In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
db = pd.read_csv("diabetic_data.csv")
print(db.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]
```

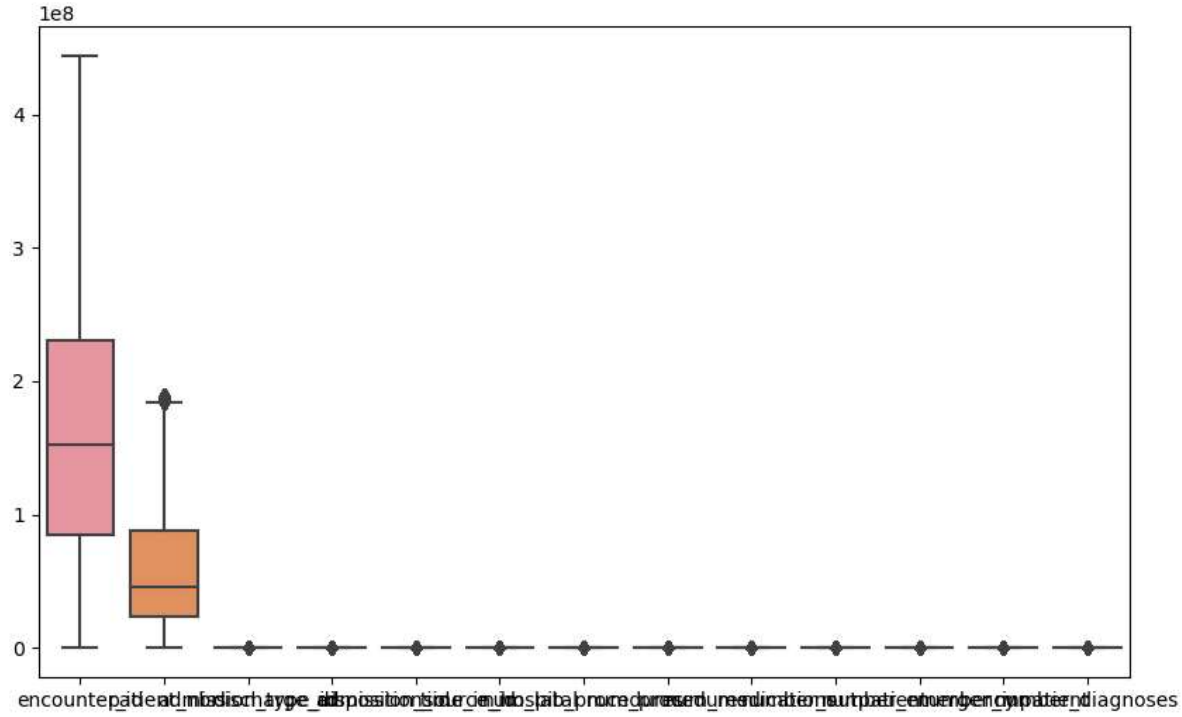
In [2]:
```
print(db.info)
print(db.describe)
import seaborn as sns
db.hist(bins=50,figsize=(20,15))
plt.show()
sns.pairplot(db)
plt.show()
```

```
<bound method DataFrame.info of         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)      ?
...              ...          ...              ...     ...      ...    ...
101761     443847548    100162476  AfricanAmerican    Male  [70-80)      ?
101762     443847782     74694222  AfricanAmerican  Female  [80-90)      ?
101763     443854148     41088789        Caucasian    Male  [70-80)      ?
101764     443857166     31693671        Caucasian  Female  [80-90)      ?
101765     443867222    175429310        Caucasian    Male  [70-80)      ?

        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
```

In [3]:
```
print(db.isnull())
db.fillna(db.mean(),inplace=True)
print(db.isnull().sum())
```

```
        encounter_id  patient_nbr   race  gender    age  weight  \
0              False        False  False   False  False   False
1              False        False  False   False  False   False
2              False        False  False   False  False   False
3              False        False  False   False  False   False
4              False        False  False   False  False   False
...              ...          ...    ...     ...    ...     ...
101761         False        False  False   False  False   False
101762         False        False  False   False  False   False
101763         False        False  False   False  False   False
101764         False        False  False   False  False   False
101765         False        False  False   False  False   False

        admission_type_id  discharge_disposition_id  admission_source_id  \
0                   False                     False                False
1                   False                     False                False
2                   False                     False                False
3                   False                     False                False
4                   False                     False                False
```

In [4]:
```
import numpy as np
plt.figure(figsize=(10,6))
sns.boxplot(data=db)
plt.show()
from scipy import stats
diabetes_df=db[(np.abs(stats.zscore(db))<3).all(axis=1)]
```



```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
Cell In[4], line 6
      4 plt.show()
      5 from scipy import stats
----> 6 diabetes_df=db[(np.abs(stats.zscore(db))<3).all(axis=1)]

File ~\anaconda3\lib\site-packages\scipy\stats\_stats_py.py:2713, in zscore
(a, axis, ddof, nan_policy)
   2644 def zscore(a, axis=0, ddof=0, nan_policy='propagate'):
   2645     """
   2646     Compute the z score.
   2647
   (...)
   2711            [-0.91611681, -0.89090508,  1.4983032 ,  0.88731639, -0.57
85977 ]])
   2712     """
-> 2713     return zmap(a, a, axis=axis, ddof=ddof, nan_policy=nan_policy)

File ~\anaconda3\lib\site-packages\scipy\stats\_stats_py.py:2872, in zmap(sco
res, compare, axis, ddof, nan_policy)
   2870         isconst = np.apply_along_axis(_isconst, axis, a)
   2871 else:
-> 2872     mn = a.mean(axis=axis, keepdims=True)
   2873     std = a.std(axis=axis, ddof=ddof, keepdims=True)
   2874     if axis is None:

File ~\anaconda3\lib\site-packages\numpy\core\_methods.py:182, in _mean(a, ax
is, dtype, out, keepdims, where)
    180 ret = umr_sum(arr, axis, dtype, out, keepdims, where=where)
    181 if isinstance(ret, mu.ndarray):
--> 182     ret = um.true_divide(
    183             ret, rcount, out=ret, casting='unsafe', subok=False)
    184     if is_float16_result and out is None:
    185         ret = arr.dtype.type(ret)

TypeError: unsupported operand type(s) for /: 'str' and 'int'
```

In [ ]: