CS306 - Data Analysis and Visualization

March 24, 2021

Riddhi Tanna

201801427

1 Importing the required libraries and dataset

[302]:
```
#importing the required libraries

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

[303]:
```
#reading the dataset
df = pd.read_excel('New_York_Neighborhoods.xlsx')
```

[304]:
```
df = df[1:]
df.head()
```

[304]:
```
  The Most Livable Neighborhoods in New
York\nhttp://nymag.com/realestate/neighborhoods/2010/65374/index10.html  \
1                                       Neighborhood
2                                         Park Slope
3                                    Lower East Side
4                                          Sunnyside
5                           Cobble Hill & Boerum Hill

      Unnamed: 1 Unnamed: 2           Unnamed: 3 Unnamed: 4 Unnamed: 5  \
1  Affordability    Transit  Shopping & Services      Crime       Food
2             73         76                   77         82         83
3             73         82                   83         75         83
4             83         76                   77         81         73
5             73         77                   83         76         87

  Unnamed: 6 Unnamed: 7 Unnamed: 8       Unnamed: 9  … Unnamed: 16374  \
1    Schools  Diversity   Creative  Housing Quality  …            NaN
```

```
2         81        73        83                81  …             NaN
3         76        78        84                72  …             NaN
4         80        90        72                72  …             NaN
5         77        71        81                83  …             NaN

  Unnamed: 16375 Unnamed: 16376 Unnamed: 16377 Unnamed: 16378  Unnamed: 16379  \
1            NaN            NaN            NaN            NaN             NaN
2            NaN            NaN            NaN            NaN             NaN
3            NaN            NaN            NaN            NaN             NaN
4            NaN            NaN            NaN            NaN             NaN
5            NaN            NaN            NaN            NaN             NaN

   Unnamed: 16380  Unnamed: 16381  Unnamed: 16382  Unnamed: 16383
1             NaN             NaN             NaN             NaN
2             NaN             NaN             NaN             NaN
3             NaN             NaN             NaN             NaN
4             NaN             NaN             NaN             NaN
5             NaN             NaN             NaN             NaN

[5 rows x 16384 columns]
```

[305]:
```
df = df[df.columns[:13]]
cols = np.array(df.iloc[0])
```

[306]:
```
df.columns = cols
```

[307]:
```
df = df.iloc[1:]
df.index=df['Neighborhood']
df = df[df.columns[1:]]
df
```

[307]:
```
                                       Affordability Transit  \
Neighborhood
Park Slope                                        73      76
Lower East Side                                   73      82
Sunnyside                                         83      76
Cobble Hill & Boerum Hill                         73      77
Greenpoint                                        77      76
Brooklyn Heights                                  70      83
Carroll Gardens & Gowanus                         74      78
Murray Hill                                       64      82
Prospect Heights                                  79      79
East Village                                      63      83
Astoria                                           78      74
Bay Ridge                                         83      70
Woodside                                          81      76
Tribeca                                           55      87
```

Jackson Heights	85	71
Long Island City	79	81
Midtown East	59	86
Fort Greene & Clinton Hill	78	78
Dumbo & Downtown Brooklyn	70	85
Williamsburg	76	78
Central Greenwich Village	56	89
Flushing	80	68
Battery Park City & Financial District	63	89
West Village	51	87
Flatiron & Gramercy	61	87
Chelsea	58	87
Sheepshead Bay	84	67
Soho	50	88
Nolita & Little Italy	53	87
Brighton Beach	79	67
Inwood	82	74
Corona Park	85	70
Red Hook	79	89
Midtown West	56	88
Upper East Side	60	81
Upper West Side	60	83
Washington Heights	70	76
Riverdale	82	70
Sunset Park	80	73
New Drop	85	60
West Brighton	87	55
Chinatown	66	81
St. George	88	56
Belmont	89	62
Co-op City	90	56
Morningside Heights	74	81
Roosevelt Island	71	79
Bedford Park	89	71
Parkchester	90	63
Harlem	76	78

	Shopping & Services	Crime	Food	Schools \
Neighborhood				
Park Slope	77	82	83	81
Lower East Side	83	75	83	76
Sunnyside	77	81	73	80
Cobble Hill & Boerum Hill	83	76	87	77
Greenpoint	75	78	81	92
Brooklyn Heights	82	88	78	72
Carroll Gardens & Gowanus	75	76	88	75
Murray Hill	85	87	80	88

Prospect Heights	76	73	79	71
East Village	88	75	92	69
Astoria	74	80	78	77
Bay Ridge	71	83	74	77
Woodside	70	80	75	77
Tribeca	85	88	89	86
Jackson Heights	69	77	74	77
Long Island City	79	64	83	65
Midtown East	90	87	85	88
Fort Greene & Clinton Hill	76	70	79	68
Dumbo & Downtown Brooklyn	86	60	77	63
Williamsburg	74	72	85	73
Central Greenwich Village	84	78	84	87
Flushing	72	82	77	77
Battery Park City & Financial District	79	86	72	85
West Village	83	79	94	84
Flatiron & Gramercy	84	82	78	79
Chelsea	85	75	75	91
Sheepshead Bay	66	80	72	83
Soho	84	84	92	85
Nolita & Little Italy	94	74	95	83
Brighton Beach	75	68	76	75
Inwood	76	64	69	72
Corona Park	68	71	67	80
Red Hook	79	62	89	54
Midtown West	89	78	77	75
Upper East Side	80	87	77	82
Upper West Side	76	85	71	79
Washington Heights	71	73	67	68
Riverdale	66	77	62	73
Sunset Park	66	78	69	69
New Drop	65	88	67	70
West Brighton	74	78	67	66
Chinatown	77	78	76	83
St. George	66	78	66	69
Belmont	68	57	78	74
Co-op City	65	80	62	75
Morningside Heights	77	63	68	70
Roosevelt Island	67	80	62	69
Bedford Park	64	67	62	71
Parkchester	66	75	65	66
Harlem	73	62	70	63

	Diversity	Creative	Housing Quality	\
Neighborhood				
Park Slope	73	83	81	
Lower East Side	78	84	72	

Sunnyside	90	72	72
Cobble Hill & Boerum Hill	71	81	83
Greenpoint	74	78	80
Brooklyn Heights	65	81	86
Carroll Gardens & Gowanus	75	82	76
Murray Hill	69	77	85
Prospect Heights	84	79	67
East Village	78	90	75
Astoria	83	74	76
Bay Ridge	80	70	77
Woodside	88	69	73
Tribeca	63	87	91
Jackson Heights	85	67	76
Long Island City	80	80	74
Midtown East	63	78	82
Fort Greene & Clinton Hill	81	81	77
Dumbo & Downtown Brooklyn	77	95	82
Williamsburg	74	82	75
Central Greenwich Village	64	91	88
Flushing	85	67	73
Battery Park City & Financial District	67	79	83
West Village	66	86	89
Flatiron & Gramercy	70	79	84
Chelsea	66	89	82
Sheepshead Bay	66	67	77
Soho	71	89	86
Nolita & Little Italy	73	80	73
Brighton Beach	79	68	78
Inwood	73	70	60
Corona Park	79	68	70
Red Hook	70	79	71
Midtown West	78	86	79
Upper East Side	63	77	85
Upper West Side	71	80	83
Washington Heights	76	72	62
Riverdale	76	70	75
Sunset Park	81	68	71
New Drop	69	65	79
West Brighton	81	71	70
Chinatown	76	68	71
St. George	84	72	70
Belmont	71	67	62
Co-op City	80	62	71
Morningside Heights	79	78	66
Roosevelt Island	83	70	77
Bedford Park	76	67	60
Parkchester	79	63	67

Harlem	75	75	73

Neighborhood	Green Space	Wellness	Nightlife
Park Slope	84	77	87
Lower East Side	76	73	92
Sunnyside	67	72	73
Cobble Hill & Boerum Hill	82	76	84
Greenpoint	72	70	87
Brooklyn Heights	86	80	72
Carroll Gardens & Gowanus	79	71	93
Murray Hill	84	80	77
Prospect Heights	75	74	85
East Village	74	74	99
Astoria	67	82	83
Bay Ridge	76	82	67
Woodside	66	77	79
Tribeca	84	75	78
Jackson Heights	69	79	72
Long Island City	69	72	76
Midtown East	74	81	75
Fort Greene & Clinton Hill	74	70	75
Dumbo & Downtown Brooklyn	83	73	74
Williamsburg	68	68	88
Central Greenwich Village	75	75	83
Flushing	70	82	66
Battery Park City & Financial District	81	74	70
West Village	87	74	91
Flatiron & Gramercy	77	78	75
Chelsea	77	70	77
Sheepshead Bay	73	79	65
Soho	79	72	80
Nolita & Little Italy	74	69	94
Brighton Beach	88	77	74
Inwood	87	75	85
Corona Park	74	80	66
Red Hook	84	69	88
Midtown West	77	68	74
Upper East Side	80	80	78
Upper West Side	86	77	76
Washington Heights	79	76	73
Riverdale	74	84	70
Sunset Park	70	73	68
New Drop	69	85	65
West Brighton	71	76	66
Chinatown	73	72	76
St. George	72	76	65

```
Belmont                                    77      72      71
Co-op City                                 67      78      65
Morningside Heights                        77      71      68
Roosevelt Island                           84      78      65
Bedford Park                               70      73      65
Parkchester                                65      74      65
Harlem                                     75      71      79
```

[308]:

```
df = df.astype(float)
df.info()
```

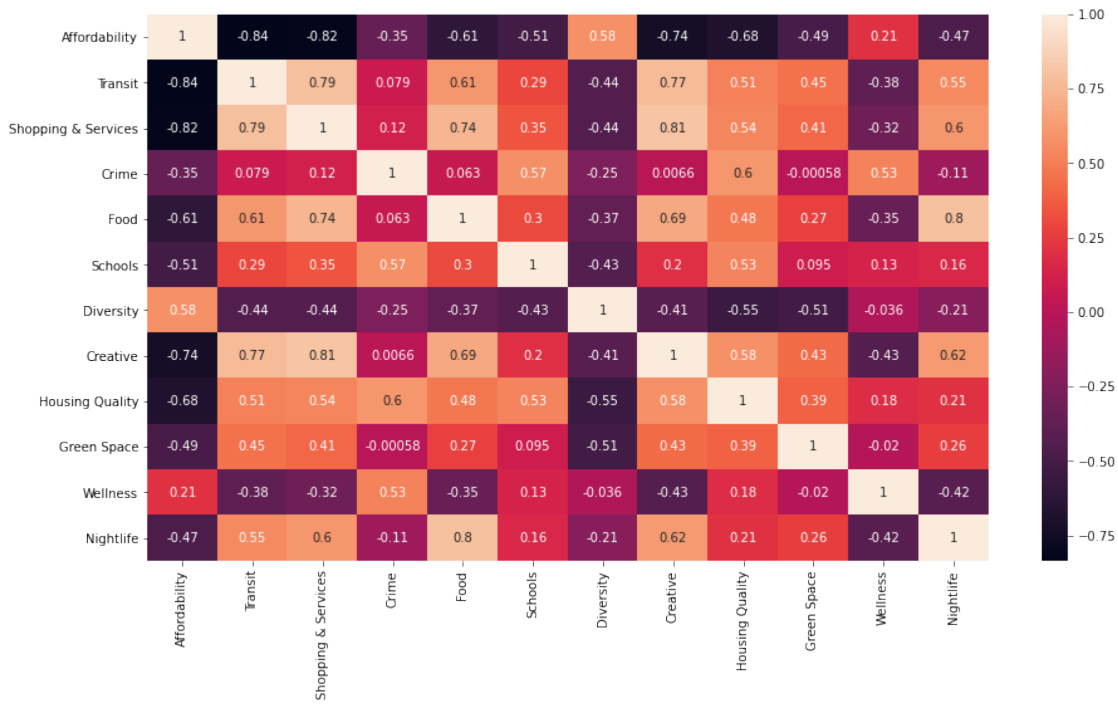
```
<class 'pandas.core.frame.DataFrame'>
Index: 50 entries, Park Slope to Harlem
Data columns (total 12 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Affordability        50 non-null     float64
 1   Transit              50 non-null     float64
 2   Shopping & Services  50 non-null     float64
 3   Crime                50 non-null     float64
 4   Food                 50 non-null     float64
 5   Schools              50 non-null     float64
 6   Diversity            50 non-null     float64
 7   Creative             50 non-null     float64
 8   Housing Quality      50 non-null     float64
 9   Green Space          50 non-null     float64
 10  Wellness             50 non-null     float64
 11  Nightlife            50 non-null     float64
dtypes: float64(12)
memory usage: 5.1+ KB
```

2 Pearson Correlation Matrix

[309]:

