D212_PA2_JWillis

January 8, 2023

0.1 D212 - Data Mining II - PA2

0.1.1 Background Info:

You are an analyst for a hospital that wants to better understand the characteristics of its patients. You have been asked to use PCA to analyze patient data to identify the principal variables of your patients, ultimately allowing better business and strategic decision-making for the hospital.

Question: "From information about previous patients who were readmitted, can we ascertain the minimum number of principal variables for our patients?"

0.1.2 Import Libraries

[1]:
```
import pandas as pd
import seaborn as sns
import numpy as np
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
%matplotlib inline
```

0.1.3 Load Data From medical_clean.csv

[2]:
```
# load data file
df = pd.read_csv('medical_clean.csv')
# quick test the data is present and see the shape
df.head()
```

[2]:

	CaseOrder	Customer_id	Interaction	UID	City	State	County	Zip
0	1	C412403	8cd49b13-f45a-4b47-a2bd-173ffa932c2f	3a83ddb66e2ae73798bdf1d705dc0932	Eva	AL	Morgan	35621
1	2	Z919181	d2450b70-0337-4406-bdbb-bc1037f1734c					
2	3	F995323	a2057123-abf5-4a2c-abad-8ffe33512562					
3	4	A879973	1dec528d-eb34-4079-adce-0d7a40e82205					
4	5	C544523	5885f56b-d6da-43a3-8760-83583af94266					

```
1  176354c5eef714957d486009feabf195      Marianna    FL       Jackson  32446
2  e19a0fa00aeda885b8a436757e889bc9   Sioux Falls    SD     Minnehaha  57110
3  cd17d7b6d152cb6f23957346d11c3f07  New Richland    MN        Waseca  56072
4  d2f0425877b10ed6bb381f3e2579424a    West Point    VA  King William  23181

        Lat       Lng  …  TotalCharge Additional_charges Item1 Item2  Item3  \
0  34.34960 -86.72508  …  3726.702860       17939.403420     3     3      2
1  30.84513 -85.22907  …  4193.190458       17612.998120     3     4      3
2  43.54321 -96.63772  …  2434.234222       17505.192460     2     4      4
3  43.89744 -93.51479  …  2127.830423       12993.437350     3     5      5
4  37.59894 -76.88958  …  2113.073274        3716.525786     2     1      3

   Item4  Item5 Item6 Item7 Item8
0      2      4     3     3     4
1      4      4     4     3     3
2      4      3     4     3     3
3      3      4     5     5     5
4      3      5     3     4     3

[5 rows x 50 columns]
```

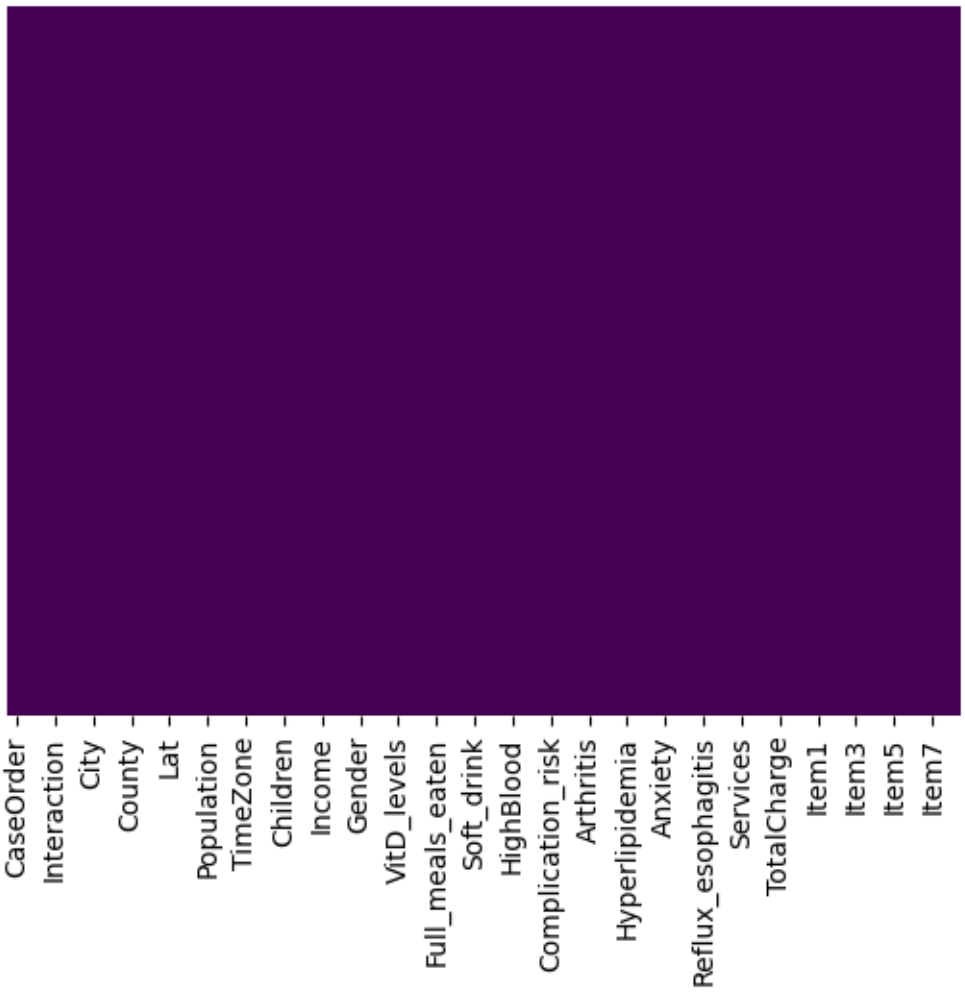
[3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 50 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   CaseOrder           10000 non-null  int64
 1   Customer_id         10000 non-null  object
 2   Interaction         10000 non-null  object
 3   UID                 10000 non-null  object
 4   City                10000 non-null  object
 5   State               10000 non-null  object
 6   County              10000 non-null  object
 7   Zip                 10000 non-null  int64
 8   Lat                 10000 non-null  float64
 9   Lng                 10000 non-null  float64
 10  Population          10000 non-null  int64
 11  Area                10000 non-null  object
 12  TimeZone            10000 non-null  object
 13  Job                 10000 non-null  object
 14  Children            10000 non-null  int64
 15  Age                 10000 non-null  int64
 16  Income              10000 non-null  float64
 17  Marital             10000 non-null  object
 18  Gender              10000 non-null  object
```

```
 19  ReAdmis              10000 non-null  object
 20  VitD_levels          10000 non-null  float64
 21  Doc_visits           10000 non-null  int64
 22  Full_meals_eaten     10000 non-null  int64
 23  vitD_supp            10000 non-null  int64
 24  Soft_drink           10000 non-null  object
 25  Initial_admin        10000 non-null  object
 26  HighBlood            10000 non-null  object
 27  Stroke               10000 non-null  object
 28  Complication_risk    10000 non-null  object
 29  Overweight           10000 non-null  object
 30  Arthritis            10000 non-null  object
 31  Diabetes             10000 non-null  object
 32  Hyperlipidemia       10000 non-null  object
 33  BackPain             10000 non-null  object
 34  Anxiety              10000 non-null  object
 35  Allergic_rhinitis    10000 non-null  object
 36  Reflux_esophagitis   10000 non-null  object
 37  Asthma               10000 non-null  object
 38  Services             10000 non-null  object
 39  Initial_days         10000 non-null  float64
 40  TotalCharge          10000 non-null  float64
 41  Additional_charges   10000 non-null  float64
 42  Item1                10000 non-null  int64
 43  Item2                10000 non-null  int64
 44  Item3                10000 non-null  int64
 45  Item4                10000 non-null  int64
 46  Item5                10000 non-null  int64
 47  Item6                10000 non-null  int64
 48  Item7                10000 non-null  int64
 49  Item8                10000 non-null  int64
dtypes: float64(7), int64(16), object(27)
memory usage: 3.8+ MB
```

0.1.4 Check for Missing Values

[4]:
```
# Mapping to view missing data...none present.
sns.heatmap(df.isnull(), yticklabels=False, cbar=False, cmap='viridis');
```



[5]:
```
df.describe()
```

[5]:

	CaseOrder	Zip	Lat	Lng	Population
count	10000.00000	10000.000000	10000.000000	10000.000000	10000.000000
mean	5000.50000	50159.323900	38.751099	-91.243080	9965.253800
std	2886.89568	27469.588208	5.403085	15.205998	14824.758614
min	1.00000	610.000000	17.967190	-174.209700	0.000000
25%	2500.75000	27592.000000	35.255120	-97.352982	694.750000
50%	5000.50000	50207.000000	39.419355	-88.397230	2769.000000
75%	7500.25000	72411.750000	42.044175	-80.438050	13945.000000
max	10000.00000	99929.000000	70.560990	-65.290170	122814.000000

	Children	Age	Income	VitD_levels	Doc_visits
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	2.097200	53.511700	40490.495160	17.964262	5.012200
std	2.163659	20.638538	28521.153293	2.017231	1.045734

```
min         0.000000     18.000000     154.080000      9.806483      1.000000
25%         0.000000     36.000000   19598.775000     16.626439      4.000000
50%         1.000000     53.000000   33768.420000     17.951122      5.000000
75%         3.000000     71.000000   54296.402500     19.347963      6.000000
max        10.000000     89.000000  207249.100000     26.394449      9.000000

       …   TotalCharge  Additional_charges         Item1         Item2  \
count  …  10000.000000        10000.000000  10000.000000  10000.000000
mean   …   5312.172769        12934.528587      3.518800      3.506700
std    …   2180.393838         6542.601544      1.031966      1.034825
min    …   1938.312067         3125.703000      1.000000      1.000000
25%    …   3179.374015         7986.487755      3.000000      3.000000
50%    …   5213.952000        11573.977735      4.000000      3.000000
75%    …   7459.699750        15626.490000      4.000000      4.000000
max    …   9180.728000        30566.070000      8.000000      7.000000

              Item3         Item4         Item5         Item6         Item7  \
count  10000.000000  10000.000000  10000.000000  10000.000000  10000.000000
mean       3.511100      3.515100      3.496900      3.522500      3.494000
std        1.032755      1.036282      1.030192      1.032376      1.021405
min        1.000000      1.000000      1.000000      1.000000      1.000000
25%        3.000000      3.000000      3.000000      3.000000      3.000000
50%        4.000000      4.000000      3.000000      4.000000      3.000000
75%        4.000000      4.000000      4.000000      4.000000      4.000000
max        8.000000      7.000000      7.000000      7.000000      7.000000

              Item8
count  10000.000000
mean       3.509700
std        1.042312
min        1.000000
25%        3.000000
50%        3.000000
75%        4.000000
max        7.000000

