Quantium Technical Test Responses

Rory Denham

22/02/2021

Contents

0.1 Background

The dataset provided is a fully de-identified set of hospital admissions data for episodes with diabetes related diagnoses. Each row represents a single encounter (inpatient episode) between a patient and one of 130 hospitals. For each encounter, multiple administratively collected variables are provided including the history of the patient's previous admissions, the diagnosis for the given admission, the results of a number of tests performed during admission, the medicine prescriptions for the patient during admission, whether they are readmitted to the hospital and some other features described in the tables below.

It won't be possible to provide the perfect answer to all of these questions. What we're looking to see is how you design your solution to the problem and how you prioritize the analysis you end up doing. The time we estimate this task to take is up to 6 hours. For each question we pose, provide a quick overview of your approach and key findings.

##`Q1.` Describe the dataset and note any points of interest or concern

0.1.1 Data Import

Table 1: Data summary

Name	diabetic_raw
Number of rows	101766
Number of columns	50
Column type frequency:	
character	5
factor	37
numeric	8
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
encounter_id	0	1.00	5	9	0	101766	0
patient_nbr	0	1.00	3	9	0	71518	0
diag_1	21	1.00	1	6	0	716	0
diag_2	358	1.00	1	6	0	748	0
diag_3	1423	0.99	1	6	0	789	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique
race	2273	0.98	FALSE	5
gender	3	1.00	FALSE	2
age	0	1.00	TRUE	10
weight	98569	0.03	TRUE	9
admission_type_id	0	1.00	FALSE	8
discharge_disposition_id	0	1.00	FALSE	26
admission_source_id	0	1.00	FALSE	17
payer_code	40256	0.60	FALSE	17
medical_specialty	49949	0.51	FALSE	72
max_glu_serum	0	1.00	FALSE	4
A1Cresult	0	1.00	FALSE	4
metformin	0	1.00	TRUE	4
repaglinide	0	1.00	TRUE	4
nateglinide	0	1.00	TRUE	4
chlorpropamide	0	1.00	TRUE	4
glimepiride	0	1.00	TRUE	4
acetohexamide	0	1.00	TRUE	2
glipizide	0	1.00	TRUE	4
glyburide	0	1.00	TRUE	4
tolbutamide	0	1.00	TRUE	2
pioglitazone	0	1.00	TRUE	4
rosiglitazone	0	1.00	TRUE	4
acarbose	0	1.00	TRUE	4
miglitol	0	1.00	TRUE	4
troglitazone	0	1.00	TRUE	2
tolazamide	0	1.00	TRUE	3
examide	0	1.00	TRUE	1
citoglipton	0	1.00	TRUE	1
insulin	0	1.00	TRUE	4
glyburide-metformin	0	1.00	TRUE	4
glipizide-metformin	0	1.00	TRUE	2
glimepiride-pioglitazone	0	1.00	TRUE	2
metformin-rosiglitazone	0	1.00	TRUE	2
metformin-pioglitazone	0	1.00	TRUE	2
change	0	1.00	FALSE	2
diabetesMed	0	1.00	FALSE	2
readmitted	0	1.00	FALSE	3

