Table of contents

Ingesting Data

We are going to Ingest Data directly from Kaggle

```
# IMPORTANT: RUN THIS CELL IN ORDER TO IMPORT YOUR KAGGLE DATA SOURCES
# TO THE CORRECT LOCATION (/kaggle/input) IN YOUR NOTEBOOK,
# THEN FEEL FREE TO DELETE THIS CELL.
# NOTE: THIS NOTEBOOK ENVIRONMENT DIFFERS FROM KAGGLE'S PYTHON
# ENVIRONMENT SO THERE MAY BE MISSING LIBRARIES USED BY YOUR
# NOTEBOOK.

import os
import sys
from tempfile import NamedTemporaryFile
from urllib.request import urlopen
from urllib.parse import unquote, urlparse
from urllib.error import HTTPError
from zipfile import ZipFile
import tarfile
import shutil

CHUNK_SIZE = 40960
DATA_SOURCE_MAPPING = 'widsdatathon2024-challenge2:https%3A%2F%2Fstorage.googleapis.com%2Fkaggle-competitions-data%2Fkaggle-v2%2F73478%2F812

KAGGLE_INPUT_PATH='/kaggle/input'
KAGGLE_WORKING_PATH='/kaggle/working'
KAGGLE_SYMLINK='kaggle'

!umount /kaggle/input/ 2> /dev/null
shutil.rmtree('/kaggle/input', ignore_errors=True)
os.makedirs(KAGGLE_INPUT_PATH, 0o777, exist_ok=True)
os.makedirs(KAGGLE_WORKING_PATH, 0o777, exist_ok=True)

try:
  os.symlink(KAGGLE_INPUT_PATH, os.path.join("..", 'input'), target_is_directory=True)
except FileExistsError:
  pass
try:
  os.symlink(KAGGLE_WORKING_PATH, os.path.join("..", 'working'), target_is_directory=True)
except FileExistsError:
  pass

for data_source_mapping in DATA_SOURCE_MAPPING.split(','):
    directory, download_url_encoded = data_source_mapping.split(':')
    download_url = unquote(download_url_encoded)
    filename = urlparse(download_url).path
    destination_path = os.path.join(KAGGLE_INPUT_PATH, directory)
    try:
        with urlopen(download_url) as fileres, NamedTemporaryFile() as tfile:
            total_length = fileres.headers['content-length']
            print(f'Downloading {directory}, {total_length} bytes compressed')
            dl = 0
            data = fileres.read(CHUNK_SIZE)
            while len(data) > 0:
                dl += len(data)
                tfile.write(data)
                done = int(50 * dl / int(total_length))
                sys.stdout.write(f"\r[{'=' * done}{' ' * (50-done)}] {dl} bytes downloaded")
                sys.stdout.flush()
                data = fileres.read(CHUNK_SIZE)
            if filename.endswith('.zip'):
              with ZipFile(tfile) as zfile:
                zfile.extractall(destination_path)
            else:
              with tarfile.open(tfile.name) as tarfile:
                tarfile.extractall(destination_path)
            print(f'\nDownloaded and uncompressed: {directory}')
    except HTTPError as e:
        print(f'Failed to load (likely expired) {download_url} to path {destination_path}')
        continue
    except OSError as e:
        print(f'Failed to load {download_url} to path {destination_path}')
        continue

print('Data source import complete.')
```

```
Downloading widsdatathon2024-challenge2, 5985935 bytes compressed
[==================================================] 5985935 bytes downloaded
Downloaded and uncompressed: widsdatathon2024-challenge2
Data source import complete.
```

```
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)

# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output when you create a version using "Save &
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current session
```

```
/kaggle/input/widsdatathon2024-challenge2/solution_template.csv
/kaggle/input/widsdatathon2024-challenge2/test.csv
/kaggle/input/widsdatathon2024-challenge2/train.csv
```

```
train_df=pd.read_csv('/kaggle/input/widsdatathon2024-challenge2/train.csv')
test_df=pd.read_csv('/kaggle/input/widsdatathon2024-challenge2/test.csv')
```

Exploring Data

```
#printing rows and columns
train_df.shape
```

```
(13173, 152)
```

```
train_df.head()
```

	patient_id int64	patient_race object	payer_type object	patient_state object	patient_zip3 int64	Region object	Division object	patient_age int64	patier
0	268700	nan	COMMERCIAL	AR	724	South	West South Central	39	F
1	484983	White	nan	IL	629	Midwest	East North Central	55	F
2	277055	nan	COMMERCIAL	CA	925	West	Pacific	59	F
3	320055	Hispanic	MEDICAID	CA	900	West	Pacific	59	F
4	190386	nan	COMMERCIAL	CA	934	West	Pacific	71	F

5 rows, showing 10 per page Page 1 of 1

```
### No of patients
train_df['patient_id'].nunique()
```

```
13173
```

```
#exploring no of patients for each race type
train_df['patient_race'].value_counts()
```

```
patient_race
White       3565
Black       1159
Hispanic     807
Other        612
Asian        373
Name: count, dtype: int64
```

```
## creating pivot table for average diagnosis period for each race type
train_df.pivot_table(index='patient_race',values='metastatic_diagnosis_period',aggfunc='mean')
```

	metastatic_diagno...
Asi...	98.19571046
Black	98.01035375
His...	82.42379182
Oth...	98.4869281
Wh...	91.68415147

