```
#Prepare Data for Analysis
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
import matplotlib.pyplot as plt
import seaborn as sns
import datetime



df = pd.read_csv('diabetic_data.csv')
df.head()
```

```
   encounter_id  patient_nbr             race  gender      age weight  
\
0       2278392      8222157        Caucasian  Female   [0-10)      ?   

1        149190     55629189        Caucasian  Female  [10-20)      ?   

2         64410     86047875  AfricanAmerican  Female  [20-30)      ?   

3        500364     82442376        Caucasian    Male  [30-40)      ?   

4         16680     42519267        Caucasian    Male  [40-50)      ?   


   admission_type_id  discharge_disposition_id  admission_source_id  \
0                  6                        25                    1   
1                  1                         1                    7   
2                  1                         1                    7   
3                  1                         1                    7   
4                  1                         1                    7   

   time_in_hospital  ... citoglipton insulin  glyburide-metformin  \
0                 1  ...          No      No                   No   
1                 3  ...          No      Up                   No   
2                 2  ...          No      No                   No   
3                 2  ...          No      Up                   No   
4                 1  ...          No  Steady                   No   

   glipizide-metformin  glimepiride-pioglitazone  metformin-
rosiglitazone  \
0                   No                        No                        
No   
1                   No                        No                        
No   
2                   No                        No                        
No   
3                   No                        No                        
```

```
No
4                   No                       No
No

   metformin-pioglitazone  change diabetesMed readmitted
0                      No      No          No         NO
1                      No      Ch         Yes        >30
2                      No      No         Yes         NO
3                      No      Ch         Yes         NO
4                      No      Ch         Yes         NO

[5 rows x 50 columns]
```

```
df.columns
```

```
Index(['encounter_id', 'patient_nbr', 'race', 'gender', 'age',
'weight',
       'admission_type_id', 'discharge_disposition_id',
'admission_source_id',
       'time_in_hospital', 'payer_code', 'medical_specialty',
       'num_lab_procedures', 'num_procedures', 'num_medications',
       'number_outpatient', 'number_emergency', 'number_inpatient',
'diag_1',
       'diag_2', 'diag_3', 'number_diagnoses', 'max_glu_serum',
'A1Cresult',
       'metformin', 'repaglinide', 'nateglinide', 'chlorpropamide',
       'glimepiride', 'acetohexamide', 'glipizide', 'glyburide',
'tolbutamide',
       'pioglitazone', 'rosiglitazone', 'acarbose', 'miglitol',
'troglitazone',
       'tolazamide', 'examide', 'citoglipton', 'insulin',
       'glyburide-metformin', 'glipizide-metformin',
       'glimepiride-pioglitazone', 'metformin-rosiglitazone',
       'metformin-pioglitazone', 'change', 'diabetesMed',
'readmitted'],
      dtype='object')
```

```
df.shape
```

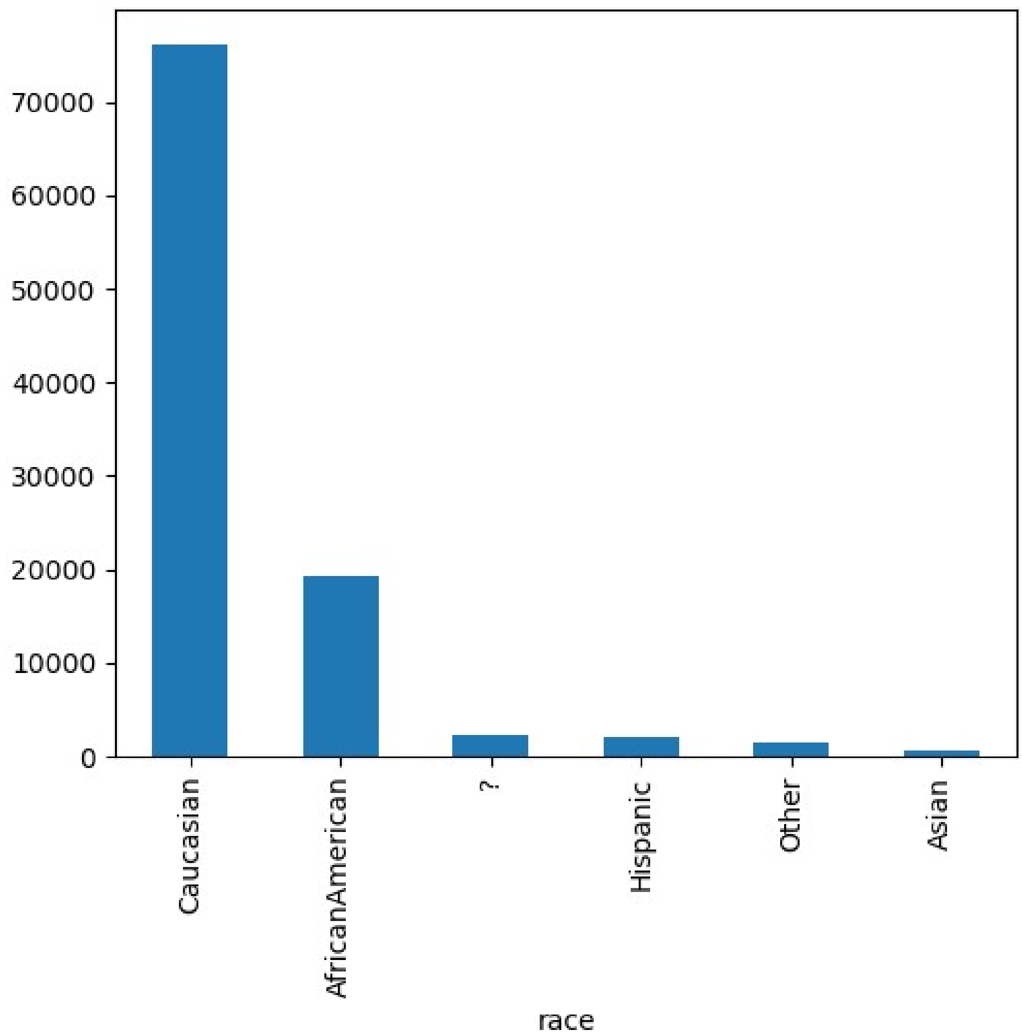
```
(101766, 50)
```

```
# Perform descriptive statistical analysis for numerical features
df.describe()
```

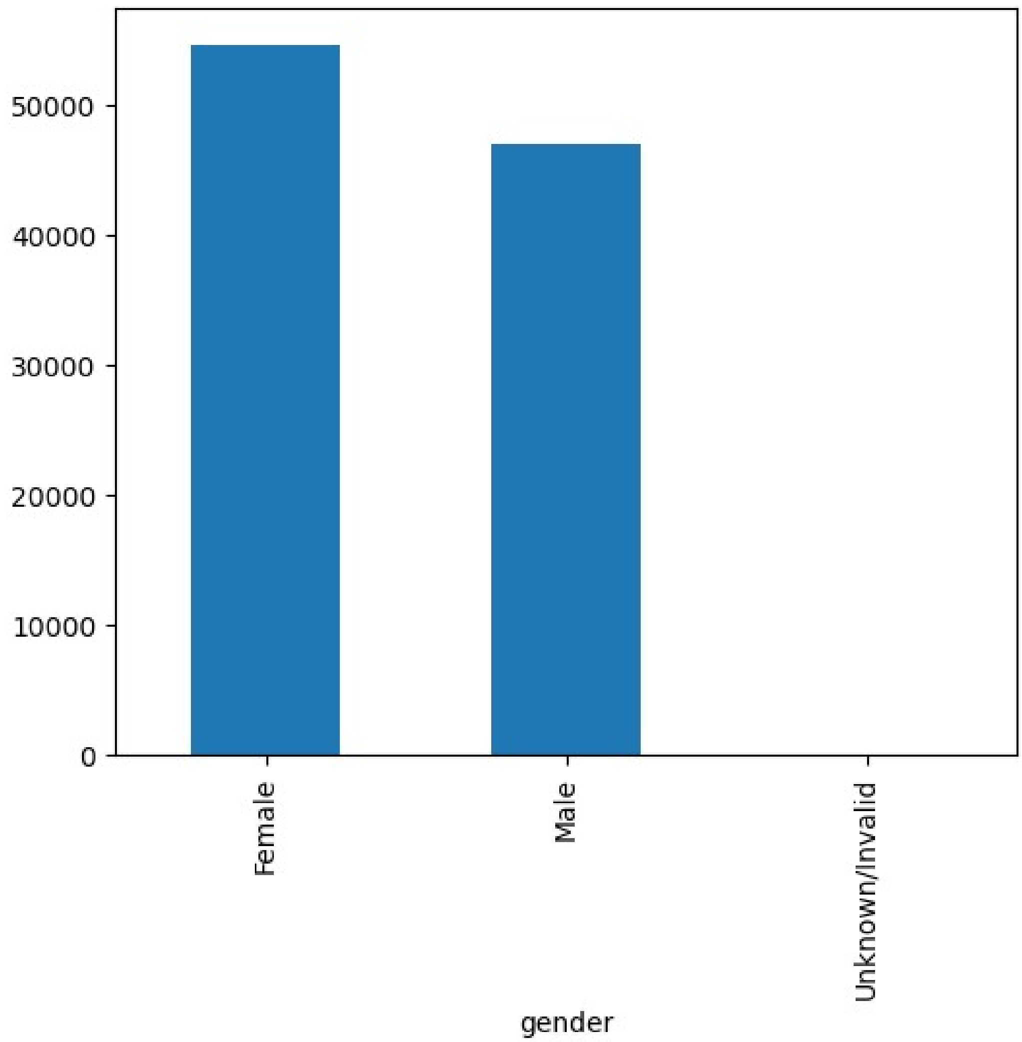
```
       encounter_id   patient_nbr  admission_type_id  \
count  1.017660e+05  1.017660e+05      101766.000000
mean   1.652016e+08  5.433040e+07           2.024006
std    1.026403e+08  3.869636e+07           1.445403
min    1.252200e+04  1.350000e+02           1.000000
25%    8.496119e+07  2.341322e+07           1.000000
50%    1.523890e+08  4.550514e+07           1.000000
```

```
75%    2.302709e+08  8.754595e+07           3.000000
max    4.438672e+08  1.895026e+08           8.000000

       discharge_disposition_id  admission_source_id  time_in_hospital  \
count             101766.000000        101766.000000     101766.000000

mean                   3.715642             5.754437          4.395987

std                    5.280166             4.064081          2.985108

min                    1.000000             1.000000          1.000000

25%                    1.000000             1.000000          2.000000

50%                    1.000000             7.000000          4.000000

75%                    4.000000             7.000000          6.000000

max                   28.000000            25.000000         14.000000


       num_lab_procedures  num_procedures  num_medications
number_outpatient  \
count       101766.000000   101766.000000    101766.000000
101766.000000
mean            43.095641        1.339730        16.021844
0.369357
std             19.674362        1.705807         8.127566
1.267265
min              1.000000        0.000000         1.000000
0.000000
25%             31.000000        0.000000        10.000000
0.000000
50%             44.000000        1.000000        15.000000
0.000000
75%             57.000000        2.000000        20.000000
0.000000
max            132.000000        6.000000        81.000000
42.000000

       number_emergency  number_inpatient  number_diagnoses
count     101766.000000     101766.000000     101766.000000
mean           0.197836          0.635566          7.422607
std            0.930472          1.262863          1.933600
min            0.000000          0.000000          1.000000
25%            0.000000          0.000000          6.000000
50%            0.000000          0.000000          8.000000
75%            0.000000          1.000000          9.000000
max           76.000000         21.000000         16.000000
```

```
#Visualize the distribution of categorical features - race and gender
plt.figure(figsize=(6,5))
df['race'].value_counts().plot(kind = 'bar')
plt.show()
```




```
plt.figure(figsize=(6,5))
df['gender'].value_counts().plot(kind = 'bar')
plt.show()
```



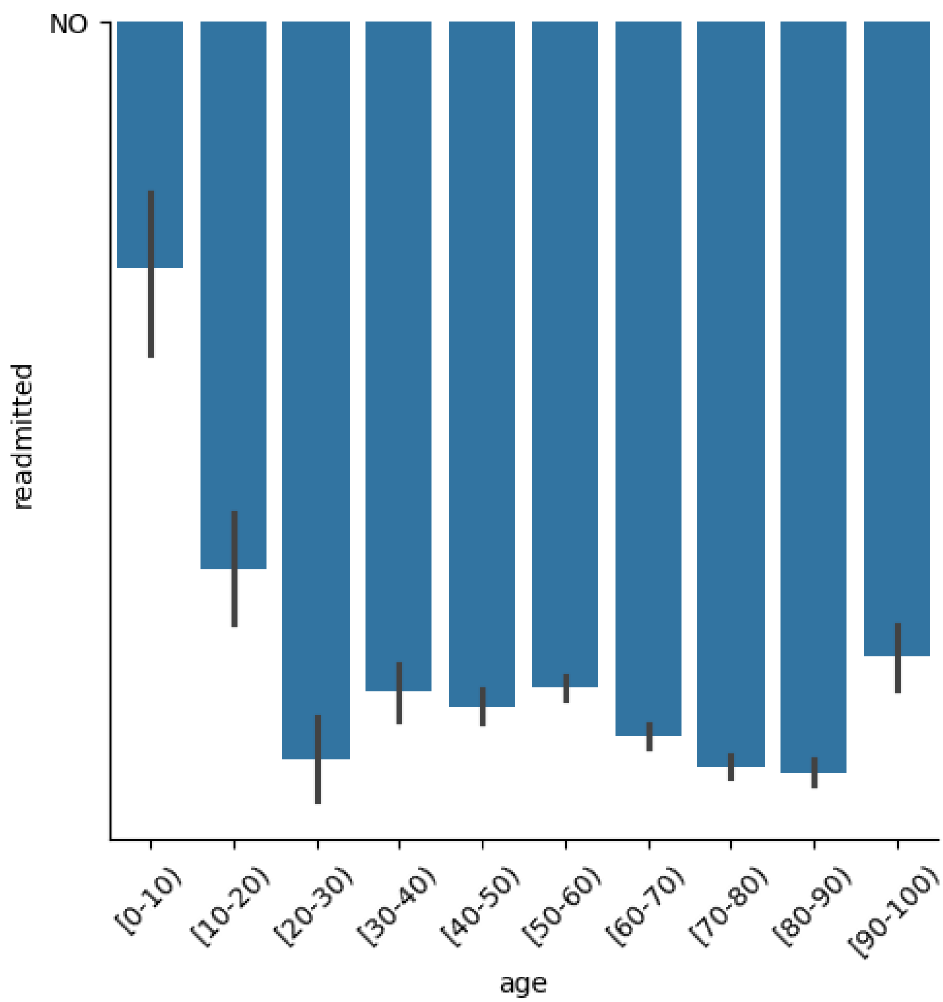

```
#Explore the relationship between readmission status and age
df['age'].value_counts()
```

```
age
[70-80)      26068
[60-70)      22483
[50-60)      17256
[80-90)      17197
[40-50)       9685
[30-40)       3775
[90-100)      2793
[20-30)       1657
[10-20)        691
[0-10)         161
Name: count, dtype: int64
```

```
plt.figure(figsize=(10,10))
sns.catplot(x='age',y='readmitted',data=df,kind='bar')
```

```
plt.xticks(rotation=45)
plt.show()
```

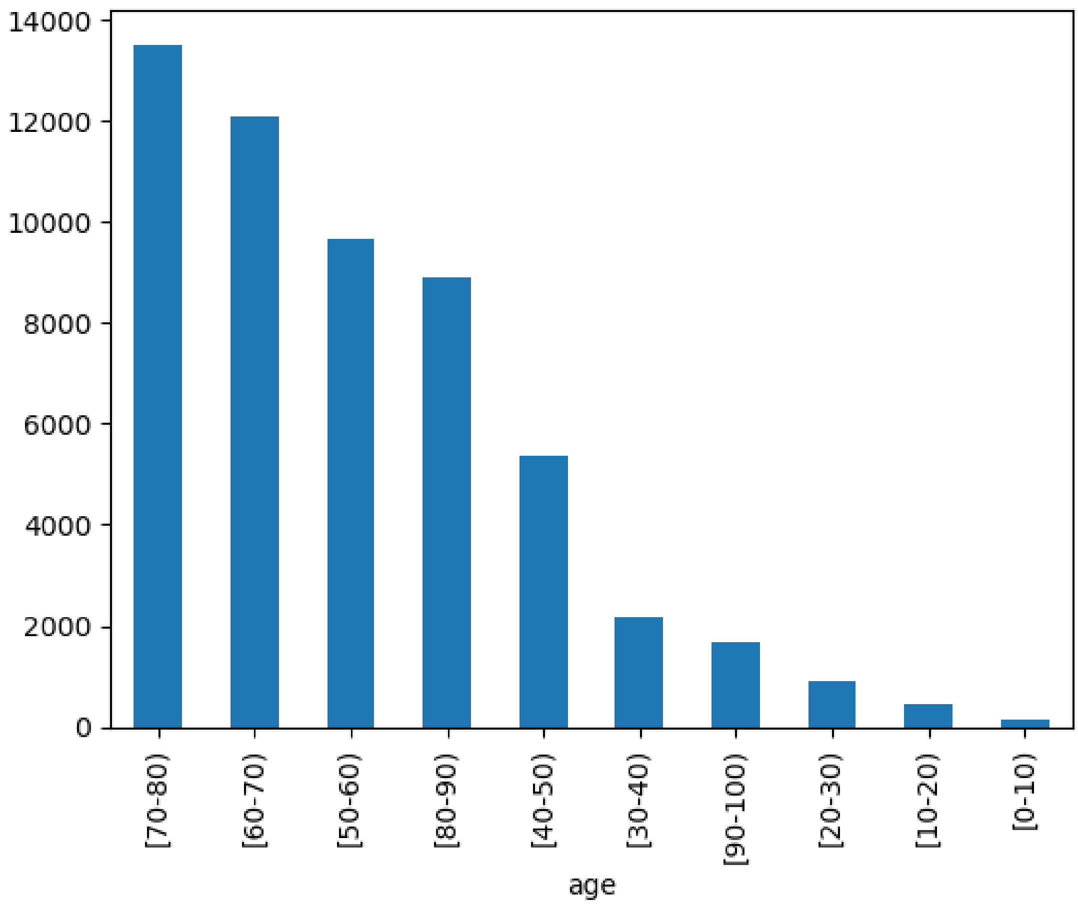
```
<Figure size 1000x1000 with 0 Axes>
```



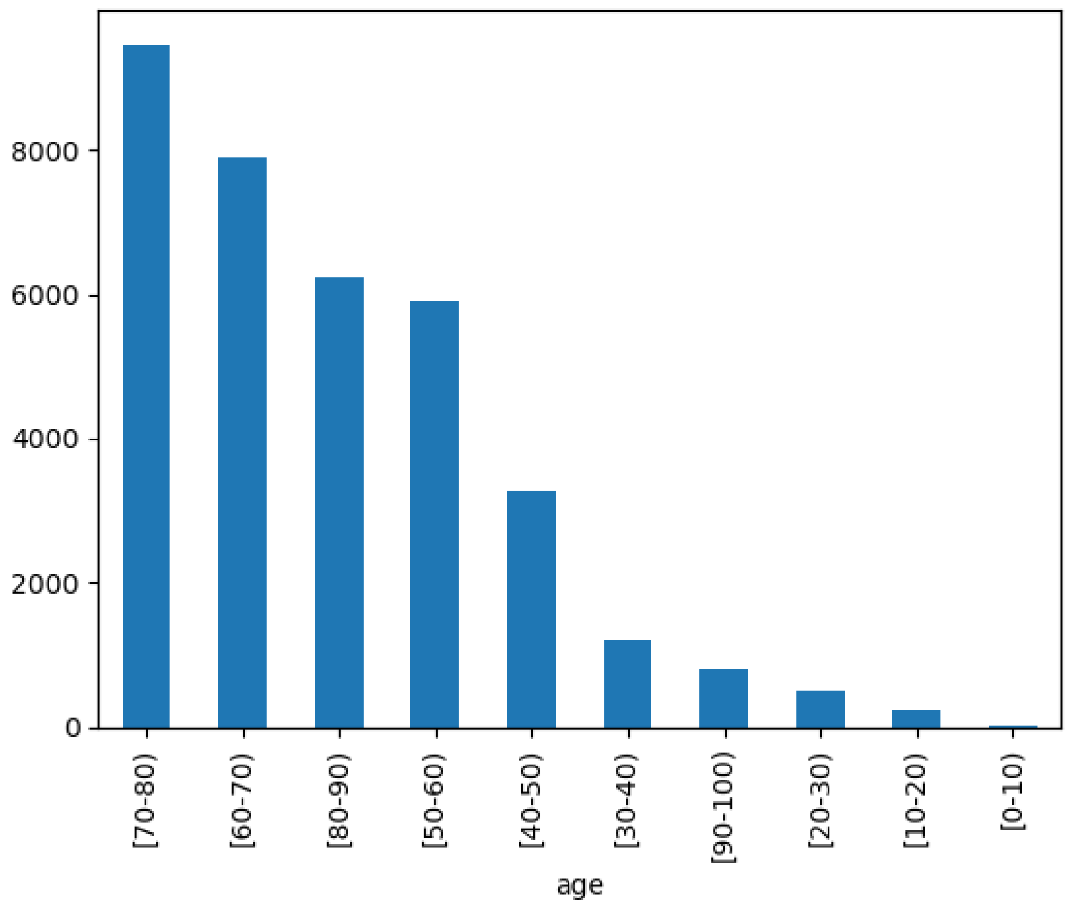

```
readmission_list=df['readmitted'].unique().tolist()
readmission_list
```

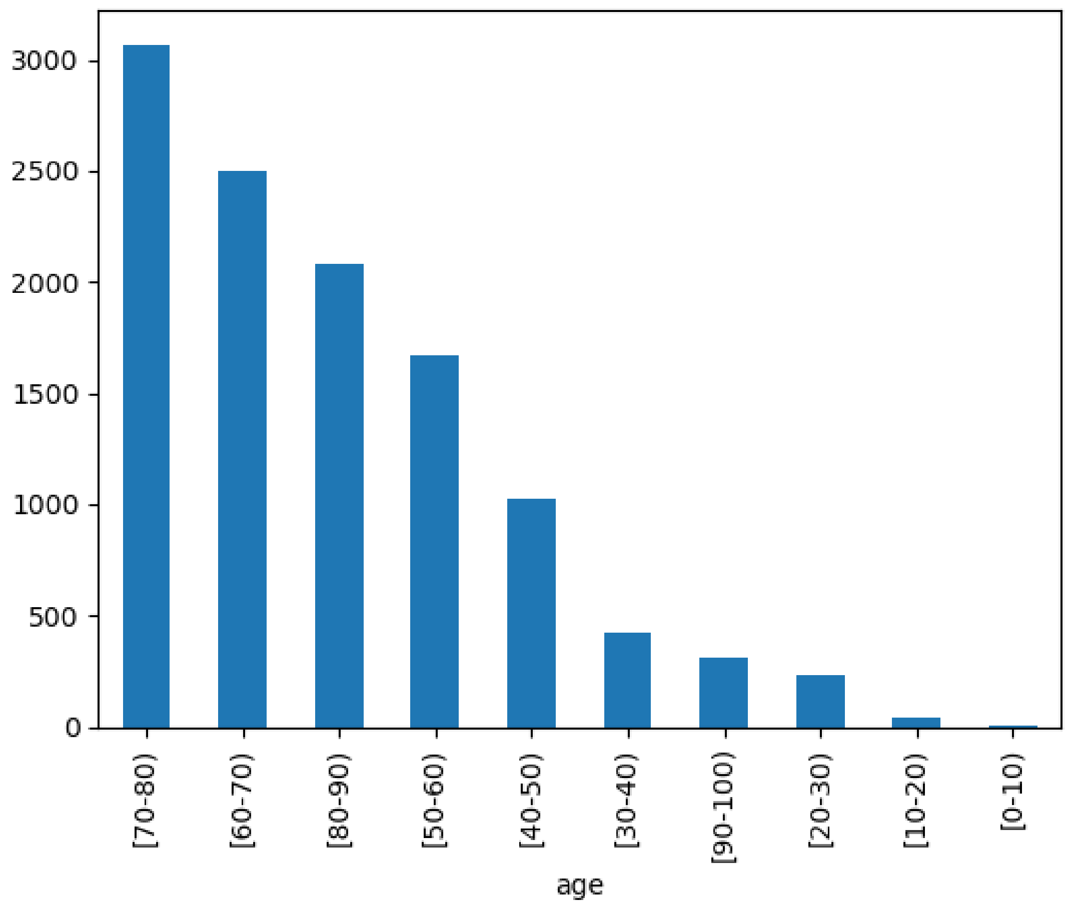
```
['NO', '>30', '<30']
```

```
import seaborn as sns
for i in readmission_list:
    df_new=df[df['readmitted']==i]
    df_new['age'].value_counts().plot(kind='bar')
    plt.show()
```



14000
12000
10000
8000
6000
4000
2000
0
[70-80)
[60-70)
[50-60)
[80-90)
[40-50)
[30-40)
[90-100)
[20-30)
[10-20)
[0-10)
age


8000
6000
4000
2000
0
[70-80)
[60-70)
[80-90)
[50-60)
[40-50)
[30-40)
[90-100)
[20-30)
[10-20)
[0-10)
age

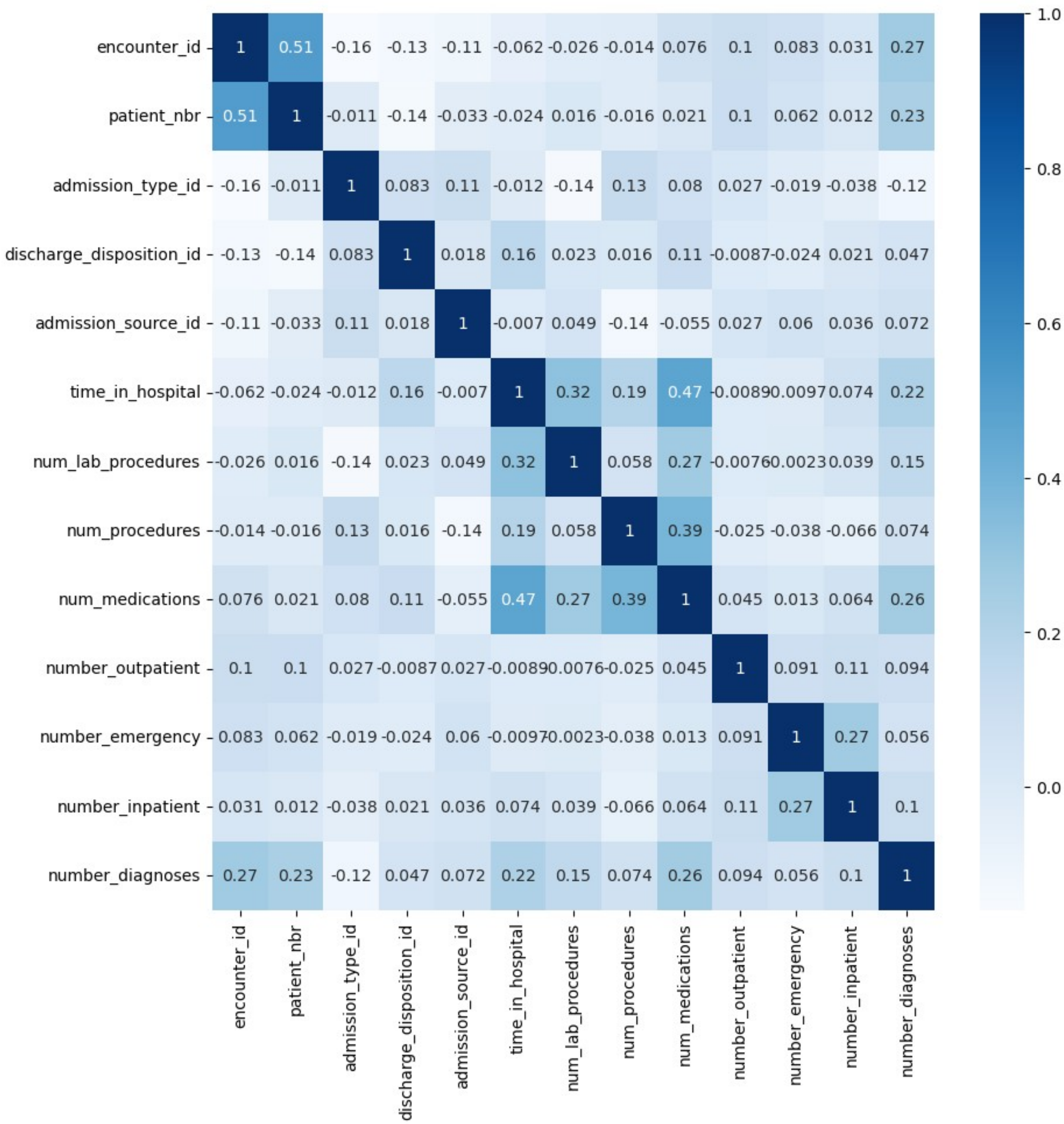



```
#Investigate correlations between numerical features
numerical_features=df.select_dtypes(include=['number']).columns
numerical_features

Index(['encounter_id', 'patient_nbr', 'admission_type_id',
       'discharge_disposition_id', 'admission_source_id',
'time_in_hospital',
       'num_lab_procedures', 'num_procedures', 'num_medications',
       'number_outpatient', 'number_emergency', 'number_inpatient',
       'number_diagnoses'],
      dtype='object')

numerical_features=numerical_features.tolist()
numerical_features

['encounter_id',
 'patient_nbr',
 'admission_type_id',
 'discharge_disposition_id',
 'admission_source_id',
 'time_in_hospital',
 'num_lab_procedures',
```