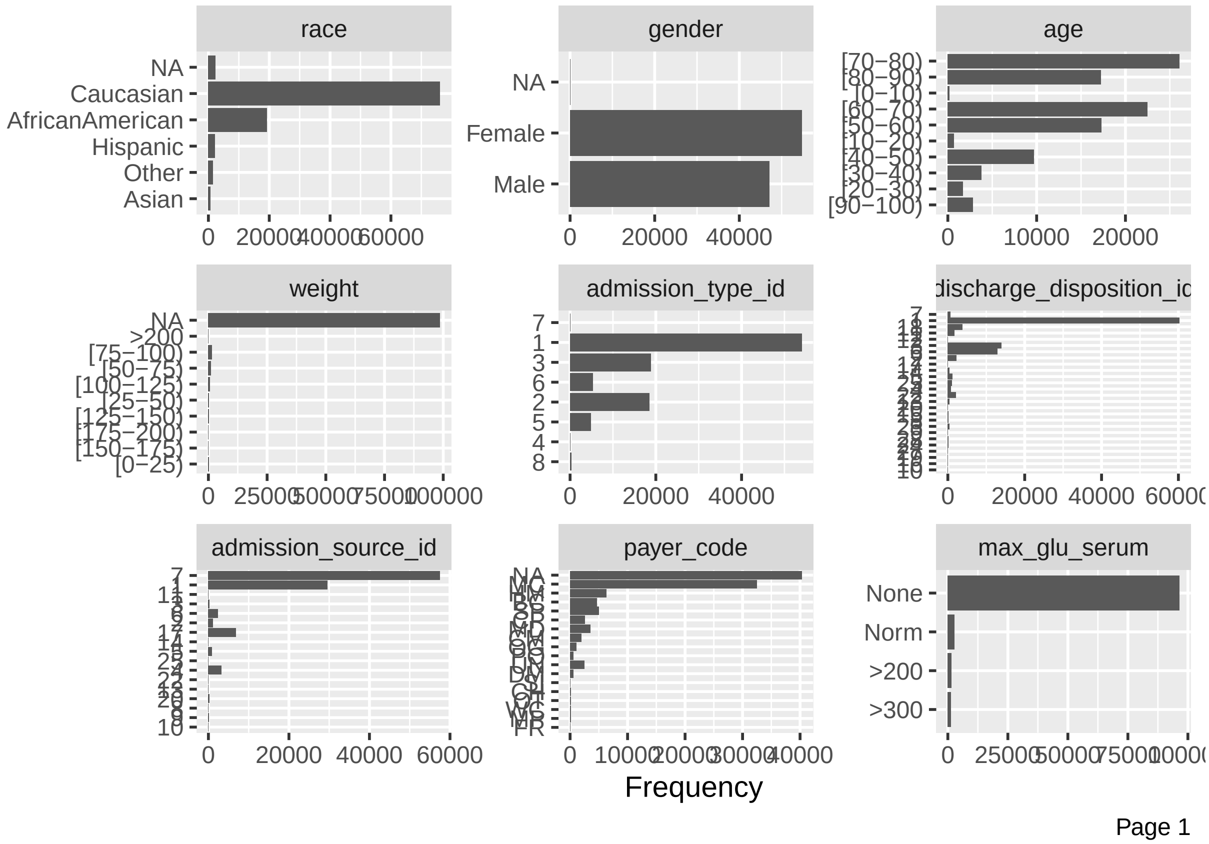
Variable type: numeric

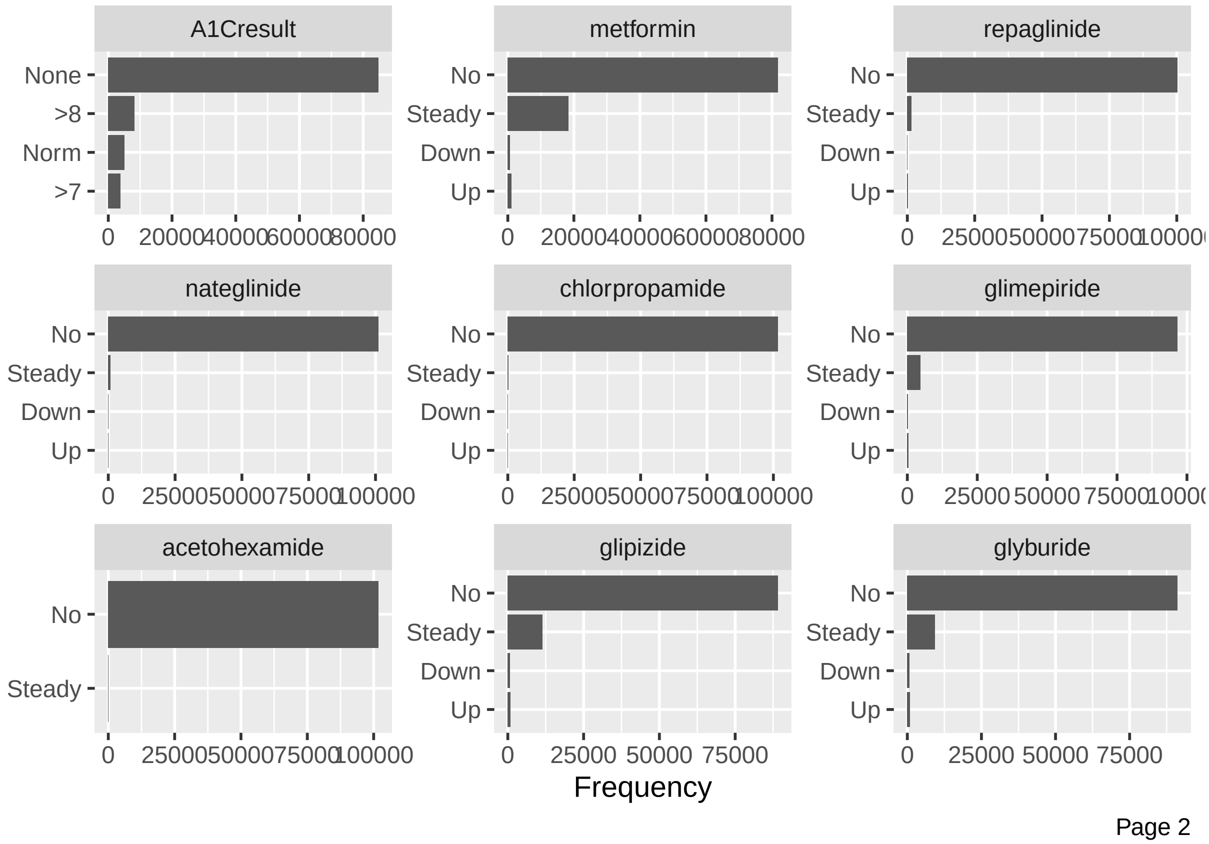
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
time_in_hospital	0	1	4.40	2.99	1	2	4	6	14
num_lab_procedures	0	1	43.10	19.67	1	31	44	57	132
num_procedures	0	1	1.34	1.71	0	0	1	2	6
num_medications	0	1	16.02	8.13	1	10	15	20	81
number_outpatient	0	1	0.37	1.27	0	0	0	0	42
number_emergency	0	1	0.20	0.93	0	0	0	0	76
number_inpatient	0	1	0.64	1.26	0	0	0	1	21
number_diagnoses	0	1	7.42	1.93	1	6	8	9	16

A large proportion of the numeric variables such as weight, age, and blood glucose concentration have been factored. This means we will only be able to obtain a broad estimate of the relevant statistics for patients within this cohort. For example, although we can determine differences between the age groups `[40-49]` and `[50-59]` respectively, however, we cannot investigate differences within groups such as `55 year olds` compared with `56 year olds`. Additionally, statistically significant differences may change if higher granularity in the variables becomes available for this cohort of patients.

