

In [2]:

```

import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA

pd.options.mode.chained_assignment = None # default='warn' ---- ignores false warning

# write csv into datafile
df = pd.read_csv('medical_clean.csv')
#print(df.head())

newdf = df[['Income', 'VitD_levels', 'HighBlood', 'Stroke', 'Overweight', 'Arthritis',
]].copy()
print(newdf.head())

print(newdf.isnull().sum())

di = {'Yes': 1, 'No': 0}
newdf = newdf.replace({'HighBlood': di, 'Stroke': di, 'Overweight': di, 'Arthritis': di, 'D
print(newdf.head())

# PCA analysis
dataTest = newdf[['Income', 'VitD_levels', 'HighBlood', 'Stroke', 'Overweight', 'Arthri
pcsNames = []
for i, col in enumerate(dataTest.columns):
    pcsNames.append('PC'+str(i+1))

print(pcsNames)

normal = (dataTest - dataTest.mean())/(dataTest.std())
normal.to_csv('standardized.csv')
print(normal)

pca = PCA(n_components=normal.shape[1])
pca.fit(normal)
data_pca = pd.DataFrame(pca.transform(dataTest), columns = pcsNames)
print(data_pca)

for pc, var in zip(pcsNames, np.cumsum(pca.explained_variance_ratio_)):
    print(pc, var)

rotation = pd.DataFrame(pca.components_.T, columns = pcsNames, index = dataTest.columns
rotation.to_csv('PCA_matrix.csv')
print(rotation)
#print(rotation.head())
print(rotation.sum(axis=0)) # Total variance in principal component analysis. Added up

PC_values = np.arange(pca.n_components_) + 1
plt.plot(PC_values, pca.explained_variance_ratio_, 'o-', linewidth=2, color='blue')
plt.title('Scree Plot')
plt.xlabel('Principal Component')
plt.ylabel('Variance Explained')
plt.savefig('scree.jpg')
plt.close()

```

	Income	VitD_levels	HighBlood	Stroke	Overweight	Arthritis	Diabetes	\
0	86575.93	19.141466	Yes	No	No	Yes	Yes	

1	46805.99	18.940352	Yes	No	Yes	No	No
2	14370.14	18.057507	Yes	No	Yes	No	Yes
3	39741.49	16.576858	No	Yes	No	Yes	No
4	1209.56	17.439069	No	No	No	No	No

	Hyperlipidemia	BackPain	Anxiety	Allergic_rhinitis	Reflux_esophagitis	Asthma
0	No	Yes	Yes	Yes	No	Yes
1	No	No	No	No	Yes	No
2	No	No	No	No	No	No
3	No	No	No	No	Yes	Yes
4	Yes	No	No	Yes	No	No

Income	0
VitD_levels	0
HighBlood	0
Stroke	0
Overweight	0
Arthritis	0
Diabetes	0
Hyperlipidemia	0
BackPain	0
Anxiety	0
Allergic_rhinitis	0
Reflux_esophagitis	0
Asthma	0

dtype: int64

	Income	VitD_levels	HighBlood	Stroke	Overweight	Arthritis	Diabetes	\
0	86575.93	19.141466	1	0	0	1	1	
1	46805.99	18.940352	1	0	1	0	0	
2	14370.14	18.057507	1	0	1	0	1	
3	39741.49	16.576858	0	1	0	1	0	
4	1209.56	17.439069	0	0	0	0	0	

	Hyperlipidemia	BackPain	Anxiety	Allergic_rhinitis	Reflux_esophagitis	\
0	0	1	1	1	0	
1	0	0	0	0	1	
2	0	0	0	0	0	
3	0	0	0	0	1	
4	1	0	0	1	0	

Asthma

0	1
1	0
2	0
3	1
4	0

['PC1', 'PC2', 'PC3', 'PC4', 'PC5', 'PC6', 'PC7', 'PC8', 'PC9', 'PC10', 'PC11', 'PC12', 'PC13']