[8 rows x 23 columns]
```

0.1.5 Describe and Explore Numeric Fields:

[6]:
```
df.describe(include = [np.number])
```

[6]:
```
         CaseOrder           Zip           Lat           Lng     Population  \
count  10000.00000  10000.000000  10000.000000  10000.000000   10000.000000
mean    5000.50000  50159.323900     38.751099    -91.243080    9965.253800
std     2886.89568  27469.588208      5.403085     15.205998   14824.758614
min        1.00000    610.000000     17.967190   -174.209700       0.000000
```

```
25%     2500.75000  27592.000000     35.255120    -97.352982     694.750000
50%     5000.50000  50207.000000     39.419355    -88.397230    2769.000000
75%     7500.25000  72411.750000     42.044175    -80.438050   13945.000000
max    10000.00000  99929.000000     70.560990    -65.290170  122814.000000

           Children           Age         Income   VitD_levels    Doc_visits  \
count  10000.000000  10000.000000   10000.000000  10000.000000  10000.000000
mean       2.097200     53.511700   40490.495160     17.964262      5.012200
std        2.163659     20.638538   28521.153293      2.017231      1.045734
min        0.000000     18.000000     154.080000      9.806483      1.000000
25%        0.000000     36.000000   19598.775000     16.626439      4.000000
50%        1.000000     53.000000   33768.420000     17.951122      5.000000
75%        3.000000     71.000000   54296.402500     19.347963      6.000000
max       10.000000     89.000000  207249.100000     26.394449      9.000000

       …   TotalCharge  Additional_charges         Item1         Item2  \
count  …  10000.000000        10000.000000  10000.000000  10000.000000
mean   …   5312.172769        12934.528587      3.518800      3.506700
std    …   2180.393838         6542.601544      1.031966      1.034825
min    …   1938.312067         3125.703000      1.000000      1.000000
25%    …   3179.374015         7986.487755      3.000000      3.000000
50%    …   5213.952000        11573.977735      4.000000      3.000000
75%    …   7459.699750        15626.490000      4.000000      4.000000
max    …   9180.728000        30566.070000      8.000000      7.000000

              Item3         Item4         Item5         Item6         Item7  \
count  10000.000000  10000.000000  10000.000000  10000.000000  10000.000000
mean       3.511100      3.515100      3.496900      3.522500      3.494000
std        1.032755      1.036282      1.030192      1.032376      1.021405
min        1.000000      1.000000      1.000000      1.000000      1.000000
25%        3.000000      3.000000      3.000000      3.000000      3.000000
50%        4.000000      4.000000      3.000000      4.000000      3.000000
75%        4.000000      4.000000      4.000000      4.000000      4.000000
max        8.000000      7.000000      7.000000      7.000000      7.000000

              Item8
count  10000.000000
mean       3.509700
std        1.042312
min        1.000000
25%        3.000000
50%        3.000000
75%        4.000000
max        7.000000

[8 rows x 23 columns]
```

Create DataFrame w/Number DataTypes Only

[7]:
```
df_num = df.select_dtypes(include='number')
df_num.head()
```

[7]:
```
   CaseOrder    Zip       Lat       Lng  Population  Children  Age    Income  \
0          1  35621  34.34960 -86.72508        2951         1   53  86575.93
1          2  32446  30.84513 -85.22907       11303         3   51  46805.99
2          3  57110  43.54321 -96.63772       17125         3   53  14370.14
3          4  56072  43.89744 -93.51479        2162         0   78  39741.49
4          5  23181  37.59894 -76.88958        5287         1   22   1209.56

   VitD_levels  Doc_visits  …  TotalCharge  Additional_charges  Item1  \
0    19.141466           6  …  3726.702860        17939.403420      3
1    18.940352           4  …  4193.190458        17612.998120      3
2    18.057507           4  …  2434.234222        17505.192460      2
3    16.576858           4  …  2127.830423        12993.437350      3
4    17.439069           5  …  2113.073274         3716.525786      2

   Item2  Item3  Item4  Item5  Item6  Item7  Item8
0      3      2      2      4      3      3      4
1      4      3      4      4      4      3      3
2      4      4      4      3      4      3      3
3      5      5      3      4      5      5      5
4      1      3      3      5      3      4      3

[5 rows x 23 columns]
```

0.1.6 Describe and Explore Categorical Fields:

[8]:
```
df.describe(exclude=[np.number])
```

[8]:
```
       Customer_id                          Interaction  \
count        10000                                10000
unique       10000                                10000
top         C412403  8cd49b13-f45a-4b47-a2bd-173ffa932c2f
freq             1                                    1

                                  UID     City  State     County   Area  \
count                           10000    10000  10000      10000  10000
unique                          10000     6072     52       1607      3
top     3a83ddb66e2ae73798bdf1d705dc0932  Houston     TX  Jefferson  Rural
freq                                1       36    553        118   3369

                TimeZone                                Job  Marital  …  \
count              10000                              10000    10000  …
unique                26                                639        5  …
top     America/New_York  Outdoor activities/education manager  Widowed  …
```

```
freq                3889                                    29     2045  …

       Overweight Arthritis Diabetes Hyperlipidemia BackPain Anxiety  \
count       10000     10000    10000          10000    10000   10000
unique          2         2        2              2        2       2
top           Yes        No       No             No       No      No
freq         7094      6426     7262           6628     5886    6785

       Allergic_rhinitis Reflux_esophagitis Asthma    Services
count              10000               10000  10000       10000
unique                 2                   2      2           4
top                   No                  No     No  Blood Work
freq                6059                5865   7107        5265

[4 rows x 27 columns]
```

Create DataFrame w/Categorical DataTypes Only

[9]:
```
df_cat = df.select_dtypes(exclude='number')
df_cat.head()
```

[9]:
```
  Customer_id                           Interaction  \
0     C412403  8cd49b13-f45a-4b47-a2bd-173ffa932c2f
1     Z919181  d2450b70-0337-4406-bdbb-bc1037f1734c
2     F995323  a2057123-abf5-4a2c-abad-8ffe33512562
3     A879973  1dec528d-eb34-4079-adce-0d7a40e82205
4     C544523  5885f56b-d6da-43a3-8760-83583af94266

                                UID          City State        County  \
0  3a83ddb66e2ae73798bdf1d705dc0932           Eva    AL        Morgan
1  176354c5eef714957d486009feabf195      Marianna    FL       Jackson
2  e19a0fa00aeda885b8a436757e889bc9   Sioux Falls    SD     Minnehaha
3  cd17d7b6d152cb6f23957346d11c3f07  New Richland    MN        Waseca
4  d2f0425877b10ed6bb381f3e2579424a    West Point    VA  King William

       Area          TimeZone                                Job   Marital  \
0  Suburban   America/Chicago  Psychologist, sport and exercise  Divorced
1     Urban   America/Chicago      Community development worker   Married
2  Suburban   America/Chicago           Chief Executive Officer   Widowed
3  Suburban   America/Chicago               Early years teacher   Married
4     Rural  America/New_York       Health promotion specialist   Widowed

   … Overweight Arthritis Diabetes Hyperlipidemia BackPain Anxiety  \
0  …         No       Yes      Yes             No      Yes     Yes
1  …        Yes        No       No             No       No      No
2  …        Yes        No      Yes             No       No      No
3  …         No       Yes       No             No       No      No
```

```
4  …          No        No       No             Yes       No      No

  Allergic_rhinitis Reflux_esophagitis Asthma     Services
0               Yes                  No    Yes   Blood Work
1                No                 Yes     No  Intravenous
2                No                  No     No   Blood Work
3                No                 Yes    Yes   Blood Work
4               Yes                  No     No      CT Scan

[5 rows x 27 columns]
```

Describe Readmissions

[10]:
```
df[['ReAdmis']].describe()
```

[10]:
```
       ReAdmis
count    10000
unique       2
top         No
freq      6331
```

Describe Columns

[11]:
```
df.columns
```

[11]:
```
Index(['CaseOrder', 'Customer_id', 'Interaction', 'UID', 'City', 'State',
       'County', 'Zip', 'Lat', 'Lng', 'Population', 'Area', 'TimeZone', 'Job',
       'Children', 'Age', 'Income', 'Marital', 'Gender', 'ReAdmis',
       'VitD_levels', 'Doc_visits', 'Full_meals_eaten', 'vitD_supp',
       'Soft_drink', 'Initial_admin', 'HighBlood', 'Stroke',
       'Complication_risk', 'Overweight', 'Arthritis', 'Diabetes',
       'Hyperlipidemia', 'BackPain', 'Anxiety', 'Allergic_rhinitis',
       'Reflux_esophagitis', 'Asthma', 'Services', 'Initial_days',
       'TotalCharge', 'Additional_charges', 'Item1', 'Item2', 'Item3', 'Item4',
       'Item5', 'Item6', 'Item7', 'Item8'],
      dtype='object')
```

0.1.7 Prep Dummies Data

[12]:
```
df_temp = df[['Age', 'Gender', 'ReAdmis', 'VitD_levels', 'Doc_visits',␣
 ↪'vitD_supp', 'Initial_admin', \
              'HighBlood', 'Stroke', 'Complication_risk', 'Overweight',␣
 ↪'Arthritis', 'Diabetes', 'Hyperlipidemia', \
              'BackPain', 'Anxiety', 'Allergic_rhinitis', 'Reflux_esophagitis',␣
 ↪'Asthma', 'Services', 'Initial_days', \
              'TotalCharge', 'Additional_charges']]
```

[13]:
```
df_dummies = pd.get_dummies(df_temp, drop_first=True)
df_dummies.head()
```

[13]:

	Age	VitD_levels	Doc_visits	vitD_supp	Initial_days	TotalCharge	\
0	53	19.141466	6	0	10.585770	3726.702860	
1	51	18.940352	4	1	15.129562	4193.190458	
2	53	18.057507	4	0	4.772177	2434.234222	
3	78	16.576858	4	0	1.714879	2127.830423	
4	22	17.439069	5	2	1.254807	2113.073274	

	Additional_charges	Gender_Male	Gender_Nonbinary	ReAdmis_Yes	…	\
0	17939.403420	1	0	0	…	
1	17612.998120	0	0	0	…	
2	17505.192460	0	0	0	…	
3	12993.437350	1	0	0	…	
4	3716.525786	0	0	0	…	

	Diabetes_Yes	Hyperlipidemia_Yes	BackPain_Yes	Anxiety_Yes	\
0	1	0	1	1	
1	0	0	0	0	
2	1	0	0	0	
3	0	0	0	0	
4	0	1	0	0	

	Allergic_rhinitis_Yes	Reflux_esophagitis_Yes	Asthma_Yes	\
0	1	0	1	
1	0	1	0	
2	0	0	0	
3	0	1	1	
4	1	0	0	

	Services_CT Scan	Services_Intravenous	Services_MRI
0	0	0	0
1	0	1	0
2	0	0	0
3	0	0	0
4	1	0	0

[5 rows x 28 columns]

[14]:
```
df_dummies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 28 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
```