```
fig, ax = plt.subplots(figsize=[15,8])
sns.heatmap(df.corr(method='pearson'), annot=True)
```

[309]: <AxesSubplot:>



The Pearson Correlation Coefficient describes how two variables are related to each other. As we can see from the above heatmap, we have multiple variables which are positively correlated, multiple variables that are negatively correlated and some that are not at all related to each other. We can use Principle Component Analysis to fully visualize and make sense of this data. Principle Component Analysis helps us reduce the number of dimensions of a dataset. The reduced number of dimensions still include factors from all the dimensions that our original dataset contains.

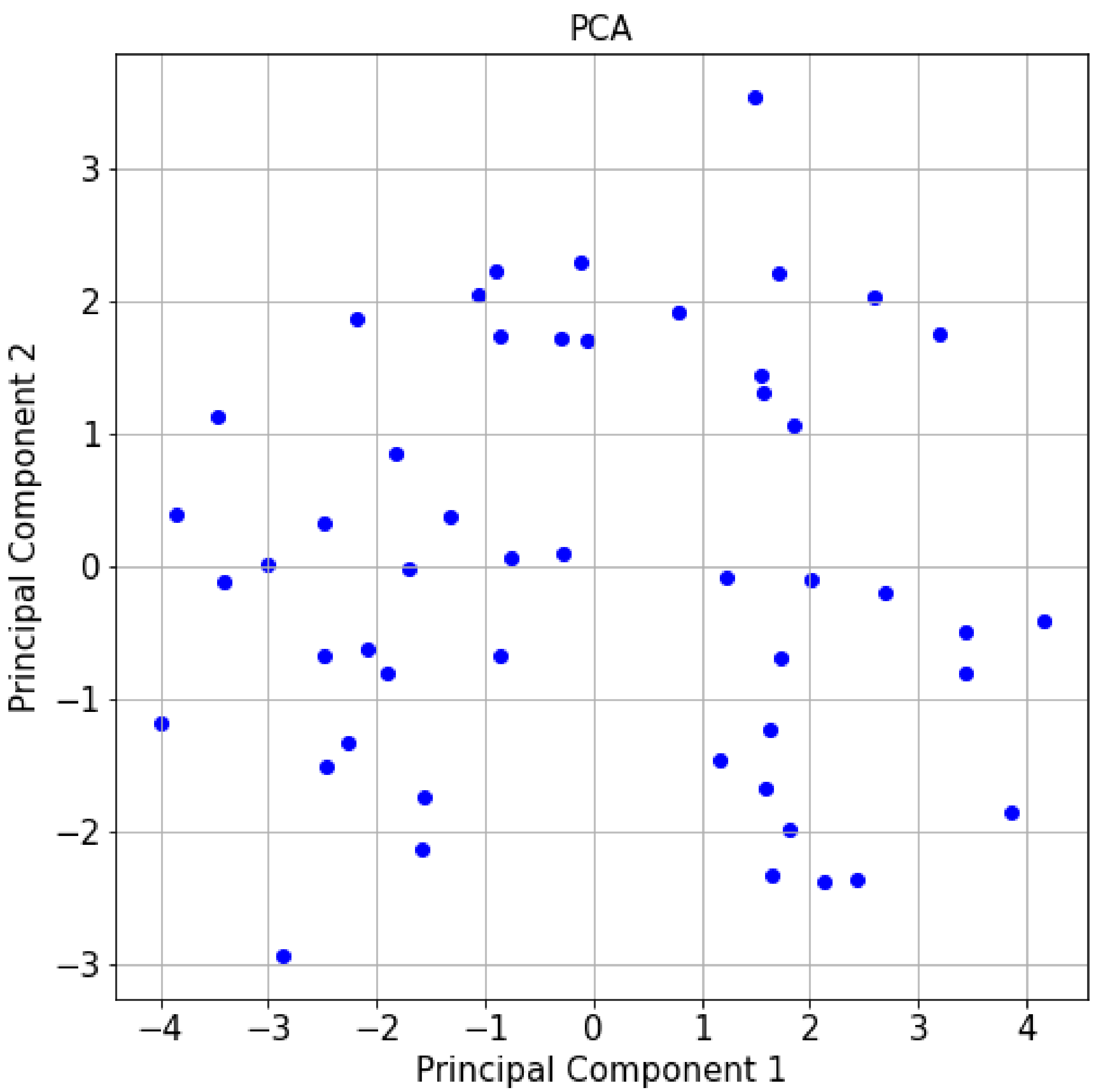
3 Principal Component Analysis

[310]:
```
from sklearn.preprocessing import StandardScaler
features = df.columns[1:]
# Separating out the features
x = df.loc[:, features].values
# Standardizing the features
x = StandardScaler().fit_transform(x)
pca = PCA(n_components=11)
principal_components = pca.fit_transform(x)
principal_df = pd.DataFrame(data = principal_components)
```

[311]:
```
print(pca.explained_variance_ratio_)
```

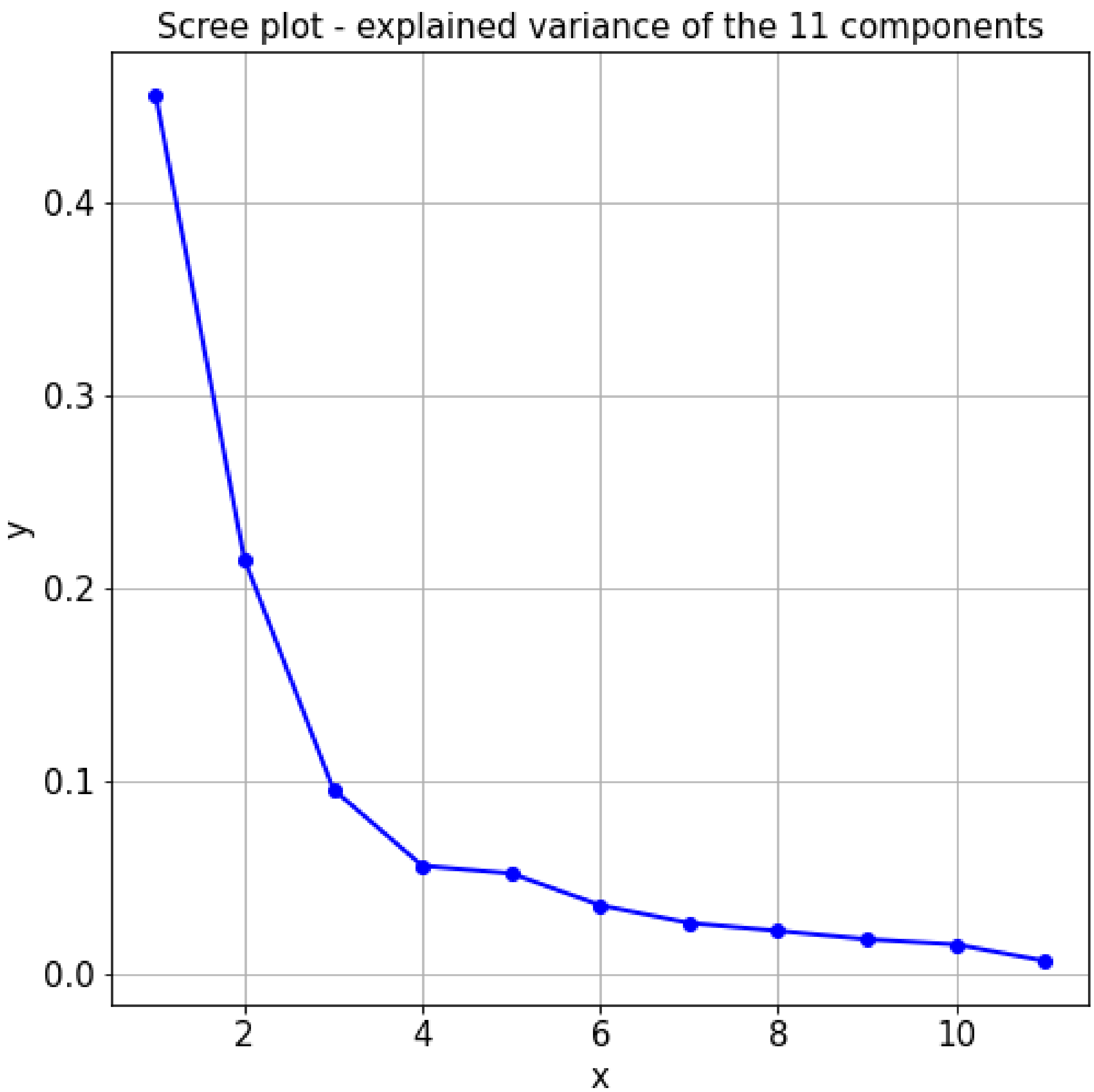
```
[0.45591043 0.21496742 0.09574508 0.05616709 0.05219889 0.03571284
 0.02654106 0.02243163 0.01809893 0.01538658 0.00684003]
```

[312]:
```
# plotting the scatter plot
fig, ax = plt.subplots(figsize=[8,8])

plt.scatter(principal_df[0], principal_df[1], color='b')

plt.grid()
plt.xlabel('Principal Component 1', fontsize = 15)
plt.ylabel('Principal Component 2', fontsize = 15)
plt.xticks(fontsize=15)
plt.yticks(fontsize=15)
plt.title('PCA', fontsize = 15)
#plt.legend()
plt.show()
```



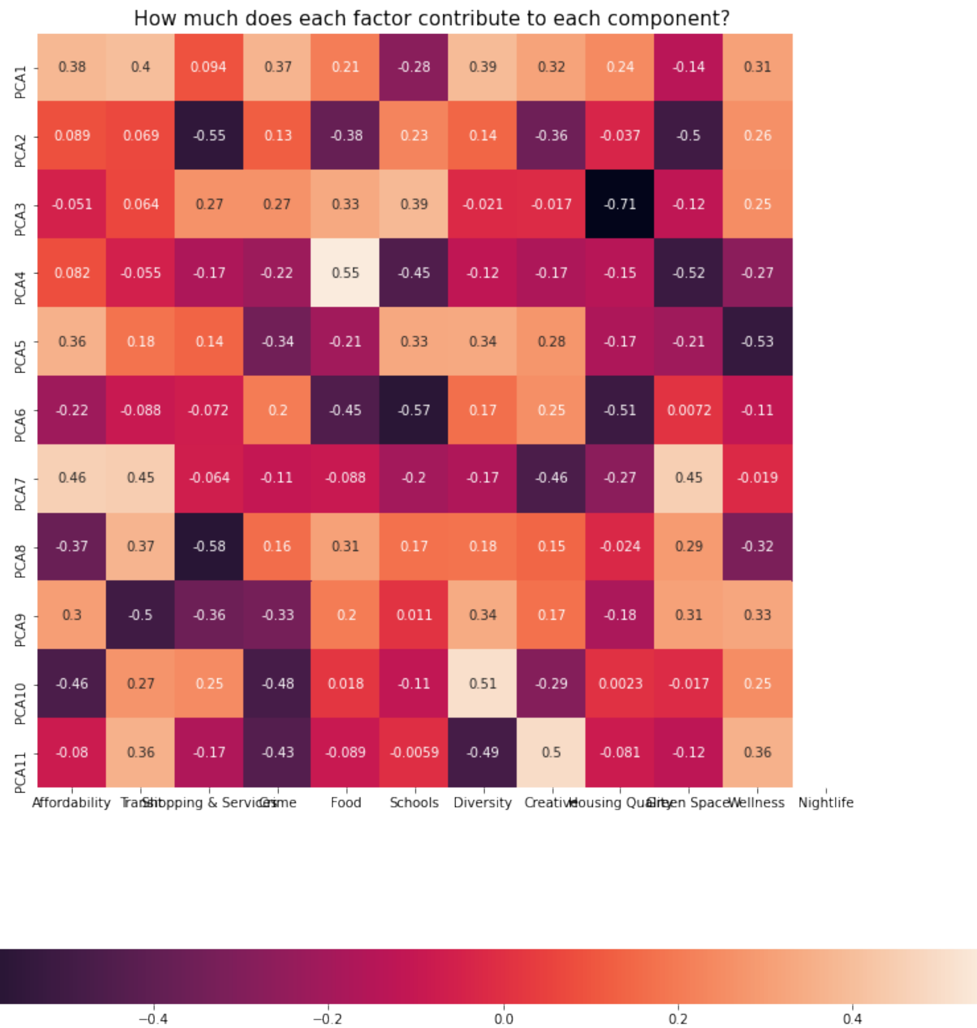
[313]:
```
fig, ax = plt.subplots(figsize=[8,8])

