SPRINT 2

Date	07-11-2022
Team ID	PNT2022TMID04531
Project Title	Analytics for Hospitals' Health-Care Data
Team Members	Haris S, Gokula Kannan G, Abishek A, hari vignesh K G

Data Cleaning and Preparation

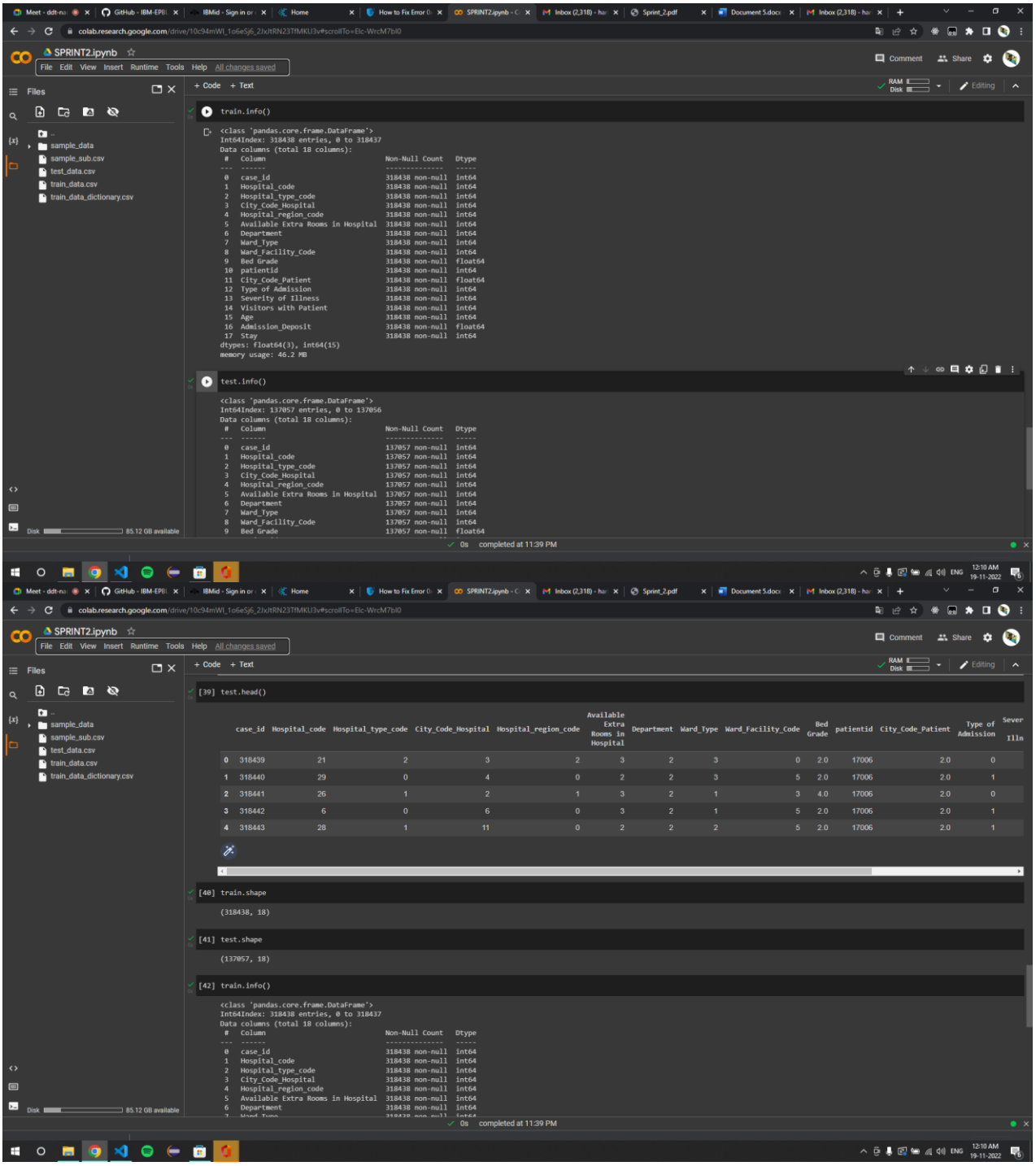
In this data set, variables “City_code_patient” and “Bed Grade” have missing values. These missing values must be treated before feeding to the algorithm as they distort the model performance.