	Income	VitD_levels	HighBlood	Stroke	Overweight	Arthritis	\
0	1.615833	0.583574	1.202016	-0.498881	-1.562342	1.340823	
1	0.221432	0.483876	1.202016	-0.498881	0.640001	-0.745736	
2	-0.915824	0.046225	1.202016	-0.498881	0.640001	-0.745736	
3	-0.026261	-0.687776	-0.831852	2.004285	-1.562342	1.340823	
4	-1.377256	-0.260353	-0.831852	-0.498881	-1.562342	-0.745736	
...	
9995	0.192037	-0.487501	1.202016	-0.498881	-1.562342	-0.745736	
9996	-0.894335	0.105470	1.202016	-0.498881	0.640001	1.340823	
9997	0.891525	-0.414029	1.202016	-0.498881	0.640001	-0.745736	
9998	-0.378252	0.964772	-0.831852	-0.498881	0.640001	-0.745736	
9999	0.778094	0.210367	-0.831852	-0.498881	0.640001	1.340823	

	Diabetes	Hyperlipidemia	BackPain	Anxiety	Allergic_rhinitis	\
0	1.628507	-0.713232	1.196069	1.452656	1.239868	
1	-0.613998	-0.713232	-0.835988	-0.688325	-0.806457	
2	1.628507	-0.713232	-0.835988	-0.688325	-0.806457	
3	-0.613998	-0.713232	-0.835988	-0.688325	-0.806457	

4	-0.613998	1.401928	-0.835988	-0.688325	1.239868
...
9995	-0.613998	-0.713232	-0.835988	1.452656	-0.806457
9996	1.628507	-0.713232	-0.835988	-0.688325	-0.806457
9997	-0.613998	-0.713232	-0.835988	1.452656	1.239868
9998	-0.613998	-0.713232	1.196069	-0.688325	-0.806457
9999	-0.613998	1.401928	-0.835988	-0.688325	1.239868

	Reflux_esophagitis	Asthma
0	-0.839619	1.567282
1	1.190898	-0.637983
2	-0.839619	-0.637983
3	1.190898	1.567282
4	-0.839619	-0.637983
...
9995	1.190898	-0.637983
9996	-0.839619	1.567282
9997	-0.839619	-0.637983
9998	-0.839619	-0.637983
9999	-0.839619	-0.637983

[10000 rows x 13 columns]

	PC1	PC2	PC3	PC4	PC5 \
0	26522.412938	32761.103303	-17745.950333	-9703.825183	5888.753006
1	14335.371754	17712.222199	-9590.759455	-5249.505389	3184.017485
2	4395.718103	5437.560003	-2940.855886	-1614.887274	979.592867
3	12171.920585	15038.864103	-8143.331321	-4457.185668	2704.274619
4	363.821270	457.470010	-242.760733	-141.004484	85.446942
...
9995	14079.507196	17394.888164	-9419.933235	-5155.367368	3127.650502
9996	4583.211086	5669.263436	-3067.163274	-1683.437029	1021.319870
9997	20192.256265	24943.909540	-13510.098336	-7389.654971	4483.584139
9998	9093.683639	11240.215229	-6083.538001	-3333.598392	2022.370286
9999	19201.244581	23719.108049	-12846.137963	-7027.628308	4263.323207

	PC6	PC7	PC8	PC9	PC10 \
0	7809.797188	-11048.373083	52968.117664	18239.567221	-27201.101005
1	4222.117614	-5971.340541	28638.466451	9863.344321	-14701.932149
2	1295.945161	-1831.118407	8795.542333	3031.357804	-4507.744630
3	3585.788837	-5069.345896	24315.990259	8375.493103	-12482.165330
4	108.315191	-152.226609	744.194103	260.064218	-370.687706
...
9995	4146.713479	-5864.732916	28125.201681	9685.627942	-14439.088022
9996	1351.347020	-1908.678394	9170.692879	3161.019358	-4700.241490
9997	5945.955139	-8412.301043	40330.053318	13887.849230	-20709.751110
9998	2678.495219	-3788.148937	18174.692815	6261.557472	-9325.813616
9999	5653.772376	-7998.835532	38351.088977	13207.772681	-19691.817867

	PC11	PC12	PC13
0	26202.657950	10736.465773	20292.602217
1	14165.976273	5806.804365	10973.889261
2	4349.987831	1787.179131	3372.896569
3	12028.539708	4930.141418	9316.696950
4	367.132237	155.165101	288.002648
...
9995	13911.954501	5702.189219	10776.746489
9996	4535.749109	1863.091464	3515.830460
9997	19950.260564	8174.521252	15451.977817
9998	8990.793590	3687.759835	6965.822662
9999	18972.100163	7773.938863	14693.591220