5 rows, showing 10 per page Page 1 of 1

```
## creating pivot table for average diagnosis period for each Division
train_df.pivot_table(index='Division',values='metastatic_diagnosis_period',aggfunc='mean')
```

	metastatic_diagno...
Eas...	96.6986711
Eas...	106.7469636
Mid...	104.404416
Mo...	90.4664723
Pac...	93.38932147
So...	96.74847561
We...	93.72089947
We...	94.59546061

8 rows, showing 10 per page Page 1 of 1

```
## creating pivot table for average diagnosis period for each Region
train_df.pivot_table(index=['Region','Division'],values='metastatic_diagnosis_period',aggfunc='mean')
```

	metastatic_diagno...
('Mi...	96.6986711
('Mi...	93.72089947
('N...	104.404416
('S...	106.7469636
('S...	96.74847561
('S...	94.59546061
('W...	90.4664723
('W...	93.38932147

8 rows, showing 10 per page Page 1 of 1

```
## creating pivot table for average diagnosis period for each type of payer
train_df.pivot_table(index='payer_type',values='metastatic_diagnosis_period',aggfunc='mean')
```

	metastatic_diagno...
CO...	102.5312053
ME...	93.6278442
ME...	98.63621922

3 rows, showing 10 per page Page 1 of 1

```
## creating pivot table for average diagnosis period for each state
train_df.pivot_table(index=['Region','patient_state'],values='metastatic_diagnosis_period',aggfunc='mean')
```

	metastatic_diagno... 47.25 - 124.654867...
('Mi...	94.27272727
('Mi...	100.5097589
('Mi...	99.85079365
('Mi...	68.87692308
('Mi...	95.51873536
('Mi...	104.2431694
('Mi...	90.98203593
('Mi...	47.25
('Mi...	65.51923077
('Mi...	94.68533333

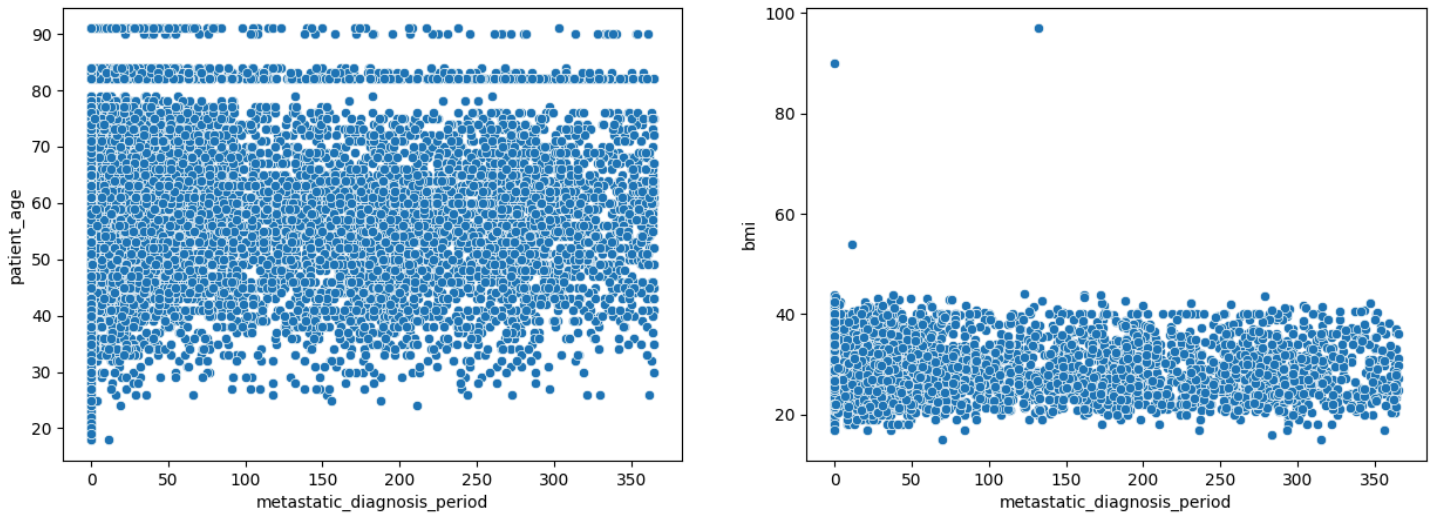
44 rows, showing 10 per page Page 1 of 5

Prelimnary insights :-

- Hispanic race patients have lowest diagnosis time
- Patients in northeast and southeastern states have higher diagnosis time that western states.
- Patients whose payment type is medicaid or medicare advantage have less diagnosis time.

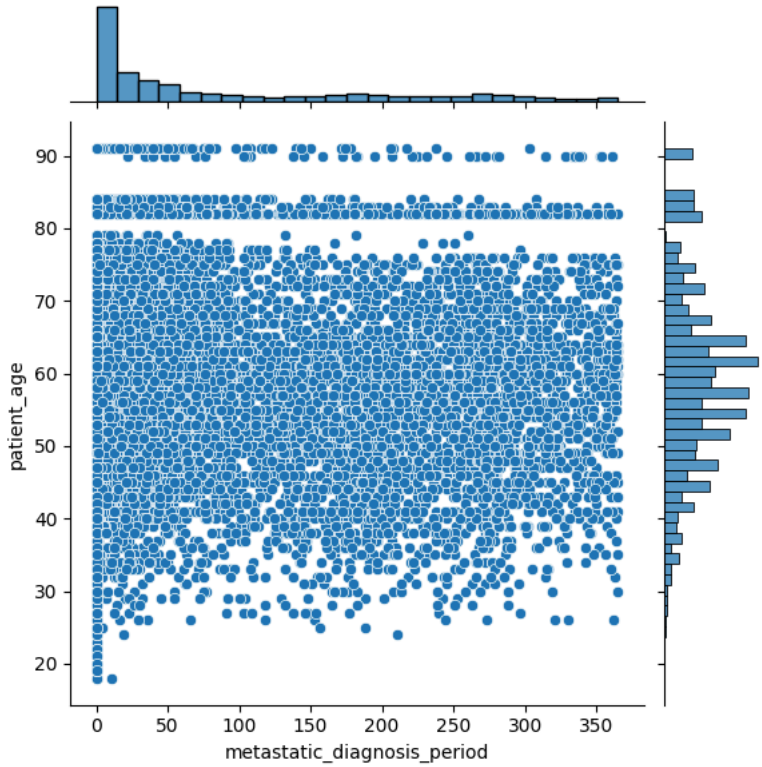
Generating plots for numerical columns like age and BMI of patients

```
import matplotlib.pyplot as plt
import seaborn as sns
fig, axes = plt.subplots(1, 2, figsize=(15, 5))
sns.scatterplot(data=train_df, y="patient_age", x="metastatic_diagnosis_period",  ax=axes[0])
sns.scatterplot(data=train_df, y="bmi", x="metastatic_diagnosis_period", ax=axes[1])
plt.show()
```



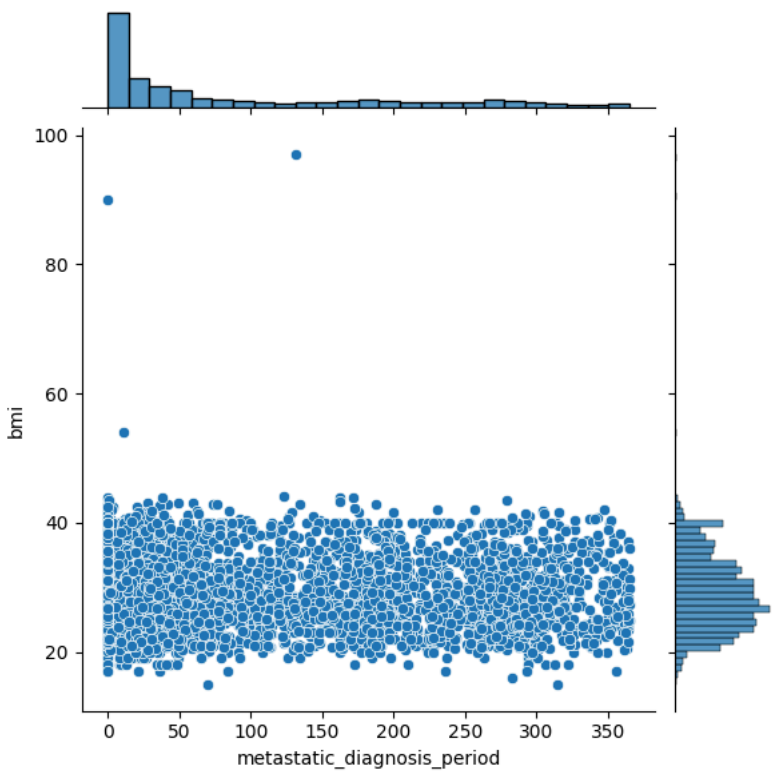
```
sns.jointplot(data=train_df, y="patient_age", x="metastatic_diagnosis_period")
```

```
<seaborn.axisgrid.JointGrid at 0x7f838cbfbf10>
```



```
sns.jointplot(data=train_df, y="bmi", x="metastatic_diagnosis_period")
```

```
<seaborn.axisgrid.JointGrid at 0x7f833b7a5100>
```



Creating New Features

New features can be created i.e

- Age group
- BMI group
- Patient age X Bmi
- removes letter and keeps numbers from 'breast_cancer_diagnosis_code'

```
train_df['patient_age'].isna().sum()
```

```
0
```

```
##define age group
def age_group(x):
  if x<=20 and x>0:
    tag='Age 0-20'
  elif x>20 and x<=40:
    tag= 'Age 20-40'
  elif x>40 and x<=60:
    tag= 'Age 40-60'
  elif x>60 and x<=80:
    tag='Age 60-80'
  else :
    tag= 'Age >80'
  return tag
```

```
train_df['Age_group']=train_df['patient_age'].apply(age_group)
```

```
train_df['Age_group'].value_counts()
```

```
Age_group
Age 40-60    6375
Age 60-80    4581
Age >80      1254
Age 20-40     947
Age 0-20       16
Name: count, dtype: int64
```

```
train_df.pivot_table(index='Age_group',values='metastatic_diagnosis_period',aggfunc='mean')
```

	metastatic_diagno...
Ag...	0.6875
Ag...	97.89440338
Ag...	102.4603922
Ag...	90.1641563
Ag...	89.6738437

5 rows, showing 10 per page Page 1 of 1

```
train_df['bmi'].isna().sum()
```

```
9071
```

```
from sklearn.impute import SimpleImputer
imputer=SimpleImputer(strategy='mean')
train_df['bmi_new']=imputer.fit_transform(train_df[['bmi']])
```

```
train_df.head()
```

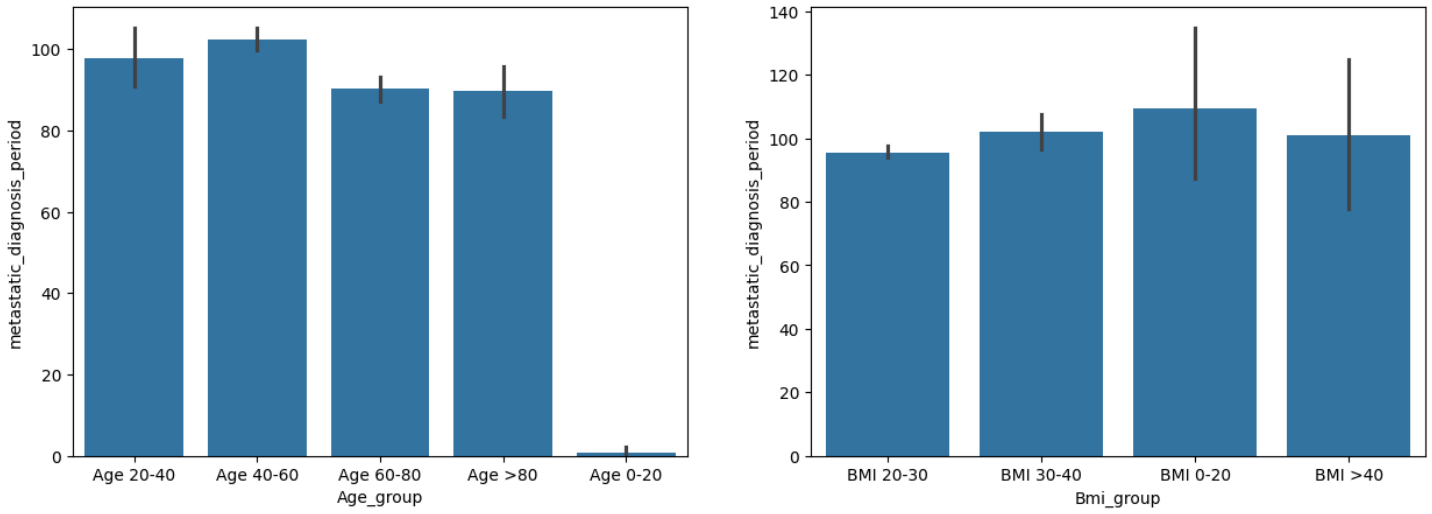
	patient_id int64	patient_race object	payer_type object	patient_state object	patient_zip3 int64	Region object	Division object	patient_age int64	patien
0	268700	nan	COMMERCIAL	AR	724	South	West South Central	39	F
1	484983	White	nan	IL	629	Midwest	East North Central	55	F
2	277055	nan	COMMERCIAL	CA	925	West	Pacific	59	F
3	320055	Hispanic	MEDICAID	CA	900	West	Pacific	59	F
4	190386	nan	COMMERCIAL	CA	934	West	Pacific	71	F

5 rows, showing 10 per page Page 1 of 1

```
##define bmi group
def bmi_group(x):
  if x<=20:
    tag='BMI 0-20'
  elif x>20 and x<=30:
    tag= 'BMI 20-30'
  elif x>30 and x<=40:
    tag= 'BMI 30-40'
  else :
    tag='BMI >40'
  return tag
```

```
train_df['Bmi_group']=train_df['bmi_new'].apply(bmi_group)
```

```
fig, axes = plt.subplots(1, 2, figsize=(15, 5))
sns.barplot(data=train_df, x="Age_group", y="metastatic_diagnosis_period",  ax=axes[0])
sns.barplot(data=train_df, x="Bmi_group", y="metastatic_diagnosis_period", ax=axes[1])
plt.show()
```




```
train_df.pivot_table(index='Bmi_group',values='metastatic_diagnosis_period',aggfunc='mean')
```

	metastatic_diagno...
BM...	109.3563218
BM...	95.63479706
BM...	101.9757653
BM...	101

4 rows, showing 10 per page Page 1 of 1

```
train_df.drop(['patient_age','bmi'],axis=1,inplace=True)
```

Handling Missing values

```
train_df.shape
```

```
(13173, 153)
```

```
train_df.iloc[:,0:8].info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13173 entries, 0 to 13172
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   patient_id      13173 non-null  int64 
 1   patient_race    6516 non-null   object
 2   payer_type      11408 non-null  object
 3   patient_state   13173 non-null  object
 4   patient_zip3    13173 non-null  int64 
 5   Region          13173 non-null  object
 6   Division        13173 non-null  object
 7   patient_gender  13173 non-null  object
dtypes: int64(2), object(6)
memory usage: 823.4+ KB
```

```
pd.concat([train_df.iloc[:,0:8],train_df.iloc[:,79:149]],axis=1).isna().sum()
```

```
patient_id              0
patient_race         6657
payer_type           1765
patient_state           0
patient_zip3            0
                     ...
Average of Aug-18      16
Average of Sep-18       7
Average of Oct-18       7
Average of Nov-18      12
Average of Dec-18      33
Length: 78, dtype: int64
```

```
train_df_cat=train_df.iloc[:,[1,2]]
```

```
from sklearn.impute import SimpleImputer
imputer=SimpleImputer(strategy='most_frequent')
train_df_cat=pd.DataFrame(imputer.fit_transform(train_df_cat),columns=['Patient_race_new','Payer_type_new'])
```

```
train_df.drop(['patient_race','payer_type'],axis=1,inplace=True)
```

```
train_df.drop(['bmi_new','patient_gender'],axis=1,inplace=True)
```

```
train_df=pd.concat([train_df,train_df_cat],axis=1)
```

```
train_df.iloc[:,0:20].head()
```

	patient_id int64	patient_state object	patient_zip3 int64	Region object	Division object	breast_cancer_dia...	breast_cancer_dia...	metastatic_cancer...	metas
0	268700	AR	724	South	West South Central	C50912	Malignant neopla...	C773	nan
1	484983	IL	629	Midwest	East North Central	C50412	Malig neoplasm o...	C773	nan
2	277055	CA	925	West	Pacific	1749	Malignant neopla...	C773	nan
3	320055	CA	900	West	Pacific	C50911	Malignant neopla...	C773	nan
4	190386	CA	934	West	Pacific	1748	Malignant neopla...	C7951	nan

5 rows, showing 10 per page Page 1 of 1

```
train_df.head()
```

	patient_id int64	patient_state object	patient_zip3 int64	Region object	Division object	breast_cancer_dia...	breast_cancer_dia...	metastatic_cancer...	metas
0	268700	AR	724	South	West South Central	C50912	Malignant neopla...	C773	nan
1	484983	IL	629	Midwest	East North Central	C50412	Malig neoplasm o...	C773	nan
2	277055	CA	925	West	Pacific	1749	Malignant neopla...	C773	nan
3	320055	CA	900	West	Pacific	C50911	Malignant neopla...	C773	nan
4	190386	CA	934	West	Pacific	1748	Malignant neopla...	C7951	nan

5 rows, showing 10 per page Page 1 of 1

```
train_df_new=train_df[['patient_id','patient_state','patient_zip3','Region','Division','Age_group','Bmi_group','Patient_race_new','Payer_typ
```

```
train_df_new.shape
```

```
(13173, 10)
```

```
# prompt: export to csv file

#train_df_new.to_csv('/content/patient_data.csv', index=False)
```

```
train_num=train_df.iloc[:,74:146]
```

```
imputer=SimpleImputer(strategy='mean')
train_num=pd.DataFrame(imputer.fit_transform(train_num),columns=train_num.columns.to_list())
```

```
train_num.head()
```

	Average of Jan-13 f..	Average of Feb-13 f..	Average of Mar-13 f..	Average of Apr-13 f..	Average of May-13 f.	Average of Jun-13 f..	Average of Jul-13 f...	Average of Aug-13 f..	Avera
0	38.55	39.88	42.75	55.16	65.17	75.98	76.75	76.45	
1	34.85	36.15	39.41	54.63	65.41	73.89	74.07	74.37	
2	53.14	55.28	64.75	67.38	73.31	79.49	84.01	83.28	
3	57.88	57.65	60.86	62.77	67.07	68.41	70.69	71.19	
4	51.08	52.29	58.31	60.43	63.65	67.41	68.21	67.95	

5 rows, showing 10 per page Page 1 of 1

```
train_num=pd.concat([train_df.iloc[:,[1,2]],train_num],axis=1)
```

```
train_num.head()
```

	patient_state object	patient_zip3 int64	Average of Jan-13 f..	Average of Feb-13 f..	Average of Mar-13 f..	Average of Apr-13 f..	Average of May-13 f.	Average of Jun-13 f..	Avera
0	AR	724	38.55	39.88	42.75	55.16	65.17	75.98	
1	IL	629	34.85	36.15	39.41	54.63	65.41	73.89	
2	CA	925	53.14	55.28	64.75	67.38	73.31	79.49	
3	CA	900	57.88	57.65	60.86	62.77	67.07	68.41	
4	CA	934	51.08	52.29	58.31	60.43	63.65	67.41	

5 rows, showing 10 per page Page 1 of 1

```
train_num.drop_duplicates(inplace=True)
```

```
train_num.head()
```

	patient_state object	patient_zip3 int64	Average of Jan-13 f..	Average of Feb-13 f..	Average of Mar-13 f..	Average of Apr-13 f..	Average of May-13 f.	Average of Jun-13 f..	Avera
0	AR	724	38.55	39.88	42.75	55.16	65.17	75.98	
1	IL	629	34.85	36.15	39.41	54.63	65.41	73.89	
2	CA	925	53.14	55.28	64.75	67.38	73.31	79.49	
3	CA	900	57.88	57.65	60.86	62.77	67.07	68.41	
4	CA	934	51.08	52.29	58.31	60.43	63.65	67.41	