So, the missing values are replaced using the “mode” of the column.

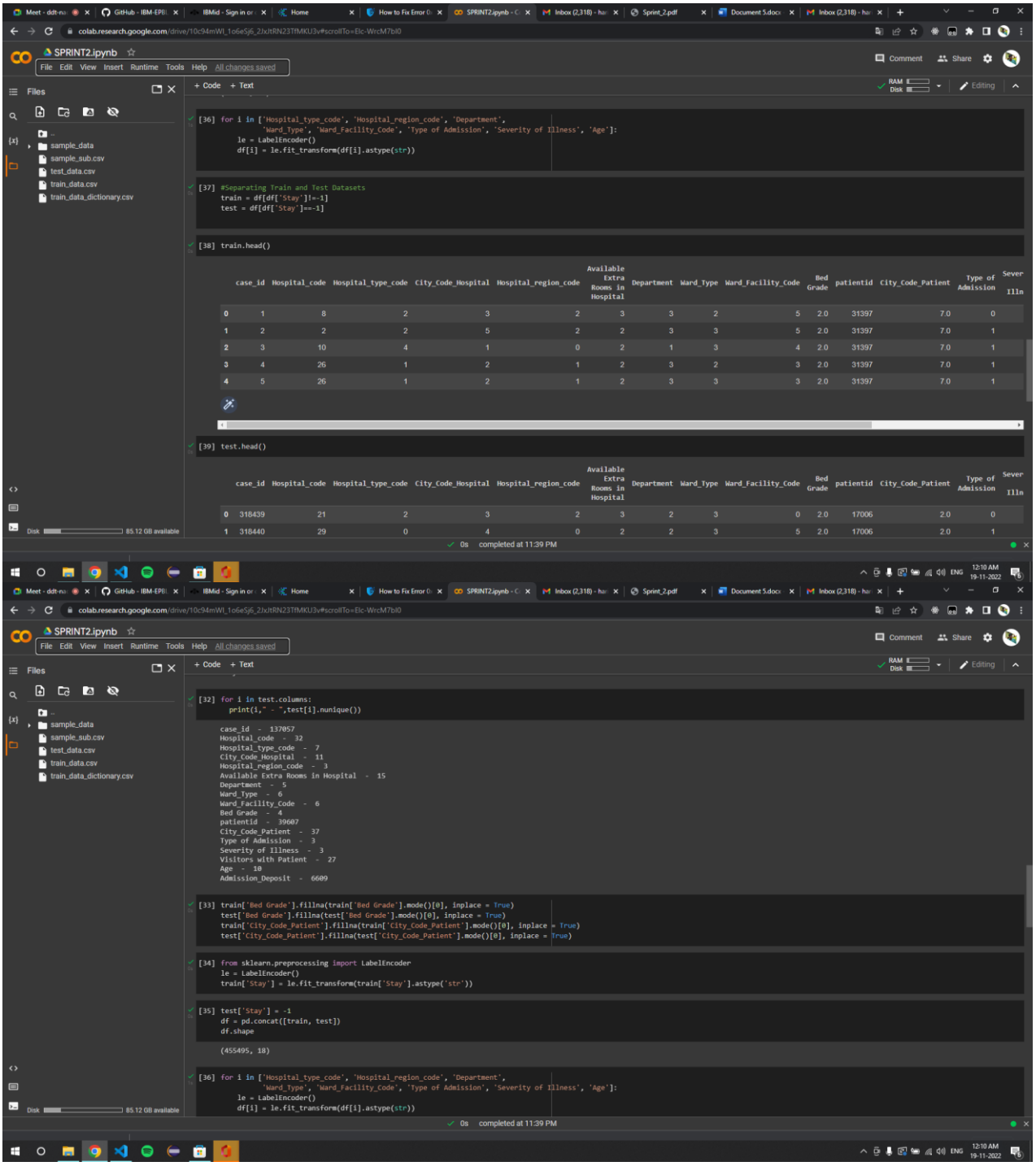
Since most of the variables in the dataset have ordinal data, we transformed them into levels by using a label encoder to perform further analysis on the data.

