SPRINT 2

Date	07-11-2022
Team ID	PNT2022TMID04305
Project Title	Analytics for Hospitals' Health-Care Data
Team Members	Abishankari S , Aravindh T, Dhivya A , Dhivya S

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.

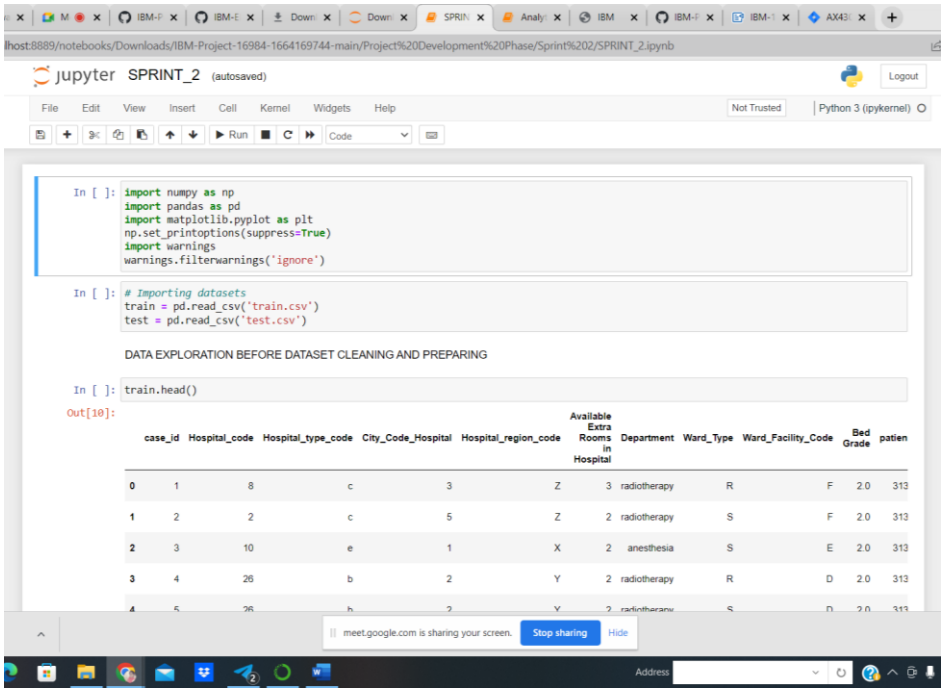
So, the missing values are replaced using the “mode” of the column.

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

Data Exploration in python



jupyter SPRINT_2 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

```
In [ ]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        np.set_printoptions(suppress=True)
        import warnings
        warnings.filterwarnings('ignore')
```

```
In [ ]: # Importing datasets
        train = pd.read_csv('train.csv')
        test = pd.read_csv('test.csv')
```

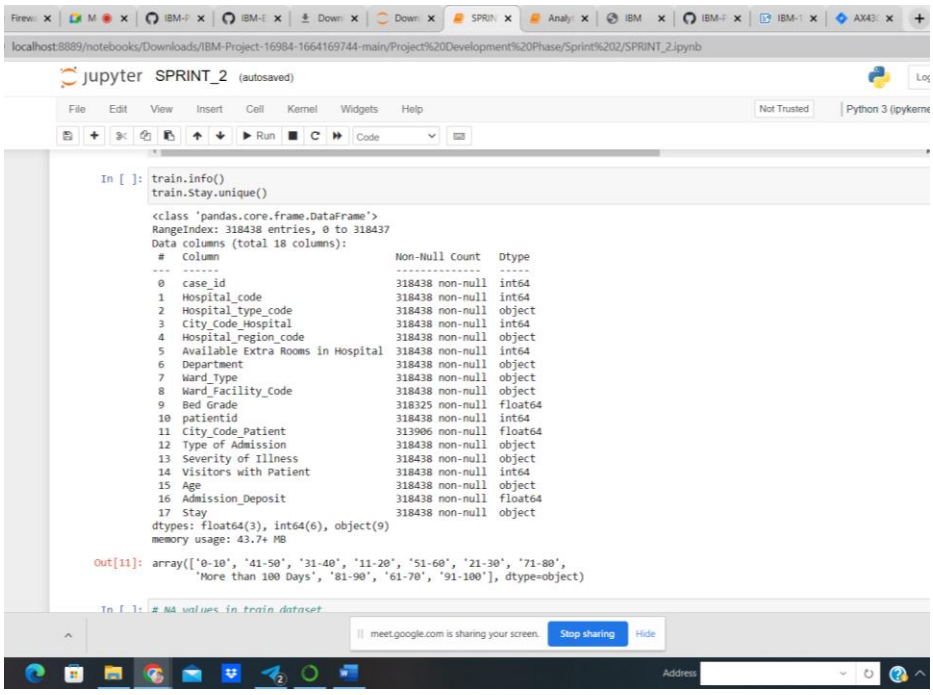
DATA EXPLORATION BEFORE DATASET CLEANING AND PREPARING

```
In [ ]: train.head()
```