5 rows, showing 10 per page Page 1 of 1

```
train_num.iloc[:,0:75].info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 753 entries, 0 to 12799
Data columns (total 74 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   patient_state      753 non-null    object
 1   patient_zip3       753 non-null    int64
 2   Average of Jan-13  753 non-null    float64
 3   Average of Feb-13  753 non-null    float64
 4   Average of Mar-13  753 non-null    float64
 5   Average of Apr-13  753 non-null    float64
 6   Average of May-13  753 non-null    float64
 7   Average of Jun-13  753 non-null    float64
 8   Average of Jul-13  753 non-null    float64
 9   Average of Aug-13  753 non-null    float64
 10  Average of Sep-13  753 non-null    float64
 11  Average of Oct-13  753 non-null    float64
 12  Average of Nov-13  753 non-null    float64
 13  Average of Dec-13  753 non-null    float64
 14  Average of Jan-14  753 non-null    float64
 15  Average of Feb-14  753 non-null    float64
 16  Average of Mar-14  753 non-null    float64
 17  Average of Apr-14  753 non-null    float64
 18  Average of May-14  753 non-null    float64
 19  Average of Jun-14  753 non-null    float64
 20  Average of Jul-14  753 non-null    float64
 21  Average of Aug-14  753 non-null    float64
 22  Average of Sep-14  753 non-null    float64
 23  Average of Oct-14  753 non-null    float64
 24  Average of Nov-14  753 non-null    float64
```

```
min_value=train_num.iloc[:,2:75].min(axis=1)
```

```
max_value=train_num.iloc[:,2:75].max(axis=1)
```

```
avg_value=train_num.iloc[:,2:75].mean(axis=1)
```

```
train_num['min_temp']=min_value
```

```
train_num['max_temp']=max_value
```

```
train_num['avg_temp']=avg_value
```

```
train_num['temp_diff']=train_num['max_temp']-train_num['min_temp']
```

train_num

	patient_state object CA 7.7% NY 6.4% 42 others 85.9%	patient_zip3 int64 100 - 995	Average of Jan-13 f.. 6.79 - 72.37	Average of Feb-13 f.. 8.93 - 71.0	Average of Mar-13 f.. 14.0 - 70.71	Average of Apr-13 f.. 29.3 - 76.73	Average of May-13 f. 43.26 - 81.45	Average of Jun-13 f.. 56.63 - 91.64	Avera 60.11
0	AR	724	38.55	39.88	42.75	55.16	65.17	75.98	
1	IL	629	34.85	36.15	39.41	54.63	65.41	73.89	
2	CA	925	53.14	55.28	64.75	67.38	73.31	79.49	
3	CA	900	57.88	57.65	60.86	62.77	67.07	68.41	
4	CA	934	51.08	52.29	58.31	60.43	63.65	67.41	
5	IN	461	29.24	30.18	34.88	51.53	65.1	71.34	
6	OH	448	28.71	26.87	33.23	47.86	63.33	69.16	
7	DE	198	34.52	32.92	39.24	52.79	62.2	72.52	
8	LA	706	53.27	55.87	57.34	65.19	72.68	81.59	
9	CA	922	49.98	54.1	65.81	69.86	77.84	85.5	

753 rows, showing 10 per page Page 1 of 76

```
train_num=train_num.iloc[:,[0,1,74,75,76,77]]
```

train_num

	patient_state object CA 7.7% NY 6.4% 42 others 85.9%	patient_zip3 int64 100 - 995	min_temp float64 -2.86 - 68.47	max_temp float64 60.4 - 106.73	avg_temp float64 38.1255555555555...	temp_diff float64 12.009999999999...
0	AR	724	31.1	82.78	59.08375	51.68
1	IL	629	25.62	79.7	56.39902778	54.08
2	CA	925	53.14	87.24	70.01125	34.1
3	CA	900	56.08	77.47	66.69486111	21.39
4	CA	934	49.49	71.37	61.84736111	21.88
5	IN	461	18.96	76.39	52.64875	57.43
6	OH	448	13.52	75.45	50.5125	61.93
7	DE	198	24.48	82.66	54.83194444	58.18
8	LA	706	44.86	84.03	67.98430556	39.17
9	CA	922	49.98	91.39	71.72472222	41.41

753 rows, showing 10 per page Page 1 of 76

```
train_temp_merge=pd.merge(train_df_new,train_num,how='left',on=['patient_state','patient_zip3'])
```

```
train_temp_merge.head()
```

	patient_id int64	patient_state object	patient_zip3 int64	Region object	Division object	Age_group object	Bmi_group object	Patient_race_new o..	Payer
0	268700	AR	724	South	West South Central	Age 20-40	BMI 20-30	White	COM
1	484983	IL	629	Midwest	East North Central	Age 40-60	BMI 30-40	White	COM
2	277055	CA	925	West	Pacific	Age 40-60	BMI 20-30	White	COM
3	320055	CA	900	West	Pacific	Age 40-60	BMI 20-30	Hispanic	MEDI
4	190386	CA	934	West	Pacific	Age 60-80	BMI 20-30	White	COM

5 rows, showing 10 per page Page 1 of 1

```
train_temp_merge.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13173 entries, 0 to 13172
Data columns (total 14 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   patient_id                   13173 non-null  int64
 1   patient_state                13173 non-null  object
 2   patient_zip3                 13173 non-null  int64
 3   Region                       13173 non-null  object
 4   Division                     13173 non-null  object
 5   Age_group                    13173 non-null  object
 6   Bmi_group                    13173 non-null  object
 7   Patient_race_new             13173 non-null  object
 8   Payer_type_new               13173 non-null  object
 9   metastatic_diagnosis_period  13173 non-null  int64
 10  min_temp                     13173 non-null  float64
 11  max_temp                     13173 non-null  float64
 12  avg_temp                     13173 non-null  float64
 13  temp_diff                    13173 non-null  float64
dtypes: float64(4), int64(3), object(7)
memory usage: 1.4+ MB
```

```
train_temp_merge.drop('patient_zip3',axis=1,inplace=True)
```

Creating Dummy variables

```
data_dummy=train_temp_merge.iloc[:,[0,1,2,3,4,5,6,7]]
```

```
data_dummy=pd.get_dummies(data_dummy,dtype=int)
```

```
data_dummy
```

	patient_id int64 100043 - 999982	patient_state_AK i... 0 - 1	patient_state_AL in... 0 - 1	patient_state_AR i... 0 - 1	patient_state_AZ i... 0 - 1	patient_state_CA i... 0 - 1	patient_state_CO i... 0 - 1	patient_state_DC i...	patier
0	268700	0	0	1	0	0	0	0	
1	484983	0	0	0	0	0	0	0	
2	277055	0	0	0	0	1	0	0	
3	320055	0	0	0	0	1	0	0	
4	190386	0	0	0	0	1	0	0	
5	559027	0	0	0	0	0	0	0	
6	293747	0	0	0	0	0	0	0	
7	517596	0	0	0	0	0	0	0	
8	533188	0	0	0	0	0	0	0	
9	639484	0	0	0	0	1	0	0	

13173 rows, showing 10 per page Page 1 of 1318

```
train_temp_merge=pd.concat([data_dummy,train_temp_merge.iloc[:,[8,9,10,11,12]]],axis=1)
```

```
train_temp_merge.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 13173 entries, 0 to 13172
Data columns (total 79 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   patient_id                      13173 non-null  int64
 1   patient_state_AK                13173 non-null  int64
 2   patient_state_AL                13173 non-null  int64
 3   patient_state_AR                13173 non-null  int64
 4   patient_state_AZ                13173 non-null  int64
 5   patient_state_CA                13173 non-null  int64
 6   patient_state_CO                13173 non-null  int64
 7   patient_state_DC                13173 non-null  int64
 8   patient_state_DE                13173 non-null  int64
 9   patient_state_FL                13173 non-null  int64
 10  patient_state_GA                13173 non-null  int64
 11  patient_state_HI                13173 non-null  int64
 12  patient_state_IA                13173 non-null  int64
 13  patient_state_ID                13173 non-null  int64
 14  patient_state_IL                13173 non-null  int64
 15  patient_state_IN                13173 non-null  int64
 16  patient_state_KS                13173 non-null  int64
 17  patient_state_KY                13173 non-null  int64
 18  patient_state_LA                13173 non-null  int64
 19  patient_state_MD                13173 non-null  int64
 20  patient_state_MI                13173 non-null  int64
 21  patient_state_MN                13173 non-null  int64
 22  patient_state_MO                13173 non-null  int64
 23  patient_state_MS                13173 non-null  int64
 24  patient_state_MT                13173 non-null  int64
```

Correlation matrix

```
matrix=train_temp_merge.corr()
```

```
m=matrix['metastatic_diagnosis_period']
```

```
# prompt: heatmap for above correlation matrix

plt.figure(figsize=(15,10))
sns.heatmap(matrix,annot=True)
plt.show()
```

Selecting the features based on correlation values

```
# prompt: sort the series in decending

sorted_m = m.sort_values(ascending=False)
print(sorted_m[1:20])
```

```
Age_group_Age 40-60                0.052835
patient_state_NY                   0.039525
Region_Northeast                   0.028067
Division_Middle Atlantic           0.028067
patient_state_KY                   0.027646
temp_diff                          0.025150
patient_state_ID                   0.024021
Division_East South Central        0.018534
Bmi_group_BMI 30-40                0.018420
patient_state_NM                   0.015971
patient_state_GA                   0.014711
max_temp                           0.013616
Patient_race_new_White             0.013042
patient_state_AZ                   0.012243
patient_state_MN                   0.011989
patient_state_NV                   0.010530
patient_state_IL                   0.009754
Bmi_group_BMI 0-20                 0.009609
Payer_type_new_MEDICARE ADVANTAGE  0.009462
Name: metastatic_diagnosis_period, dtype: float64
```

```
X=train_temp_merge.set_index('patient_id')
Y=X['metastatic_diagnosis_period']
```

```
columns=sorted_m[1:20].index.to_list()
```

```
X=X[columns]
```

```
Y.head()
```

```
patient_id
268700    191
484983     33
277055    157
320055    146
190386    286
Name: metastatic_diagnosis_period, dtype: int64
```

```
Y.info()
```

```
<class 'pandas.core.series.Series'>
Index: 13173 entries, 268700 to 379418
Series name: metastatic_diagnosis_period
Non-Null Count  Dtype
--------------  -----
13173 non-null  int64
dtypes: int64(1)
memory usage: 205.8 KB
```

Splitting the values in test and train datasets

```
# Split into training and test set
from sklearn.model_selection import train_test_split
X_train, X_val, y_train, y_val = train_test_split(X, Y, test_size=0.2)
```

```
print(X_train.shape)
print(X_val.shape)
print(y_train.shape)
print(y_val.shape)
```

```
(10538, 19)
(2635, 19)
(10538,)
(2635,)
```

Model Training with linear regression

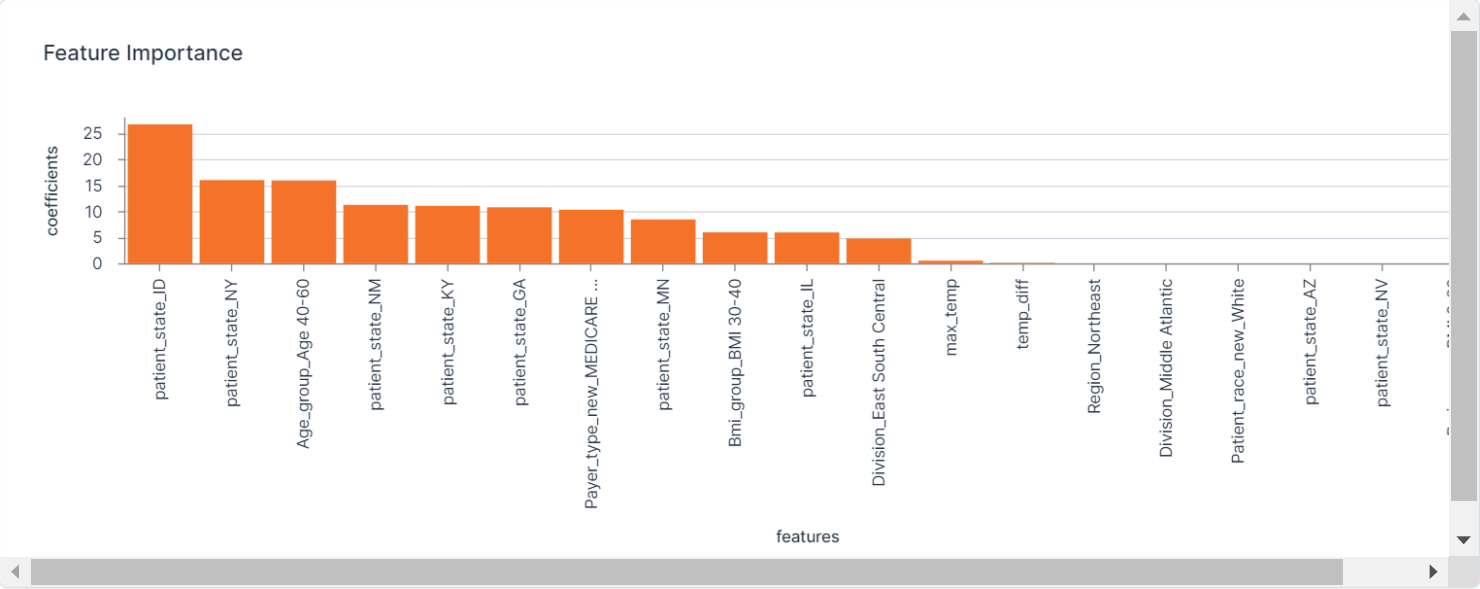
```
from sklearn.linear_model import Lasso
from sklearn.metrics import mean_squared_error

alpha = 0.1  # Regularization strength
lasso_reg = Lasso(alpha=alpha)
lasso_reg.fit(X_train, y_train)
print(X_train.columns.to_list())
print(lasso_reg.coef_)
y_pred = lasso_reg.predict(X_val)

# Calculate Mean Squared Error
mse = mean_squared_error(y_val, y_pred)
print("Mean Squared Error:", mse)
print()
```

```
['Age_group_Age 40-60', 'patient_state_NY', 'Region_Northeast', 'Division_Middle Atlantic', 'patient_state_KY', 'temp_diff', 'patient_state_ID', 'Division_East Sout
[15.91443646 15.98695251 -0.         -0.         11.03873633  0.10962569
 26.7035212   4.76151777  5.94147095 11.22007212 10.75830615  0.50365612
  0.          0.          8.42428329  0.          5.92902273  0.
 10.28892926]
Mean Squared Error: 11659.600993144202
```

```
cols=(X_train.columns.to_list())
coefficients=lasso_reg.coef_
key_value_pairs = pd.DataFrame({'features':cols,'coefficients':coefficients})
```

```
print(key_value_pairs)
```

	features	coefficients
0	Age_group_Age 40-60	15.914436
1	patient_state_NY	15.986953
2	Region_Northeast	-0.000000
3	Division_Middle Atlantic	-0.000000
4	patient_state_KY	11.038736
5	temp_diff	0.109626
6	patient_state_ID	26.703521
7	Division_East South Central	4.761518
8	Bmi_group_BMI 30-40	5.941471
9	patient_state_NM	11.220072
10	patient_state_GA	10.758306
11	max_temp	0.503656
12	Patient_race_new_White	0.000000
13	patient_state_AZ	0.000000
14	patient_state_MN	8.424283
15	patient_state_NV	0.000000
16	patient_state_IL	5.929023
17	Bmi_group_BMI 0-20	0.000000
18	Payer_type_new_MEDICARE ADVANTAGE	10.288929




```
print('Root Mean Squared Error',np.sqrt(mse))
```

```
Root Mean Squared Error 107.97963230695038
```