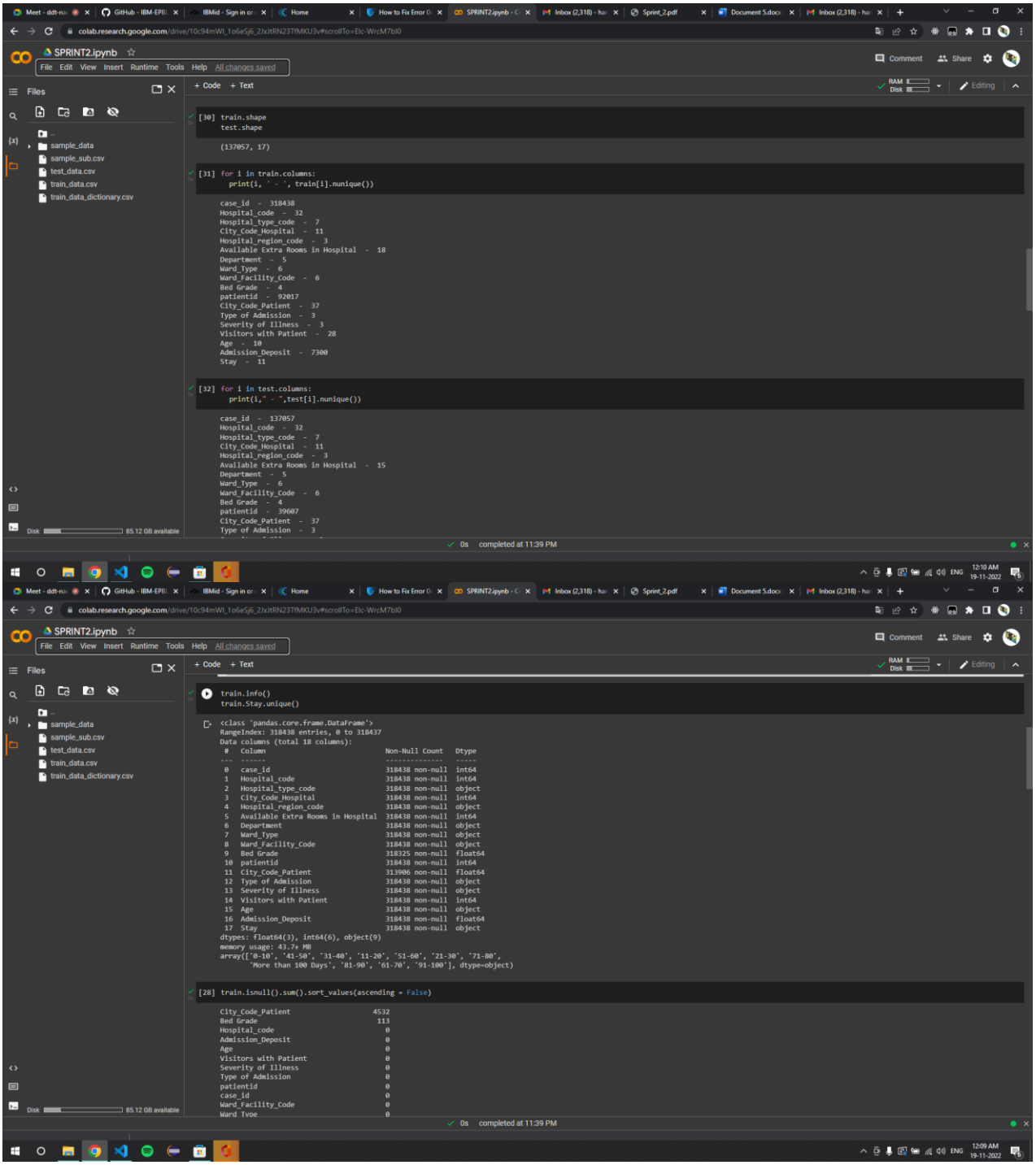
Distinct Observations of Ordinal Data

Variables	Number of distinct observations
Hospital_type_code	7
Hospital_region_code	3
Department	5
Ward_Type	6
Ward_Facility_Code	6
Type of Admission	3
Severity of Illness	3
Age	10
Stay	11