plt.plot(np.arange(1,12),pca.explained_variance_ratio_,'b-o', lw=2)
plt.grid()
plt.xlabel('x', fontsize = 15)
plt.ylabel('y', fontsize = 15)
plt.xticks(fontsize=15)
plt.yticks(fontsize=15)
plt.title('Scree plot - explained variance of the 11 components', fontsize = 15)
```

[313]:
```
Text(0.5, 1.0, 'Scree plot - explained variance of the 11 components')
```



[314]:

```python
fig, ax1 = plt.subplots(figsize=[15,15])
ax = sns.heatmap(pca.components_,
                 yticklabels=[ "PCA"+str(x) for x in range(1,pca.
 ↪n_components_+1)],
                 xticklabels=list(df.columns),
                 cbar_kws={"orientation": "horizontal"}, annot = True)
plt.title('How much does each factor contribute to each component?',␣
 ↪fontsize=15)
ax.set_aspect("equal")
```

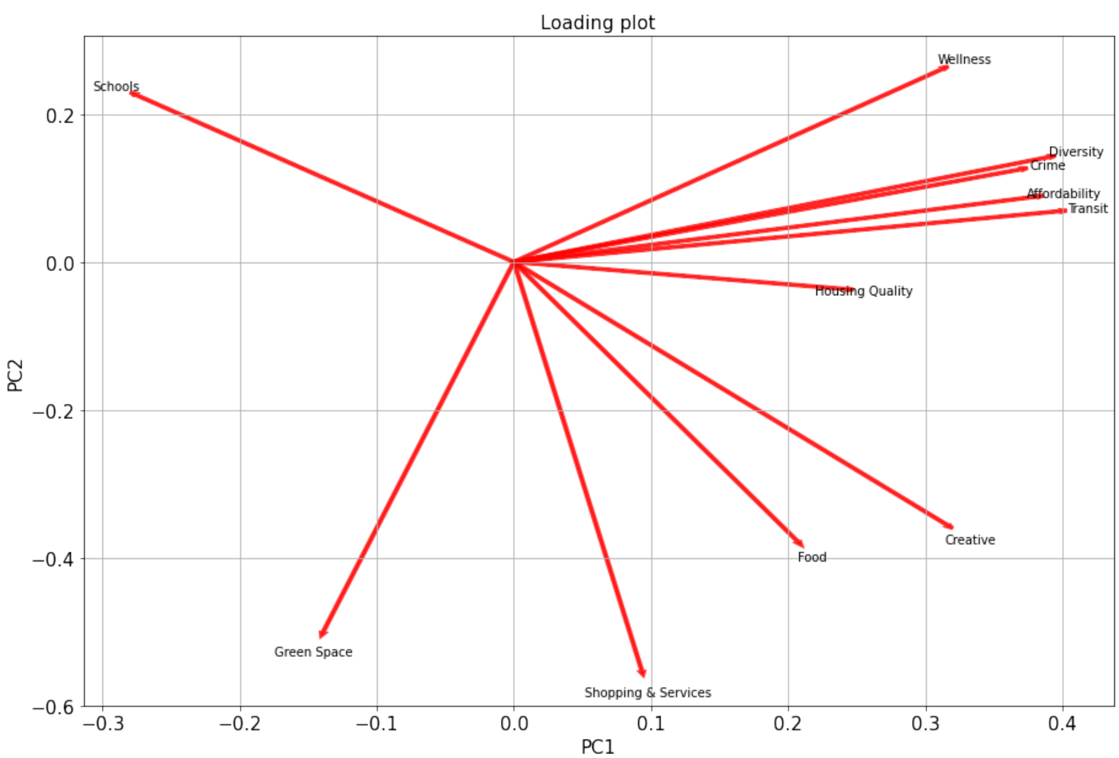


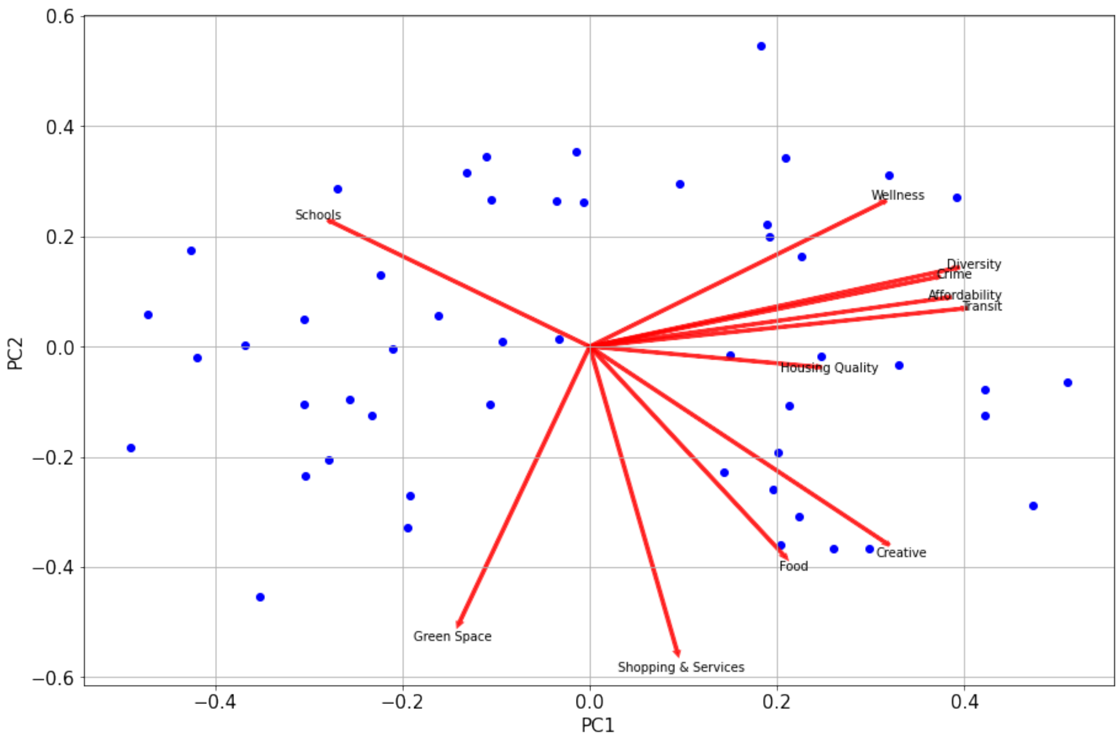
4 Plotting the biplot - score plot + loadings plot

[346]:
```
def loading_plot(coeff,labels=None):
    """
    score = projections of the points on the various principle components
    coeff = coefficients of the linear combinations for the principle␣
↪components
    """
    fig , ax = plt.subplots(figsize = [15,10])
    n = coeff.shape[0] #number of principle components
    for i in range(n):
        plt.arrow(0, 0, coeff[i,0], coeff[i,1],color = 'r',alpha = 0.8, lw=3)␣
↪#for the loading plot
        if labels is None:
            plt.text(coeff[i,0]*1.05, coeff[i,1]*1.05, "Var"+str(i+1), color =␣
↪'green', ha='center', va='center')
        else:
            plt.text(coeff[i,0]*1.05, coeff[i,1]*1.05, labels[i], color = 'k',␣
↪ha='center', va='center')

    plt.title('Loading plot', fontsize=15)
    plt.xlabel("PC{}".format(1), fontsize = 15)
    plt.ylabel("PC{}".format(2), fontsize = 15)
    plt.xticks(fontsize = 15)
    plt.yticks(fontsize = 15)
    plt.grid()
```

[347]:
```
def biplot(score,coeff,labels=None):
    """
    score = projections of the points on the various principle components
    coeff = coefficients of the linear combinations for the principle␣
↪components
    """
    xs = score[:,0] #x values - PC1
    ys = score[:,1] #y values - PC2
    n = coeff.shape[0] #number of principle components
    scalex = 1.0/(xs.max() - xs.min())
    scaley = 1.0/(ys.max() - ys.min())
    fig , ax = plt.subplots(figsize = [15,10])
    plt.scatter(xs*scalex,ys*scaley, color='b') #for the scatter plot -␣
↪plotting the scaled values
    for i in range(n):
        plt.arrow(0, 0, coeff[i,0], coeff[i,1],color = 'r',alpha = 0.8, lw=3)␣
↪#for the loading plot
        if labels is None:
```

```
            plt.text(coeff[i,0]*1.05, coeff[i,1]*1.05, "Var"+str(i+1), color =␣
 ↪'green', ha='center', va='center')
        else:
            plt.text(coeff[i,0]*1.05, coeff[i,1]*1.05, labels[i], color = 'k',␣
 ↪ha='center', va='center')

    plt.xlabel("PC{}".format(1), fontsize = 15)
    plt.ylabel("PC{}".format(2), fontsize = 15)
    plt.xticks(fontsize = 15)
    plt.yticks(fontsize = 15)
    plt.grid()

"""
principle_components[:,0:2] returns the projections of all the points on the␣
 ↪first two principle components - needed to
plot the scatter plot

pca.components_ returns the coefficients for the linear combinations
"""
loading_plot(np.transpose(pca.components_[0:2,:]),list(df_new.columns))
biplot(principle_components[:,0:2],np.transpose(pca.components_[0:2,:]),list(df.
 ↪columns))
```





5 Introducing two rows of outliers in the data

[342]:
```
df_new = df.copy()
df_new.loc[len(df_new)]=[70,70,700,80,83,71,600,70,65,900,45,800]
df_new.loc[len(df_new)]=[77,60,72,82,800,73,65,900,62,75,-500,80]
```

[317]:
```
df_new
```

[317]:
```
                                    Affordability  Transit  \
Neighborhood
Park Slope                                   73.0     76.0
Lower East Side                              73.0     82.0
Sunnyside                                    83.0     76.0
Cobble Hill & Boerum Hill                    73.0     77.0
Greenpoint                                   77.0     76.0
Brooklyn Heights                             70.0     83.0
Carroll Gardens & Gowanus                    74.0     78.0
Murray Hill                                  64.0     82.0
Prospect Heights                             79.0     79.0
East Village                                 63.0     83.0
Astoria                                      78.0     74.0
Bay Ridge                                    83.0     70.0
```

Woodside	81.0	76.0
Tribeca	55.0	87.0
Jackson Heights	85.0	71.0
Long Island City	79.0	81.0
Midtown East	59.0	86.0
Fort Greene & Clinton Hill	78.0	78.0
Dumbo & Downtown Brooklyn	70.0	85.0
Williamsburg	76.0	78.0
Central Greenwich Village	56.0	89.0
Flushing	80.0	68.0
Battery Park City & Financial District	63.0	89.0
West Village	51.0	87.0
Flatiron & Gramercy	61.0	87.0
Chelsea	58.0	87.0
Sheepshead Bay	84.0	67.0
Soho	50.0	88.0
Nolita & Little Italy	53.0	87.0
Brighton Beach	79.0	67.0
Inwood	82.0	74.0
Corona Park	85.0	70.0
Red Hook	79.0	89.0
Midtown West	56.0	88.0
Upper East Side	60.0	81.0
Upper West Side	60.0	83.0
Washington Heights	70.0	76.0
Riverdale	82.0	70.0
Sunset Park	80.0	73.0
New Drop	85.0	60.0
West Brighton	87.0	55.0
Chinatown	66.0	81.0
St. George	88.0	56.0
Belmont	89.0	62.0
Co-op City	90.0	56.0
Morningside Heights	74.0	81.0
Roosevelt Island	71.0	79.0
Bedford Park	89.0	71.0
Parkchester	90.0	63.0
Harlem	76.0	78.0
50	70.0	70.0
51	77.0	60.0

	Shopping & Services	Crime	Food \
Neighborhood			
Park Slope	77.0	82.0	83.0
Lower East Side	83.0	75.0	83.0
Sunnyside	77.0	81.0	73.0
Cobble Hill & Boerum Hill	83.0	76.0	87.0

Greenpoint	75.0	78.0	81.0
Brooklyn Heights	82.0	88.0	78.0
Carroll Gardens & Gowanus	75.0	76.0	88.0
Murray Hill	85.0	87.0	80.0
Prospect Heights	76.0	73.0	79.0
East Village	88.0	75.0	92.0
Astoria	74.0	80.0	78.0
Bay Ridge	71.0	83.0	74.0
Woodside	70.0	80.0	75.0
Tribeca	85.0	88.0	89.0
Jackson Heights	69.0	77.0	74.0
Long Island City	79.0	64.0	83.0
Midtown East	90.0	87.0	85.0
Fort Greene & Clinton Hill	76.0	70.0	79.0
Dumbo & Downtown Brooklyn	86.0	60.0	77.0
Williamsburg	74.0	72.0	85.0
Central Greenwich Village	84.0	78.0	84.0
Flushing	72.0	82.0	77.0
Battery Park City & Financial District	79.0	86.0	72.0
West Village	83.0	79.0	94.0
Flatiron & Gramercy	84.0	82.0	78.0
Chelsea	85.0	75.0	75.0
Sheepshead Bay	66.0	80.0	72.0
Soho	84.0	84.0	92.0
Nolita & Little Italy	94.0	74.0	95.0
Brighton Beach	75.0	68.0	76.0
Inwood	76.0	64.0	69.0
Corona Park	68.0	71.0	67.0
Red Hook	79.0	62.0	89.0
Midtown West	89.0	78.0	77.0
Upper East Side	80.0	87.0	77.0
Upper West Side	76.0	85.0	71.0
Washington Heights	71.0	73.0	67.0
Riverdale	66.0	77.0	62.0
Sunset Park	66.0	78.0	69.0
New Drop	65.0	88.0	67.0
West Brighton	74.0	78.0	67.0
Chinatown	77.0	78.0	76.0
St. George	66.0	78.0	66.0
Belmont	68.0	57.0	78.0
Co-op City	65.0	80.0	62.0
Morningside Heights	77.0	63.0	68.0
Roosevelt Island	67.0	80.0	62.0
Bedford Park	64.0	67.0	62.0
Parkchester	66.0	75.0	65.0
Harlem	73.0	62.0	70.0
50	700.0	80.0	83.0

51	72.0	82.0	800.0

	Schools	Diversity	Creative \
Neighborhood			
Park Slope	81.0	73.0	83.0
Lower East Side	76.0	78.0	84.0
Sunnyside	80.0	90.0	72.0
Cobble Hill & Boerum Hill	77.0	71.0	81.0
Greenpoint	92.0	74.0	78.0
Brooklyn Heights	72.0	65.0	81.0
Carroll Gardens & Gowanus	75.0	75.0	82.0
Murray Hill	88.0	69.0	77.0
Prospect Heights	71.0	84.0	79.0
East Village	69.0	78.0	90.0
Astoria	77.0	83.0	74.0
Bay Ridge	77.0	80.0	70.0
Woodside	77.0	88.0	69.0
Tribeca	86.0	63.0	87.0
Jackson Heights	77.0	85.0	67.0
Long Island City	65.0	80.0	80.0
Midtown East	88.0	63.0	78.0
Fort Greene & Clinton Hill	68.0	81.0	81.0
Dumbo & Downtown Brooklyn	63.0	77.0	95.0
Williamsburg	73.0	74.0	82.0
Central Greenwich Village	87.0	64.0	91.0
Flushing	77.0	85.0	67.0
Battery Park City & Financial District	85.0	67.0	79.0
West Village	84.0	66.0	86.0
Flatiron & Gramercy	79.0	70.0	79.0
Chelsea	91.0	66.0	89.0
Sheepshead Bay	83.0	66.0	67.0
Soho	85.0	71.0	89.0
Nolita & Little Italy	83.0	73.0	80.0
Brighton Beach	75.0	79.0	68.0
Inwood	72.0	73.0	70.0
Corona Park	80.0	79.0	68.0
Red Hook	54.0	70.0	79.0
Midtown West	75.0	78.0	86.0
Upper East Side	82.0	63.0	77.0
Upper West Side	79.0	71.0	80.0
Washington Heights	68.0	76.0	72.0
Riverdale	73.0	76.0	70.0
Sunset Park	69.0	81.0	68.0
New Drop	70.0	69.0	65.0
West Brighton	66.0	81.0	71.0
Chinatown	83.0	76.0	68.0
St. George	69.0	84.0	72.0

Belmont	74.0	71.0	67.0
Co-op City	75.0	80.0	62.0
Morningside Heights	70.0	79.0	78.0
Roosevelt Island	69.0	83.0	70.0
Bedford Park	71.0	76.0	67.0
Parkchester	66.0	79.0	63.0
Harlem	63.0	75.0	75.0
50	71.0	600.0	70.0
51	73.0	65.0	900.0

	Housing Quality	Green Space \
Neighborhood		
Park Slope	81.0	84.0
Lower East Side	72.0	76.0
Sunnyside	72.0	67.0
Cobble Hill & Boerum Hill	83.0	82.0
Greenpoint	80.0	72.0
Brooklyn Heights	86.0	86.0
Carroll Gardens & Gowanus	76.0	79.0
Murray Hill	85.0	84.0
Prospect Heights	67.0	75.0
East Village	75.0	74.0
Astoria	76.0	67.0
Bay Ridge	77.0	76.0
Woodside	73.0	66.0
Tribeca	91.0	84.0
Jackson Heights	76.0	69.0
Long Island City	74.0	69.0
Midtown East	82.0	74.0
Fort Greene & Clinton Hill	77.0	74.0
Dumbo & Downtown Brooklyn	82.0	83.0
Williamsburg	75.0	68.0
Central Greenwich Village	88.0	75.0
Flushing	73.0	70.0
Battery Park City & Financial District	83.0	81.0
West Village	89.0	87.0
Flatiron & Gramercy	84.0	77.0
Chelsea	82.0	77.0
Sheepshead Bay	77.0	73.0
Soho	86.0	79.0
Nolita & Little Italy	73.0	74.0
Brighton Beach	78.0	88.0
Inwood	60.0	87.0
Corona Park	70.0	74.0
Red Hook	71.0	84.0
Midtown West	79.0	77.0
Upper East Side	85.0	80.0

Upper West Side	83.0	86.0
Washington Heights	62.0	79.0
Riverdale	75.0	74.0
Sunset Park	71.0	70.0
New Drop	79.0	69.0
West Brighton	70.0	71.0
Chinatown	71.0	73.0
St. George	70.0	72.0
Belmont	62.0	77.0
Co-op City	71.0	67.0
Morningside Heights	66.0	77.0
Roosevelt Island	77.0	84.0
Bedford Park	60.0	70.0
Parkchester	67.0	65.0
Harlem	73.0	75.0
50	65.0	900.0
51	62.0	75.0

	Wellness	Nightlife
Neighborhood		
Park Slope	77.0	87.0
Lower East Side	73.0	92.0
Sunnyside	72.0	73.0
Cobble Hill & Boerum Hill	76.0	84.0
Greenpoint	70.0	87.0
Brooklyn Heights	80.0	72.0
Carroll Gardens & Gowanus	71.0	93.0
Murray Hill	80.0	77.0
Prospect Heights	74.0	85.0
East Village	74.0	99.0
Astoria	82.0	83.0
Bay Ridge	82.0	67.0
Woodside	77.0	79.0
Tribeca	75.0	78.0
Jackson Heights	79.0	72.0
Long Island City	72.0	76.0
Midtown East	81.0	75.0
Fort Greene & Clinton Hill	70.0	75.0
Dumbo & Downtown Brooklyn	73.0	74.0
Williamsburg	68.0	88.0
Central Greenwich Village	75.0	83.0
Flushing	82.0	66.0
Battery Park City & Financial District	74.0	70.0
West Village	74.0	91.0
Flatiron & Gramercy	78.0	75.0
Chelsea	70.0	77.0
Sheepshead Bay	79.0	65.0

```
Soho                                        72.0       80.0
Nolita & Little Italy                       69.0       94.0
Brighton Beach                              77.0       74.0
Inwood                                      75.0       85.0
Corona Park                                 80.0       66.0
Red Hook                                    69.0       88.0
Midtown West                                68.0       74.0
Upper East Side                             80.0       78.0
Upper West Side                             77.0       76.0
Washington Heights                          76.0       73.0
Riverdale                                   84.0       70.0
Sunset Park                                 73.0       68.0
New Drop                                    85.0       65.0
West Brighton                               76.0       66.0
Chinatown                                   72.0       76.0
St. George                                  76.0       65.0
Belmont                                     72.0       71.0
Co-op City                                  78.0       65.0
Morningside Heights                         71.0       68.0
Roosevelt Island                            78.0       65.0
Bedford Park                                73.0       65.0
Parkchester                                 74.0       65.0
Harlem                                      71.0       79.0
50                                          45.0      800.0
51                                        -500.0       80.0
```

6 Principal Component Analysis and Biplot after introducing two new rows

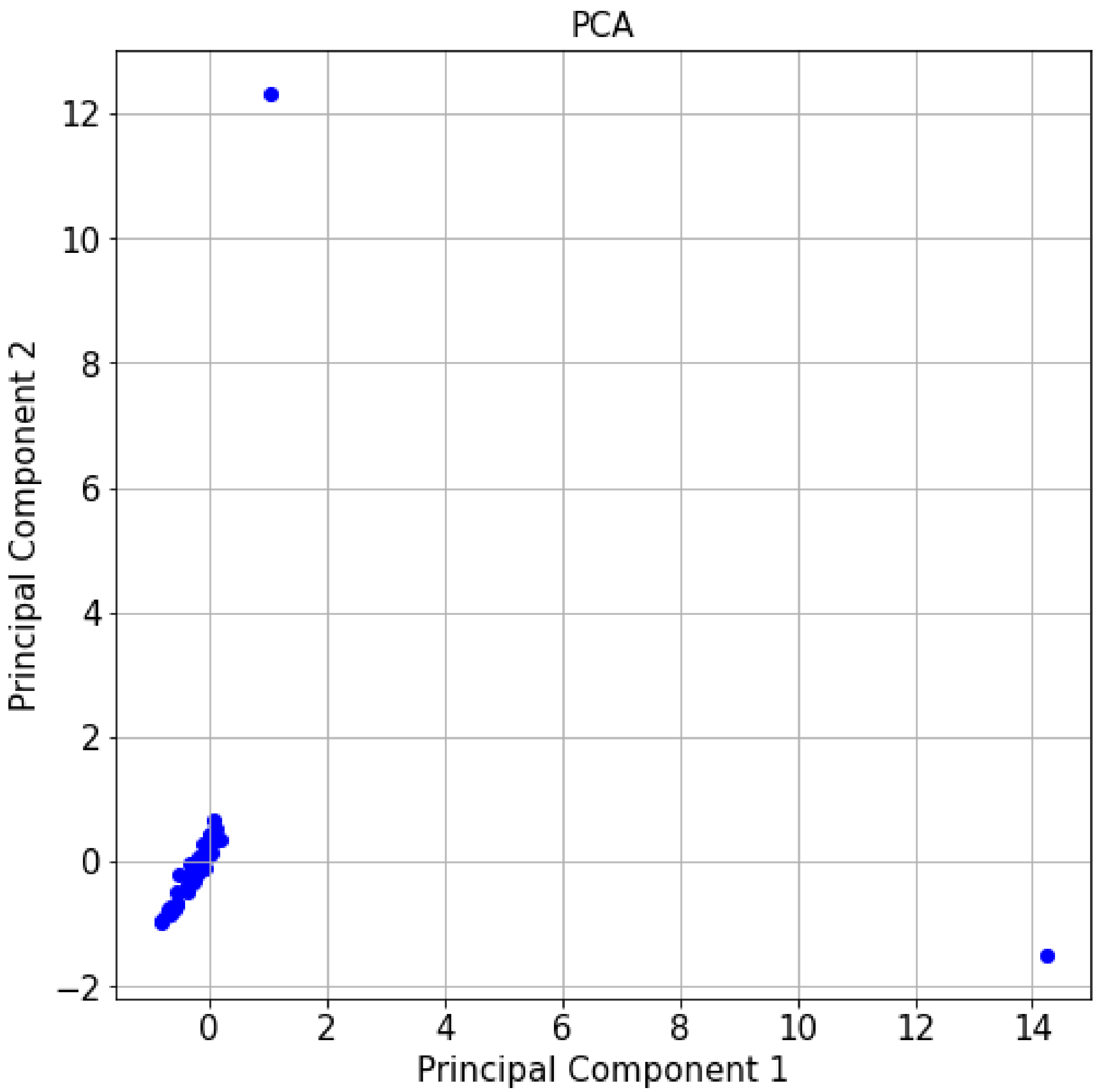
[327]:
```
features = df_new.columns[1:]
# Separating out the features
x_new = df_new.loc[:, features].values
# Standardizing the features
x_new = StandardScaler().fit_transform(x_new)
pca_new = PCA(n_components=11)
principal_components_new = pca_new.fit_transform(x_new)
principal_df_new = pd.DataFrame(data = principal_components_new)
```

[328]:
```
print(pca_new.explained_variance_ratio_)
```

```
[3.69647883e-01 2.89222657e-01 1.96209640e-01 8.22571341e-02
 4.11511393e-02 2.01991331e-02 5.37377791e-04 4.05780952e-04
 1.79918385e-04 1.12172123e-04 7.71645000e-05]
```

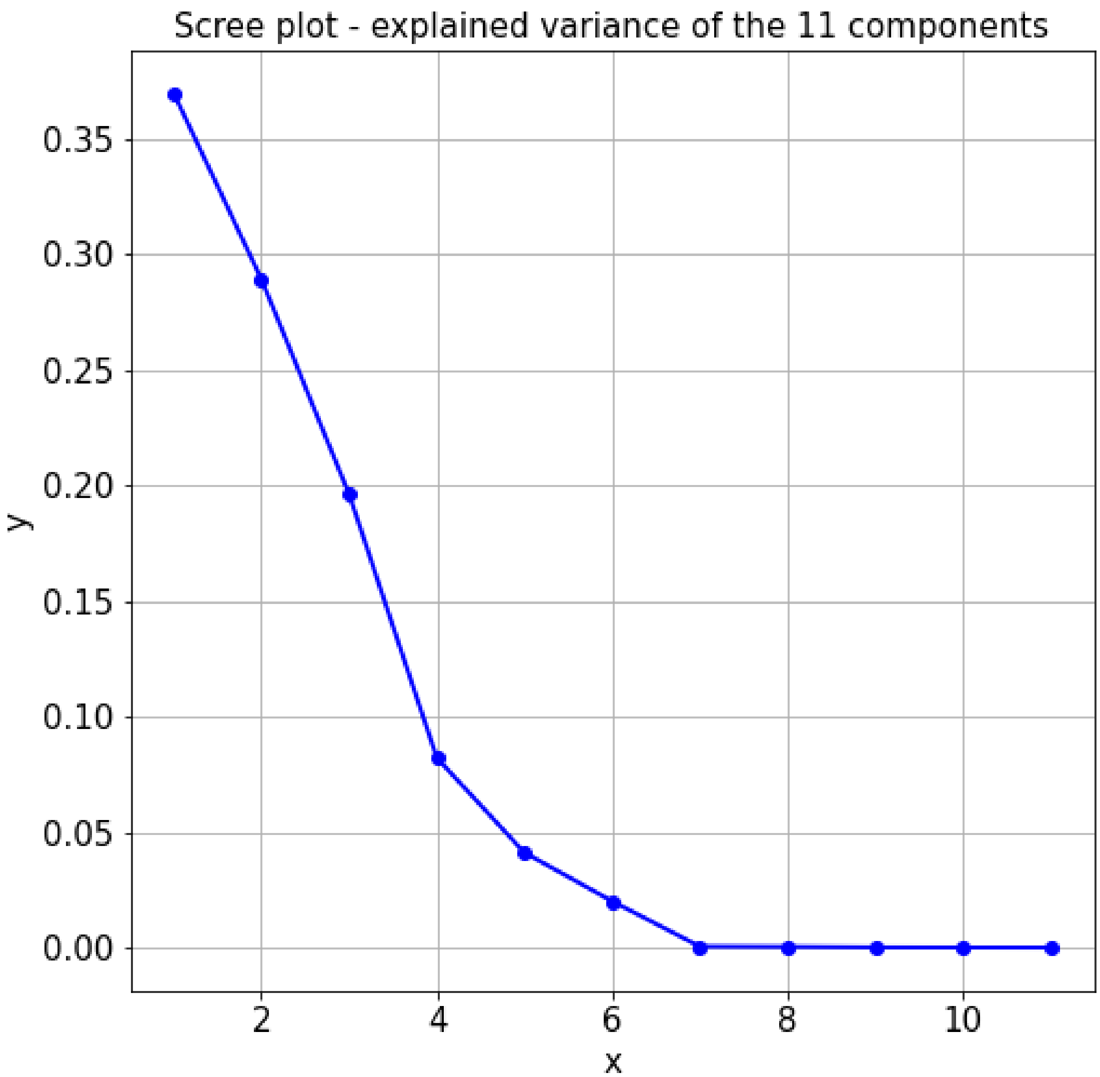
[329]:
```
# plotting the scatter plot
fig, ax = plt.subplots(figsize=[8,8])
```

```
plt.scatter(principal_df_new[0], principal_df_new[1], color='b')