```
 0   Age                                  10000 non-null  int64
 1   VitD_levels                          10000 non-null  float64
 2   Doc_visits                           10000 non-null  int64
 3   vitD_supp                            10000 non-null  int64
 4   Initial_days                         10000 non-null  float64
 5   TotalCharge                          10000 non-null  float64
 6   Additional_charges                   10000 non-null  float64
 7   Gender_Male                          10000 non-null  uint8
 8   Gender_Nonbinary                     10000 non-null  uint8
 9   ReAdmis_Yes                          10000 non-null  uint8
 10  Initial_admin_Emergency Admission    10000 non-null  uint8
 11  Initial_admin_Observation Admission  10000 non-null  uint8
 12  HighBlood_Yes                        10000 non-null  uint8
 13  Stroke_Yes                           10000 non-null  uint8
 14  Complication_risk_Low                10000 non-null  uint8
 15  Complication_risk_Medium             10000 non-null  uint8
 16  Overweight_Yes                       10000 non-null  uint8
 17  Arthritis_Yes                        10000 non-null  uint8
 18  Diabetes_Yes                         10000 non-null  uint8
 19  Hyperlipidemia_Yes                   10000 non-null  uint8
 20  BackPain_Yes                         10000 non-null  uint8
 21  Anxiety_Yes                          10000 non-null  uint8
 22  Allergic_rhinitis_Yes                10000 non-null  uint8
 23  Reflux_esophagitis_Yes               10000 non-null  uint8
 24  Asthma_Yes                           10000 non-null  uint8
 25  Services_CT Scan                     10000 non-null  uint8
 26  Services_Intravenous                 10000 non-null  uint8
 27  Services_MRI                         10000 non-null  uint8
dtypes: float64(4), int64(3), uint8(21)
memory usage: 752.1 KB
```

[15]:
```
df_dummies.columns
```

[15]:
```
Index(['Age', 'VitD_levels', 'Doc_visits', 'vitD_supp', 'Initial_days',
       'TotalCharge', 'Additional_charges', 'Gender_Male', 'Gender_Nonbinary',
       'ReAdmis_Yes', 'Initial_admin_Emergency Admission',
       'Initial_admin_Observation Admission', 'HighBlood_Yes', 'Stroke_Yes',
       'Complication_risk_Low', 'Complication_risk_Medium', 'Overweight_Yes',
       'Arthritis_Yes', 'Diabetes_Yes', 'Hyperlipidemia_Yes', 'BackPain_Yes',
       'Anxiety_Yes', 'Allergic_rhinitis_Yes', 'Reflux_esophagitis_Yes',
       'Asthma_Yes', 'Services_CT Scan', 'Services_Intravenous',
       'Services_MRI'],
      dtype='object')
```

0.1.8 Keep Only Necessary Columns

[16]:
```
# Start pruning non-relavent features
# Create target and predictor series
pca_df_target = df_dummies['ReAdmis_Yes']
pca_df_pred = df_dummies.drop(['ReAdmis_Yes'], axis=1);
print('-----'*5)
print('ReAdmis_Yes as Target: ' + str(pca_df_target.info()))
print('-----'*5)
print("Predictor Variables: " + str(pca_df_pred.columns))
print('-----'*5)
```

```
-------------------------
<class 'pandas.core.series.Series'>
RangeIndex: 10000 entries, 0 to 9999
Series name: ReAdmis_Yes
Non-Null Count  Dtype
--------------  -----
10000 non-null  uint8
dtypes: uint8(1)
memory usage: 9.9 KB
ReAdmis_Yes as Target: None
-------------------------
Predictor Variables: Index(['Age', 'VitD_levels', 'Doc_visits', 'vitD_supp',
'Initial_days',
       'TotalCharge', 'Additional_charges', 'Gender_Male', 'Gender_Nonbinary',
       'Initial_admin_Emergency Admission',
       'Initial_admin_Observation Admission', 'HighBlood_Yes', 'Stroke_Yes',
       'Complication_risk_Low', 'Complication_risk_Medium', 'Overweight_Yes',
       'Arthritis_Yes', 'Diabetes_Yes', 'Hyperlipidemia_Yes', 'BackPain_Yes',
       'Anxiety_Yes', 'Allergic_rhinitis_Yes', 'Reflux_esophagitis_Yes',
       'Asthma_Yes', 'Services_CT Scan', 'Services_Intravenous',
       'Services_MRI'],
      dtype='object')
-------------------------
```

[17]:
```
pca_df_pred.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 27 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Age                                   10000 non-null  int64
 1   VitD_levels                           10000 non-null  float64
 2   Doc_visits                            10000 non-null  int64
 3   vitD_supp                             10000 non-null  int64
 4   Initial_days                          10000 non-null  float64
```

```
 5   TotalCharge                          10000 non-null  float64
 6   Additional_charges                   10000 non-null  float64
 7   Gender_Male                          10000 non-null  uint8
 8   Gender_Nonbinary                     10000 non-null  uint8
 9   Initial_admin_Emergency Admission    10000 non-null  uint8
 10  Initial_admin_Observation Admission  10000 non-null  uint8
 11  HighBlood_Yes                        10000 non-null  uint8
 12  Stroke_Yes                           10000 non-null  uint8
 13  Complication_risk_Low                10000 non-null  uint8
 14  Complication_risk_Medium             10000 non-null  uint8
 15  Overweight_Yes                       10000 non-null  uint8
 16  Arthritis_Yes                        10000 non-null  uint8
 17  Diabetes_Yes                         10000 non-null  uint8
 18  Hyperlipidemia_Yes                   10000 non-null  uint8
 19  BackPain_Yes                         10000 non-null  uint8
 20  Anxiety_Yes                          10000 non-null  uint8
 21  Allergic_rhinitis_Yes                10000 non-null  uint8
 22  Reflux_esophagitis_Yes               10000 non-null  uint8
 23  Asthma_Yes                           10000 non-null  uint8
 24  Services_CT Scan                     10000 non-null  uint8
 25  Services_Intravenous                 10000 non-null  uint8
 26  Services_MRI                         10000 non-null  uint8
dtypes: float64(4), int64(3), uint8(20)
memory usage: 742.3 KB
```

[18]:
```
pca_df_pred.head()
```

[18]:

	Age	VitD_levels	Doc_visits	vitD_supp	Initial_days	TotalCharge	\
0	53	19.141466	6	0	10.585770	3726.702860	
1	51	18.940352	4	1	15.129562	4193.190458	
2	53	18.057507	4	0	4.772177	2434.234222	
3	78	16.576858	4	0	1.714879	2127.830423	
4	22	17.439069	5	2	1.254807	2113.073274	

	Additional_charges	Gender_Male	Gender_Nonbinary	\
0	17939.403420	1	0	
1	17612.998120	0	0	
2	17505.192460	0	0	
3	12993.437350	1	0	
4	3716.525786	0	0	

	Initial_admin_Emergency Admission	…	Diabetes_Yes	Hyperlipidemia_Yes	\
0	1	…	1	0	
1	1	…	0	0	
2	0	…	1	0	
3	0	…	0	0	
4	0	…	0	1	

```
   BackPain_Yes  Anxiety_Yes  Allergic_rhinitis_Yes  Reflux_esophagitis_Yes  \
0             1            1                      1                       0
1             0            0                      0                       1
2             0            0                      0                       0
3             0            0                      0                       1
4             0            0                      1                       0

   Asthma_Yes  Services_CT Scan  Services_Intravenous  Services_MRI
0           1                 0                     0             0
1           0                 0                     1             0
2           0                 0                     0             0
3           1                 0                     0             0
4           0                 1                     0             0

[5 rows x 27 columns]
```

[19]:
```
print('pca_df_target: ' + str(pca_df_target.shape))
print('------'*5)
print('pca_df_pred: ' + str(pca_df_pred.shape))
```

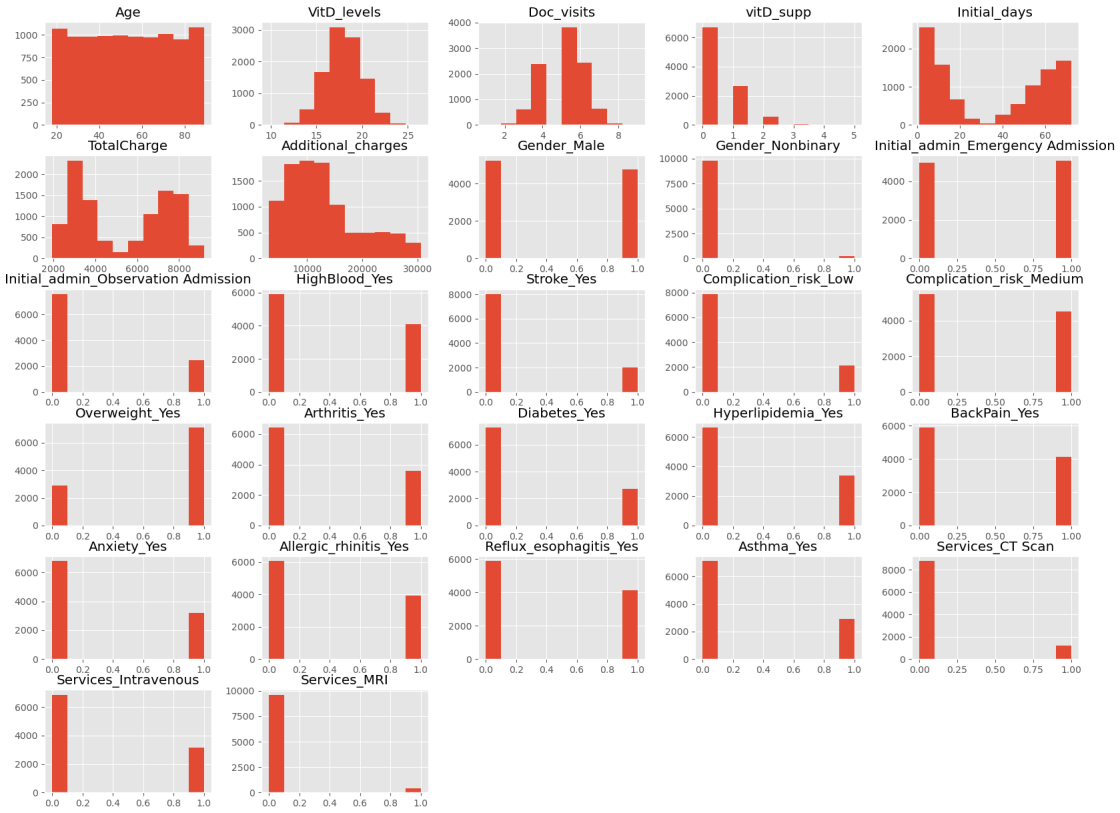
```
pca_df_target: (10000,)
------------------------------
pca_df_pred: (10000, 27)
```

[20]:
```
print('pca_df_target: ' + str(pca_df_target.info()))
print('-------'*10)
print('pca_df_pred: ' + str(pca_df_pred.info()))
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 10000 entries, 0 to 9999
Series name: ReAdmis_Yes
Non-Null Count  Dtype
--------------  -----
10000 non-null  uint8
dtypes: uint8(1)
memory usage: 9.9 KB
pca_df_target: None
----------------------------------------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 27 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Age                                   10000 non-null  int64
 1   VitD_levels                           10000 non-null  float64
 2   Doc_visits                            10000 non-null  int64
 3   vitD_supp                             10000 non-null  int64
```