SPRINT2.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Files
+ Code + Text
RAM
Disk
Editing
sample_data
sample_sub.csv
test_data.csv
train_data.csv
train_data_dictionary.csv
train.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 318438 entries, 0 to 318437
Data columns (total 18 columns):
Column Non-Null Count Dtype
0 case_id 318438 non-null int64
1 Hospital_code 318438 non-null int64
2 Hospital_type_code 318438 non-null int64
3 City_Code_Hospital 318438 non-null int64
4 Hospital_region_code 318438 non-null int64
5 Available Extra Rooms in Hospital 318438 non-null int64
6 Department 318438 non-null int64
7 Ward_Type 318438 non-null int64
8 Ward_Facility_Code 318438 non-null int64
9 Bed Grade 318438 non-null float64
10 patientid 318438 non-null int64
11 City_Code_Patient 318438 non-null float64
12 Type of Admission 318438 non-null int64
13 Severity of Illness 318438 non-null int64
14 Visitors with Patient 318438 non-null int64
15 Age 318438 non-null int64
16 Admission_Deposit 318438 non-null float64
17 Stay 318438 non-null int64
dtypes: float64(3), int64(15)
memory usage: 46.2 MB
test.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 137057 entries, 0 to 137056
Data columns (total 18 columns):
Column Non-Null Count Dtype
0 case_id 137057 non-null int64
1 Hospital_code 137057 non-null int64
2 Hospital_type_code 137057 non-null int64
3 City_Code_Hospital 137057 non-null int64
4 Hospital_region_code 137057 non-null int64
5 Available Extra Rooms in Hospital 137057 non-null int64
6 Department 137057 non-null int64
7 Ward_Type 137057 non-null int64
8 Ward_Facility_Code 137057 non-null int64
9 Bed Grade 137057 non-null float64
Disk 85.12 GB available
0s completed at 11:39 PM

SPRINT2.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
Files
+ Code + Text
RAM
Disk
Editing
sample_data
sample_sub.csv
test_data.csv
train_data.csv
train_data_dictionary.csv
[39] test.head()
case_id Hospital_code Hospital_type_code City_Code_Hospital Hospital_region_code Available Extra Rooms in Hospital Department Ward_Type Ward_Facility_Code Bed Grade patientid City_Code_Patient Type of Admission Sever Illn
0 318439 21 2 3 2 3 2 3 0 2.0 17006 2.0 0
1 318440 29 0 4 0 2 2 3 5 2.0 17006 2.0 1
2 318441 26 1 2 1 3 2 1 3 4.0 17006 2.0 0
3 318442 6 0 6 0 3 2 1 5 2.0 17006 2.0 1
4 318443 28 1 11 0 2 2 2 5 2.0 17006 2.0 1
[40] train.shape
(318438, 18)
[41] test.shape
(137057, 18)
[42] train.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 318438 entries, 0 to 318437
Data columns (total 18 columns):
Column Non-Null Count Dtype
0 case_id 318438 non-null int64
1 Hospital_code 318438 non-null int64
2 Hospital_type_code 318438 non-null int64
3 City_Code_Hospital 318438 non-null int64
4 Hospital_region_code 318438 non-null int64
5 Available Extra Rooms in Hospital 318438 non-null int64
6 Department 318438 non-null int64
Disk 85.12 GB available
0s completed at 11:39 PM


SPRINT2.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
Files
sample_data
sample_sub.csv
test_data.csv
train_data.csv
train_data_dictionary.csv
Disk 85.12 GB available
[36] for i in ['Hospital_type_code', 'Hospital_region_code', 'Department',
'Ward_Type', 'Ward_Facility_Code', 'Type of Admission', 'Severity of Illness', 'Age']:
le = LabelEncoder()
df[i] = le.fit_transform(df[i].astype(str))
[37] #Separating Train and Test Datasets
train = df[df['Stay']!=-1]
test = df[df['Stay']==-1]
[38] train.head()
case_id Hospital_code Hospital_type_code City_Code_Hospital Hospital_region_code Available Extra Rooms in Hospital Department Ward_Type Ward_Facility_Code Bed Grade patientid City_Code_Patient Type of Admission Sever Illn
0 1 8 2 3 2 3 2 5 2.0 31397 7.0 0
1 2 2 2 5 2 2 3 3 5 2.0 31397 7.0 1
2 3 10 4 1 0 2 1 3 4 2.0 31397 7.0 1
3 4 26 1 2 1 2 3 2 3 2.0 31397 7.0 1
4 5 26 1 2 1 2 3 3 3 2.0 31397 7.0 1
[39] test.head()
case_id Hospital_code Hospital_type_code City_Code_Hospital Hospital_region_code Available Extra Rooms in Hospital Department Ward_Type Ward_Facility_Code Bed Grade patientid City_Code_Patient Type of Admission Sever Illn
0 318439 21 2 3 2 3 2 3 0 2.0 17006 2.0 0
1 318440 29 0 4 0 2 2 3 5 2.0 17006 2.0 1
0s completed at 11:39 PM
12:10 AM 19-11-2022
[32] for i in test.columns:
print(i," - ",test[i].nunique())
case_id - 137057
Hospital_code - 32
Hospital_type_code - 7
City_Code_Hospital - 11
Hospital_region_code - 3
Available Extra Rooms in Hospital - 15
Department - 5
Ward_Type - 6
Ward_Facility_Code - 6
Bed Grade - 4
patientid - 39607
City_Code_Patient - 37
Type of Admission - 3
Severity of Illness - 3
Visitors with Patient - 27
Age - 10
Admission_Deposit - 6609
[33] train['Bed Grade'].fillna(train['Bed Grade'].mode()[0], inplace = True)
test['Bed Grade'].fillna(test['Bed Grade'].mode()[0], inplace = True)
train['City_Code_Patient'].fillna(train['City_Code_Patient'].mode()[0], inplace = True)
test['City_Code_Patient'].fillna(test['City_Code_Patient'].mode()[0], inplace = True)
[34] from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
train['Stay'] = le.fit_transform(train['Stay'].astype('str'))
[35] test['Stay'] = -1
df = pd.concat([train, test])
df.shape
(455495, 18)
[36] for i in ['Hospital_type_code', 'Hospital_region_code', 'Department',
'Ward_Type', 'Ward_Facility_Code', 'Type of Admission', 'Severity of Illness', 'Age']:
le = LabelEncoder()
df[i] = le.fit_transform(df[i].astype(str))
0s completed at 11:39 PM
12:10 AM 19-11-2022



```
[30] train.shape
     test.shape

     (137057, 17)

[31] for i in train.columns:
       print(i, ' - ', train[i].nunique())

     case_id  -  318438
     Hospital_code  -  32
     Hospital_type_code  -  7
     City_Code_Hospital  -  11
     Hospital_region_code  -  3
     Available Extra Rooms in Hospital  -  18
     Department  -  5
     Ward_Type  -  6
     Ward_Facility_Code  -  6
     Bed Grade  -  4
     patientid  -  92017
     City_Code_Patient  -  37
     Type of Admission  -  3
     Severity of Illness  -  3
     Visitors with Patient  -  28
     Age  -  10
     Admission_Deposit  -  7300
     Stay  -  11

[32] for i in test.columns:
       print(i," - ",test[i].nunique())

     case_id  -  137057
     Hospital_code  -  32
     Hospital_type_code  -  7
     City_Code_Hospital  -  11
     Hospital_region_code  -  3
     Available Extra Rooms in Hospital  -  15
     Department  -  5
     Ward_Type  -  6
     Ward_Facility_Code  -  6
     Bed Grade  -  4
     patientid  -  39607
     City_Code_Patient  -  37
     Type of Admission  -  3
```

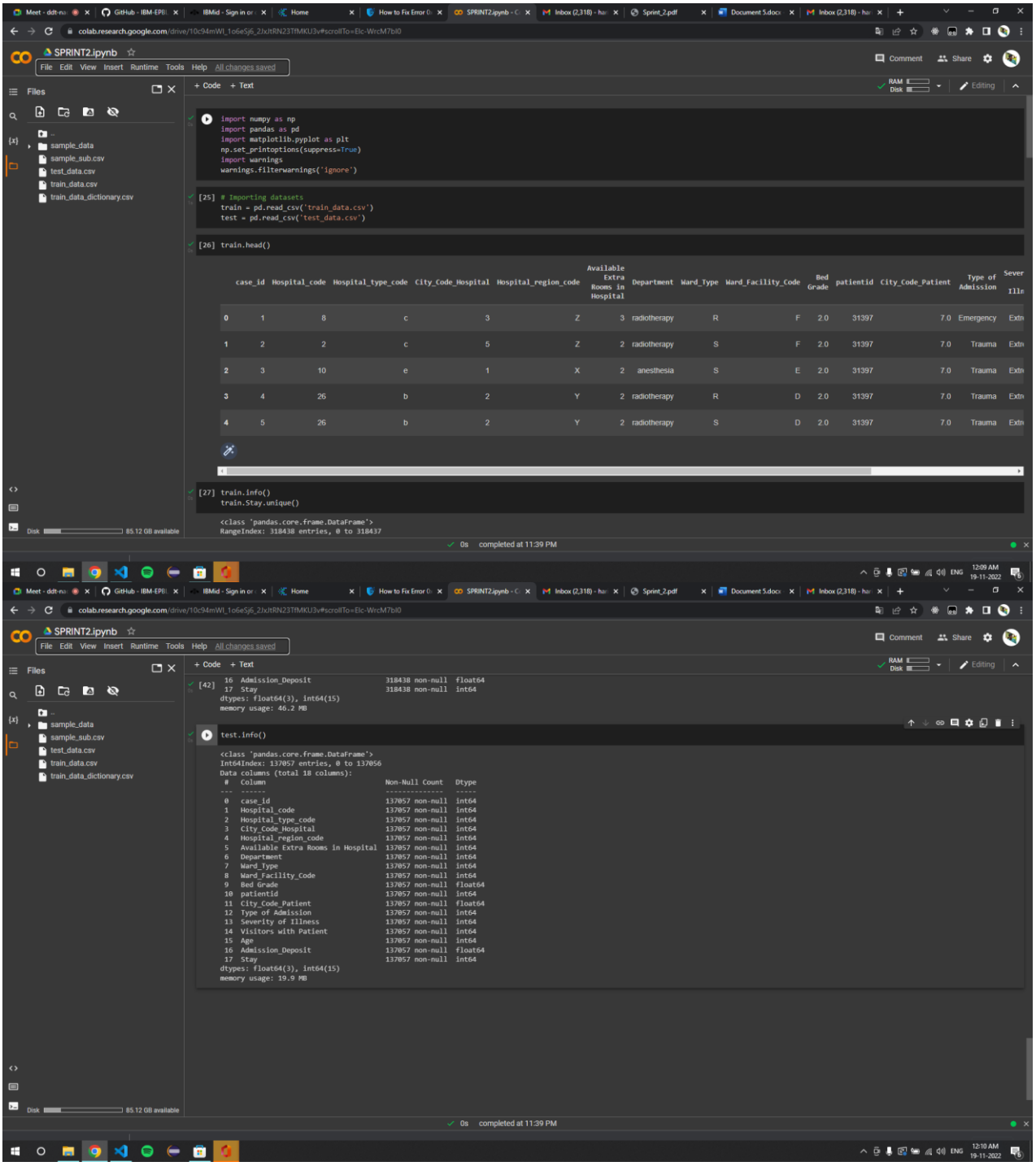


```
train.info()
train.Stay.unique()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 318438 entries, 0 to 318437
Data columns (total 18 columns):
 #   Column                             Non-Null Count   Dtype
---  ------                             --------------   -----
 0   case_id                            318438 non-null  int64
 1   Hospital_code                      318438 non-null  int64
 2   Hospital_type_code                 318438 non-null  object
 3   City_Code_Hospital                 318438 non-null  int64
 4   Hospital_region_code               318438 non-null  object
 5   Available Extra Rooms in Hospital  318438 non-null  int64
 6   Department                         318438 non-null  object
 7   Ward_Type                          318438 non-null  object
 8   Ward_Facility_Code                 318438 non-null  object
 9   Bed Grade                          318325 non-null  float64
 10  patientid                          318438 non-null  int64
 11  City_Code_Patient                  313906 non-null  float64
 12  Type of Admission                  318438 non-null  object
 13  Severity of Illness                318438 non-null  object
 14  Visitors with Patient              318438 non-null  int64
 15  Age                                318438 non-null  object
 16  Admission_Deposit                  318438 non-null  float64
 17  Stay                               318438 non-null  object
dtypes: float64(3), int64(6), object(9)
memory usage: 43.7+ MB
array(['0-10', '41-50', '31-40', '11-20', '51-60', '21-30', '71-80',
       'More than 100 Days', '81-90', '61-70', '91-100'], dtype=object)

[28] train.isnull().sum().sort_values(ascending = False)

     City_Code_Patient                  4532
     Bed Grade                           113
     Hospital_code                         0
     Admission_Deposit                     0
     Age                                   0
     Visitors with Patient                 0
     Severity of Illness                   0
     Type of Admission                     0
     patientid                             0
     case_id                               0
     Ward_Facility_Code                    0
     Ward_Type                             0
```


SPRINT2.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
Files
sample_data
sample_sub.csv
test_data.csv
train_data.csv
train_data_dictionary.csv
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
np.set_printoptions(suppress=True)
import warnings
warnings.filterwarnings('ignore')
[25] # importing datasets
train = pd.read_csv('train_data.csv')
test = pd.read_csv('test_data.csv')
[26] train.head()
case_id Hospital_code Hospital_type_code City_Code_Hospital Hospital_region_code Available Extra Rooms in Hospital Department Ward_Type Ward_Facility_Code Bed Grade patientid City_Code_Patient Type of Admission Sever Illn
0 1 8 c 3 Z 3 radiotherapy R F 2.0 31397 7.0 Emergency Extn
1 2 2 c 5 Z 2 radiotherapy S F 2.0 31397 7.0 Trauma Extn
2 3 10 e 1 X 2 anesthesia S E 2.0 31397 7.0 Trauma Extn
3 4 26 b 2 Y 2 radiotherapy R D 2.0 31397 7.0 Trauma Extn
4 5 26 b 2 Y 2 radiotherapy S D 2.0 31397 7.0 Trauma Extn
[27] train.info()
train.Stay.unique()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 318438 entries, 0 to 318437
Disk 85.12 GB available
0s completed at 11:39 PM

SPRINT2.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Editing
Files
sample_data
sample_sub.csv
test_data.csv
train_data.csv
train_data_dictionary.csv
[42] 16 Admission_Deposit 318438 non-null float64
17 Stay 318438 non-null int64
dtypes: float64(3), int64(15)
memory usage: 46.2 MB
test.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 137057 entries, 0 to 137056
Data columns (total 18 columns):
Column Non-Null Count Dtype
0 case_id 137057 non-null int64
1 Hospital_code 137057 non-null int64
2 Hospital_type_code 137057 non-null int64
3 City_Code_Hospital 137057 non-null int64
4 Hospital_region_code 137057 non-null int64
5 Available Extra Rooms in Hospital 137057 non-null int64
6 Department 137057 non-null int64
7 Ward_Type 137057 non-null int64
8 Ward_Facility_Code 137057 non-null int64
9 Bed Grade 137057 non-null float64
10 patientid 137057 non-null int64
11 City_Code_Patient 137057 non-null float64
12 Type of Admission 137057 non-null int64
13 Severity of Illness 137057 non-null int64
14 Visitors with Patient 137057 non-null int64
15 Age 137057 non-null int64
16 Admission_Deposit 137057 non-null float64
17 Stay 137057 non-null int64
dtypes: float64(3), int64(15)
memory usage: 19.9 MB
Disk 85.12 GB available
0s completed at 11:39 PM