[10000 rows x 13 columns]

PC1 0.08178216061897818
 PC2 0.16203381565930702
 PC3 0.2416694121235673

PC4 0.3208133797227713
 PC5 0.39895685065519887
 PC6 0.4765973081090132
 PC7 0.5533973957615896
 PC8 0.6295754306950531
 PC9 0.7052162183054738
 PC10 0.7801168455252615
 PC11 0.854210275749585
 PC12 0.9272710379103936
 PC13 1.0

	PC1	PC2	PC3	PC4	PC5	\
Income	0.306448	0.378410	-0.205021	-0.112021	0.067971	
VitD_levels	-0.404637	-0.002931	0.314193	-0.332087	0.197455	
HighBlood	-0.399148	0.104290	-0.210879	0.049559	-0.366709	
Stroke	-0.083427	0.281594	0.318053	0.272694	0.029718	
Overweight	-0.456725	0.050338	0.040764	0.202643	-0.464484	
Arthritis	-0.063670	-0.421179	-0.419038	-0.263374	-0.243198	
Diabetes	0.173372	-0.343394	-0.172530	0.608320	0.103289	
Hyperlipidemia	0.384541	-0.086776	0.088298	0.125344	-0.193065	
BackPain	-0.033502	0.574501	-0.222087	0.081489	0.001927	
Anxiety	-0.193902	-0.038238	-0.403231	-0.225855	0.541115	
Allergic_rhinitis	-0.174581	-0.101449	-0.342537	0.157044	-0.018637	
Reflux_esophagitis	0.267817	0.278121	-0.308827	-0.234516	-0.349070	
Asthma	-0.211123	0.205169	-0.272361	0.408469	0.288433	

	PC6	PC7	PC8	PC9	PC10	\
Income	0.090211	-0.127652	0.611759	0.210616	-0.314307	
VitD_levels	-0.020786	0.166320	0.224992	0.292001	0.496173	
HighBlood	0.111817	-0.084928	0.446545	-0.472049	0.050223	
Stroke	0.619829	0.000518	0.048140	-0.187376	0.247815	
Overweight	-0.295997	0.150226	-0.036779	0.093597	-0.243600	
Arthritis	0.208759	0.239033	0.013037	0.127555	0.060219	
Diabetes	0.155993	0.189098	0.007021	-0.117470	0.125091	
Hyperlipidemia	-0.513335	0.018013	0.362168	-0.137776	0.522531	
BackPain	-0.279159	-0.052149	-0.412235	-0.132812	0.198883	
Anxiety	-0.109530	0.062927	0.025848	-0.474395	0.064048	
Allergic_rhinitis	0.075054	-0.742416	-0.050961	0.357126	0.310176	
Reflux_esophagitis	0.267339	0.319027	-0.191259	0.103364	0.315135	
Asthma	-0.093104	0.416418	0.175752	0.416787	0.016113	

	PC11	PC12	PC13
Income	0.302638	0.123935	0.234335
VitD_levels	0.055315	0.324644	0.263915
HighBlood	-0.348503	0.247748	-0.145252
Stroke	0.374018	-0.314010	-0.125793
Overweight	0.314178	-0.288382	0.410960
Arthritis	0.498975	0.091190	-0.369291
Diabetes	0.069298	0.394732	0.437413
Hyperlipidemia	0.138310	-0.245616	-0.131668
BackPain	0.319954	0.428474	-0.138268
Anxiety	0.126650	-0.362795	0.247167
Allergic_rhinitis	-0.029838	-0.157542	0.089549
Reflux_esophagitis	-0.319935	-0.202532	0.339855
Asthma	-0.225237	-0.174917	-0.347639

PC1 -0.888536
 PC2 0.878456
 PC3 -1.795204
 PC4 0.737709
 PC5 -0.405254
 PC6 0.217091
 PC7 0.554436
 PC8 1.224028
 PC9 0.079169
 PC10 1.848499
 PC11 1.275822

```
PC12    -0.135071  
PC13     0.765284  
dtype: float64
```

In []: