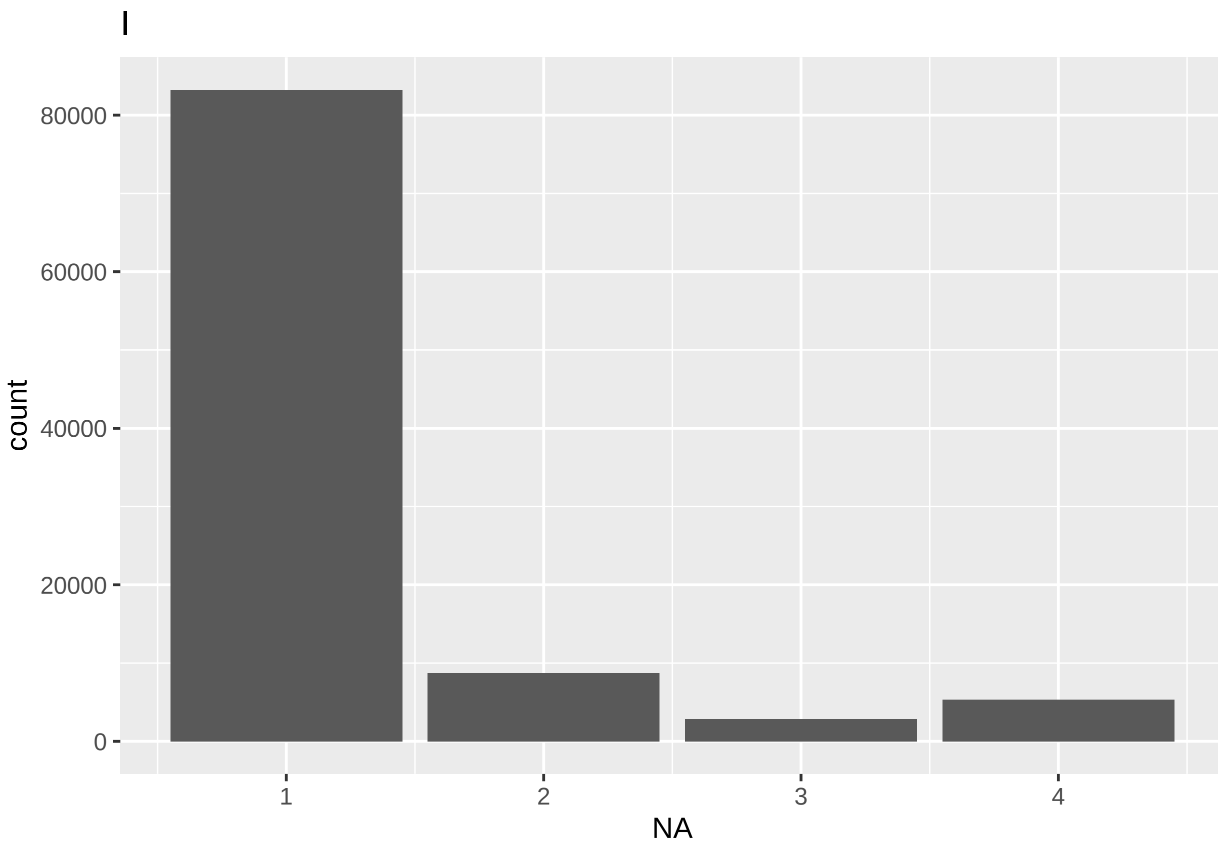

F
count
75000
50000
25000
0
0
20
40
60
80
Number of emergency visits

G

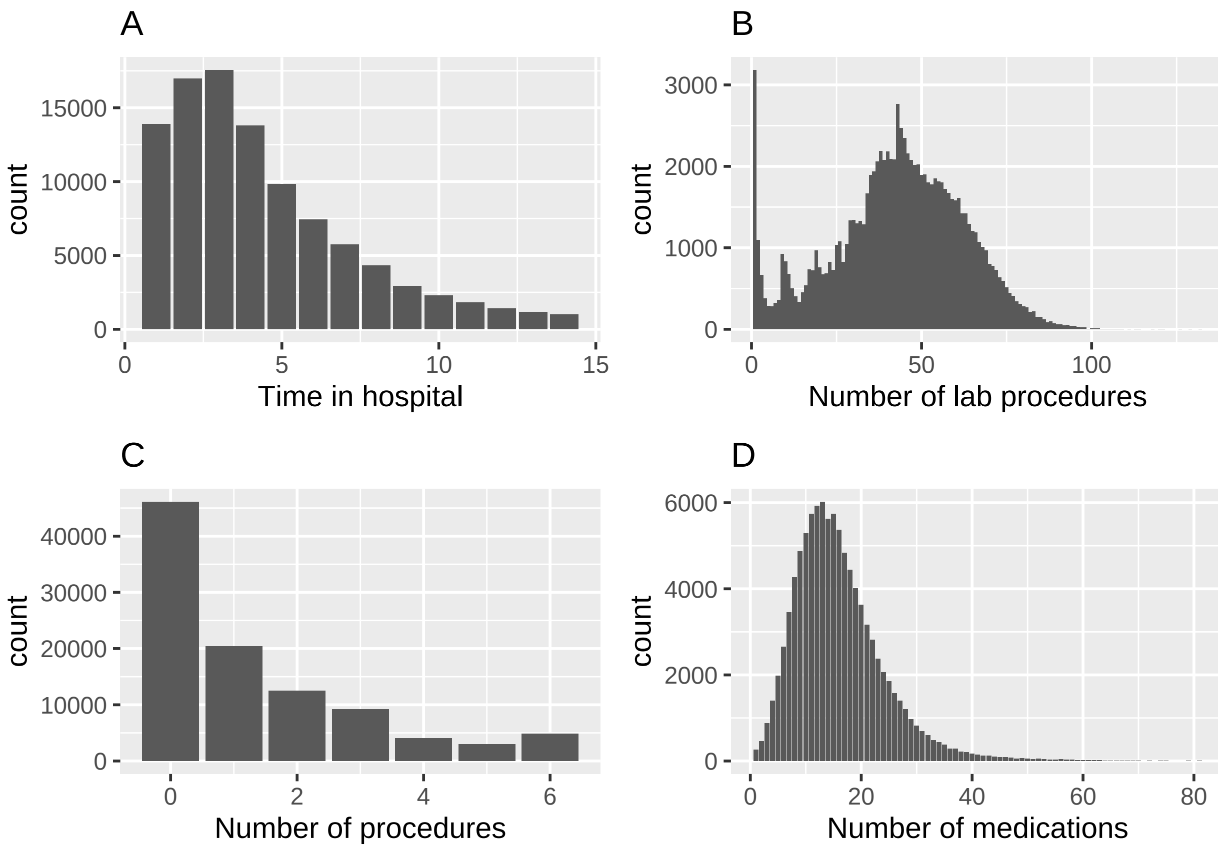


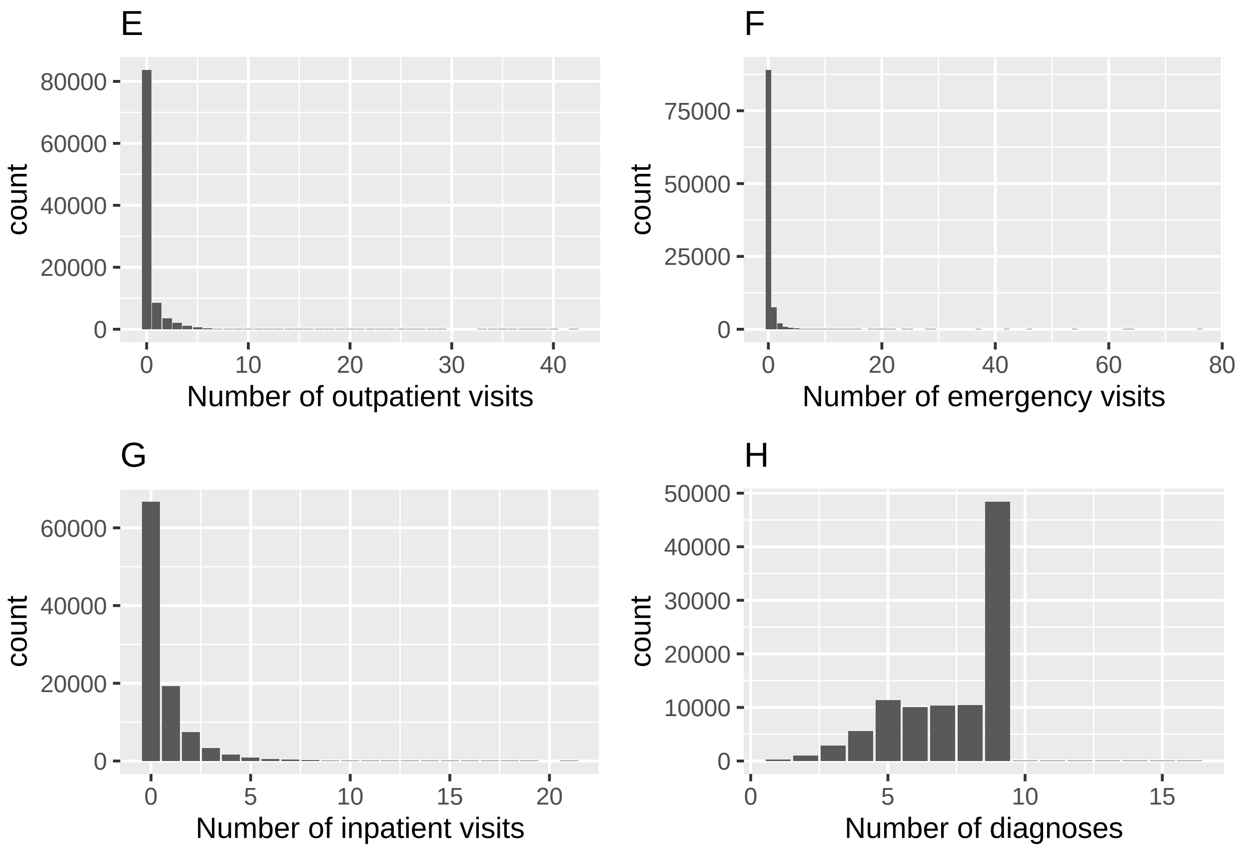
60000
40000
20000
0
count
0
5
10
15
20
Number of inpatient visits


H
count
50000
40000
30000
20000
10000
0
0
5
10
15
Number of diagnoses



I
count
80000
60000
40000
20000
0
1
2
3
4
NA

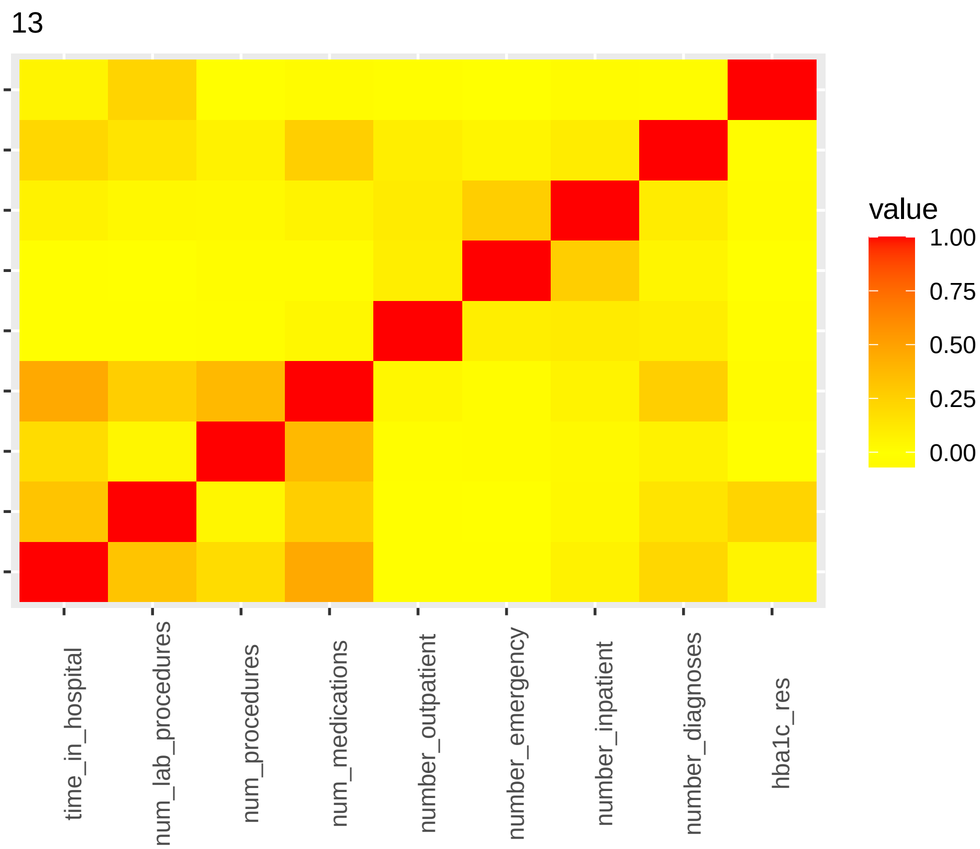

A
count
15000
10000
5000
0
0
5
10
15
Time in hospital
B
count
3000
2000
1000
0
0
50
100
Number of lab procedures
C
count
40000
30000
20000
10000
0
0
2
4
6
Number of procedures
D
count
6000
4000
2000
0
0
20
40
60
80
Number of medications