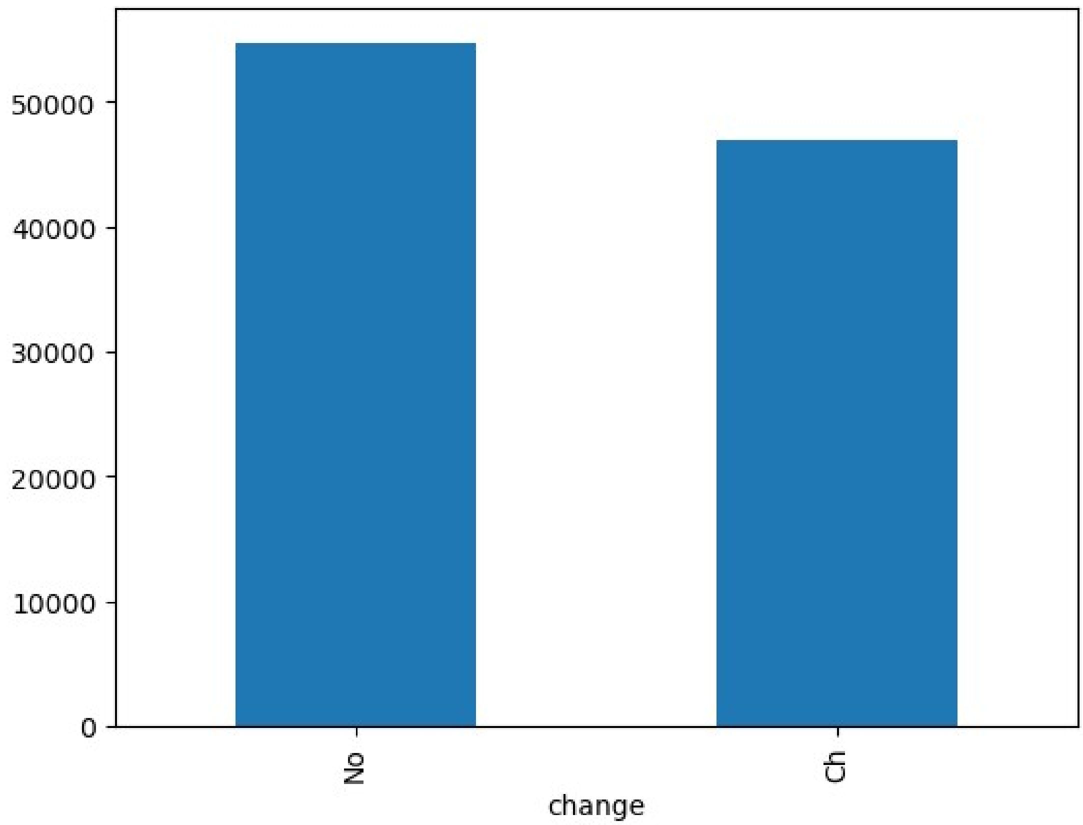
```
 'num_procedures',
 'num_medications',
 'number_outpatient',
 'number_emergency',
 'number_inpatient',
 'number_diagnoses']

df[numerical_features].corr()

                          encounter_id  patient_nbr  admission_type_id  \
encounter_id                  1.000000     0.512028          -0.158961

patient_nbr                   0.512028     1.000000          -0.011128

admission_type_id            -0.158961    -0.011128           1.000000

discharge_disposition_id     -0.132876    -0.136814           0.083483

admission_source_id          -0.112402    -0.032568           0.106654

time_in_hospital             -0.062221    -0.024092          -0.012500

num_lab_procedures           -0.026062     0.015946          -0.143713

num_procedures               -0.014225    -0.015570           0.129888

num_medications               0.076113     0.020665           0.079535

number_outpatient             0.103756     0.103379           0.026511

number_emergency              0.082803     0.062352          -0.019116

number_inpatient              0.030962     0.012480          -0.038161

number_diagnoses              0.265149     0.226847          -0.117126


                          discharge_disposition_id
admission_source_id  \
encounter_id                             -0.132876             -
0.112402
patient_nbr                              -0.136814             -
0.032568
admission_type_id                         0.083483
0.106654
discharge_disposition_id                  1.000000
0.018193
admission_source_id                       0.018193
1.000000
time_in_hospital                          0.162748             -
```

```
0.006965
num_lab_procedures                          0.023415
0.048885
num_procedures                              0.015921             -
0.135400
num_medications                             0.108753             -
0.054533
number_outpatient                          -0.008715
0.027244
number_emergency                           -0.024471
0.059892
number_inpatient                            0.020787
0.036314
number_diagnoses                            0.046891
0.072114

                          time_in_hospital  num_lab_procedures  \
encounter_id                     -0.062221           -0.026062
patient_nbr                      -0.024092            0.015946
admission_type_id                -0.012500           -0.143713
discharge_disposition_id          0.162748            0.023415
admission_source_id              -0.006965            0.048885
time_in_hospital                  1.000000            0.318450
num_lab_procedures                0.318450            1.000000
num_procedures                    0.191472            0.058066
num_medications                   0.466135            0.268161
number_outpatient                -0.008916           -0.007602
number_emergency                 -0.009681           -0.002279
number_inpatient                  0.073623            0.039231
number_diagnoses                  0.220186            0.152773

                          num_procedures  num_medications
number_outpatient  \
encounter_id                   -0.014225         0.076113
0.103756
patient_nbr                    -0.015570         0.020665
0.103379
admission_type_id               0.129888         0.079535
0.026511
discharge_disposition_id        0.015921         0.108753           -
0.008715
admission_source_id            -0.135400        -0.054533
0.027244
time_in_hospital                0.191472         0.466135           -
0.008916
num_lab_procedures              0.058066         0.268161           -
0.007602
num_procedures                  1.000000         0.385767           -
0.024819
```

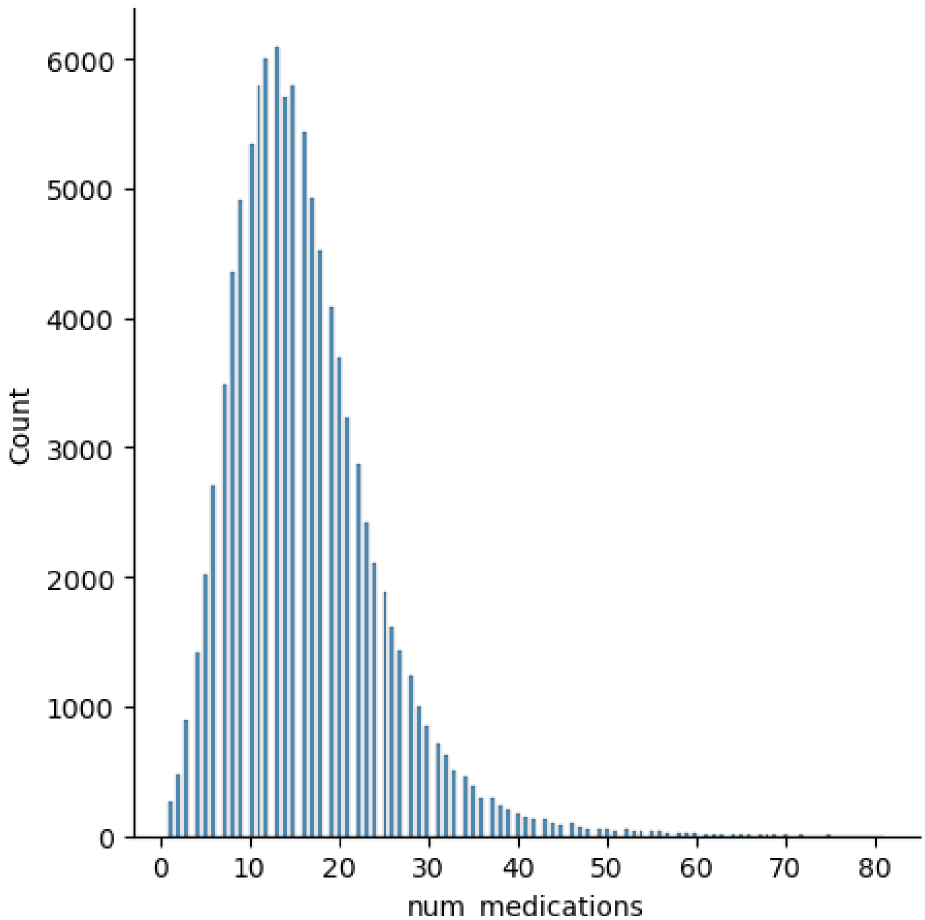
```
num_medications                        0.385767         1.000000
0.045197
number_outpatient                     -0.024819         0.045197
1.000000
number_emergency                      -0.038179         0.013180
0.091459
number_inpatient                      -0.066236         0.064194
0.107338
number_diagnoses                       0.073734         0.261526
0.094152

                          number_emergency  number_inpatient
number_diagnoses
encounter_id                      0.082803          0.030962
0.265149
patient_nbr                       0.062352          0.012480
0.226847
admission_type_id                -0.019116         -0.038161          -
0.117126
discharge_disposition_id         -0.024471          0.020787
0.046891
admission_source_id               0.059892          0.036314
0.072114
time_in_hospital                 -0.009681          0.073623
0.220186
num_lab_procedures               -0.002279          0.039231
0.152773
num_procedures                   -0.038179         -0.066236
0.073734
num_medications                   0.013180          0.064194
0.261526
number_outpatient                 0.091459          0.107338
0.094152
number_emergency                  1.000000          0.266559
0.055539
number_inpatient                  0.266559          1.000000
0.104710
number_diagnoses                  0.055539          0.104710
1.000000

import seaborn as sns
import matplotlib.pyplot as plt
plt.figure(figsize=(10,10))
sns.heatmap(df[numerical_features].corr(),annot=True,cmap='Blues')
plt.show()
```



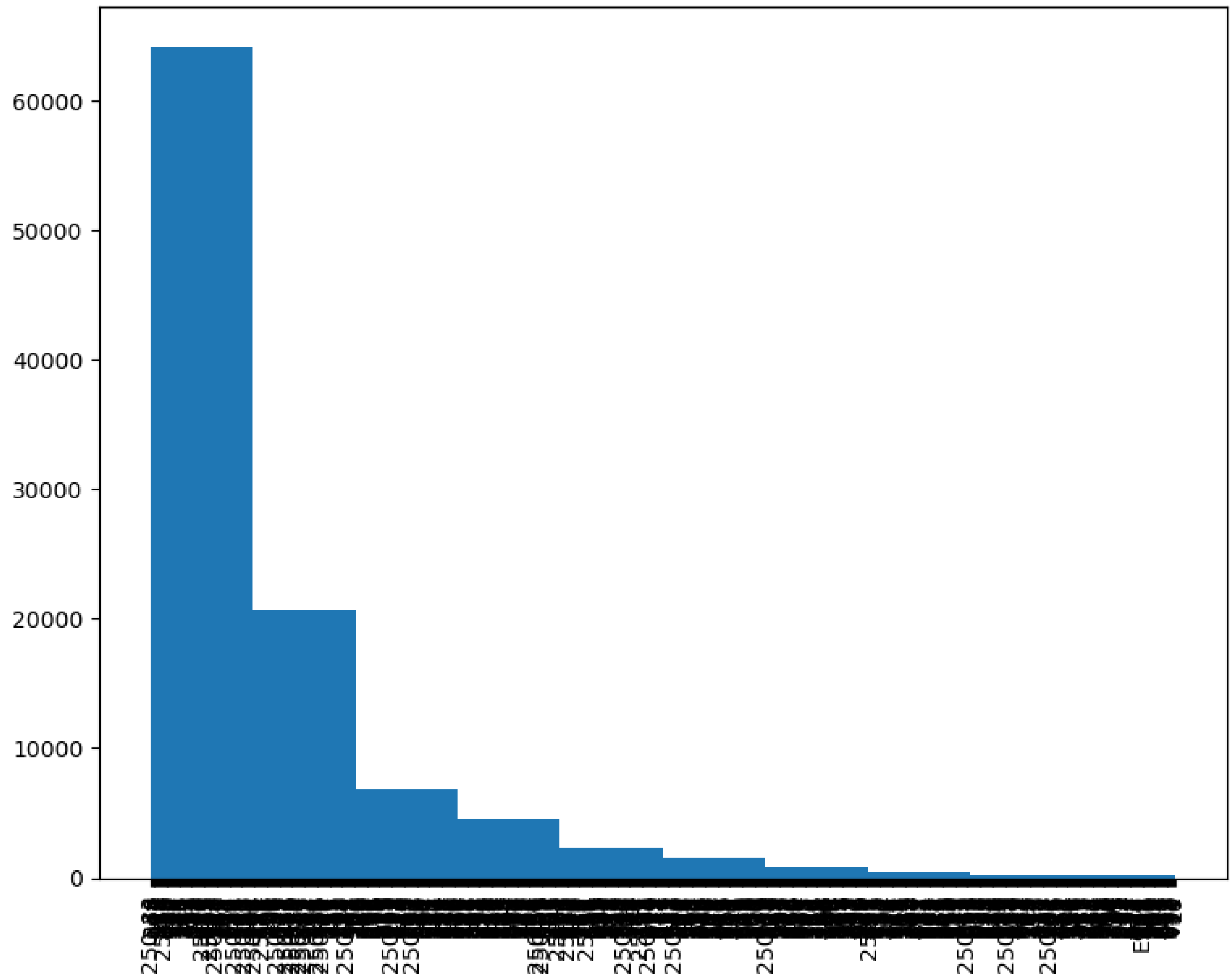

```
#Analyze the distribution of medication changes and total medications
taken
df['change'].value_counts().plot(kind='bar')
plt.show()
```



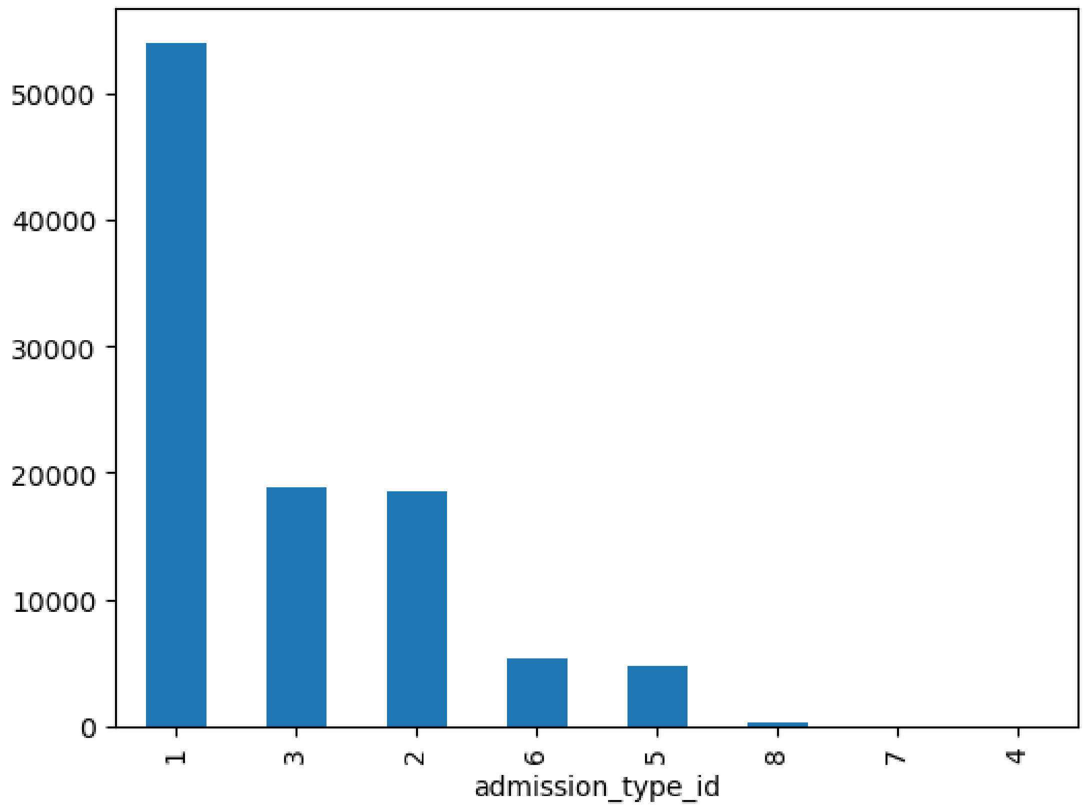

```
#number of medications
sns.displot(df['num_medications'])
plt.show()
```



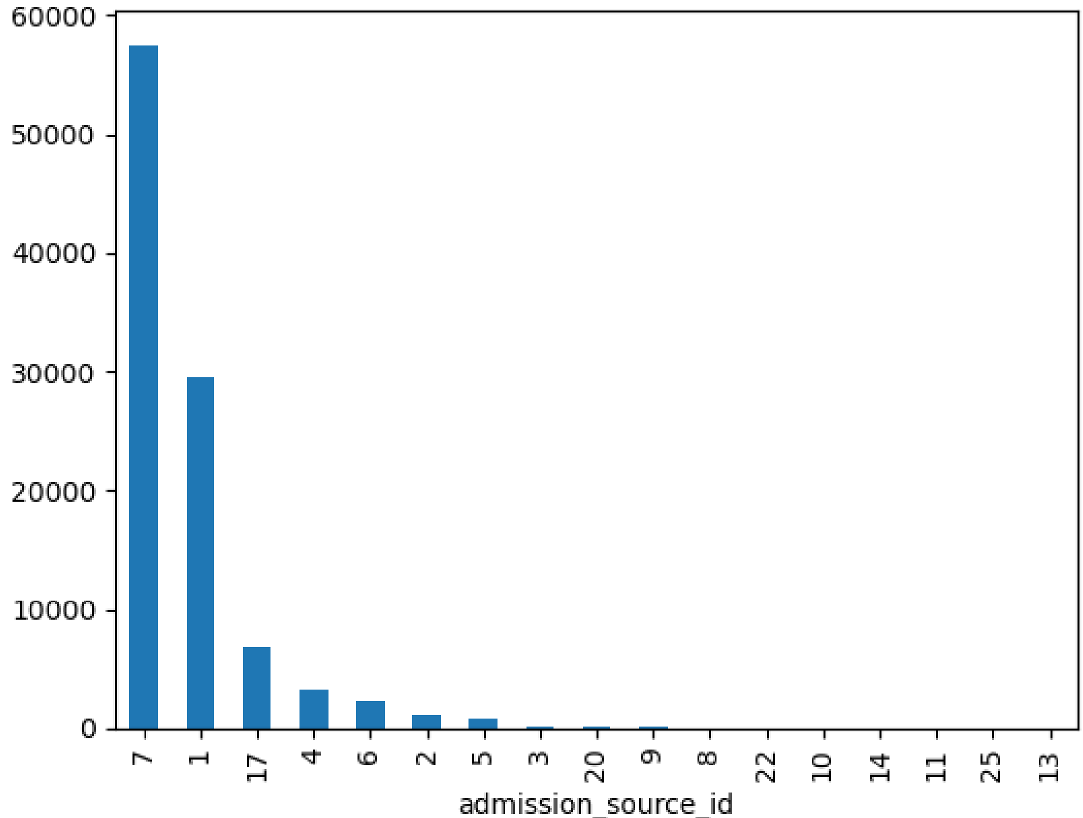

```
#Examine the distribution of diagnoses categories
plt.figure(figsize=(9,7))
plt.hist(df['diag_1'],bins=10)
plt.xticks(rotation=90)
plt.show()
```



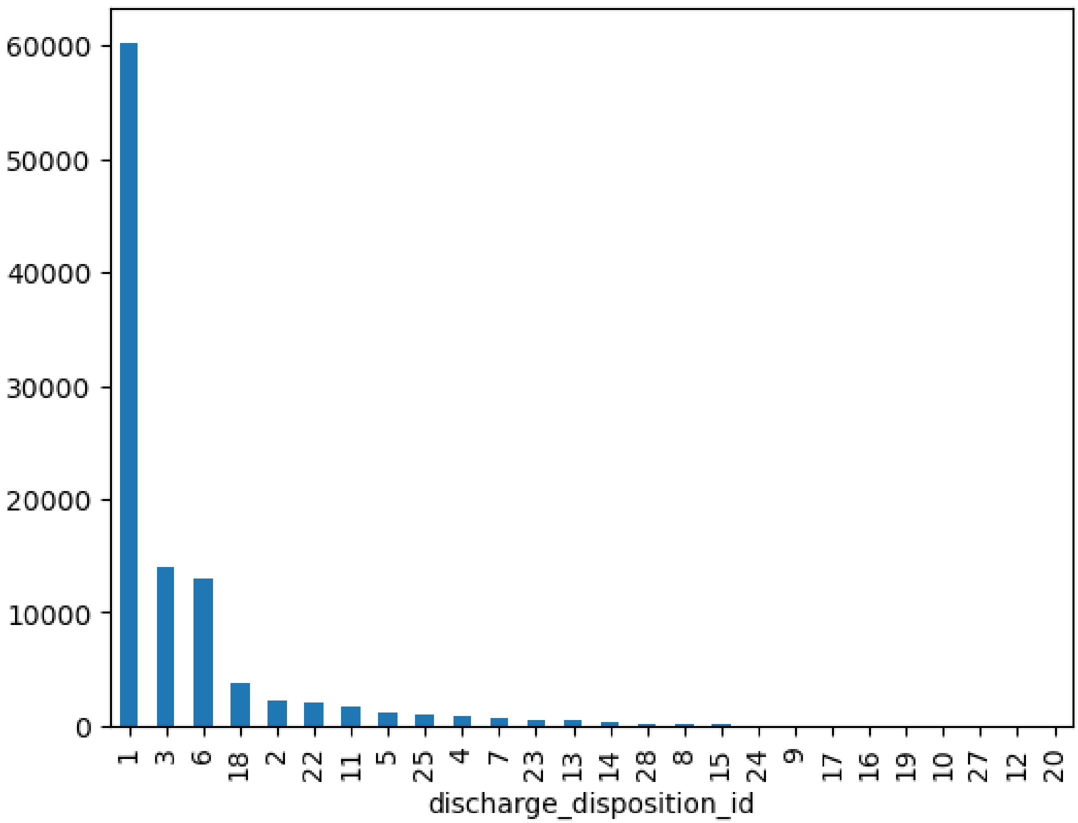

```
#Explore the distribution of patients across admission types, sources, 
and discharge dispositions
df['admission_type_id'].value_counts().plot(kind='bar')
plt.show()
```




```
df['admission_source_id'].value_counts().plot(kind='bar')
plt.show()
```



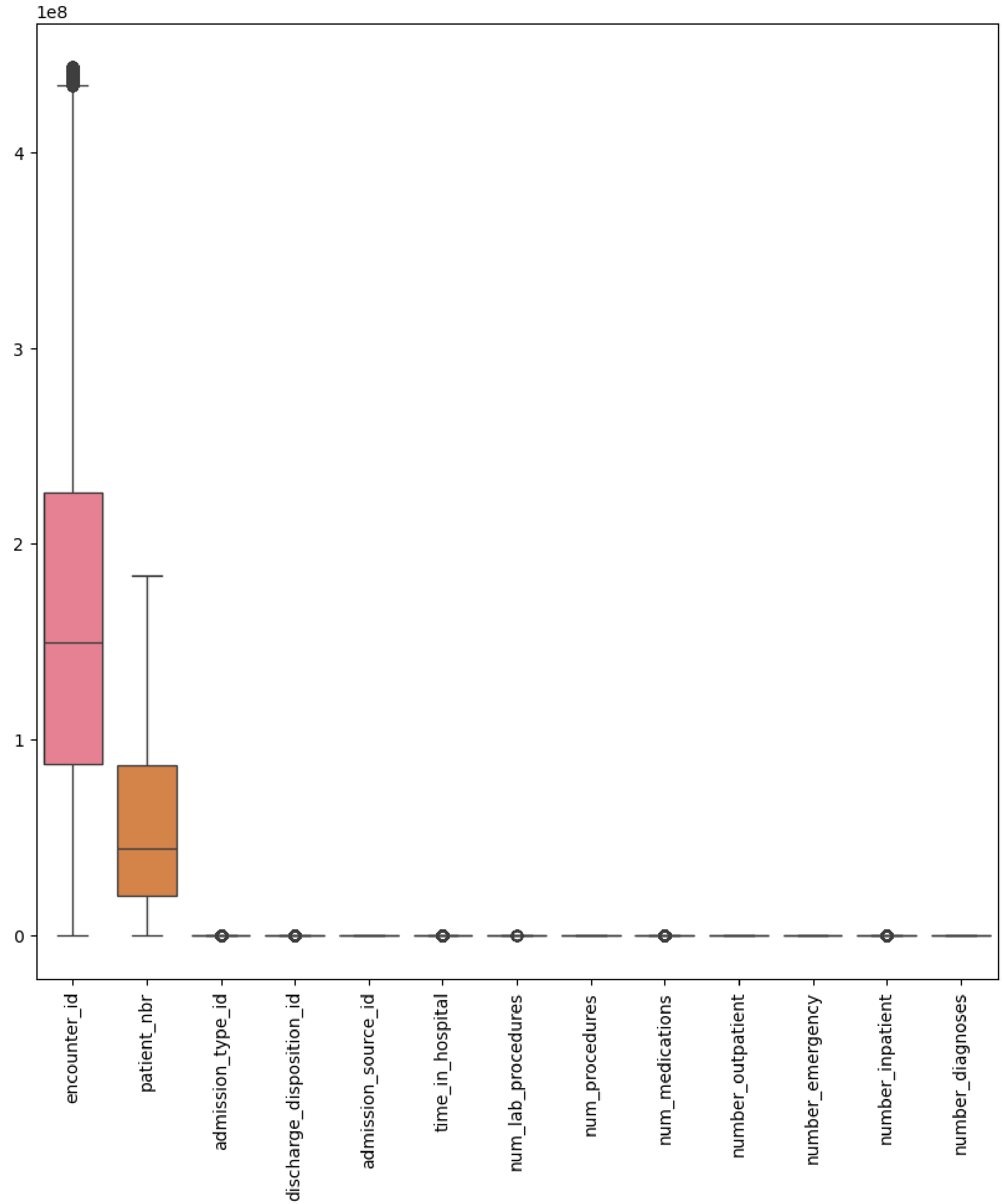

```
df['discharge_disposition_id'].value_counts().plot(kind='bar')
plt.show()
```




```
#removing outliers from numerical  features
for i in numerical_features:
    q1 = df[i].quantile(0.25)
    q3 = df[i].quantile(0.75)
    iqr=q3 - q1
    lower_bound = q1 - 1.5 * iqr
    upper_bound = q3 + 1.5 * iqr
    df = df[(df[i] >=lower_bound) & (df[i] <= upper_bound)]

df.shape
```

```
(56269, 50)
```

```
#boxplots for numerical features
import seaborn as sns
plt.figure(figsize=(10,10))
sns.boxplot(data=df[numerical_features])
plt.xticks(rotation=90)
plt.show()
```




```
# Healthcare EDA Report - Diabetic Data

# 1. Load Libraries
import pandas as pd
```

```python
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import warnings
warnings.filterwarnings('ignore')

# 2. Load the Data
df = pd.read_csv('diabetic_data.csv')
print(df.head())

# 3. Descriptive Statistical Analysis (Numerical Features)
numerical_features = [
    'time_in_hospital', 'num_lab_procedures', 'num_procedures',
    'num_medications', 'number_outpatient', 'number_emergency',
'number_inpatient'
]
print(df[numerical_features].describe())

# 4. Distribution of Categorical Features: Race and Gender
# Race Distribution
plt.figure(figsize=(8,6))
sns.countplot(y='race', data=df,
order=df['race'].value_counts().index)
plt.title('Distribution of Race')
plt.show()

# Gender Distribution
plt.figure(figsize=(6,4))
sns.countplot(x='gender', data=df)
plt.title('Distribution of Gender')
plt.show()

# 5. Relationship between Readmission Status and Age
plt.figure(figsize=(10,6))
sns.countplot(x='age', hue='readmitted', data=df,
order=sorted(df['age'].unique()))
plt.title('Readmission Status across Age Groups')
plt.xticks(rotation=45)
plt.show()

# 6. Correlations between Numerical Features
plt.figure(figsize=(10,8))
corr_matrix = df[numerical_features].corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix - Numerical Features')
plt.show()

# 7. Distribution of Medication Changes and Total Medications
if 'change' in df.columns:
```

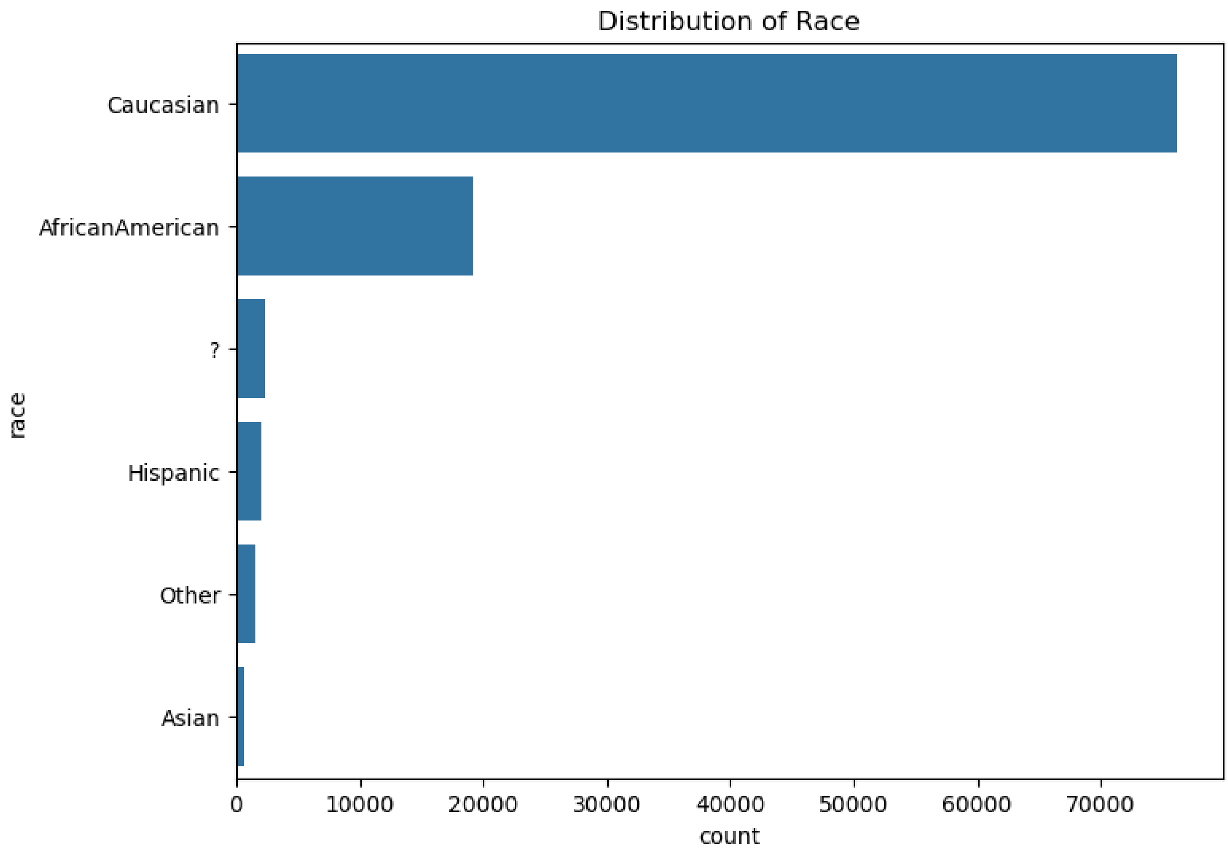
```
    plt.figure(figsize=(6,4))
    sns.countplot(x='change', data=df)
    plt.title('Medication Change Distribution')
    plt.show()

plt.figure(figsize=(8,6))
sns.histplot(df['num_medications'], bins=30, kde=True)
plt.title('Distribution of Number of Medications')
plt.show()

# 8. Distribution of Diagnoses Categories
diag_cols = ['diag_1', 'diag_2', 'diag_3']

for col in diag_cols:
    plt.figure(figsize=(10,5))
    df[col].value_counts().head(10).plot(kind='bar')
    plt.title(f'Top 10 Diagnoses in {col}')
    plt.xlabel('Diagnosis Code')
    plt.ylabel('Count')
    plt.show()

# 9. Patients Across Admission Types, Sources, and Discharge
Dispositions
plt.figure(figsize=(8,5))
sns.countplot(x='admission_type_id', data=df)
plt.title('Distribution of Admission Types')
plt.show()

plt.figure(figsize=(8,5))
sns.countplot(x='admission_source_id', data=df)
plt.title('Distribution of Admission Sources')
plt.show()

plt.figure(figsize=(8,5))
sns.countplot(x='discharge_disposition_id', data=df)
plt.title('Distribution of Discharge Dispositions')
plt.show()

# 10. Outlier Detection in Numerical Features
for feature in numerical_features:
    plt.figure(figsize=(8,4))
    sns.boxplot(x=df[feature])
    plt.title(f'Outliers in {feature}')
    plt.show()

# 11. EDA Analysis Report Summary
print('''
---
# EDA Report - Key Findings:
```

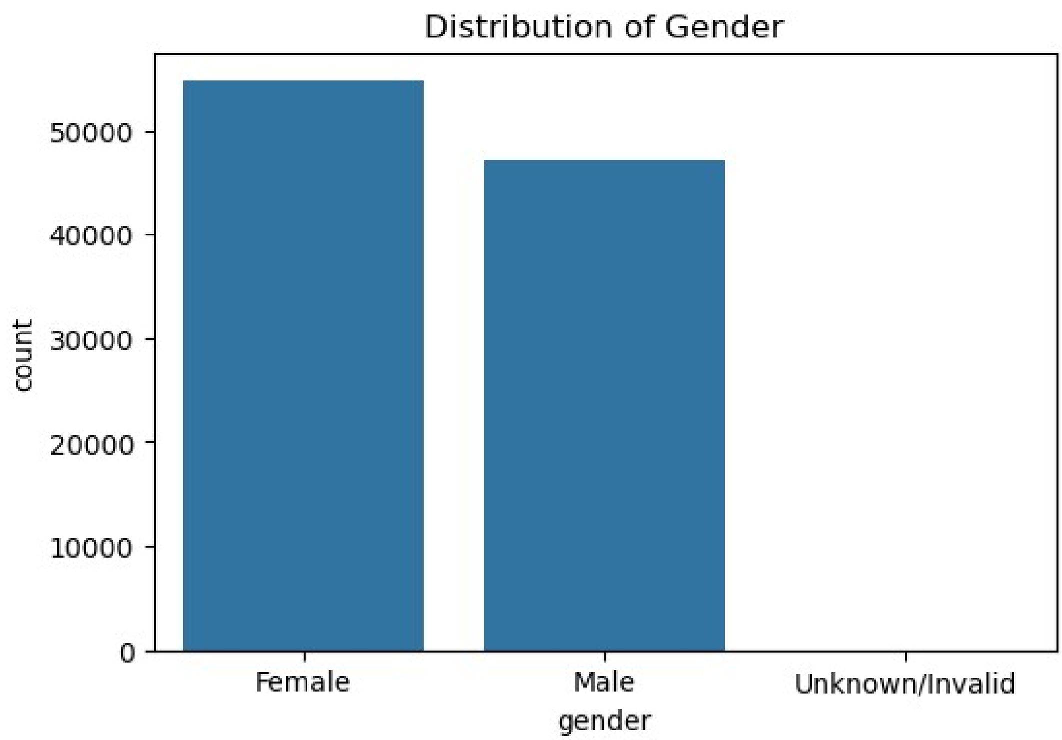
```
- Race: Predominantly Caucasian patients; other races are less
represented.
- Gender: Fairly balanced but slightly more females.
- Readmission vs Age: Higher readmission rates observed among elderly
groups (70+ years).
- Correlations:
    - Number of inpatient visits correlates moderately with time in
hospital.
    - Emergency visits show some correlation with hospitalization.
- Medication Analysis:
    - Most patients had minor or no changes in medications.
    - Medication counts are right-skewed; some patients take a very
high number of meds.
- Diagnoses:
    - Frequent diagnosis codes indicate common chronic diseases like
diabetes complications.
- Admission, Source, Discharge:
    - Most admissions are emergency-based or urgent.
    - Discharge disposition has a variety of categories; important for
detecting suspicious activities.
- Outliers:
    - Several numerical features show extreme values, especially
number of medications and time in hospital.
    - These cases might be flagged for potential fraud investigation.

Note:
- Outliers and abnormal patterns could hint at fraudulent claims.
- Specific focus on discharge types (like discharged to home health)
might be important for fraud detection.
- Patients with frequent readmissions and high medication use should
be analyzed deeper.
---
''')
```

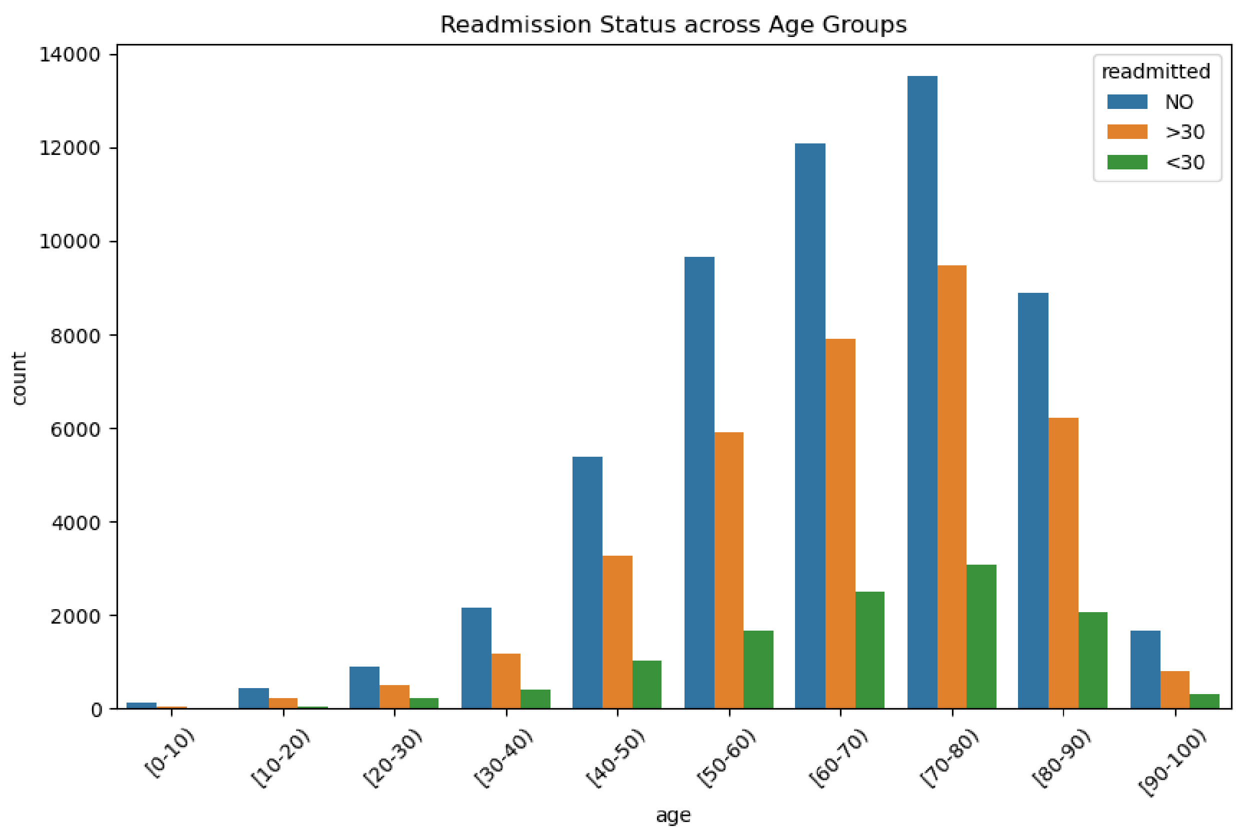
```
   encounter_id  patient_nbr             race  gender      age weight
\
0       2278392      8222157        Caucasian  Female   [0-10)      ?

1        149190     55629189        Caucasian  Female  [10-20)      ?

2         64410     86047875  AfricanAmerican  Female  [20-30)      ?

3        500364     82442376        Caucasian    Male  [30-40)      ?

4         16680     42519267        Caucasian    Male  [40-50)      ?


   admission_type_id  discharge_disposition_id  admission_source_id  \
0                  6                        25                    1
1                  1                         1                    7
```

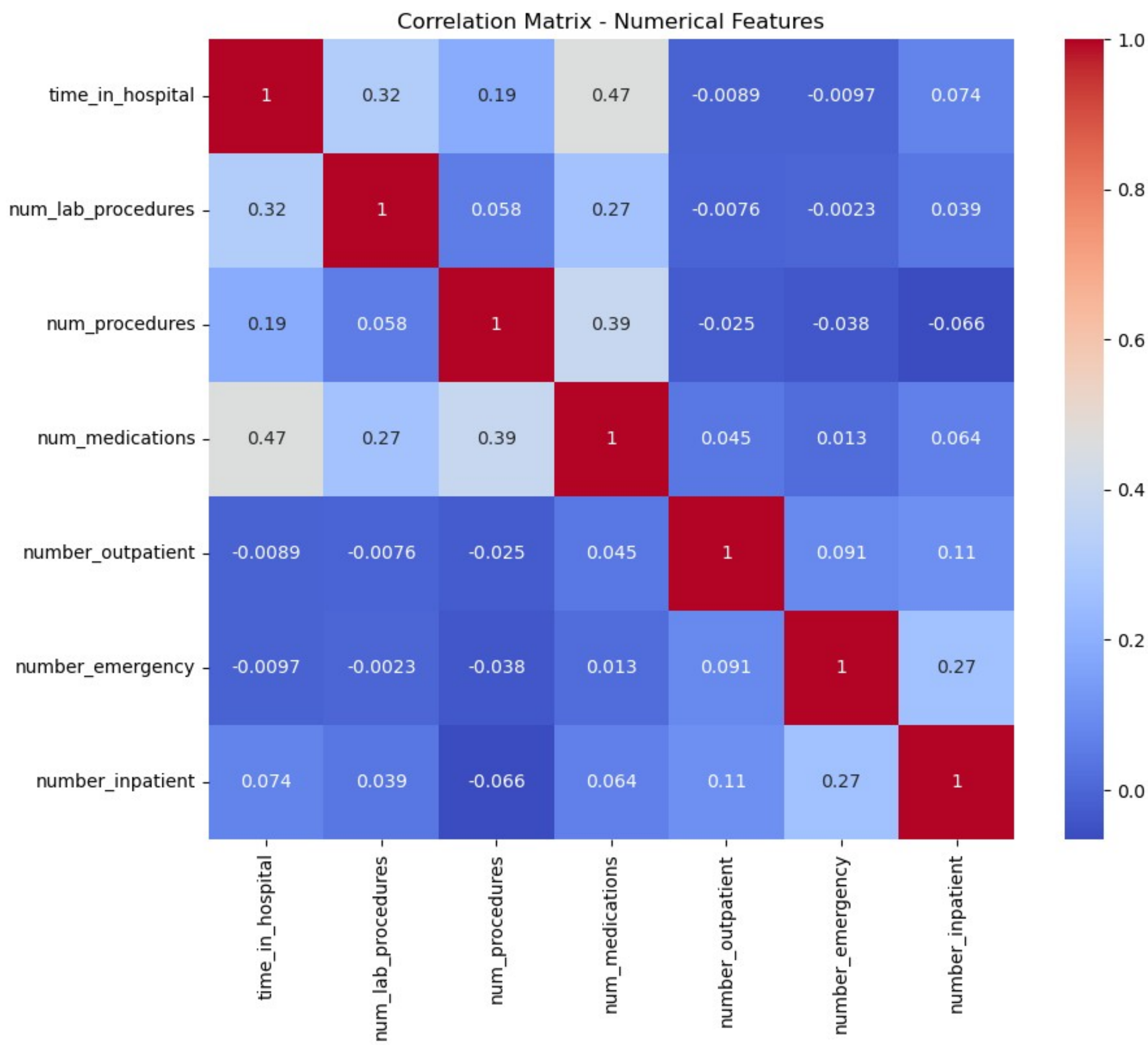
```
2                  1                         1                    7
3                  1                         1                    7
4                  1                         1                    7

   time_in_hospital  ... citoglipton insulin  glyburide-metformin  \
0                 1  ...          No      No                   No
1                 3  ...          No      Up                   No
2                 2  ...          No      No                   No
3                 2  ...          No      Up                   No
4                 1  ...          No  Steady                   No

   glipizide-metformin  glimepiride-pioglitazone  metformin-
rosiglitazone  \
0                   No                        No
No
1                   No                        No
No
2                   No                        No
No
3                   No                        No
No
4                   No                        No
No

   metformin-pioglitazone  change diabetesMed readmitted
0                      No      No          No         NO
1                      No      Ch         Yes        >30
2                      No      No         Yes         NO
3                      No      Ch         Yes         NO
4                      No      Ch         Yes         NO

[5 rows x 50 columns]
       time_in_hospital  num_lab_procedures  num_procedures
num_medications  \
count     101766.000000       101766.000000   101766.000000
101766.000000
mean           4.395987           43.095641        1.339730
16.021844
std            2.985108           19.674362        1.705807
8.127566
min            1.000000            1.000000        0.000000
1.000000
25%            2.000000           31.000000        0.000000
10.000000
50%            4.000000           44.000000        1.000000
15.000000
75%            6.000000           57.000000        2.000000
20.000000
max           14.000000          132.000000        6.000000
81.000000
```

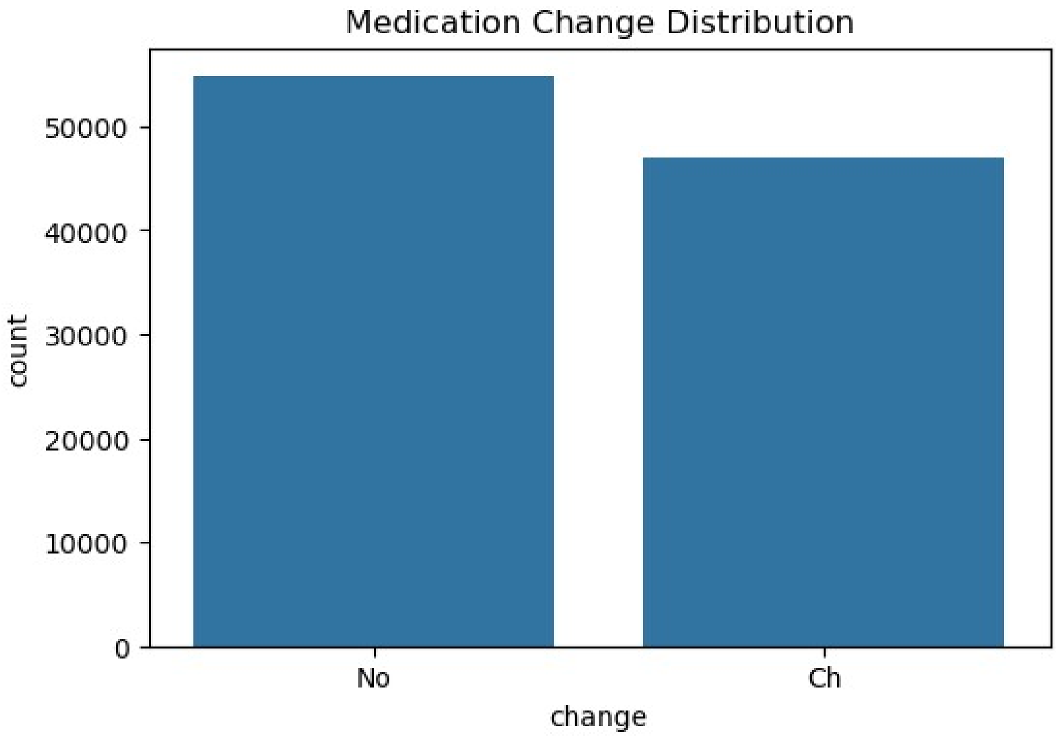
	number_outpatient	number_emergency	number_inpatient
count	101766.000000	101766.000000	101766.000000
mean	0.369357	0.197836	0.635566
std	1.267265	0.930472	1.262863
min	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000
75%	0.000000	0.000000	1.000000
max	42.000000	76.000000	21.000000



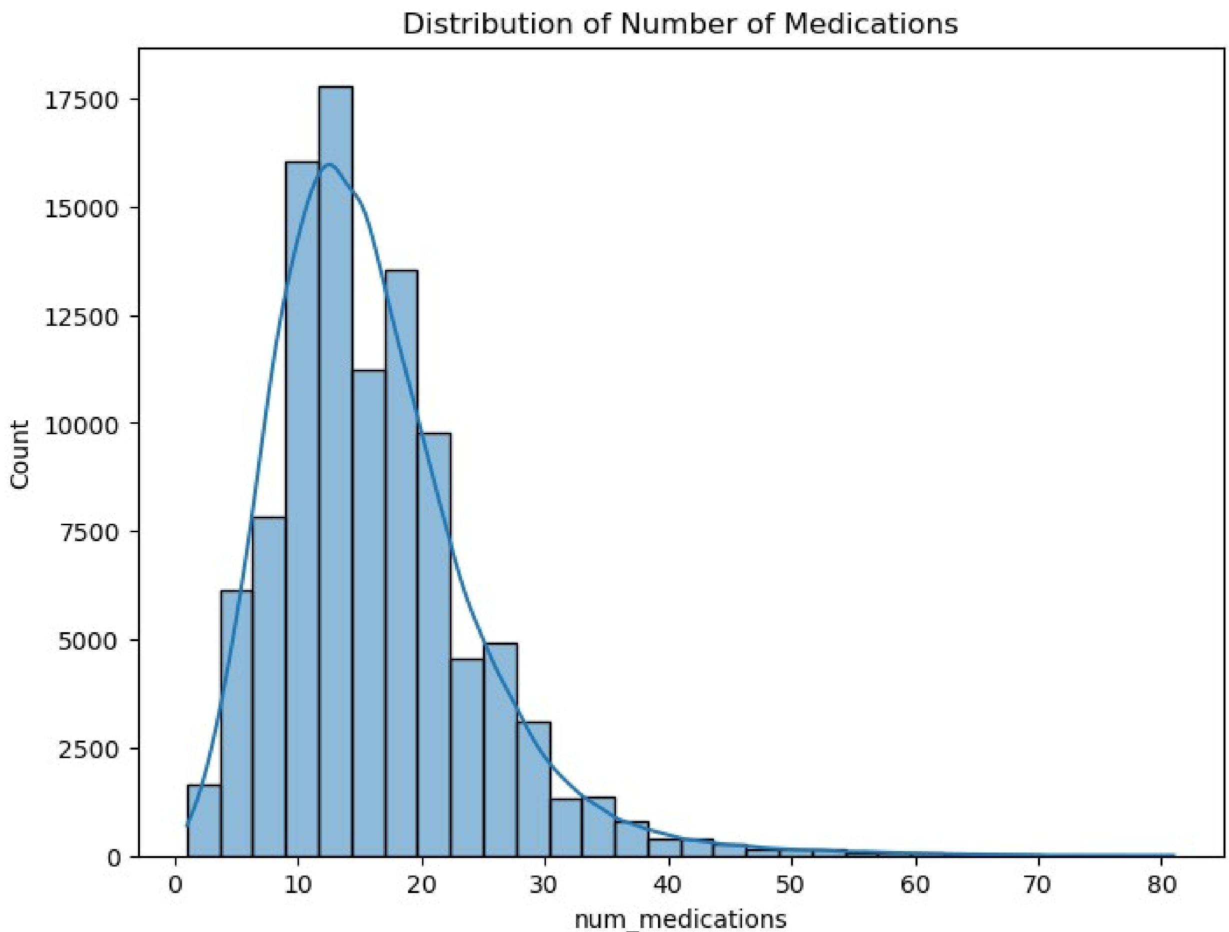

Distribution of Gender
count
50000
40000
30000
20000
10000
0
Female
Male
Unknown/Invalid
gender

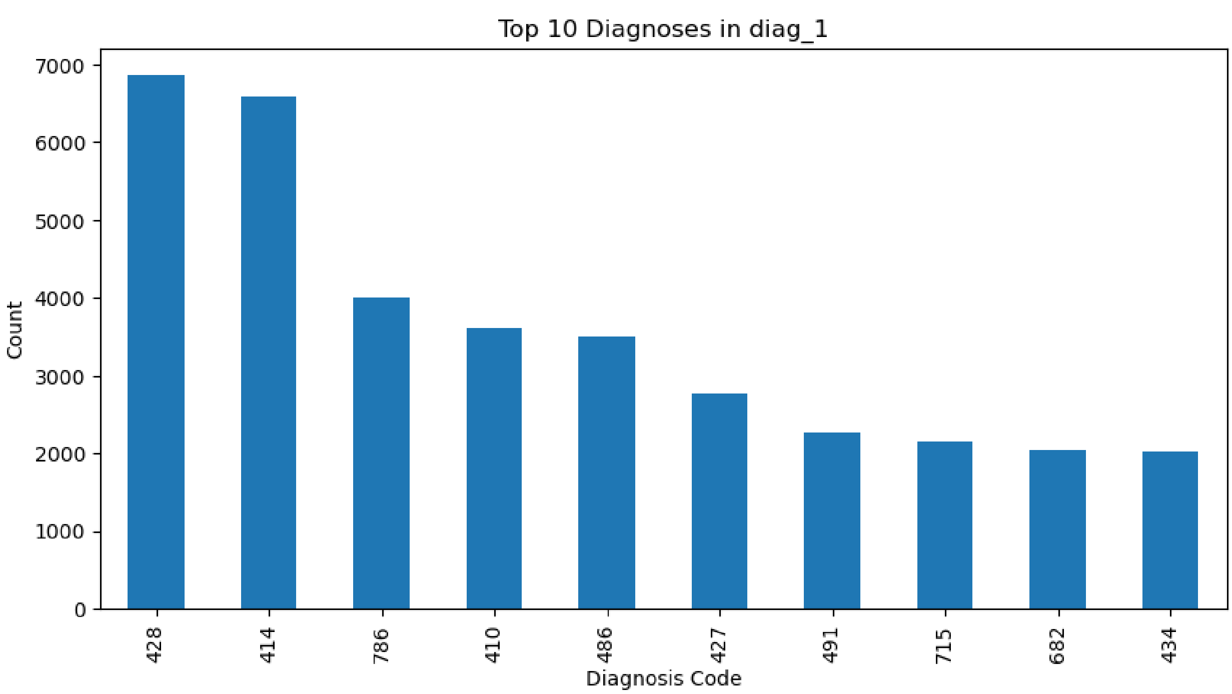

Readmission Status across Age Groups
count
14000
12000
10000
8000
6000
4000
2000
0
readmitted
NO
>30
<30
[0-10)
[10-20)
[20-30)
[30-40)
[40-50)
[50-60)
[60-70)
[70-80)
[80-90)
[90-100)
age

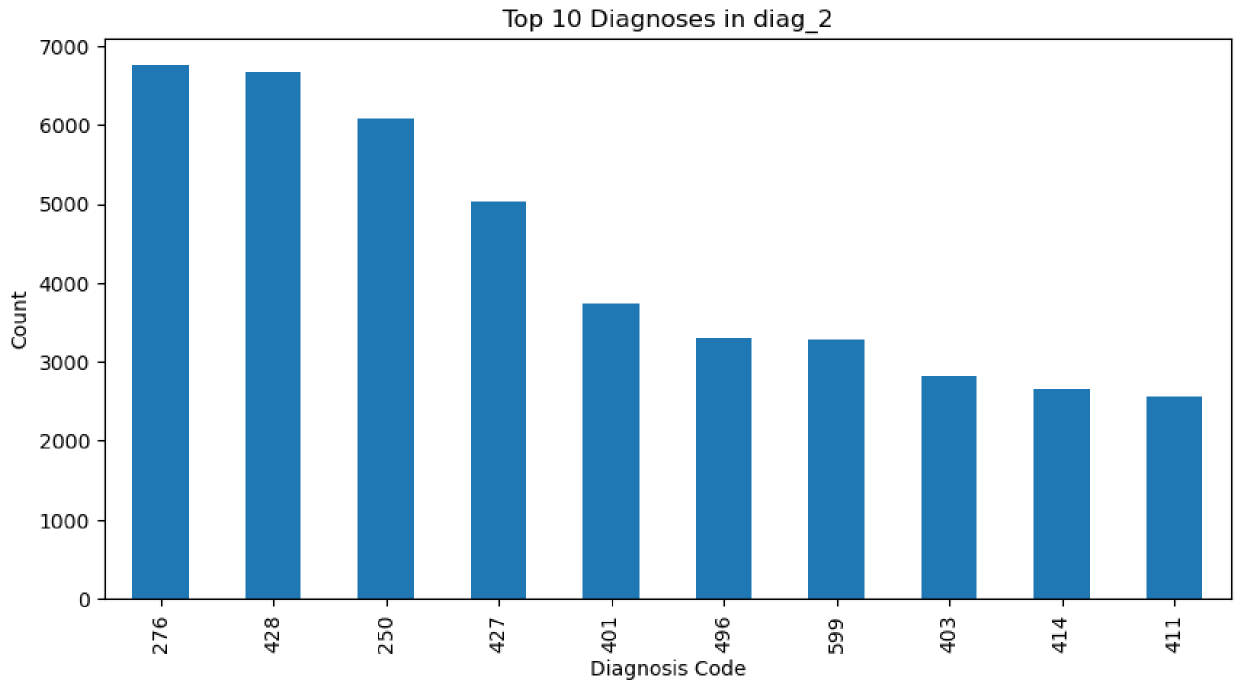

Correlation Matrix - Numerical Features
time_in_hospital
num_lab_procedures
num_procedures
num_medications
number_outpatient
number_emergency
number_inpatient
1 0.32 0.19 0.47 -0.0089 -0.0097 0.074
0.32 1 0.058 0.27 -0.0076 -0.0023 0.039
0.19 0.058 1 0.39 -0.025 -0.038 -0.066
0.47 0.27 0.39 1 0.045 0.013 0.064
-0.0089 -0.0076 -0.025 0.045 1 0.091 0.11
-0.0097 -0.0023 -0.038 0.013 0.091 1 0.27
0.074 0.039 -0.066 0.064 0.11 0.27 1
1.0
0.8
0.6
0.4
0.2
0.0

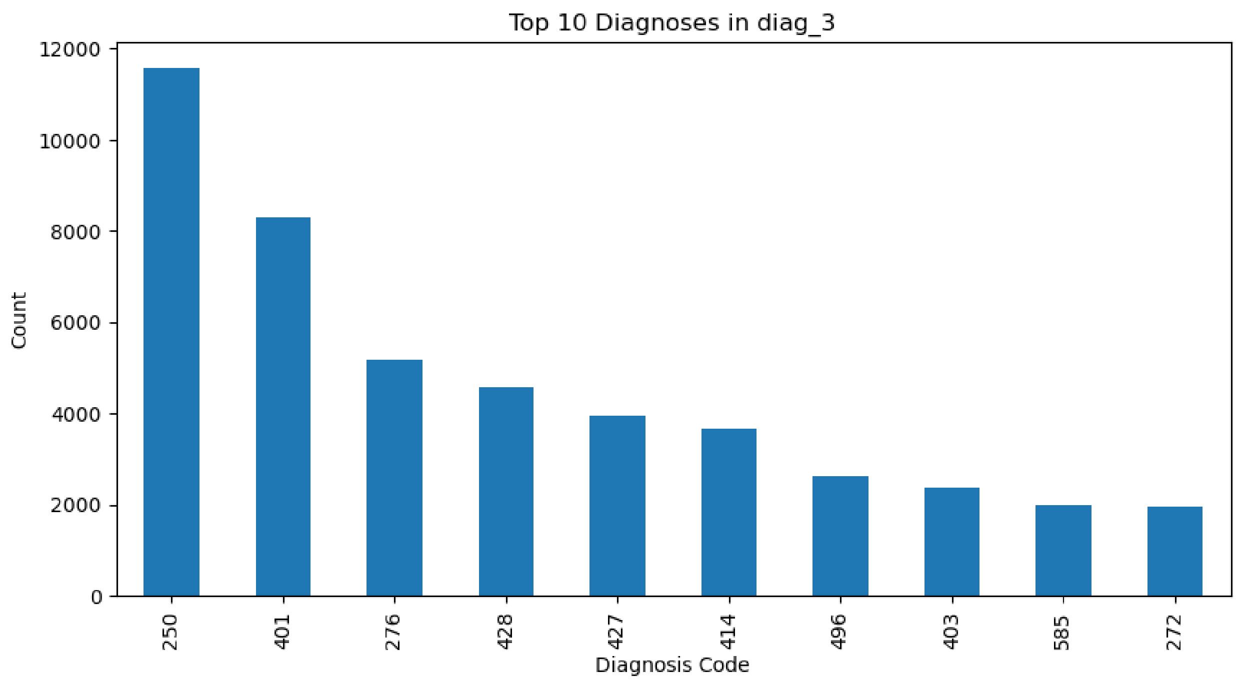


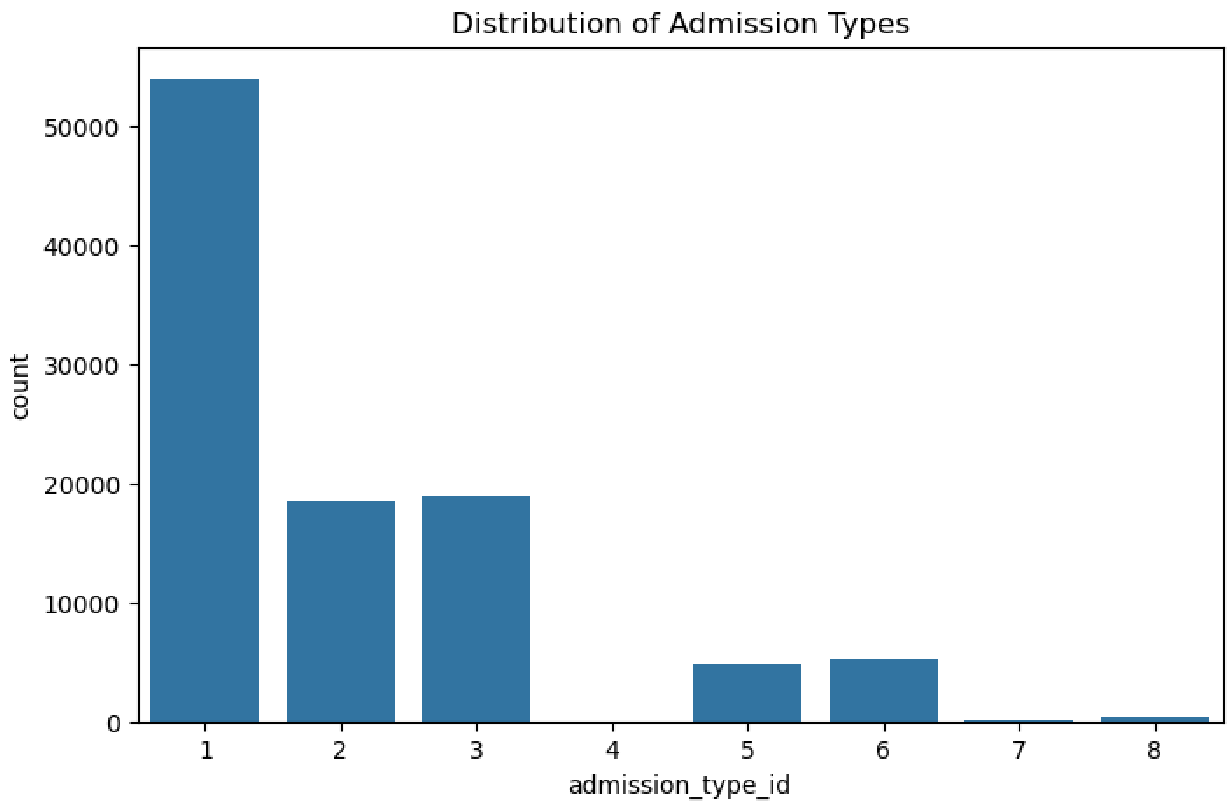
Medication Change Distribution
50000
40000
30000
20000
10000
0
count
No
Ch
change

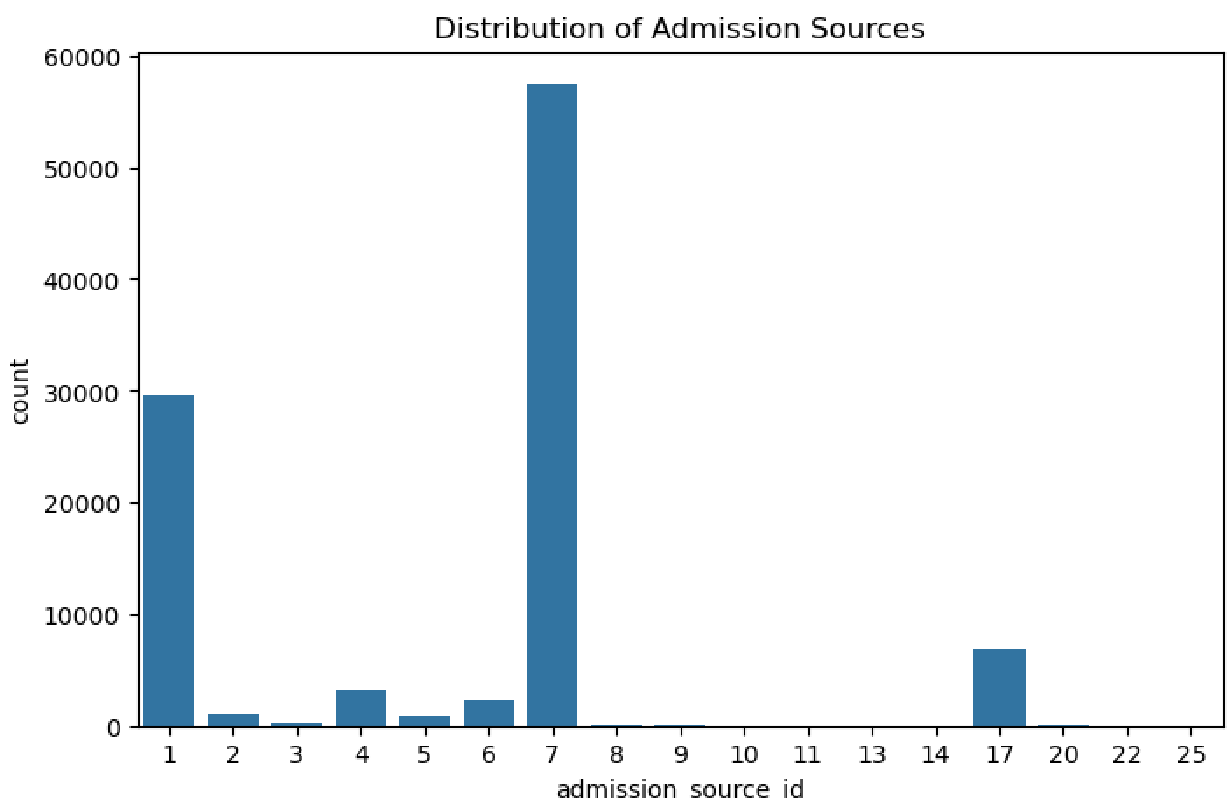

Distribution of Number of Medications
Count
num_medications
0
2500
5000
7500
10000
12500
15000
17500
0
10
20
30
40
50
60
70
80

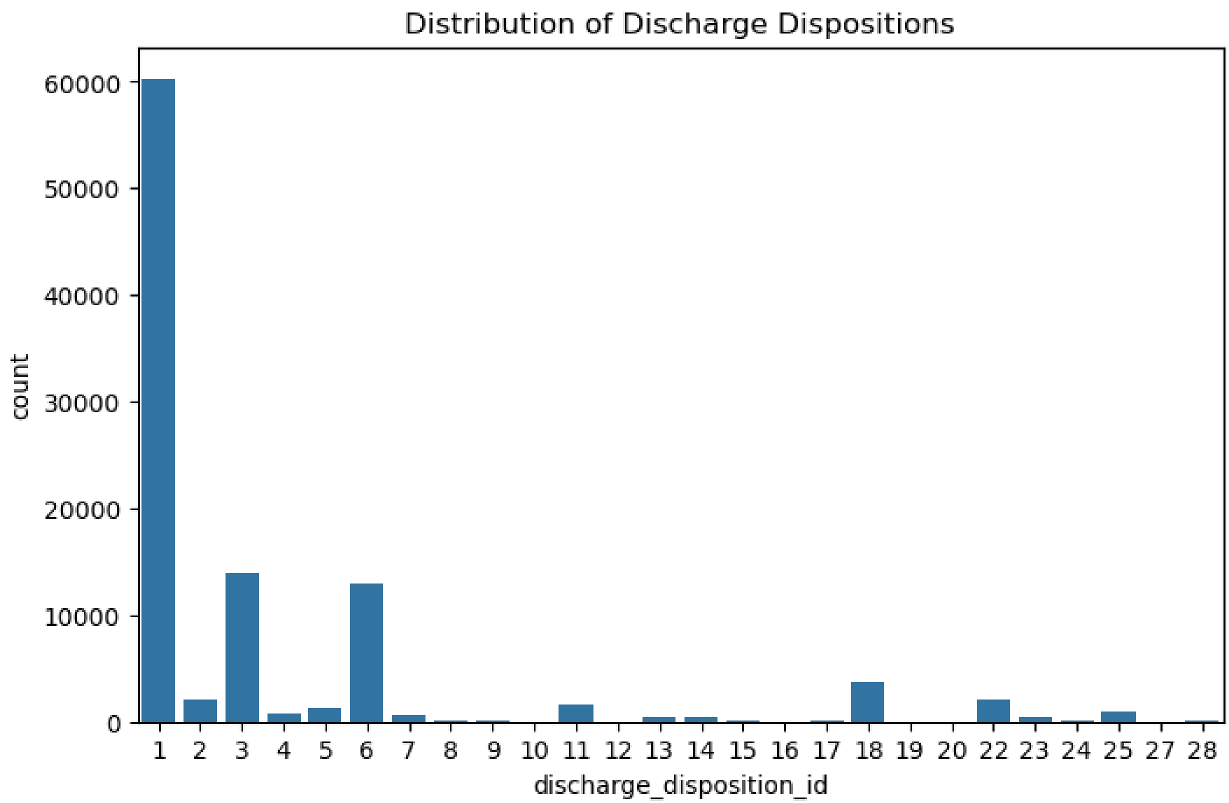

Top 10 Diagnoses in diag_1
Count
Diagnosis Code
0
1000
2000
3000
4000
5000
6000
7000
428
414
786
410
486
427
491
715
682
434


Top 10 Diagnoses in diag_2
Count
Diagnosis Code
0
1000
2000
3000
4000
5000
6000
7000
276
428
250
427
401
496
599
403
414
411

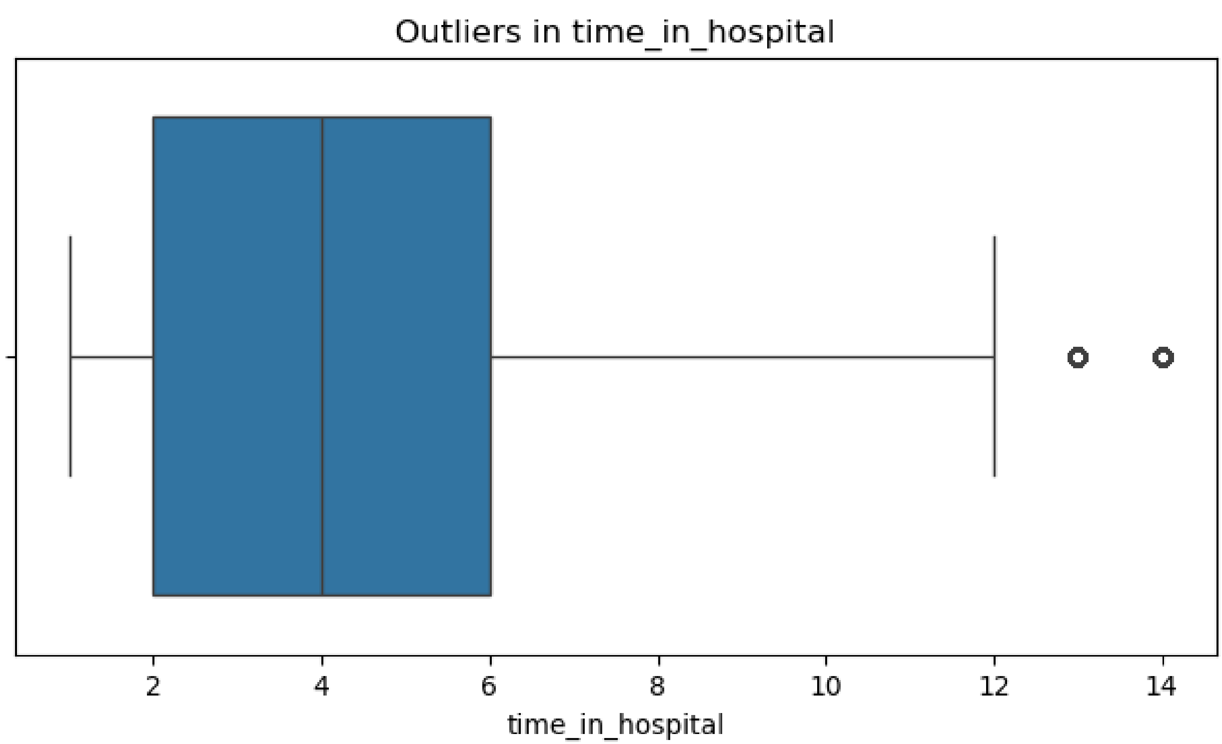

Top 10 Diagnoses in diag_3
Count
Diagnosis Code
0
2000
4000
6000
8000
10000
12000
250
401
276
428
427
414
496
403
585
272


Distribution of Admission Types
count
50000
40000
30000
20000
10000
0
1
2
3
4
5
6
7
8
admission_type_id

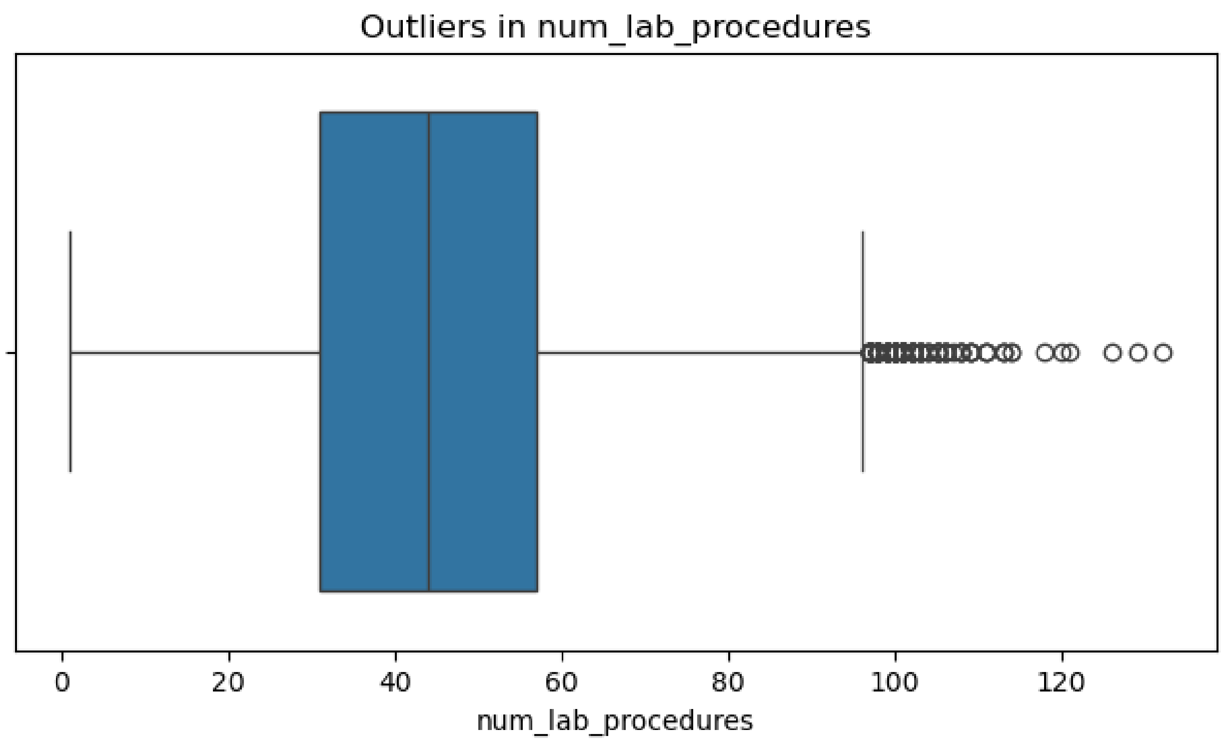

Distribution of Admission Sources
count
60000
50000
40000
30000
20000
10000
0
1
2
3
4
5
6
7
8
9
10
11
13
14
17
20
22
25
admission_source_id

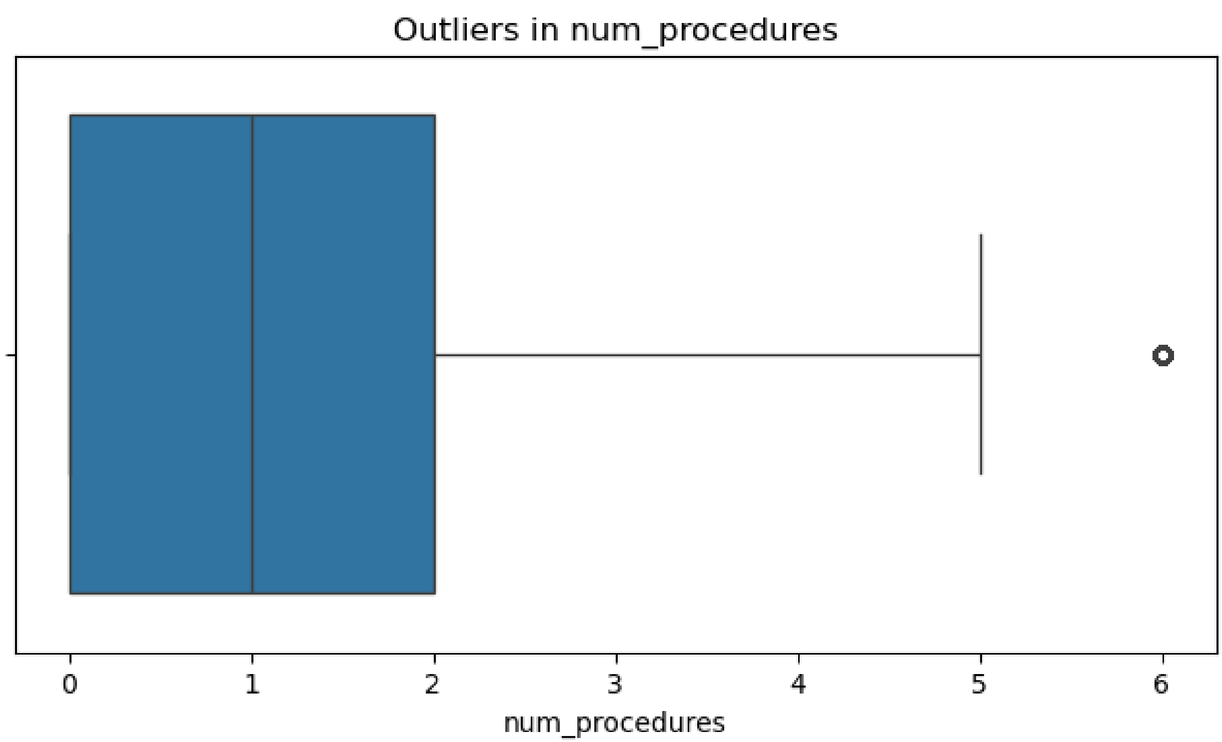


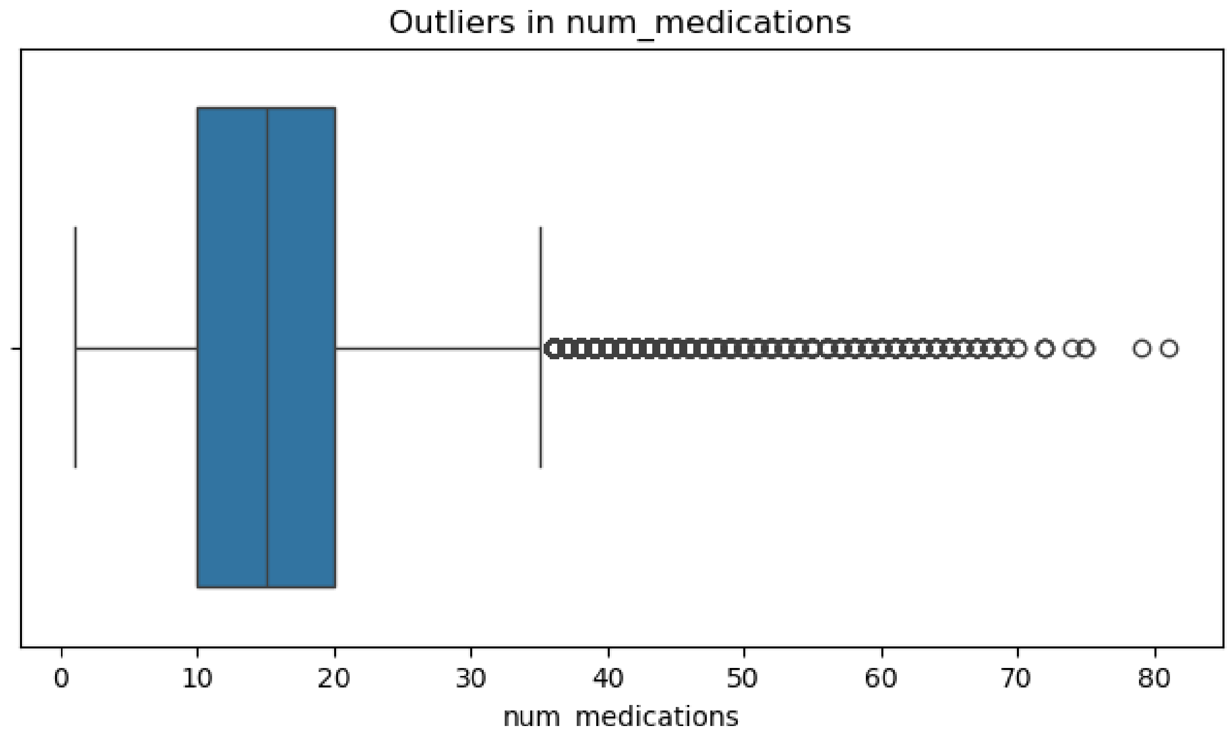
Distribution of Discharge Dispositions
60000
50000
40000
30000
20000
10000
0
count
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 22 23 24 25 27 28
discharge_disposition_id



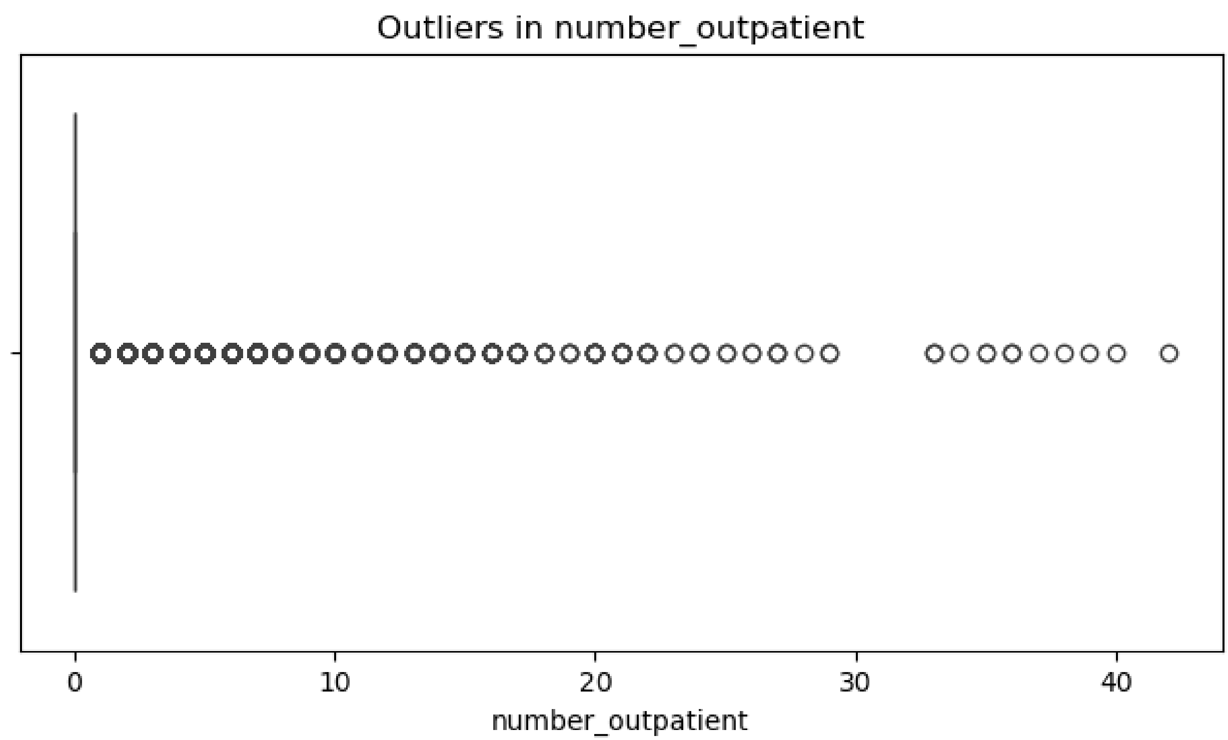
Outliers in time_in_hospital
2 4 6 8 10 12 14
time_in_hospital


Outliers in num_lab_procedures
0
20
40
60
80
100
120
num_lab_procedures

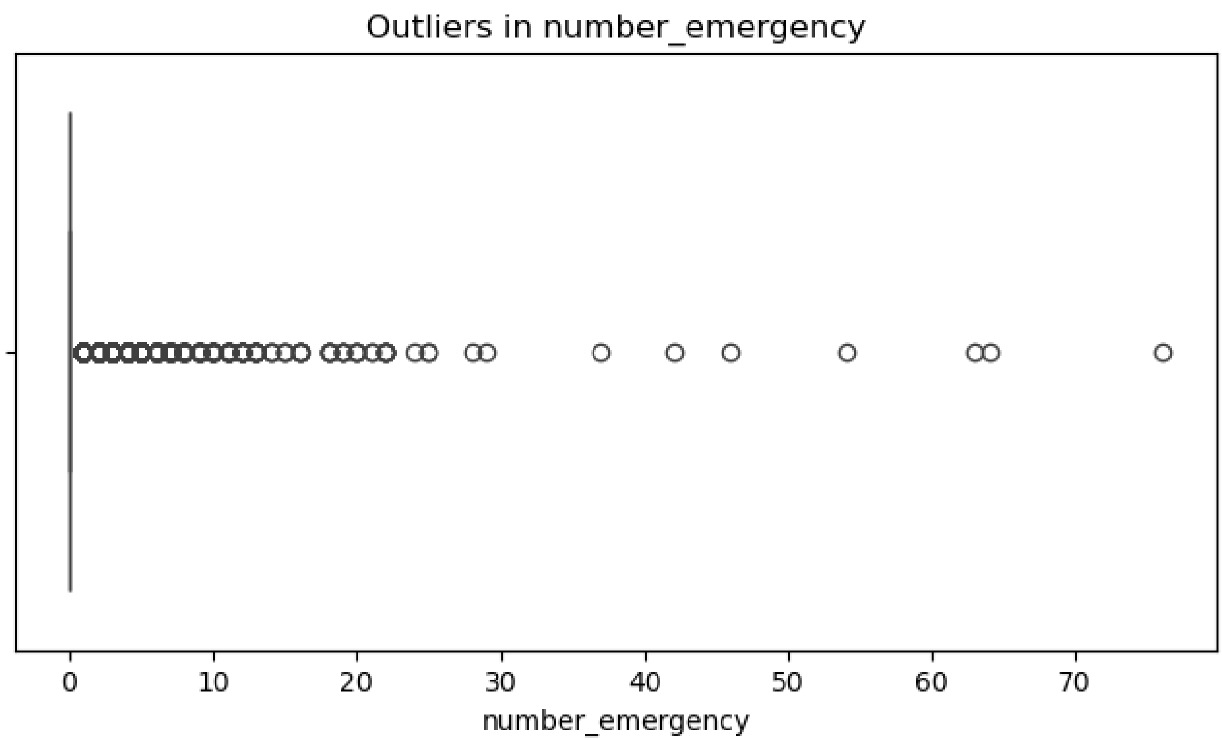

Outliers in num_procedures
0
1
2
3
4
5
6
num_procedures

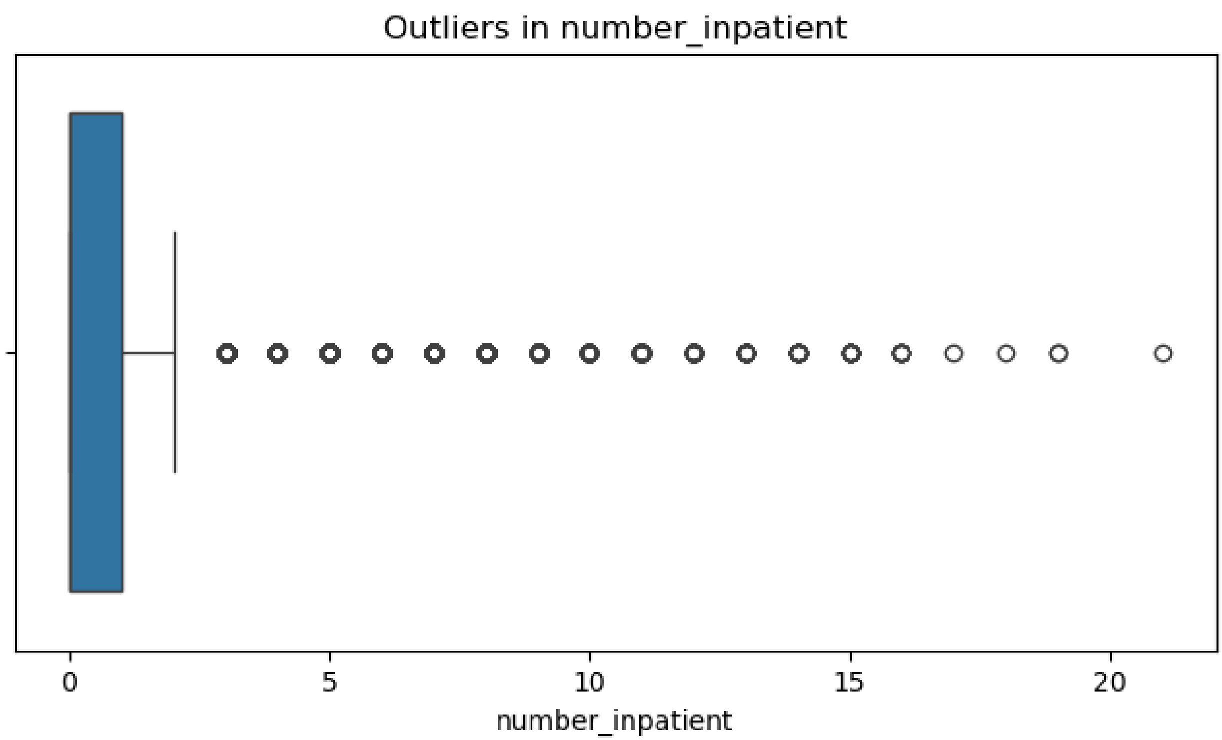


Outliers in num_medications
0
10
20
30
40
50
60
70
80
num_medications



Outliers in number_outpatient
0
10
20
30
40
number_outpatient






```
---
# EDA Report - Key Findings:
```

```
- Race: Predominantly Caucasian patients; other races are less
represented.
- Gender: Fairly balanced but slightly more females.
- Readmission vs Age: Higher readmission rates observed among elderly
groups (70+ years).
- Correlations:
    - Number of inpatient visits correlates moderately with time in
hospital.
    - Emergency visits show some correlation with hospitalization.
- Medication Analysis:
    - Most patients had minor or no changes in medications.
    - Medication counts are right-skewed; some patients take a very
high number of meds.
- Diagnoses:
    - Frequent diagnosis codes indicate common chronic diseases like
diabetes complications.
- Admission, Source, Discharge:
    - Most admissions are emergency-based or urgent.
    - Discharge disposition has a variety of categories; important for
detecting suspicious activities.
- Outliers:
    - Several numerical features show extreme values, especially
number of medications and time in hospital.
    - These cases might be flagged for potential fraud investigation.

Note:
- Outliers and abnormal patterns could hint at fraudulent claims.
- Specific focus on discharge types (like discharged to home health)
might be important for fraud detection.
- Patients with frequent readmissions and high medication use should
be analyzed deeper.
---
```