Out[10]:

	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	patien
0	1	8	c	3	Z	3	radiotherapy	R	F	2.0	313
1	2	2	c	5	Z	2	radiotherapy	S	F	2.0	313
2	3	10	e	1	X	2	anesthesia	S	E	2.0	313
3	4	26	b	2	Y	2	radiotherapy	R	D	2.0	313
4	5	26	b	2	Y	2	radiotherapy	S	D	2.0	313

meet.google.com is sharing your screen. Stop sharing Hide



jupyter SPRINT_2 (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Not Trusted | Python 3 (ipykernel)

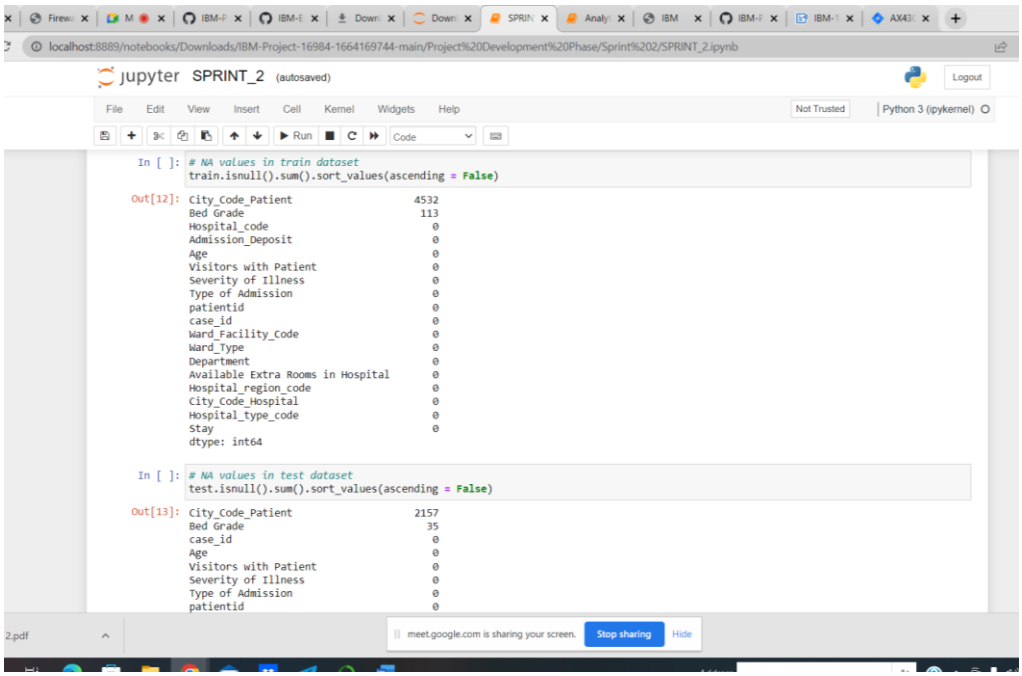
```
In [ ]: train.info()
        train.Stay.unique()
```

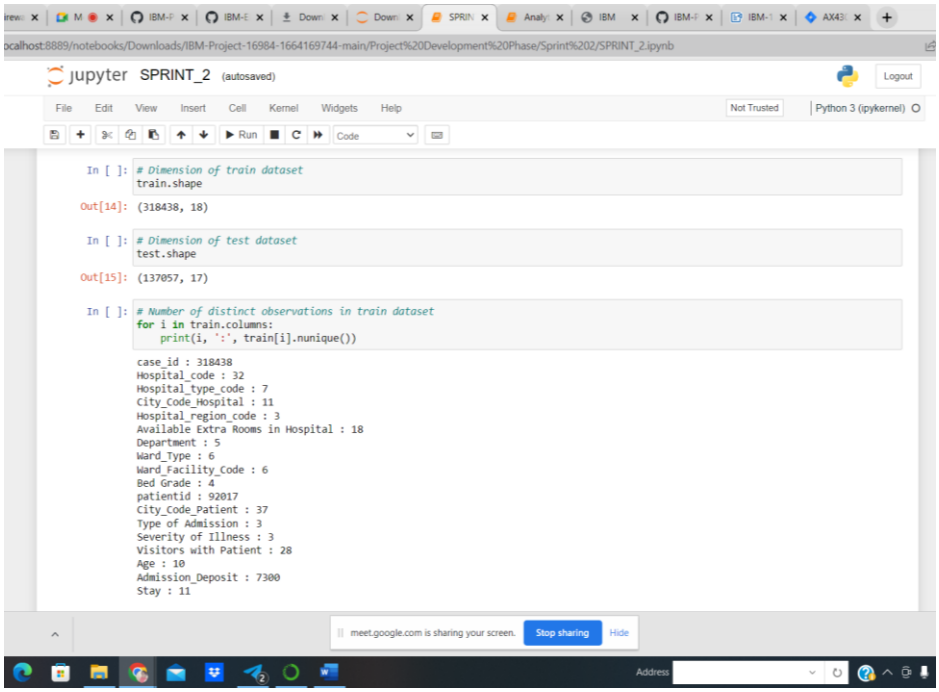
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 318438 entries, 0 to 318437
Data columns (total 18 columns):
 #   Column                             Non-Null Count   Dtype
---  ------                             --------------   -----
 0   case_id                            318438 non-null  int64
 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
```

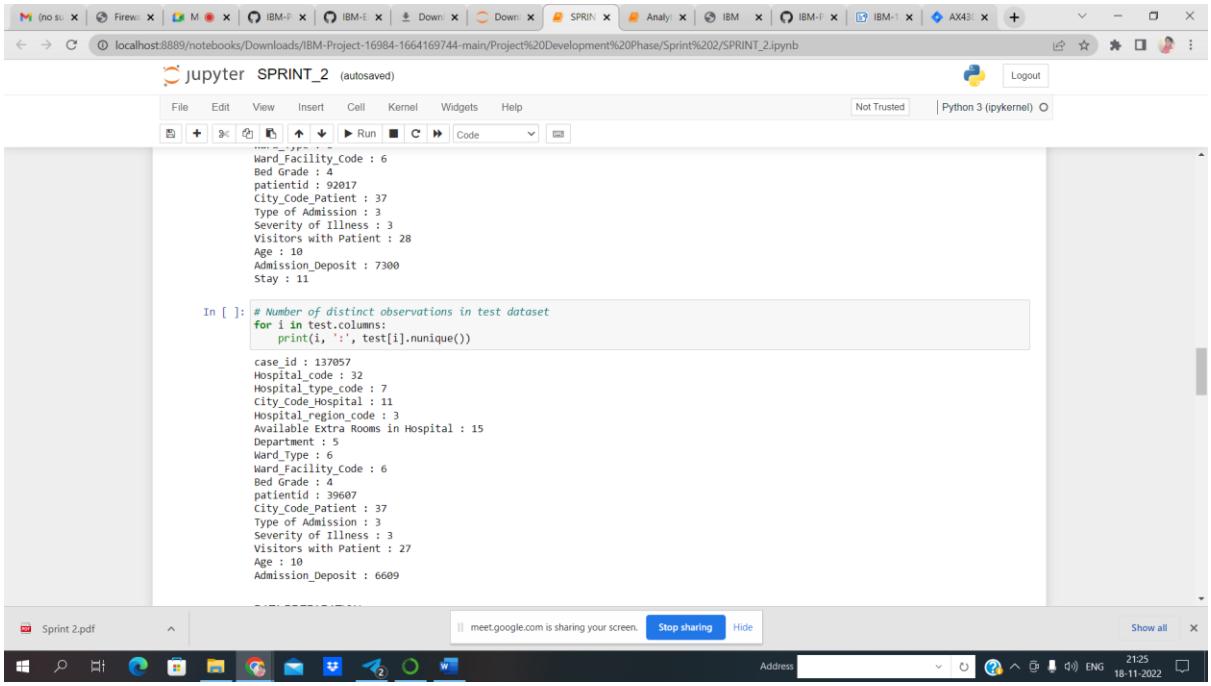
Out[11]:
```
array(['0-10', '41-50', '31-40', '11-20', '51-60', '21-30', '71-80',
       'More than 100 Days', '81-90', '61-70', '91-100'], dtype=object)
```

```
In [ ]: # NA values in train dataset
```

meet.google.com is sharing your screen. Stop sharing Hide

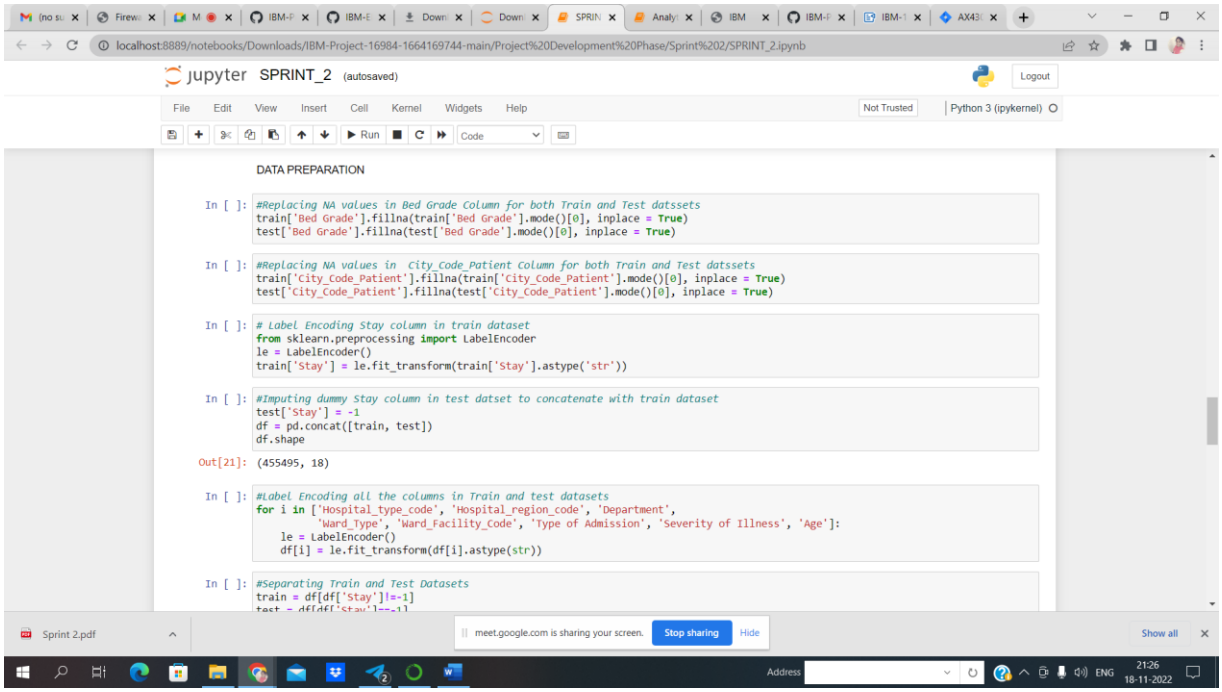

jupyter SPRINT_2 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [ ]:
# NA values in train dataset
train.isnull().sum().sort_values(ascending = False)
Out[12]:
City_Code_Patient 4532
Bed Grade 113
Hospital_code 0
Admission_Deposit 0
Age 0
Visitors with Patient 0
Severity of Illness 0
Type of Admission 0
patientid 0
case_id 0
Ward_Facility_Code 0
Ward_Type 0
Department 0
Available Extra Rooms in Hospital 0
Hospital_region_code 0
City_Code_Hospital 0
Hospital_type_code 0
Stay 0
dtype: int64
In [ ]:
# NA values in test dataset
test.isnull().sum().sort_values(ascending = False)
Out[13]:
City_Code_Patient 2157
Bed Grade 35
case_id 0
Age 0