```
 4   Initial_days                          10000 non-null  float64
 5   TotalCharge                           10000 non-null  float64
 6   Additional_charges                    10000 non-null  float64
 7   Gender_Male                           10000 non-null  uint8
 8   Gender_Nonbinary                      10000 non-null  uint8
 9   Initial_admin_Emergency Admission     10000 non-null  uint8
 10  Initial_admin_Observation Admission   10000 non-null  uint8
 11  HighBlood_Yes                         10000 non-null  uint8
 12  Stroke_Yes                            10000 non-null  uint8
 13  Complication_risk_Low                 10000 non-null  uint8
 14  Complication_risk_Medium              10000 non-null  uint8
 15  Overweight_Yes                        10000 non-null  uint8
 16  Arthritis_Yes                         10000 non-null  uint8
 17  Diabetes_Yes                          10000 non-null  uint8
 18  Hyperlipidemia_Yes                    10000 non-null  uint8
 19  BackPain_Yes                          10000 non-null  uint8
 20  Anxiety_Yes                           10000 non-null  uint8
 21  Allergic_rhinitis_Yes                 10000 non-null  uint8
 22  Reflux_esophagitis_Yes                10000 non-null  uint8
 23  Asthma_Yes                            10000 non-null  uint8
 24  Services_CT Scan                      10000 non-null  uint8
 25  Services_Intravenous                  10000 non-null  uint8
 26  Services_MRI                          10000 non-null  uint8
dtypes: float64(4), int64(3), uint8(20)
memory usage: 742.3 KB
pca_df_pred: None
```

[21]:
```
# https://www.datacamp.com/community/tutorials/
  ↪preprocessing-in-data-science-part-1-centering-scaling-and-knn
plt.style.use('ggplot')
# df = pd.read_csv('http://archive.ics.uci.edu/ml/machine-learning-databases/
  ↪wine-quality/winequality-red.csv ' , sep = ';')
X = pca_df_pred.values # drop target variable
y = pca_df_target.values
pd.DataFrame.hist(pca_df_pred, figsize = [20,15]);
```

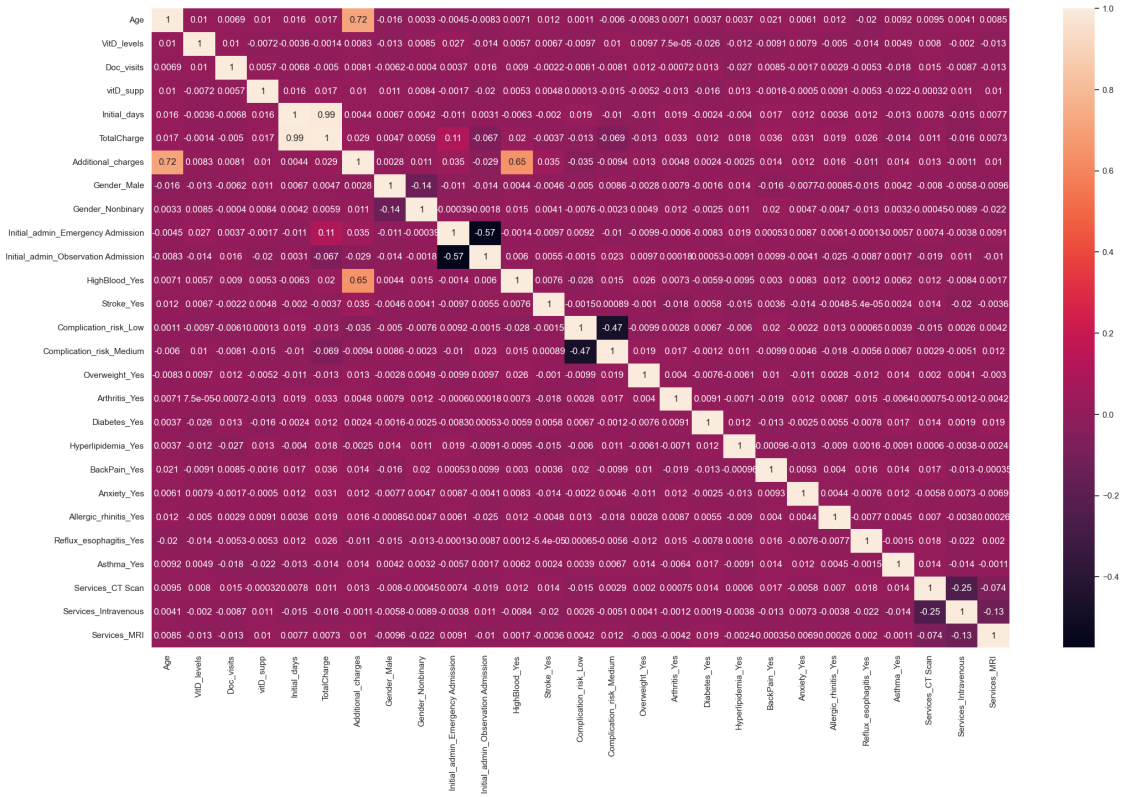


0.2 Correlation Data and Matrix

0.2.1 Correlation Matrix of Predictor Values

[22]:
```
# Correlation of PCA Predictor Values

sns.set(rc = {'figure.figsize':(25,15)})
sns.heatmap(pca_df_pred.corr(), annot=True);
```



[23]:
```
df_dummies.corr()
```

[23]:

	Age	VitD_levels	Doc_visits	\
Age	1.000000	0.010315	0.006898	
VitD_levels	0.010315	1.000000	0.010210	
Doc_visits	0.006898	0.010210	1.000000	
vitD_supp	0.010014	-0.007203	0.005681	
Initial_days	0.016264	-0.003642	-0.006754	
TotalCharge	0.016876	-0.001403	-0.005043	
Additional_charges	0.716854	0.008290	0.008072	
Gender_Male	-0.016297	-0.013205	-0.006159	
Gender_Nonbinary	0.003265	0.008457	-0.000404	
ReAdmis_Yes	0.015810	0.004083	0.000246	
Initial_admin_Emergency Admission	-0.004538	0.027322	0.003686	
Initial_admin_Observation Admission	-0.008336	-0.013690	0.015658	
HighBlood_Yes	0.007147	0.005660	0.008967	
Stroke_Yes	0.012035	0.006721	-0.002230	
Complication_risk_Low	0.001085	-0.009669	-0.006061	
Complication_risk_Medium	-0.006021	0.010341	-0.008091	
Overweight_Yes	-0.008292	0.009689	0.011890	
Arthritis_Yes	0.007110	0.000075	-0.000719	
Diabetes_Yes	0.003694	-0.025834	0.012781	

```
Hyperlipidemia_Yes                  0.003736    -0.011870   -0.026730
BackPain_Yes                        0.021081    -0.009080    0.008514
Anxiety_Yes                         0.006130     0.007875   -0.001684
Allergic_rhinitis_Yes               0.012092    -0.005035    0.002920
Reflux_esophagitis_Yes             -0.019609    -0.014419   -0.005330
Asthma_Yes                          0.009229     0.004937   -0.017989
Services_CT Scan                    0.009506     0.008048    0.014600
Services_Intravenous                0.004142    -0.001984   -0.008700
Services_MRI                        0.008529    -0.012840   -0.012822

                                    vitD_supp  Initial_days  TotalCharge  \
Age                                  0.010014      0.016264     0.016876
VitD_levels                         -0.007203     -0.003642    -0.001403
Doc_visits                           0.005681     -0.006754    -0.005043
vitD_supp                            1.000000      0.015974     0.016924
Initial_days                         0.015974      1.000000     0.987640
TotalCharge                          0.016924      0.987640     1.000000
Additional_charges                   0.010327      0.004409     0.029256
Gender_Male                          0.011164      0.006704     0.004748
Gender_Nonbinary                     0.008395      0.004196     0.005890
ReAdmis_Yes                          0.011039      0.850862     0.843726
Initial_admin_Emergency Admission   -0.001729     -0.011349     0.106985
Initial_admin_Observation Admission -0.020284      0.003085    -0.066870
HighBlood_Yes                        0.005340     -0.006333     0.019910
Stroke_Yes                           0.004777     -0.002043    -0.003694
Complication_risk_Low                0.000131      0.019029    -0.013344
Complication_risk_Medium            -0.014653     -0.010313    -0.068781
Overweight_Yes                      -0.005185     -0.011077    -0.012782
Arthritis_Yes                       -0.012839      0.018907     0.032932
Diabetes_Yes                        -0.015768     -0.002411     0.011524
Hyperlipidemia_Yes                   0.012759     -0.003974     0.017565
BackPain_Yes                        -0.001641      0.017344     0.035828
Anxiety_Yes                         -0.000499      0.011908     0.031199
Allergic_rhinitis_Yes                0.009096      0.003635     0.018919
Reflux_esophagitis_Yes              -0.005316      0.012237     0.026284
Asthma_Yes                          -0.021763     -0.013496    -0.014290
Services_CT Scan                    -0.000317      0.007786     0.010561
Services_Intravenous                 0.011475     -0.015430    -0.016170
Services_MRI                         0.010334      0.007692     0.007341

                                    Additional_charges  Gender_Male  \
Age                                           0.716854    -0.016297
VitD_levels                                   0.008290    -0.013205
Doc_visits                                    0.008072    -0.006159
vitD_supp                                     0.010327     0.011164
Initial_days                                  0.004409     0.006704
TotalCharge                                   0.029256     0.004748
```

Additional_charges	1.000000	0.002757
Gender_Male	0.002757	1.000000
Gender_Nonbinary	0.010869	-0.141169
ReAdmis_Yes	0.013620	0.009813
Initial_admin_Emergency Admission	0.034762	-0.011056
Initial_admin_Observation Admission	-0.029231	-0.013753
HighBlood_Yes	0.654316	0.004434
Stroke_Yes	0.035140	-0.004642
Complication_risk_Low	-0.035234	-0.004992
Complication_risk_Medium	-0.009418	0.008567
Overweight_Yes	0.012771	-0.002831
Arthritis_Yes	0.004788	0.007903
Diabetes_Yes	0.002450	-0.001562
Hyperlipidemia_Yes	-0.002475	0.014073
BackPain_Yes	0.014245	-0.015687
Anxiety_Yes	0.011666	-0.007679
Allergic_rhinitis_Yes	0.016154	-0.000848
Reflux_esophagitis_Yes	-0.011405	-0.015274
Asthma_Yes	0.014083	0.004247
Services_CT Scan	0.013137	-0.007988
Services_Intravenous	-0.001095	-0.005779
Services_MRI	0.010134	-0.009617

	Gender_Nonbinary	ReAdmis_Yes	…	\
Age	0.003265	0.015810	…	
VitD_levels	0.008457	0.004083	…	
Doc_visits	-0.000404	0.000246	…	
vitD_supp	0.008395	0.011039	…	
Initial_days	0.004196	0.850862	…	
TotalCharge	0.005890	0.843726	…	
Additional_charges	0.010869	0.013620	…	
Gender_Male	-0.141169	0.009813	…	
Gender_Nonbinary	1.000000	0.006428	…	
ReAdmis_Yes	0.006428	1.000000	…	
Initial_admin_Emergency Admission	-0.000393	0.019707	…	
Initial_admin_Observation Admission	-0.001820	-0.011972	…	
HighBlood_Yes	0.014721	0.002270	…	
Stroke_Yes	0.004065	0.000918	…	
Complication_risk_Low	-0.007559	0.001186	…	
Complication_risk_Medium	-0.002310	0.002799	…	
Overweight_Yes	0.004853	-0.008586	…	
Arthritis_Yes	0.012280	0.007663	…	
Diabetes_Yes	-0.002469	-0.003058	…	
Hyperlipidemia_Yes	0.011458	0.004307	…	
BackPain_Yes	0.019604	0.013313	…	
Anxiety_Yes	0.004733	0.002406	…	
Allergic_rhinitis_Yes	-0.004719	-0.004651	…	

Reflux_esophagitis_Yes	-0.013315	0.005422	…
Asthma_Yes	0.003185	-0.017133	…
Services_CT Scan	-0.000453	0.024395	…
Services_Intravenous	-0.008914	-0.020313	…
Services_MRI	-0.022162	0.009309	…

	Diabetes_Yes	Hyperlipidemia_Yes	\
Age	0.003694	0.003736	
VitD_levels	-0.025834	-0.011870	
Doc_visits	0.012781	-0.026730	
vitD_supp	-0.015768	0.012759	
Initial_days	-0.002411	-0.003974	
TotalCharge	0.011524	0.017565	
Additional_charges	0.002450	-0.002475	
Gender_Male	-0.001562	0.014073	
Gender_Nonbinary	-0.002469	0.011458	
ReAdmis_Yes	-0.003058	0.004307	
Initial_admin_Emergency Admission	-0.008266	0.018941	
Initial_admin_Observation Admission	0.000535	-0.009077	
HighBlood_Yes	-0.005858	-0.009529	
Stroke_Yes	0.005792	-0.014847	
Complication_risk_Low	0.006675	-0.005972	
Complication_risk_Medium	-0.001241	0.010995	
Overweight_Yes	-0.007575	-0.006102	
Arthritis_Yes	0.009097	-0.007130	
Diabetes_Yes	1.000000	0.011739	
Hyperlipidemia_Yes	0.011739	1.000000	
BackPain_Yes	-0.013405	-0.000963	
Anxiety_Yes	-0.002529	-0.013178	
Allergic_rhinitis_Yes	0.005486	-0.009049	
Reflux_esophagitis_Yes	-0.007816	0.001580	
Asthma_Yes	0.016765	-0.009106	
Services_CT Scan	0.014087	0.000600	
Services_Intravenous	0.001937	-0.003848	
Services_MRI	0.018715	-0.002363	

	BackPain_Yes	Anxiety_Yes	\
Age	0.021081	0.006130	
VitD_levels	-0.009080	0.007875	
Doc_visits	0.008514	-0.001684	
vitD_supp	-0.001641	-0.000499	
Initial_days	0.017344	0.011908	
TotalCharge	0.035828	0.031199	
Additional_charges	0.014245	0.011666	
Gender_Male	-0.015687	-0.007679	
Gender_Nonbinary	0.019604	0.004733	
ReAdmis_Yes	0.013313	0.002406	

Initial_admin_Emergency Admission	0.000535	0.008655
Initial_admin_Observation Admission	0.009861	-0.004077
HighBlood_Yes	0.003048	0.008303
Stroke_Yes	0.003602	-0.013801
Complication_risk_Low	0.019759	-0.002192
Complication_risk_Medium	-0.009920	0.004640
Overweight_Yes	0.010083	-0.011186
Arthritis_Yes	-0.018804	0.012045
Diabetes_Yes	-0.013405	-0.002529
Hyperlipidemia_Yes	-0.000963	-0.013178
BackPain_Yes	1.000000	0.009289
Anxiety_Yes	0.009289	1.000000
Allergic_rhinitis_Yes	0.004023	0.004368
Reflux_esophagitis_Yes	0.016036	-0.007566
Asthma_Yes	0.014261	0.011758
Services_CT Scan	0.016757	-0.005771
Services_Intravenous	-0.013446	0.007251
Services_MRI	-0.000353	-0.006909

	Allergic_rhinitis_Yes \
Age	0.012092
VitD_levels	-0.005035
Doc_visits	0.002920
vitD_supp	0.009096
Initial_days	0.003635
TotalCharge	0.018919
Additional_charges	0.016154
Gender_Male	-0.000848
Gender_Nonbinary	-0.004719
ReAdmis_Yes	-0.004651
Initial_admin_Emergency Admission	0.006080
Initial_admin_Observation Admission	-0.025280
HighBlood_Yes	0.011709
Stroke_Yes	-0.004837
Complication_risk_Low	0.013276
Complication_risk_Medium	-0.017744
Overweight_Yes	0.002819
Arthritis_Yes	0.008748
Diabetes_Yes	0.005486
Hyperlipidemia_Yes	-0.009049
BackPain_Yes	0.004023
Anxiety_Yes	0.004368
Allergic_rhinitis_Yes	1.000000
Reflux_esophagitis_Yes	-0.007731
Asthma_Yes	0.004454
Services_CT Scan	0.007008
Services_Intravenous	-0.003766

```
Services_MRI                                    0.000259

                                  Reflux_esophagitis_Yes  Asthma_Yes  \
Age                                            -0.019609    0.009229
VitD_levels                                    -0.014419    0.004937
Doc_visits                                     -0.005330   -0.017989
vitD_supp                                      -0.005316   -0.021763
Initial_days                                    0.012237   -0.013496
TotalCharge                                     0.026284   -0.014290
Additional_charges                             -0.011405    0.014083
Gender_Male                                    -0.015274    0.004247
Gender_Nonbinary                               -0.013315    0.003185
ReAdmis_Yes                                     0.005422   -0.017133
Initial_admin_Emergency Admission              -0.000126   -0.005672
Initial_admin_Observation Admission            -0.008650    0.001678
HighBlood_Yes                                   0.001150    0.006174
Stroke_Yes                                     -0.000054    0.002443
Complication_risk_Low                           0.000652    0.003902
Complication_risk_Medium                       -0.005622    0.006750
Overweight_Yes                                 -0.012240    0.013943
Arthritis_Yes                                   0.014894   -0.006423
Diabetes_Yes                                   -0.007816    0.016765
Hyperlipidemia_Yes                              0.001580   -0.009106
BackPain_Yes                                    0.016036    0.014261
Anxiety_Yes                                    -0.007566    0.011758
Allergic_rhinitis_Yes                          -0.007731    0.004454
Reflux_esophagitis_Yes                          1.000000   -0.001458
Asthma_Yes                                     -0.001458    1.000000
Services_CT Scan                                0.017628    0.013862
Services_Intravenous                           -0.022007   -0.013559
Services_MRI                                    0.001986   -0.001077

                                  Services_CT Scan  Services_Intravenous  \
Age                                       0.009506              0.004142
VitD_levels                               0.008048             -0.001984
Doc_visits                                0.014600             -0.008700
vitD_supp                                -0.000317              0.011475
Initial_days                              0.007786             -0.015430
TotalCharge                               0.010561             -0.016170
Additional_charges                        0.013137             -0.001095
Gender_Male                              -0.007988             -0.005779
Gender_Nonbinary                         -0.000453             -0.008914
ReAdmis_Yes                               0.024395             -0.020313
Initial_admin_Emergency Admission         0.007412             -0.003787
Initial_admin_Observation Admission      -0.019476              0.010817
HighBlood_Yes                             0.011772             -0.008408
Stroke_Yes                                0.013635             -0.019871
```

Complication_risk_Low	-0.015145	0.002570
Complication_risk_Medium	0.002861	-0.005122
Overweight_Yes	0.002005	0.004074
Arthritis_Yes	0.000754	-0.001198
Diabetes_Yes	0.014087	0.001937
Hyperlipidemia_Yes	0.000600	-0.003848
BackPain_Yes	0.016757	-0.013446
Anxiety_Yes	-0.005771	0.007251
Allergic_rhinitis_Yes	0.007008	-0.003766
Reflux_esophagitis_Yes	0.017628	-0.022007
Asthma_Yes	0.013862	-0.013559
Services_CT Scan	1.000000	-0.252196
Services_Intravenous	-0.252196	1.000000
Services_MRI	-0.074259	-0.134152

	Services_MRI
Age	0.008529
VitD_levels	-0.012840
Doc_visits	-0.012822
vitD_supp	0.010334
Initial_days	0.007692
TotalCharge	0.007341
Additional_charges	0.010134
Gender_Male	-0.009617
Gender_Nonbinary	-0.022162
ReAdmis_Yes	0.009309
Initial_admin_Emergency Admission	0.009122
Initial_admin_Observation Admission	-0.010440
HighBlood_Yes	0.001681
Stroke_Yes	-0.003580
Complication_risk_Low	0.004155
Complication_risk_Medium	0.011933
Overweight_Yes	-0.002963
Arthritis_Yes	-0.004160
Diabetes_Yes	0.018715
Hyperlipidemia_Yes	-0.002363
BackPain_Yes	-0.000353
Anxiety_Yes	-0.006909
Allergic_rhinitis_Yes	0.000259
Reflux_esophagitis_Yes	0.001986
Asthma_Yes	-0.001077
Services_CT Scan	-0.074259
Services_Intravenous	-0.134152
Services_MRI	1.000000

[28 rows x 28 columns]

[24]:
```
# https://realpython.com/knn-python/
# Correlations with Arbitrary Target?
correlation_matrix = df_dummies.corr()

print(correlation_matrix["ReAdmis_Yes"] > 0.5)
```

```
Age                                      False
VitD_levels                              False
Doc_visits                               False
vitD_supp                                False
Initial_days                              True
TotalCharge                               True
Additional_charges                       False
Gender_Male                              False
Gender_Nonbinary                         False
ReAdmis_Yes                               True
Initial_admin_Emergency Admission        False
Initial_admin_Observation Admission      False
HighBlood_Yes                            False
Stroke_Yes                               False
Complication_risk_Low                    False
Complication_risk_Medium                 False
Overweight_Yes                           False
Arthritis_Yes                            False
Diabetes_Yes                             False
Hyperlipidemia_Yes                       False
BackPain_Yes                             False
Anxiety_Yes                              False
Allergic_rhinitis_Yes                    False
Reflux_esophagitis_Yes                   False
Asthma_Yes                               False
Services_CT Scan                         False
Services_Intravenous                     False
Services_MRI                             False
Name: ReAdmis_Yes, dtype: bool
```

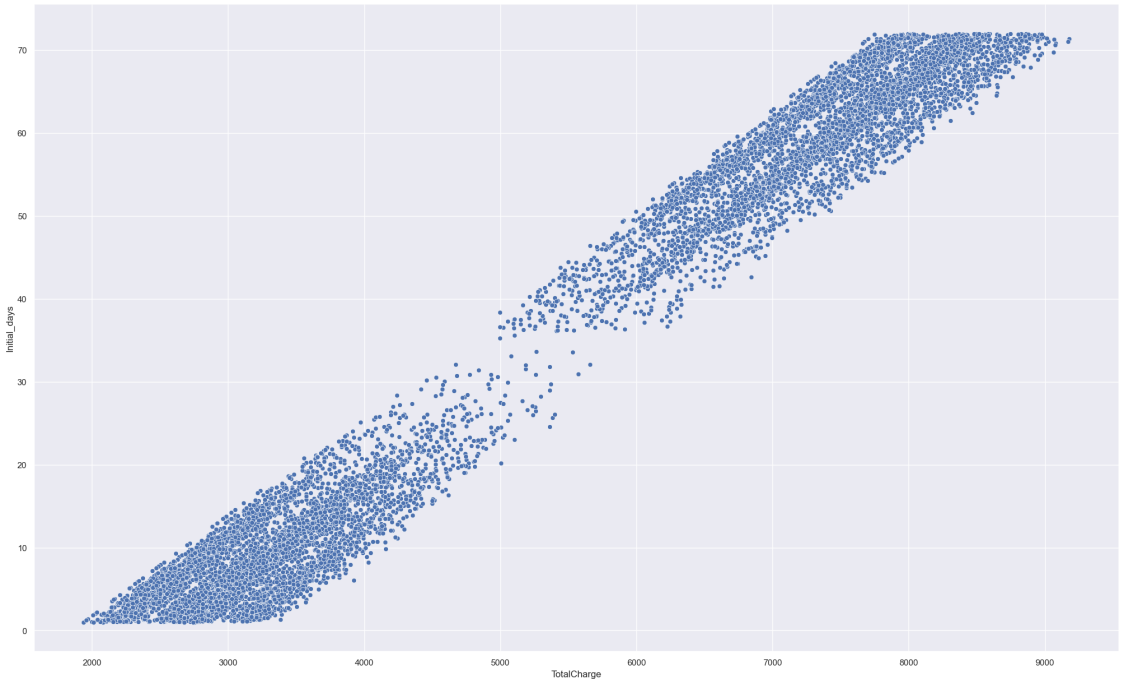
[25]:
```
# Focused features from correlation matrix

#pruned_df = df_dummies[['Initial_days', 'TotalCharge', 'ReAdmis_Yes']]
```

[26]:
```
#pruned_df.shape
```

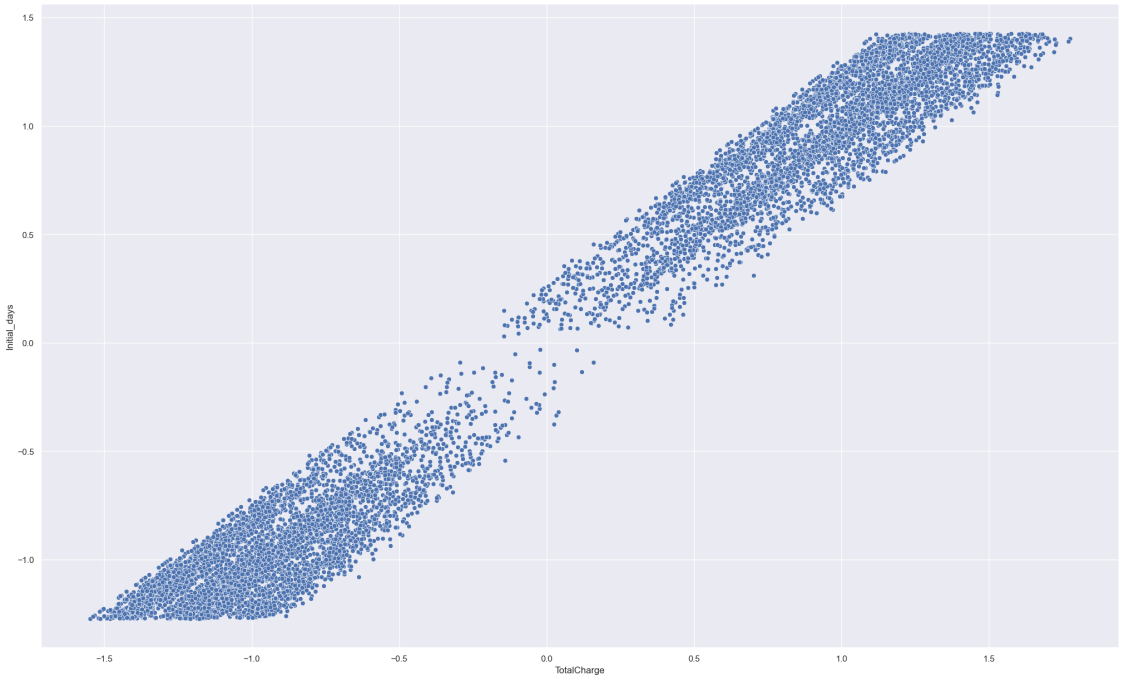
0.3 Intro to PCA

[27]:
```
sns.scatterplot(data=df_dummies, x='TotalCharge', y='Initial_days');
```



[28]:
```
scaler = StandardScaler()
df_std = pd.DataFrame(scaler.fit_transform(df_dummies), columns = df_dummies.
↪columns)
```

[29]:
```
sns.scatterplot(data=df_std, x='TotalCharge', y='Initial_days');
```



0.3.1 Calculating Principle Components

[30]:
```
# Standardizing Continuous Dataset
scaler = StandardScaler()
std_df = scaler.fit_transform(df_dummies)

pca = PCA()
print(pca.fit_transform(std_df))
```

```
[[-1.22345671e+00  1.38470718e+00 -8.99594122e-01 … -8.02858050e-02
   2.95935943e-02 -4.74066999e-10]
 [-1.07180686e+00  1.05141455e+00 -1.10734035e+00 …  5.92369348e-02
   5.00476368e-02 -1.58767057e-09]
 [-1.94004422e+00  1.10048864e+00  5.73019886e-01 … -2.23728403e-01
   2.00594989e-02  3.79674353e-09]
 …
 [ 2.26175452e+00  4.45697534e-01  4.88310313e-01 … -8.21186227e-02
   9.77371257e-02  2.67002221e-08]
 [ 1.97986293e+00 -1.19281366e+00 -7.54375873e-01 … -1.68552210e-01
  -9.31021652e-02  1.12305099e-07]
 [ 2.16046478e+00 -4.67578843e-01  1.54701307e+00 … -5.68823370e-02
   7.47586277e-02  2.00799114e-07]]
```

[31]:
```
# Note: X and y were set on the ggplot graph above
X = pca_df_pred # drop .value
y = pca_df_target

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

0.3.2 Principal Component Explained Variance Ratio

[32]:
```
pca.fit(std_df)
print(pca.explained_variance_ratio_)
```

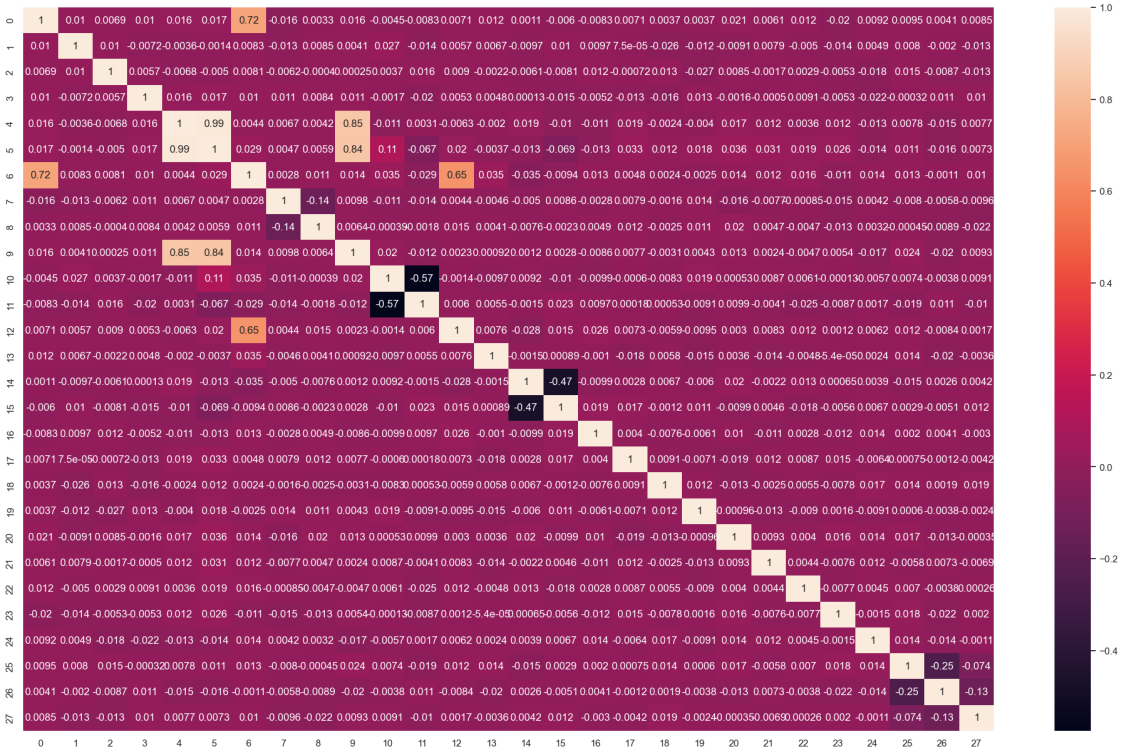
```
[1.00118221e-01 7.07693086e-02 5.65669575e-02 5.25829888e-02
 4.52843961e-02 4.10973459e-02 3.83912689e-02 3.73819320e-02
 3.71499909e-02 3.68721612e-02 3.66664449e-02 3.64041305e-02
 3.63028553e-02 3.62048502e-02 3.56723878e-02 3.53076877e-02
 3.47577197e-02 3.44899748e-02 3.43474241e-02 3.37296808e-02
 3.33302010e-02 3.02798083e-02 2.41038248e-02 1.90005142e-02
 1.51483195e-02 6.92287086e-03 1.11673424e-03 2.55980038e-16]
```

0.3.3 Export Cleaned and Standardized Dataset

[33]:
```
pd.DataFrame(std_df).to_csv('std_df.csv', index=False)
```

0.3.4 Correlation Matrix of All Principle Components (Standardized)

[34]:
```
sns.set(rc = {'figure.figsize':(25,15)})
sns.heatmap(pd.DataFrame(std_df).corr(), annot=True);
```



[35]:
```
print(pca.explained_variance_ratio_.cumsum())
```

```
[0.10011822 0.17088753 0.22745449 0.28003748 0.32532187 0.36641922
 0.40481049 0.44219242 0.47934241 0.51621457 0.55288102 0.58928515
 0.625588   0.66179285 0.69746524 0.73277293 0.76753065 0.80202062
 0.83636805 0.87009773 0.90342793 0.93370774 0.95781156 0.97681208
 0.99196039 0.99888327 1.         1.        ]
```

[36]:
```
print(pca.components_)
```

```
[[ 3.04575870e-02  5.00336346e-04 -3.31746085e-03  1.51748444e-02
   5.83242471e-01  5.86650185e-01  3.58096885e-02  5.73009813e-03
   5.56203989e-03  5.52181242e-01  5.19919824e-02 -4.28250381e-02
   1.79823753e-02 -1.01369514e-03  9.07304720e-03 -2.83097490e-02
  -1.11070503e-02  1.91858523e-02  1.74716739e-03  6.39558426e-03
```

```
  2.21687649e-02  1.53770805e-02  7.87333771e-03  1.45506974e-02
 -1.37897245e-02  1.71131933e-02 -2.00208969e-02  9.17357715e-03]
[ 5.16596935e-01  1.78835567e-02  1.44967793e-02  1.41977419e-02
 -4.14287377e-02 -2.07425129e-02  7.01130740e-01 -8.01346295e-03
  1.87091310e-02 -3.21820045e-02  4.77168157e-02 -4.88453158e-02
  4.74658866e-01  3.50428608e-02 -4.81547157e-02  2.05756116e-02
  1.93906471e-02  8.15052746e-03  4.19980344e-04 -4.52176252e-03
  2.05610045e-02  1.52595291e-02  2.43011018e-02 -1.88755012e-02
  1.99131851e-02  2.40366360e-02 -1.05552074e-02  1.14013049e-02]
[ 3.28487784e-02 -3.98450257e-02  1.44563295e-02 -2.21919739e-02
  7.52314936e-02 -9.45562035e-03  2.99949161e-02  1.06363567e-03
  2.57482284e-03  5.73157128e-02 -6.84011888e-01  6.85377281e-01
  5.17858620e-02  2.04837608e-02 -1.32974179e-01  1.48030986e-01
  3.13841350e-02  7.79274780e-03  8.00508386e-03 -3.03091805e-02
  8.17234841e-03 -1.00876073e-02 -4.17407230e-02 -1.23770592e-02
  7.89080114e-03 -3.14540746e-02  2.86342211e-02 -2.03763773e-02]
[ 7.06419792e-02 -4.12260179e-02  7.94535461e-03  1.91612358e-02
 -2.48263216e-04 -4.81248177e-03  3.93803985e-02 -2.34039002e-02
 -9.85740105e-04 -2.19701825e-02 -1.38972924e-01  1.34530320e-01
 -7.48260903e-03  4.12556981e-03  6.85922697e-01 -6.85368400e-01
 -3.52049151e-02 -2.24294268e-02  1.32236846e-02 -3.22430763e-02
  4.73114685e-02 -8.75479737e-03  3.91279858e-02  5.46012047e-04
 -1.43834771e-03 -5.41379436e-02  5.00837924e-02 -1.80301283e-02]
[-2.60654805e-02  2.55204424e-04  4.97585194e-02 -3.73953885e-02
 -1.71983055e-02 -1.88379350e-02 -1.88130226e-02 -3.92405968e-02
  2.62747199e-02 -5.25692255e-03 -4.76076653e-02  3.26258242e-02
  4.24187193e-03  8.80724380e-02  4.18973148e-02 -4.18151754e-02
 -5.55411911e-04 -1.04736459e-02  4.58036241e-02 -9.43329317e-03
  8.92892452e-02 -4.47144960e-02  2.10969415e-02  1.06172065e-01
  7.58798053e-02  6.41559617e-01 -7.12015452e-01  1.71070819e-01]
[-2.59130622e-04  1.32449810e-01  5.86019132e-02 -3.11209736e-02
 -1.34371177e-03  4.08102704e-03 -2.32836268e-02 -6.69565228e-01
  6.74160976e-01 -2.01790930e-03  2.03663791e-02  8.86430990e-03
 -2.25985889e-02  3.15208489e-02 -1.94224163e-02  4.76154562e-03
  5.92871162e-02  5.98637105e-03 -5.62548138e-02 -4.60093380e-02
  1.60790005e-01  7.87578553e-02 -2.17820609e-02 -5.19540783e-04
  1.35669691e-02  2.51893593e-02  6.65002005e-02 -1.60241754e-01]
[ 9.75067607e-02 -2.82606013e-01 -2.54536500e-01  1.00331831e-01
 -1.93929376e-03 -3.66280764e-03  9.38387575e-03 -2.01015476e-01
  1.38339628e-01 -7.84087200e-03 -1.04214886e-02  3.52240265e-03
 -9.37021243e-02 -3.62322529e-02 -5.05903922e-03  3.17213166e-02
 -1.91439009e-01 -4.86289668e-02  1.87313849e-01  2.23166635e-01
 -2.32951239e-02 -8.97186780e-02 -5.33878268e-02  4.60862861e-02
 -6.72148555e-02 -3.22204477e-01 -1.26035589e-01  7.09313138e-01]
[ 5.92410877e-02  2.95629269e-02  3.36480622e-02  4.81090718e-01
 -1.84740310e-03 -1.50720601e-02  5.83932911e-03  3.16801311e-02
  2.22674788e-02  1.41358458e-02 -1.32072612e-02  5.95664223e-03
 -6.17545907e-02  2.41942976e-01 -3.28627408e-02 -1.19336181e-02
```

```
 -1.19570109e-01 -4.20545466e-01 -3.29896925e-01  2.56459765e-01
  6.07840639e-02 -3.16799156e-01 -1.73384475e-01 -2.05736599e-02
 -4.09819291e-01  8.76871949e-02  4.89155663e-03 -1.46027372e-01]
[-1.03224227e-01  3.21804096e-01  2.89904220e-01  7.62930051e-02
  1.10226435e-02 -4.15592446e-03 -1.87320447e-02 -2.95393638e-02
 -7.83495961e-02  1.48373101e-02  1.50279190e-02  8.46260655e-04
  8.17725090e-02  1.60245614e-01  9.68320856e-03 -5.12655441e-03
  3.10798554e-01 -1.95678184e-01 -2.32418635e-01 -4.97160046e-01
  2.52088817e-02  4.52836495e-03  5.43139745e-02 -2.31985563e-01
 -3.74860662e-02 -2.18763811e-01 -8.88815500e-02  4.52582031e-01]
[ 1.08101633e-01  5.14745528e-02 -3.31902879e-01 -2.20977933e-01
  1.00369430e-02  3.54114830e-03 -3.83308717e-03  5.14826327e-02
 -7.32008264e-02  1.31880153e-02  1.66087883e-02 -1.24610694e-02
 -1.41164499e-01  2.74155489e-01 -8.20775505e-04  2.37260330e-02
  7.45048734e-02 -4.75064707e-01 -5.71489915e-03  7.07890118e-02
  3.64235170e-01  7.91321496e-04 -1.55013992e-01 -9.89202436e-02
  5.59778082e-01 -3.83836014e-02  5.60791651e-02 -3.93222140e-02]
[ 2.21721039e-01  5.14986684e-02  2.66846543e-01  2.37473680e-02
  8.95036624e-03 -3.22855353e-03 -1.00830795e-02 -2.37294061e-02
  4.05609206e-02  1.76356292e-02 -1.58169022e-03 -6.76634550e-03
 -2.66040032e-01  2.22008740e-01 -2.08920165e-02  1.11838368e-02
 -1.45281806e-01 -9.74155604e-02  5.40588662e-01  1.10043186e-02
 -2.84811335e-01 -7.95893836e-02  1.29252362e-01 -5.58628088e-01
  1.19936731e-02  8.41774045e-02  2.00630140e-02 -7.88711908e-02]
[ 2.62473241e-01  3.27177428e-01 -1.58700516e-01  7.42733728e-02
 -3.02996524e-03 -1.31878352e-02 -1.87352193e-03  1.15344333e-02
 -6.67101438e-02 -1.08264123e-02 -3.34118179e-02  3.20375403e-02
 -2.80501644e-01 -1.15783095e-01 -1.10934722e-02  1.66592997e-02
 -5.23672905e-01  6.20822931e-02 -3.08137237e-01 -1.11399282e-01
 -2.35793612e-02  5.34272770e-01 -2.84394121e-02 -1.06765585e-01
 -3.36472851e-02  3.34083511e-02 -7.50044480e-02  4.80844206e-02]
[ 4.05008149e-02 -3.74157457e-01  2.44816535e-01  3.53543343e-01
 -1.08721270e-02  5.52540916e-03 -2.04238684e-02  4.39291360e-02
 -2.71613881e-02 -2.29883522e-02 -1.11088001e-02 -9.34054683e-03
 -6.84874239e-02 -2.23894606e-01 -5.77746724e-02  4.47338482e-02
  4.26087628e-02 -1.54534798e-01  5.53506422e-02 -9.20877379e-03
  5.02076458e-01  2.78723552e-01  4.93043611e-01 -4.82309761e-02
  5.67399958e-02  1.75501880e-02  3.60973721e-02 -2.11122100e-02]
[-9.93858469e-02  1.97997396e-01 -3.25303343e-01  2.01174958e-01
  3.07475851e-03  6.21375539e-04 -9.60066161e-03  7.87856402e-02
  1.34886659e-01  6.64917859e-03 -3.35300464e-02  2.52981178e-02
  1.07745883e-01 -3.32507961e-01  4.55941842e-02 -2.98626929e-02
  3.72807346e-01  4.78115615e-02 -1.67872871e-01  4.04275878e-01
 -1.89443241e-01  4.33112224e-02  1.69798106e-01 -4.72322557e-01
  1.31149056e-01  6.47279219e-02 -9.06824345e-02  3.13564240e-02]
[-1.93497931e-01 -1.18157796e-01 -4.15475635e-01  2.42673571e-01
  4.33289294e-03  5.04418194e-03 -8.88740214e-03 -2.81021828e-02
  6.86441471e-02 -7.72891297e-03 -3.12730228e-02 -1.80198409e-02
```

```
  1.81301909e-01  5.07087855e-01 -1.89792641e-02  2.40519440e-03
 -1.48483791e-01  4.91910732e-02 -4.48872588e-02 -2.39433787e-01
 -2.69292129e-01  4.47144614e-02  4.83384279e-01  8.72235623e-02
  1.15390139e-01  1.12409970e-02  2.75634604e-02 -7.69896474e-02]
[-3.52822676e-01 -1.83278287e-01  9.25623702e-02  1.69714928e-01
 -4.64230943e-03  8.31225385e-03  5.77467153e-03  3.46883970e-02
  6.63970621e-02  4.45918101e-04  1.93637022e-02 -2.67655766e-03
  3.86948021e-01  2.05606350e-02  1.63539340e-02 -4.22963759e-02
 -9.28922025e-02 -2.71131921e-01  2.09304356e-01  8.68478550e-03
 -1.48872418e-01  5.18276429e-01 -4.64426452e-01 -1.03788079e-01
  3.62180614e-02  3.03969394e-02 -1.88801555e-02 -1.93241606e-02]
[ 1.60608429e-01 -1.67092810e-03  7.63374290e-03  4.19751944e-01
 -9.35185785e-03 -3.63348449e-03 -1.65133766e-02  1.09842192e-01
  1.44282738e-02 -1.85065912e-02  3.94454654e-03 -1.96891930e-02
 -2.12838753e-01  3.81776647e-01  2.77390219e-02  8.53439664e-03
  3.38835671e-01  5.38902304e-01  8.32588798e-02  5.64736915e-02
  1.64405487e-01  1.73152663e-01 -3.24668030e-01  4.43657978e-02
  1.05900526e-01 -2.35952945e-02  1.44767718e-02  3.07744389e-02]
[-1.33586541e-02  4.39934501e-01  2.27325296e-01  3.93917748e-01
  2.17488087e-03  5.27527590e-03 -1.41997983e-03 -4.36834574e-02
 -2.66620899e-02 -1.69639053e-04 -3.02417744e-02  2.10424307e-02
  1.83091524e-02 -1.86944447e-01 -1.15994199e-02 -8.70660219e-03
 -1.14019213e-01 -1.23997562e-01  1.98660381e-01  1.17613071e-01
 -1.89120904e-01 -1.23466901e-01  6.25168998e-02  4.49867598e-01
  4.61791449e-01 -5.20792332e-02  5.57317385e-02  4.33837316e-02]
[-1.38519639e-01  4.05628857e-01  6.60301473e-02 -2.42990896e-01
 -1.68074003e-02  1.23139126e-02 -8.71280938e-03  3.99458269e-02
 -6.90027638e-02 -1.59261364e-02 -1.01383964e-02  3.90254363e-02
  1.30456063e-01  3.28265108e-01  1.79189124e-02 -4.11001685e-03
  3.65363826e-02 -1.36523314e-02  2.05254003e-01  4.78463828e-01
  2.18239269e-01  2.87948372e-01  2.72808780e-01  1.42307116e-01
 -3.24069208e-01 -9.47973011e-02  9.36235444e-03  8.43799984e-02]
[-2.68217204e-01  1.11229300e-01  1.19994324e-01  3.58477652e-02
 -7.93266615e-03  1.71896137e-03 -9.91812236e-03  1.77862315e-01
  1.41591007e-01 -4.85252453e-03  1.27634977e-02  7.27529665e-03
  2.82902851e-01  4.06554205e-02  6.77272274e-03  7.76741885e-03
 -4.90989381e-01  3.36092001e-01 -5.27837037e-02  3.88816933e-02
  3.98327885e-01 -3.14059990e-01 -5.14594047e-02 -3.19218518e-01
  1.94485956e-01 -7.72903790e-02 -7.47280016e-03  4.57829280e-02]
[-3.68413352e-02  2.89145881e-01 -4.67249173e-01  1.96893397e-01
 -5.31649534e-03 -4.28778518e-03  2.25225273e-03  3.04149709e-04
 -5.89709335e-02 -7.05516408e-03 -1.28996895e-02 -7.21561013e-03
  5.31010159e-02 -2.37671688e-01 -7.28986791e-03  1.18258031e-02
  4.44778680e-02 -4.83897384e-02  4.85243522e-01 -3.64238191e-01
  2.97726526e-01 -1.10817049e-01 -8.77588300e-02 -5.25405657e-03
 -3.13605051e-01  9.67483314e-02  4.30352632e-02 -6.31561713e-02]
[ 9.05872061e-02  3.30552658e-02  1.22378891e-02 -8.62510219e-02
 -2.86247803e-03 -8.45147173e-03  4.15859398e-03  6.63083193e-01
```

```
  6.65551884e-01  8.66854328e-05 -2.54263477e-02 -2.32260480e-02
 -1.06726983e-01 -2.35339929e-02 -4.30848141e-03 -1.20089131e-02
  3.68371537e-02 -1.32587272e-01  4.40750964e-02 -1.17325034e-01
 -5.00792844e-02  3.33426298e-02  3.89243847e-02  1.65444472e-01
 -6.29713519e-02 -9.73429741e-02 -7.87661991e-02  1.69582397e-02]
[-1.71901449e-02  3.37120987e-03  1.35881874e-02 -4.38604014e-02
 -5.11921170e-03 -5.99846535e-03 -7.43542678e-03  7.49931591e-02
  8.11953692e-02  1.09494137e-02 -1.59406549e-02  5.37312286e-04
  3.85320693e-03  2.28523587e-02  1.70616099e-03 -1.45539558e-02
 -1.03682860e-02  3.67148904e-03 -5.73033951e-02  1.17469206e-02
 -7.27109276e-03  8.44613005e-03 -5.66718173e-04  1.27754711e-02
  3.32493175e-03  6.08661728e-01  6.55396687e-01  4.24581175e-01]
[-1.89476811e-02  5.54168969e-03 -1.71545901e-02 -2.09122443e-02
 -3.67644321e-02  7.63865579e-02  6.90516863e-03 -1.46523761e-03
 -1.55066957e-02 -6.19883264e-02 -6.27627427e-02 -3.62533190e-02
 -3.24910013e-02  7.43791447e-04 -7.01034170e-01 -6.97991139e-01
  1.34502944e-02  2.94379428e-02  3.42698143e-03  7.11778386e-03
  1.62355778e-02  5.01942479e-04 -1.04913253e-02 -1.30740139e-02
  1.31849811e-02 -1.76099023e-02 -8.85909376e-03  1.97963222e-02]
[-2.02477906e-03  1.46618933e-02  2.52540771e-02 -2.56221203e-02
  3.90977265e-02 -5.02354291e-02 -1.21298891e-02 -3.35560844e-02
 -1.17741058e-02  2.60387704e-02 -6.97276439e-01 -7.07197866e-01
  2.38194271e-02 -3.92037411e-03  4.43794126e-02  4.54856544e-02
 -2.79191636e-03  2.18509502e-04 -8.21322002e-03  1.62866266e-02
  1.37424361e-02  6.35343679e-03 -2.12748772e-02 -1.05734800e-02
 -4.45422148e-03 -1.50366819e-02  2.36771119e-03 -3.79979252e-03]
[ 8.35505599e-03  6.32740244e-03  6.76742494e-03 -4.41313391e-03
  4.01875325e-01  3.84501224e-01 -8.71565169e-03  4.42297858e-03
  3.46986310e-03 -8.26549099e-01 -2.65549023e-02 -2.91666957e-03
  3.14291791e-03  2.65264162e-03  3.46247215e-02  6.04383509e-02
  1.26574486e-03 -1.84825076e-02 -8.20672074e-03 -2.20624245e-03
 -1.25677924e-02 -1.77280282e-02 -1.44873328e-02 -1.25328720e-02
 -4.67241643e-03  1.78116680e-02  6.30250903e-04  3.54848065e-03]
[ 5.20772361e-01 -2.99369503e-03 -2.40398119e-03 -8.99785950e-04
 -1.35961368e-02  5.43211427e-03 -7.08360648e-01  9.01928724e-03
  1.73139899e-04  7.15062013e-03  2.60115176e-02 -3.42325613e-03
  4.74002207e-01  1.58383652e-02 -2.30448336e-02 -2.15261355e-02
  1.58735059e-03 -3.27384369e-03  2.77192719e-03  2.33776895e-04
 -2.18910758e-03  1.33803448e-03 -6.47882396e-04  1.76441288e-03
  2.58065009e-03 -1.55138753e-03  1.48100003e-03  2.20536099e-03]
[ 3.50529730e-09  7.22039805e-10  3.21327252e-10 -1.90607019e-10
 -6.97566206e-01  7.05546888e-01 -5.91468235e-09  8.75808728e-11
  2.47578440e-10  2.19108572e-09 -8.28887163e-02  9.29960992e-11
 -1.78705308e-02  4.05685308e-10  5.47377496e-02  6.65911248e-02
  6.05747279e-11 -1.11582552e-02 -1.08516073e-02 -1.43790396e-02
 -1.35587225e-02 -1.30152800e-02 -9.57973436e-03 -9.51074934e-03
 -5.44887871e-10  2.47588205e-10  7.10860208e-10  7.75015187e-10]]
```