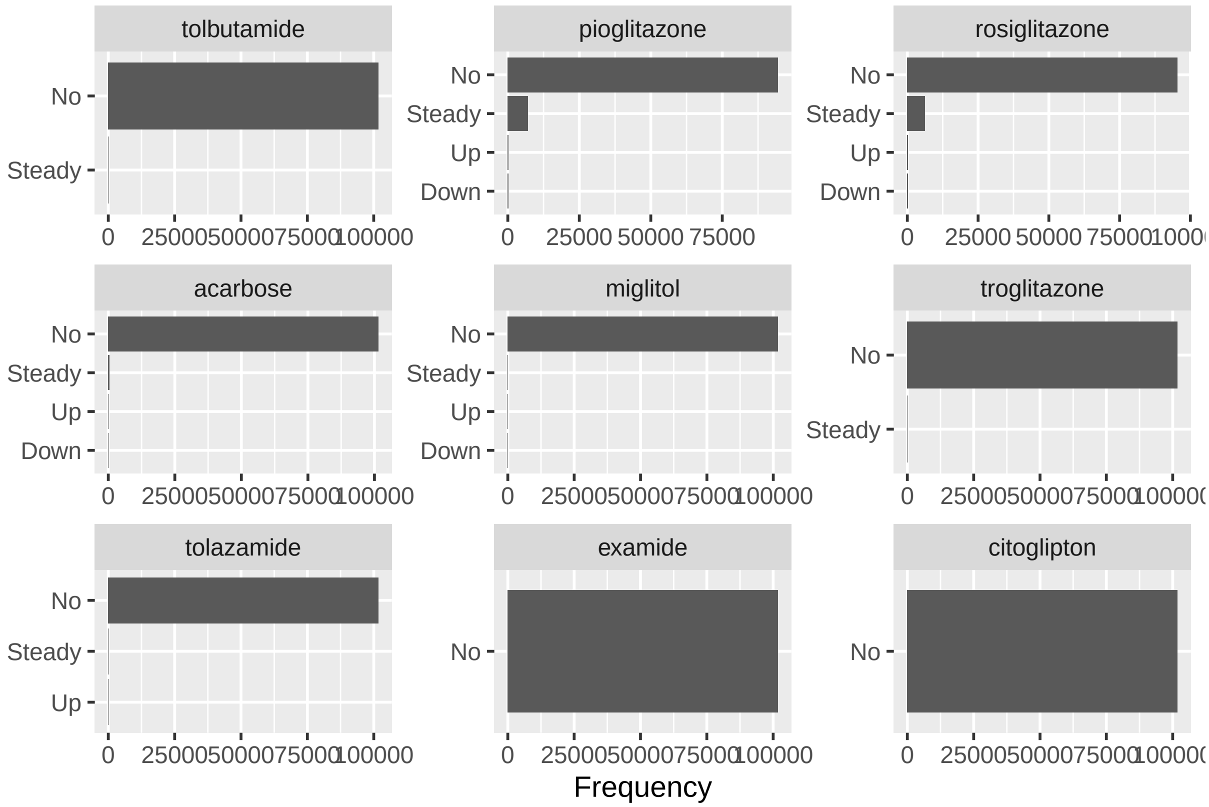
The majority of the variables do not contain NULL or missing values. However, variables `payer_code` (0.604%), `medical_specialty` (0.509%), and especially `weight` (0.031%) exhibit very low completion rates.

```
## 6 columns ignored with more than 50 categories.
## encounter_id: 101766 categories
## patient_nbr: 71518 categories
## medical_specialty: 73 categories
## diag_1: 717 categories
## diag_2: 749 categories
## diag_3: 790 categories
```

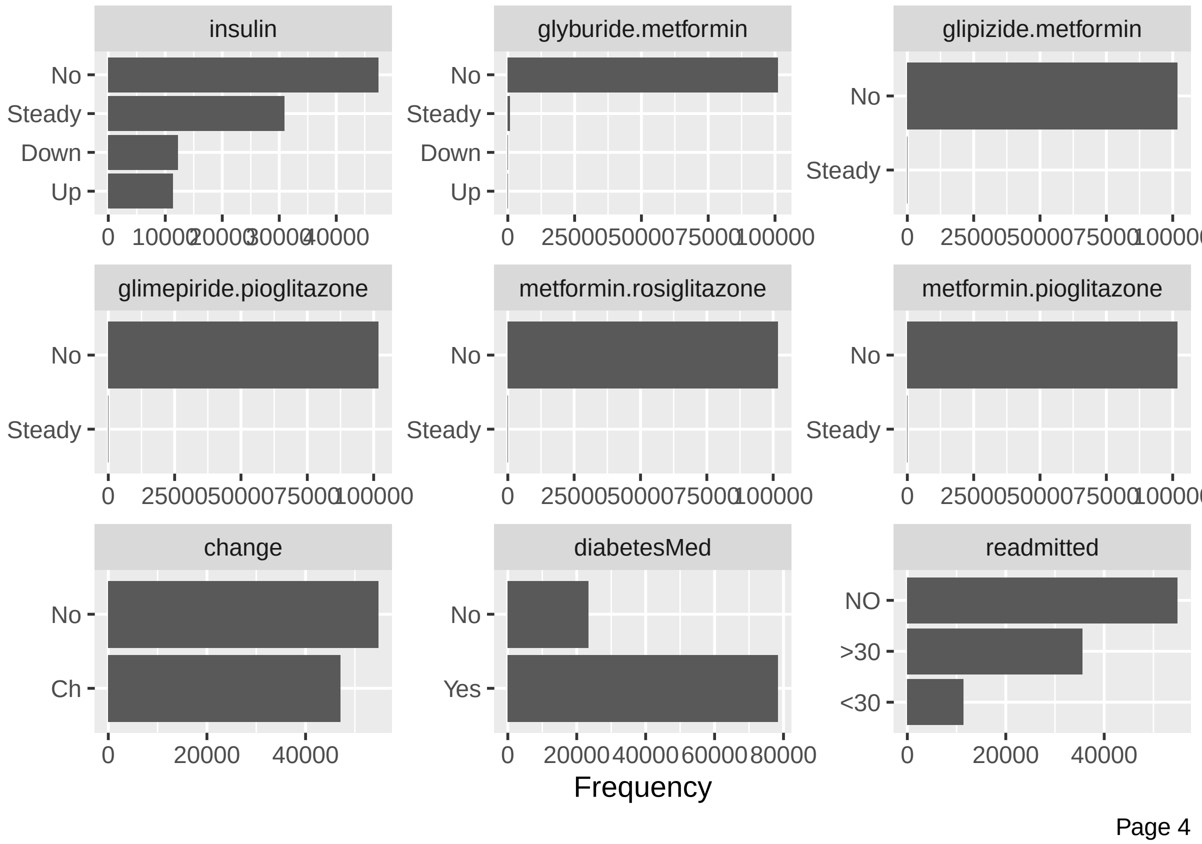

race
NA
Caucasian
AfricanAmerican
Hispanic
Other
Asian
0
20000
40000
60000
gender
NA
Female
Male
0
20000
40000
age
[70−80)
[80−90)
[0−10)
[60−70)
[50−60)
[10−20)
[40−50)
[30−40)
[20−30)
[90−100)
0
10000
20000
weight
NA
>200
[75−100)
[50−75)
[100−125)
[25−50)
[125−150)
[175−200)
[150−175)
[0−25)
0
25000
50000
75000
100000
admission_type_id
7
1
3
6
2
5
4
8
0
20000
40000
discharge_disposition_i
0
20000
40000
admission_source_id
7
10
0
20000
40000
60000
payer_code
0
10000
20000
30000
40000
max_glu_serum
None
Norm
>200
>300
0
25000
50000
75000
Frequency
Page 1

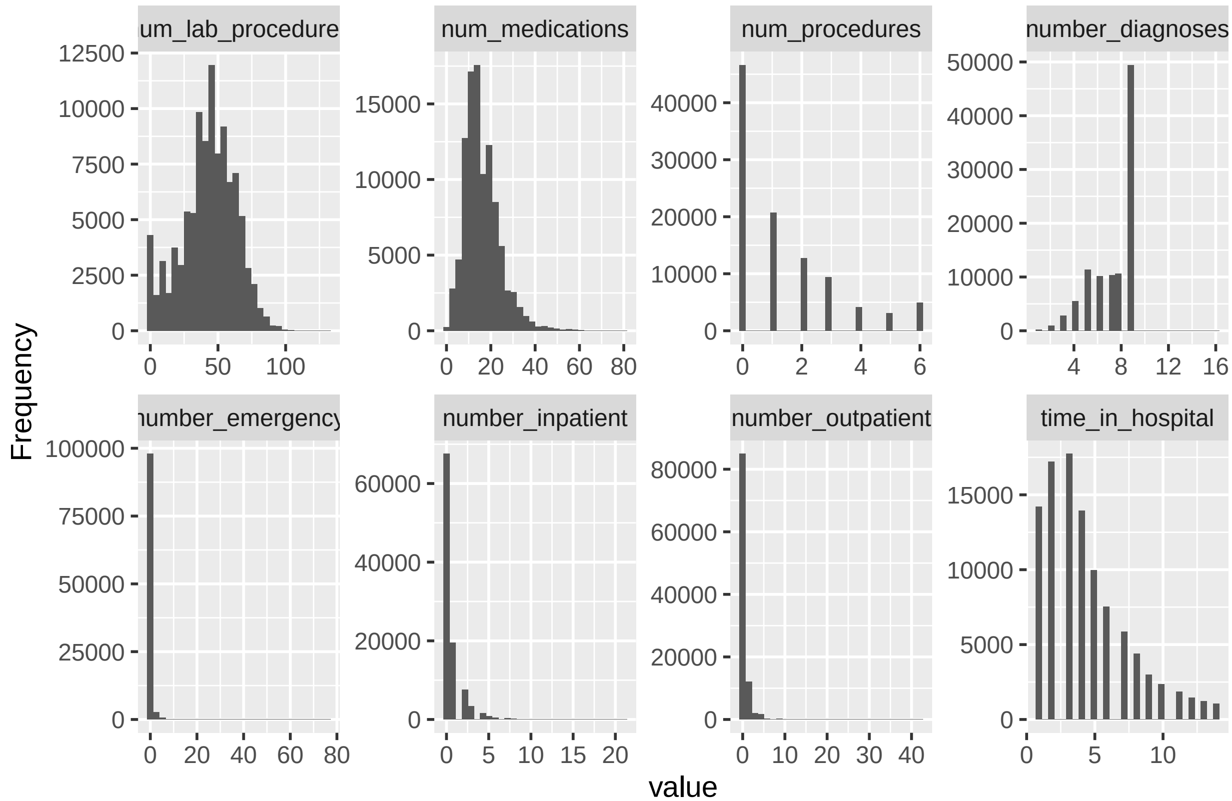


A1Cresult
None
>8
Norm
>7
0
20000
40000
60000
80000
metformin
No
Steady
Down
Up
0
20000
40000
60000
80000
repaglinide
No
Steady
Down
Up
0
25000
50000
75000
nateglinide
No
Steady
Down
Up
0
25000
50000
75000
100000
chlorpropamide
No
Steady
Down
Up
0
25000
50000
75000
100000
glimepiride
No
Steady
Down
Up
0
25000
50000
75000
acetohexamide
No
Steady
0
25000
50000
75000
100000
glipizide
No
Steady
Down
Up
0
25000
50000
75000
glyburide
No
Steady
Down
Up
0
25000
50000
75000
Frequency
Page 2


tolbutamide
No
Steady
pioglitazone
No
Steady
Up
Down
rosiglitazone
No
Steady
Up
Down
acarbose
No
Steady
Up
Down
miglitol
No
Steady
Up
Down
troglitazone
No
Steady
tolazamide
No
Steady
Up
examide
No
citoglipton
No
0
25000
50000
75000
100000
Frequency

Page 3


insulin
No
Steady
Down
Up
0 10000 20000 30000 40000
glyburide.metformin
No
Steady
Down
Up
0 25000 50000 75000 100000
glipizide.metformin
No
Steady
0 25000 50000 75000 100000
glimepiride.pioglitazone
No
Steady
0 25000 50000 75000 100000
metformin.rosiglitazone
No
Steady
0 25000 50000 75000 100000
metformin.pioglitazone
No
Steady
0 25000 50000 75000 100000
change
No
Ch
0 20000 40000
diabetesMed
No
Yes
0 20000 40000 60000 80000
readmitted
NO
>30
<30
0 20000 40000
Frequency
Page 4



Looking at the frequencies of patients within each categorical variable subgroup, there are no obvious anomalies.

Within the continuous variables the only variable which stands out is `number of diagnoses`. There are a very small number of encounters with greater than 8 total diagnoses for a given encounter. This may be a limitation of the Electronic Medical Record system used for data entry.

By quickly inspecting the above graphs our patient demographic appears to be adults (no pediatric hospitals), predominately Caucasian, with slightly more females. A large proportion of the patients are elderly and most patients were prescribed diabetes medications.

0.2 `Q2.` What's the average number of days spent in hospital by admission type?

```
## # A tibble: 8 x 3
##   admission_type_id `Number of Observations` `Mean Length of Stay`
##   <fct>                                  <int>                 <dbl>
## 1 7                                         21                  4.86
## 2 2                                      18480                  4.61
## 3 6                                       5291                  4.58
## 4 1                                      53990                  4.38
## 5 3                                      18869                  4.32
## 6 5                                       4785                  3.95
## 7 4                                         10                  3.2
## 8 8                                        320                  3.06
```

`Admission type 7` has the longest mean length of stay at 4.86 days (to 2 dp). Althought there are a very small number of observations for this group.

0.3 Q3. Given a patient stays in hospital at least 3 days, how much longer do they generally stay?

Mean additional time spent in hospital (all patients):", 1.4

```
## # A tibble: 8 x 4
##   admission_type_id `Number of Observa~ `Mean Length of S~ `Mean Additional Len~
##   <fct>                           <int>              <dbl>                 <dbl>
## 1 7                                  21               4.86                1.86
## 2 2                               18480               4.61                1.61
## 3 6                                5291               4.58                1.58
## 4 1                               53990               4.38                1.38
## 5 3                               18869               4.32                1.32
## 6 5                                4785               3.95                0.947
## 7 4                                  10               3.2                 0.2
## 8 8                                 320               3.06                0.0625
```

Overall, patients stay approximately 1.4 additional days after the 3 day period. Following from the previous question, this additional time in hospital is far less for admission types 4 and 8.

0.4 Q4. Is there a significant different in time spent in hospital for patients over 60 compared to all other patients?

```
##
##  Welch Two Sample t-test
##
## data:  LOS_over_60yo and LOS_all_patients
## t = 12.631, df = 145712, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1586678 0.2169516
## sample estimates:
## mean of x mean of y
##  4.583797  4.395987
```

A simple T-test is one way of determining if there is a statistically significant difference between over 60 patients compared to other patients.

Note that the question specifies `over 60` but the subgroup `[60-70)` will include 60 year olds. This cannot be fixed using the provided data.

```
##
## Call:
## glm(formula = time_in_hospital ~ over_60, family = gaussian,
##     data = diabetic_raw)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5838  -2.0085  -0.5838   1.4162   9.9915
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.00855    0.01631  245.78   <2e-16 ***
## over_60TRUE  0.57525    0.01987   28.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for gaussian family taken to be 8.83819)
##
##     Null deviance: 906815  on 101765  degrees of freedom
## Residual deviance: 899410  on 101764  degrees of freedom
## AIC: 510560
##
## Number of Fisher Scoring iterations: 2
```

A linear model is another method of investigating the differences between groups. The univariate results displayed above reflect the T-test results.

```
##
## Call:
## glm(formula = time_in_hospital ~ over_60 + gender + admission_type_id,
##     family = gaussian, data = diabetic_raw)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2504  -2.1354  -0.6903   1.4125  10.8082
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.28303    0.04390  97.573  < 2e-16 ***
## over_60TRUE         0.57156    0.01985  28.796  < 2e-16 ***
## genderMale         -0.16431    0.01867  -8.801  < 2e-16 ***
## admission_type_id1 -0.21398    0.04275  -5.005 5.58e-07 ***
## admission_type_id2  0.01668    0.04627   0.360    0.719
## admission_type_id3 -0.26711    0.04616  -5.787 7.21e-09 ***
## admission_type_id4 -1.36810    0.93925  -1.457    0.145
## admission_type_id5 -0.66348    0.05920 -11.207  < 2e-16 ***
## admission_type_id8 -1.49845    0.17083  -8.772  < 2e-16 ***
## admission_type_id7  0.39584    0.64882   0.610    0.542
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 8.805052)
##
##     Null deviance: 906778  on 101762  degrees of freedom
## Residual deviance: 895940  on 101753  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 510171
##
## Number of Fisher Scoring iterations: 2
```

`confidence intervals`

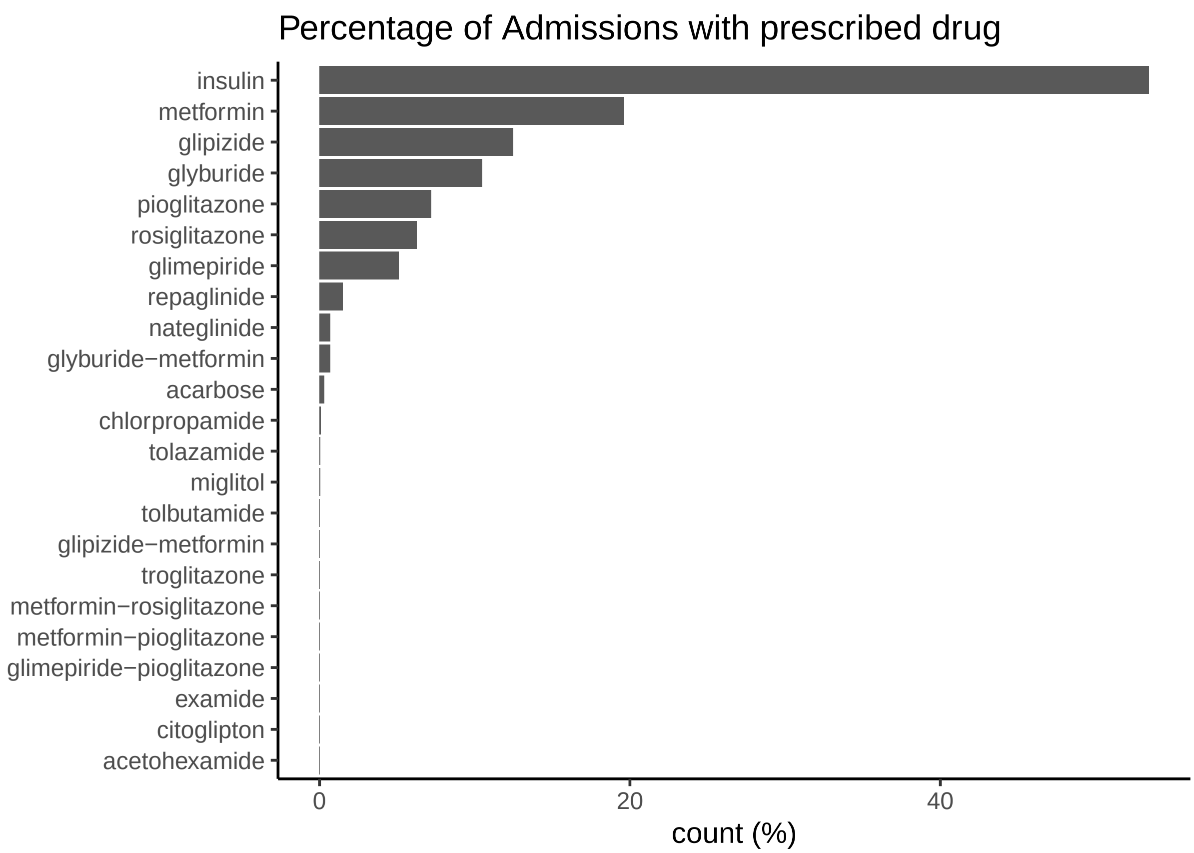
```
## Waiting for profiling to be done...
```

```
##                         2.5 %     97.5 %
## (Intercept)         4.1969927  4.3690605
## over_60TRUE         0.5326600  0.6104653
## genderMale         -0.2009025 -0.1277195
## admission_type_id1 -0.2977700 -0.1301926
## admission_type_id2 -0.0740085  0.1073627
## admission_type_id3 -0.3575863 -0.1766379
## admission_type_id4 -3.2089944  0.4727889
```

```
## admission_type_id5 -0.7795216 -0.5474437
## admission_type_id8 -1.8332665 -1.1636433
## admission_type_id7 -0.8758300  1.6675011
```

A multivariate approach incorporating possible confounder variables will allow for a more accurate estimate of the effect sizes and P values. When investigating underlying causes for increased length of stay this is also critical. Note that the list of covariates in the above multivariate model could be expanded based on existing research and investigating causal diagrams with more time spent on this exercise. However, looking at the multivariate analysis we can see that patients who are over >60 spend 0.68

0.5 Q5. People have very different sets of medicines provided during an episode of care for diabetes. Describe / group the different types of admissions in terms of the types of medicines patients are taking.



Tabulated results

```
## # A tibble: 16 x 2
##    medication         `count (%)`
##    <chr>                    <dbl>
##  1 insulin                  53.4
##  2 metformin                19.6
##  3 glipizide                12.5
##  4 glyburide                10.5
##  5 pioglitazone              7.2
##  6 rosiglitazone             6.25
##  7 glimepiride               5.1
##  8 repaglinide               1.51
```

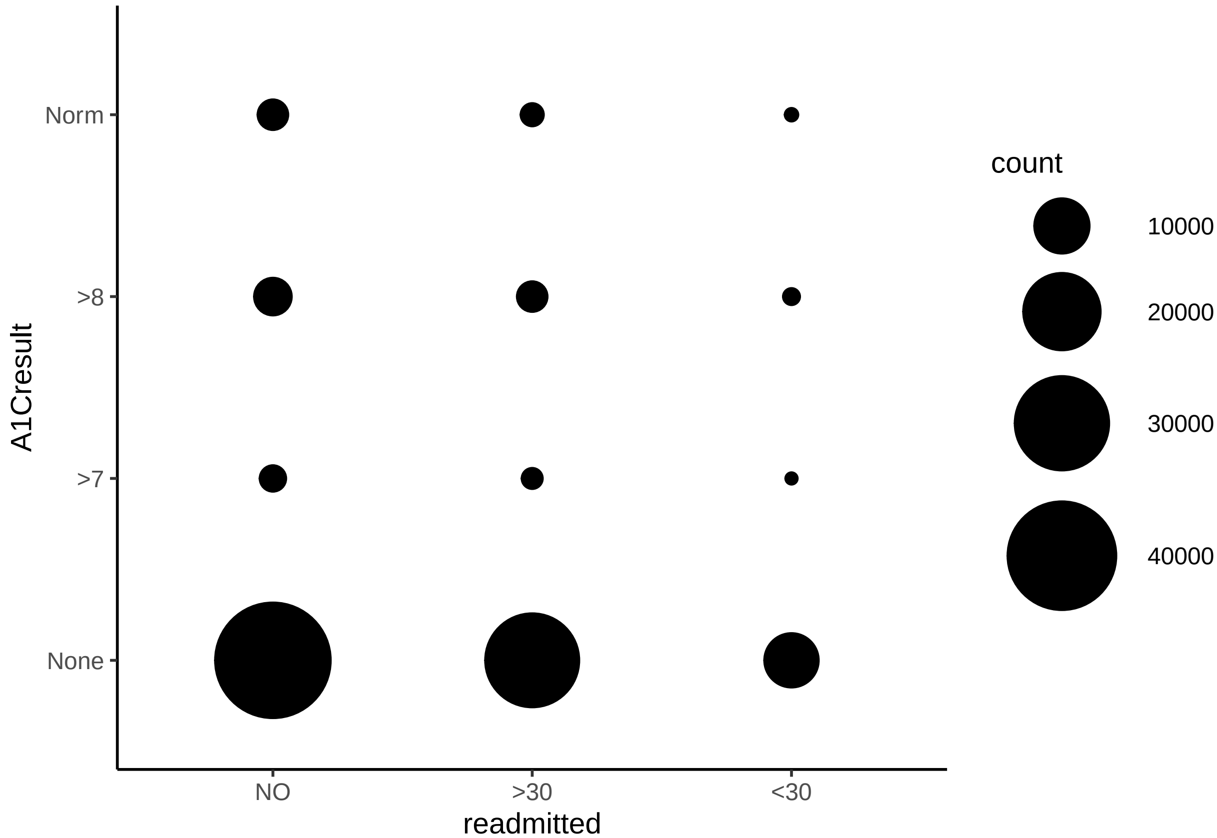
```
##   9 glyburide-metformin        0.69
##  10 nateglinide                0.69
##  11 acarbose                   0.3
##  12 chlorpropamide             0.08
##  13 miglitol                   0.04
##  14 tolazamide                 0.04
##  15 tolbutamide                0.02
##  16 glipizide-metformin        0.01
```

0.6 Q6. All else being equal, is HBA1c testing a useful measure for identifying whether a patient is likely to be readmitted? Explain your approach and the key information which drives your decision.

```
## `summarise()` has grouped output by 'A1Cresult'. You can override using the `.groups` argument.
```

```
## # A tibble: 4 x 4
## # Groups:   A1Cresult [4]
##   A1Cresult    NO `>30` `<30`
##   <fct>     <int> <int> <int>
## 1 None      45322 29745  9681
## 2 >7         2129  1300   383
## 3 >8         4504  2901   811
## 4 Norm       2909  1599   482
```

```
## # A tibble: 12 x 3
## # Groups:   A1Cresult [4]
##    A1Cresult readmitted count
##    <fct>     <fct>      <int>
##  1 None      NO         45322
##  2 None      >30        29745
##  3 None      <30         9681
##  4 >7        NO          2129
##  5 >7        >30         1300
##  6 >7        <30          383
##  7 >8        NO          4504
##  8 >8        >30         2901
##  9 >8        <30          811
## 10 Norm      NO          2909
## 11 Norm      >30         1599
## 12 Norm      <30          482
```



Using a similar approach to question 4, I would use generalized linear modelling to determine if HBA1C results are predictive of readmission to hospital. This could be followed up with multilevel modelling techniques to further understand more complex relationships and trends within this dataset. We can see in this univariate analysis that HBA1C is statistically significant and negatively associated with readmission under 30 days.

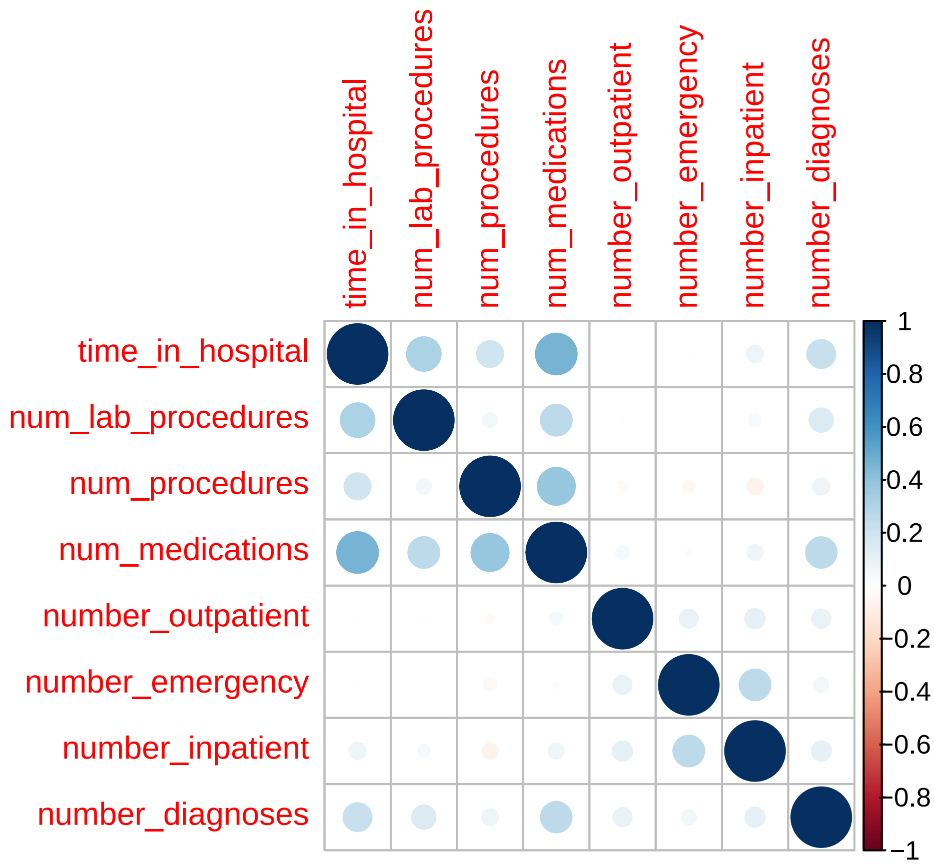
0.7 `Q7.` What other analysis would you do if you had more time on any of the previous questions?

Generally I would prefer to spend longer exploring relationships between the variables, using more graphical techniques to group the data differently and look for underlying trends. With a large number of features principle component analysis (PCA) can be a helpful technique to summarize which variables account for the majority of the variation in key metrics such as readmission. It's then useful to look at the specific scenarios/subgroups which could be particularly vulnerable to readmission for diabetic patients.

Incorporation of the ICD10 codes in the analysis was not explored and would also be something I'd investigate with more time. For example, specific diagnoses may be associated with higher HBA1c results or readmission.

As the tasks included here are predominately descriptive rather than predictive, I have primarily focused on traditional statistical techniques such as generalized linear modelling. However, I have found machine learning techniques very interesting and useful for predictive tasks.

0.8 Q8. 8. Are there any other interesting findings from the data you would like to share which don't fall into the above questions?



The number of medications prescribed is positively associated with time spent in hospital.

0.9 Q9. 9. What other information would you request or try to collect to make a better decision?

In Australia PBS (pharmaceutical benefits scheme) data can provide information surrounding whether the patient has been purchasing their medication outside of the hospital (presumably less likely to readmit to hospital) (Note however that PBS dispensing data does not necessarily indicate the patient has administered the drug), information about their residence and nearest hospital location (difficulty returning to hospital), patient mobility (are they able to travel to hospital even if they wanted to).

Essentially any data which can provide an indication about the patient's level of adherence to the medical advice after discharge ie. prescription dispensing data, remote blood sugar level readings throughout the day, exercise levels, diet information.

From here confounding information which may bias our outcome variable is also useful, eg. A patient living a block away from the hospital would presumably be more likely to readmit compared to a patient living in the countryside with a long distance between the hospital and the patient's residence.