Visitors with Patient 0
Severity of Illness 0
Type of Admission 0
patientid 0
meet.google.com is sharing your screen.
Stop sharing
Hide

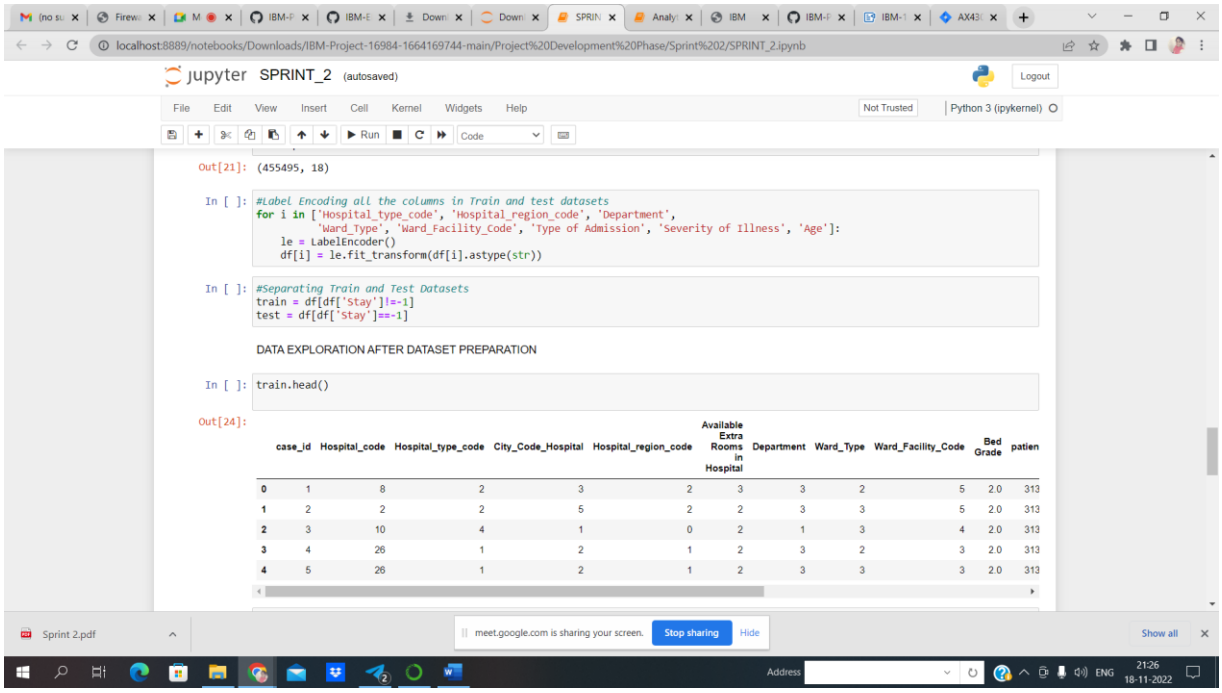

jupyter SPRINT_2 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [ ]:
Dimension of train dataset
train.shape
Out[14]:
(318438, 18)
In [ ]:
Dimension of test dataset
test.shape
Out[15]:
(137057, 17)
In [ ]:
# Number of distinct observations in train dataset
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
meet.google.com is sharing your screen.
Stop sharing
Hide

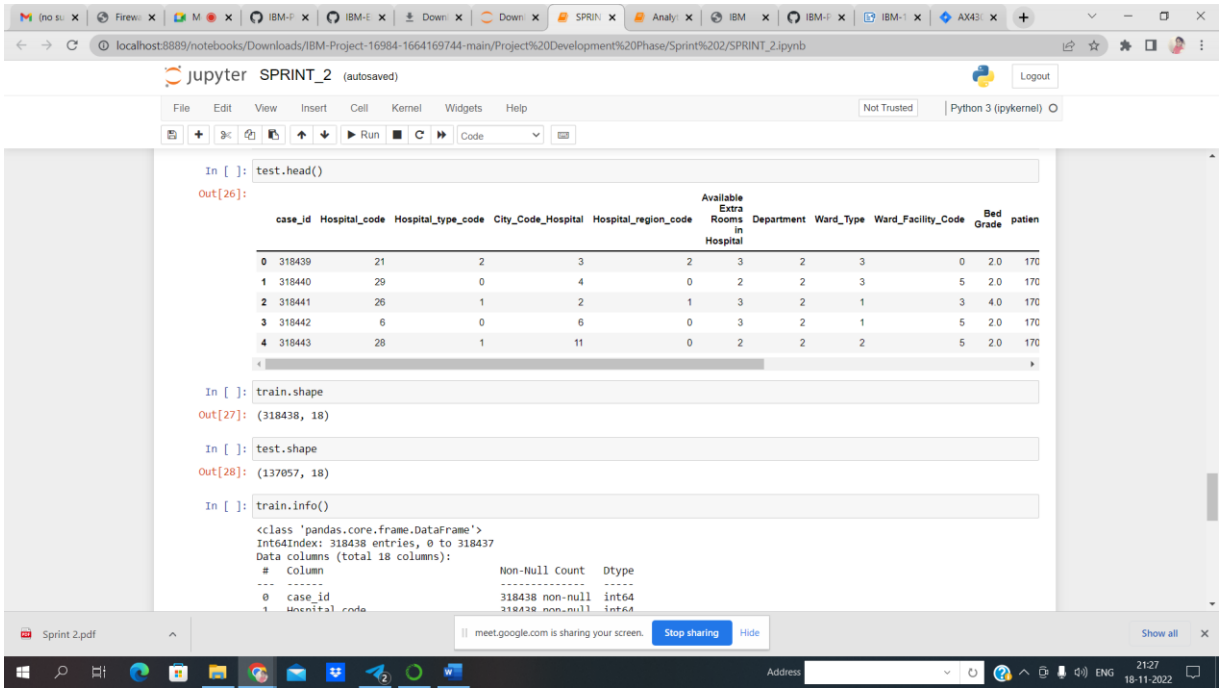

jupyter SPRINT_2 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Ward_Facility_Code : 6
Bed Grade : 4
patientid : 92017
City_Code_Patient : 37
Type of Admission : 3
Severity of Illness : 3
Visitors with Patient : 28
Age : 10
Admission_Deposit : 7300
Stay : 11
In [ ]:
# Number of distinct observations in test dataset
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
Sprint 2.pdf
meet.google.com is sharing your screen.
Stop sharing
Hide
Show all
Address
ENG
21:25
18-11-2022

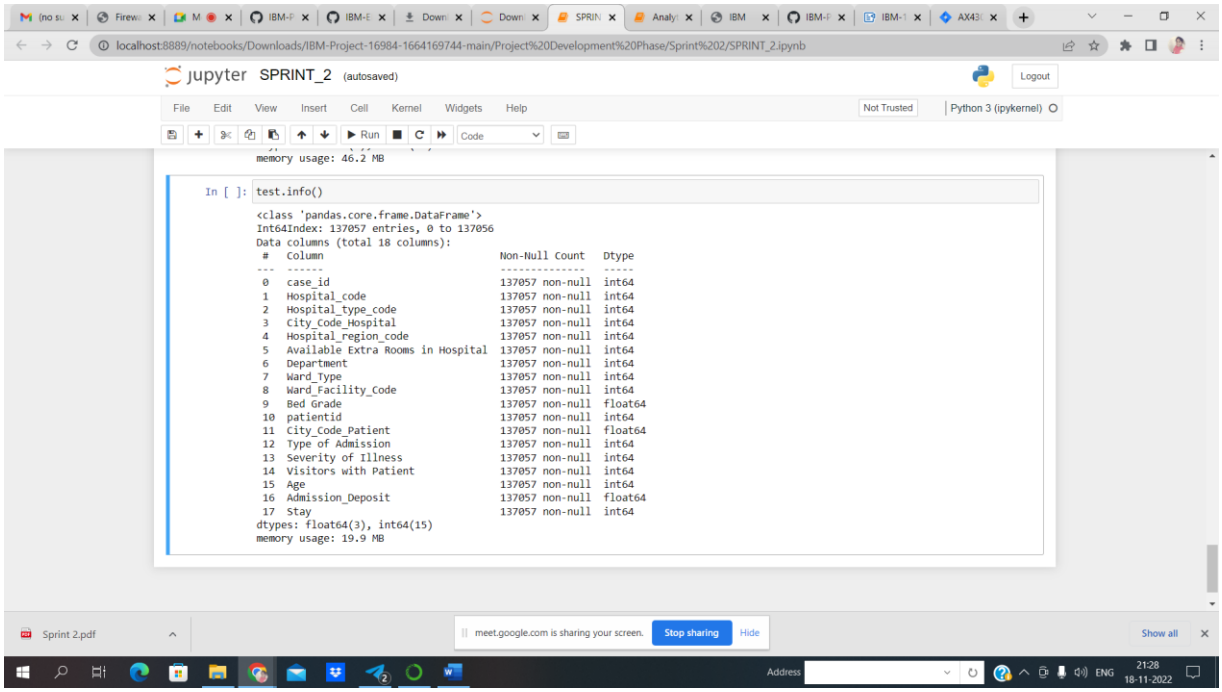
DATA PREPARATION

Data exploration after preparing:






localhost:8889/notebooks/Downloads/IBM-Project-16984-1664169744-main/Project%20Development%20Phase/Sprint%202/SPRINT_2.ipynb
jupyter SPRINT_2 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
In [ ]: test.head()
Out[26]:
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 patien
0 318439 21 2 3 2 3 2 3 0 2.0 170
1 318440 29 0 4 0 2 2 3 5 2.0 170
2 318441 26 1 2 1 3 2 1 3 4.0 170
3 318442 6 0 6 0 3 2 1 5 2.0 170
4 318443 28 1 11 0 2 2 2 5 2.0 170
In [ ]: train.shape
Out[27]: (318438, 18)
In [ ]: test.shape
Out[28]: (137057, 18)
In [ ]: 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
Sprint 2.pdf
meet.google.com is sharing your screen.
Stop sharing
Hide
Show all
Address
ENG
21:27
18-11-2022


localhost:8889/notebooks/Downloads/IBM-Project-16984-1664169744-main/Project%20Development%20Phase/Sprint%202/SPRINT_2.ipynb
jupyter SPRINT_2 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
memory usage: 46.2 MB
In [ ]: 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
Sprint 2.pdf
meet.google.com is sharing your screen.
Stop sharing
Hide
Show all
Address
ENG
21:28
18-11-2022

Sprint 2 completed successfully

“Uploaded ipynb file in the sprint 2 folder in github.”