0.3.5 Plotting Standardized Data

[37]:
```
# Takes a bit
# sns.pairplot(pd.DataFrame(std_df))
# plt.show()
```

0.3.6 PCA in a Pipeline

[38]:
```
from sklearn.pipeline import Pipeline

pipe = Pipeline([
    ('scaler', StandardScaler()),
    ('reducer', PCA())])

pc = pipe.fit_transform(pd.DataFrame(std_df))

print(pc[:,:2])
```

```
[[-1.22345671  1.38470718]
 [-1.07180686  1.05141455]
 [-1.94004422  1.10048864]
 …
 [ 2.26175452  0.44569753]
 [ 1.97986293 -1.19281366]
 [ 2.16046478 -0.46757884]]
```

0.3.7 PCA In a Model Pipeline

[39]:
```
from sklearn.ensemble import RandomForestClassifier
pipe = Pipeline([
    ('scaler', StandardScaler()),
    ('reducer', PCA(n_components=3)),
    ('classifier', RandomForestClassifier())])
print(pipe['reducer'])
```

```
PCA(n_components=3)
```

[40]:
```
pipe.fit(X_train, y_train)
pipe['reducer'].explained_variance_ratio_
```

[40]:
```
array([0.07576005, 0.07235497, 0.05732626])
```

[41]:
```
pipe['reducer'].explained_variance_ratio_.sum()
```

[41]:
```
0.20544128485576596
```

[42]:
```
print(pipe.score(X_test, y_test))
```

0.9566666666666667

0.3.8 Setting an Explained Variance Threshold

[43]:
```
pipe = Pipeline([
    ('scaler', StandardScaler()),
    ('reducer', PCA(n_components=0.9))])
# Fit the pipe to the data
pipe.fit(pd.DataFrame(std_df))

print(len(pipe['reducer'].components_))
```

21

0.3.9 Optimal Number of Components

[44]:
```
pipe.fit(pd.DataFrame(std_df))

var = pipe['reducer'].explained_variance_ratio_

plt.plot(var)

plt.xlabel('Principal Component Index')
plt.ylabel('Explained Variance Ratio')
plt.show()
```