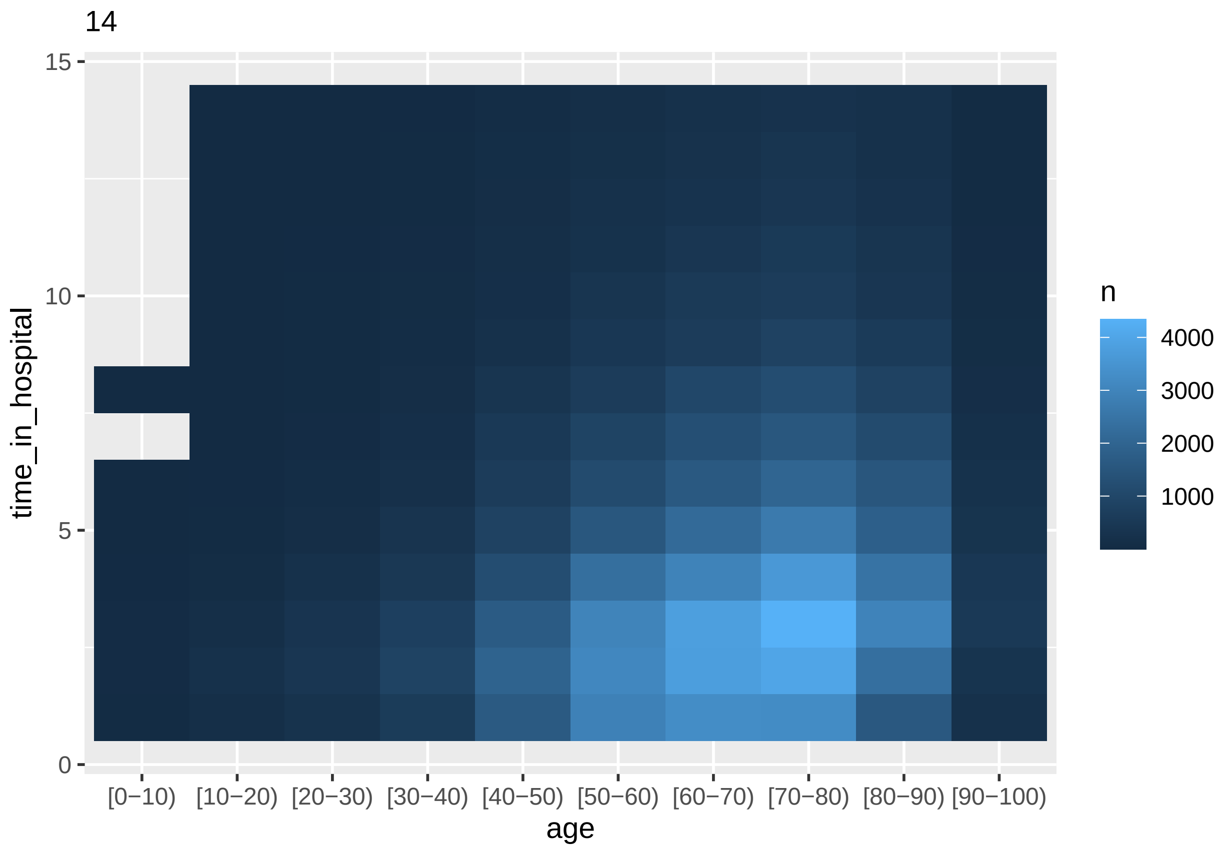


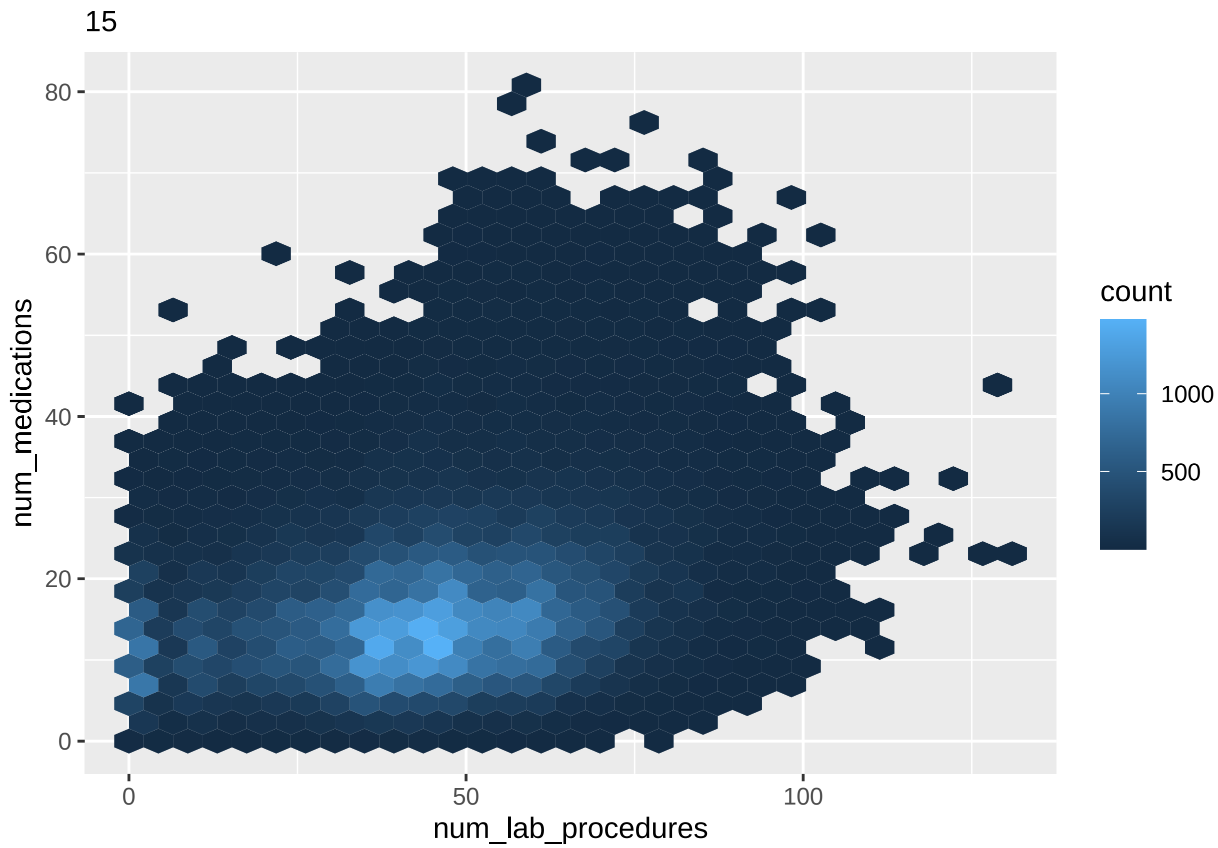
Looking at the results, depicted in figures 11 and 12, shows that not all of the attributes are evenly distributed. Normalization might be necessary to correct the distributions of namely B, which makes a couple of weird spikes, E, F, and G, which are all -but B- number of visits of an encounter. We might want to introduce a new attribute, total_visits, which accounts for all visits made by one encounter. The method of normalization is still up for debate.

2.5 Correlation - numeric attributes

To discover any correlation between the numeric data in our dataset, we constructed a heatmap, which is depicted below. We discuss a possible correlation between age and time spent in the hospital and the potential relationship between the num_lab_procedures and num_medications.




14
15
10
5
0
time_in_hospital
[0−10)
[10−20)
[20−30)
[30−40)
[40−50)
[50−60)
[60−70)
[70−80)
[80−90)
[90−100)
age
n
4000
3000
2000
1000



Looking at our heatmap (shown in figure 13), we determine a strong correlation as red, a small correlation as orange, and no correlation as yellow. The attributes that stand out are, for example, num_medications-time_in_hospital, num_procedures-num_medications, and hba1c_res-num_medications. As we can see, it is mostly the num_medication attribute that has some correlation. Others do not stand out that much. We can conclude that most of the attributes are non-correlated, at least for the numeric attributes.

Looking at figure 14, we observe that the age groups above [40-50) increases in total counts; consequently, the total time spent in the hospital before release also increases. We can, therefore, conclude that the time_in_hospital attribute is dependent on a patient's age. Finally, we zoom in on two attributes that had a (small) correlation depicted in figure 13. Figure 15 shows this the result of zooming in on num_medications and num_lab_procedures, the result shows that most encounters have a valuation between 15-20 medications and 40-60 amount of lab procedures. The result also shows a lot of outliers, which normalization could solve.

2.6 End of the EDA.

3 References

References

[1] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014. DOI: https://doi.org/10.1155/2014/781670

[2] Centers for Disease Control and Prevention, National Center for Health Statistics, ICD-9, https://www.cdc.gov/nchs/icd/icd9.htm, November 6, 2015.

4 Week 3: A Clean Dataset