plt.grid()
plt.xlabel('Principal Component 1', fontsize = 15)
plt.ylabel('Principal Component 2', fontsize = 15)
plt.xticks(fontsize=15)
plt.yticks(fontsize=15)
plt.title('PCA', fontsize = 15)
#plt.legend()
plt.show()
```

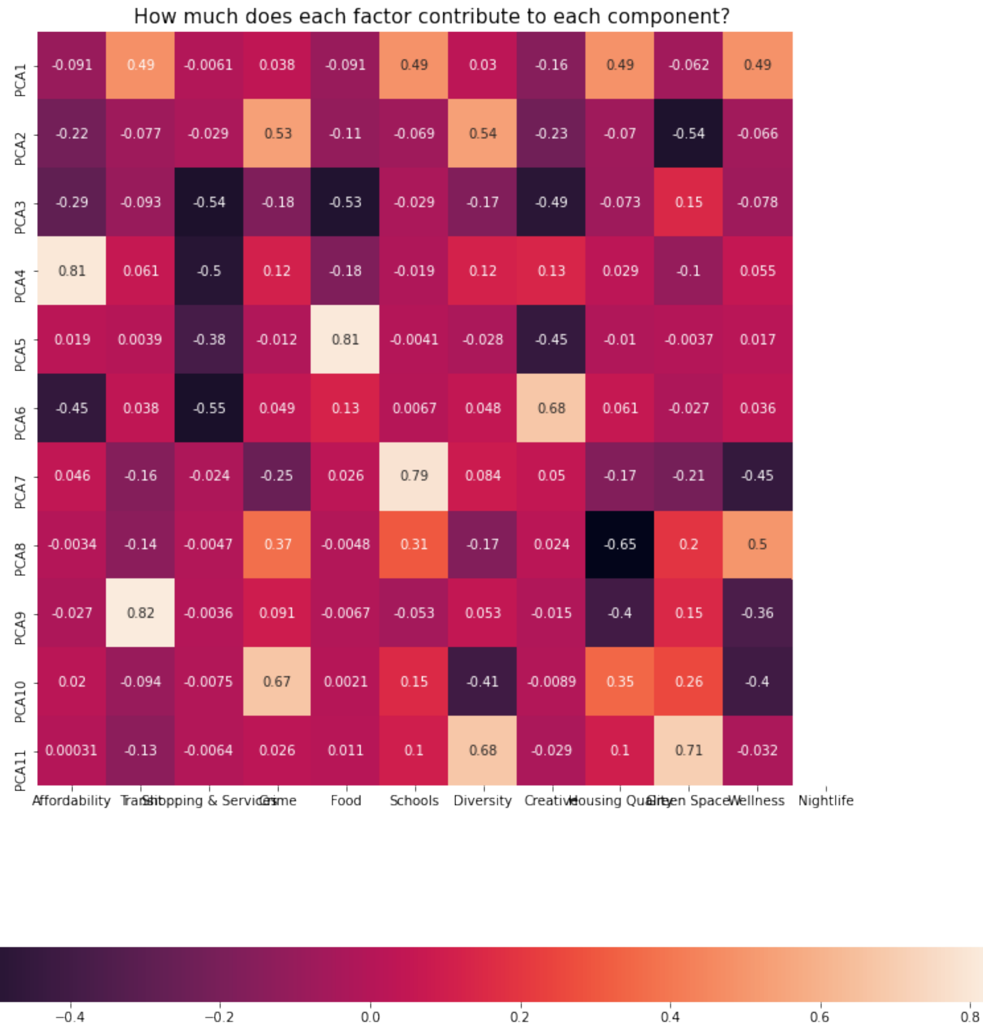


[330]:
```
fig, ax = plt.subplots(figsize=[8,8])

plt.plot(np.arange(1,12),pca_new.explained_variance_ratio_,'b-o', lw=2)
plt.grid()
plt.xlabel('x', fontsize = 15)
plt.ylabel('y', fontsize = 15)
plt.xticks(fontsize=15)
plt.yticks(fontsize=15)
plt.title('Scree plot - explained variance of the 11 components', fontsize = 15)
```

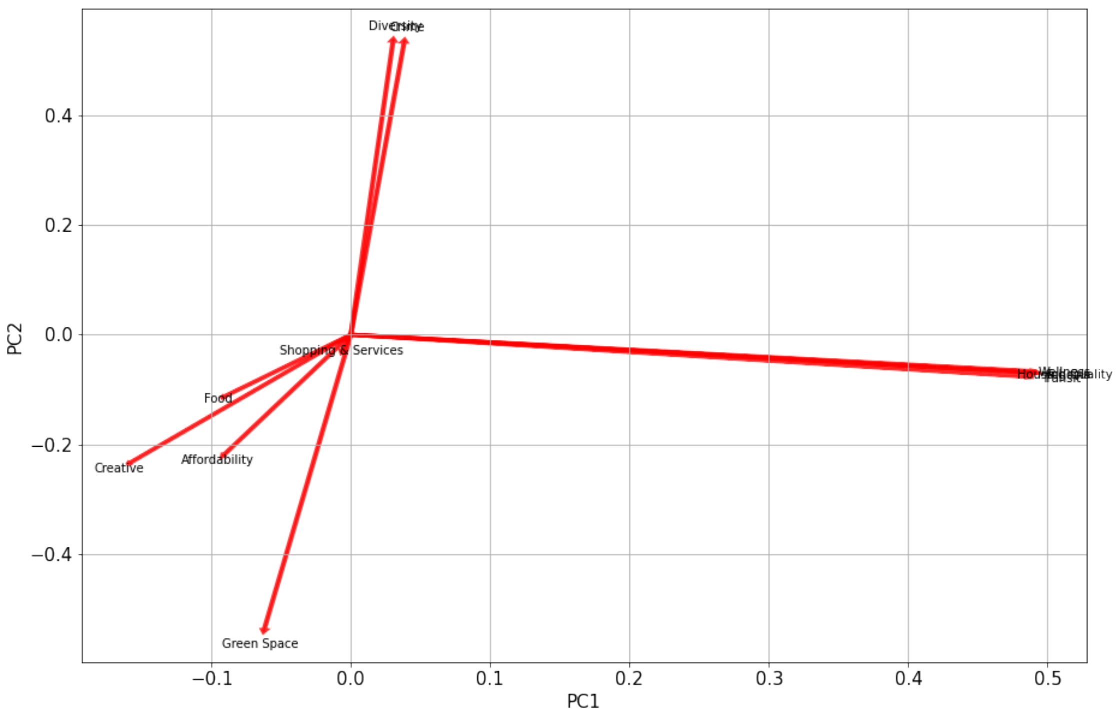
[330]: Text(0.5, 1.0, 'Scree plot - explained variance of the 11 components')



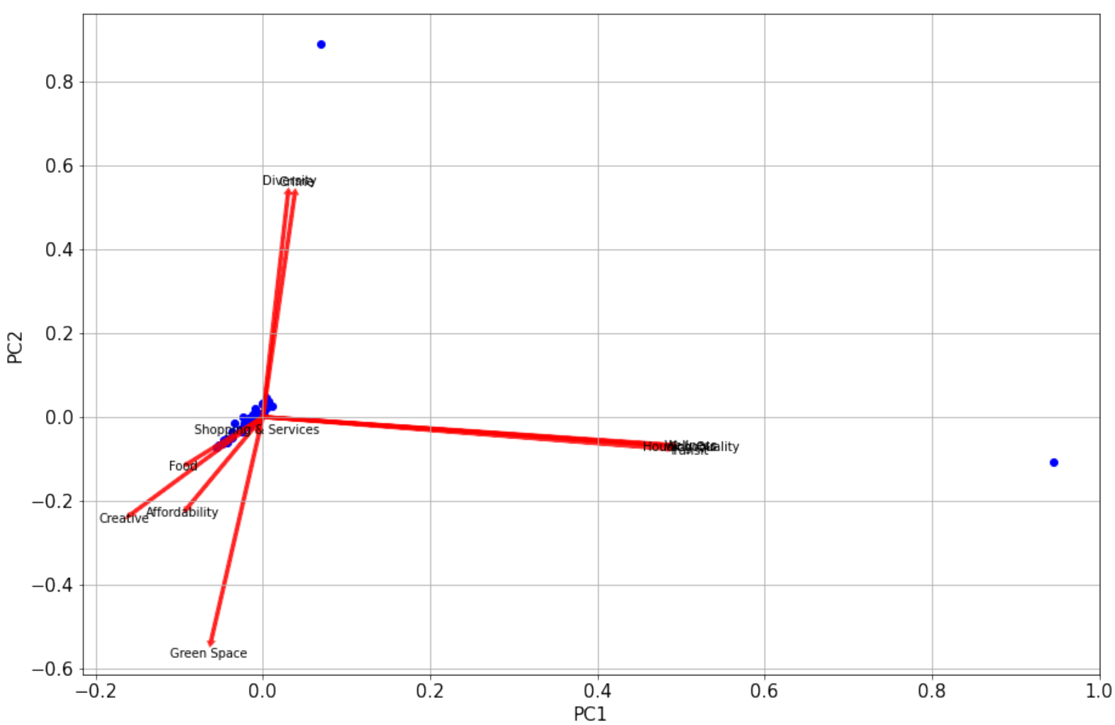
```
[334]: fig, ax1 = plt.subplots(figsize=[15,15])
ax = sns.heatmap(pca_new.components_,
                 yticklabels=[ "PCA"+str(x) for x in range(1,pca_new.
 ↪n_components_+1)],
                 xticklabels=list(df.columns),
                 cbar_kws={"orientation": "horizontal"}, annot=True)
plt.title('How much does each factor contribute to each component?',␣
 ↪fontsize=15)
ax.set_aspect("equal")
```



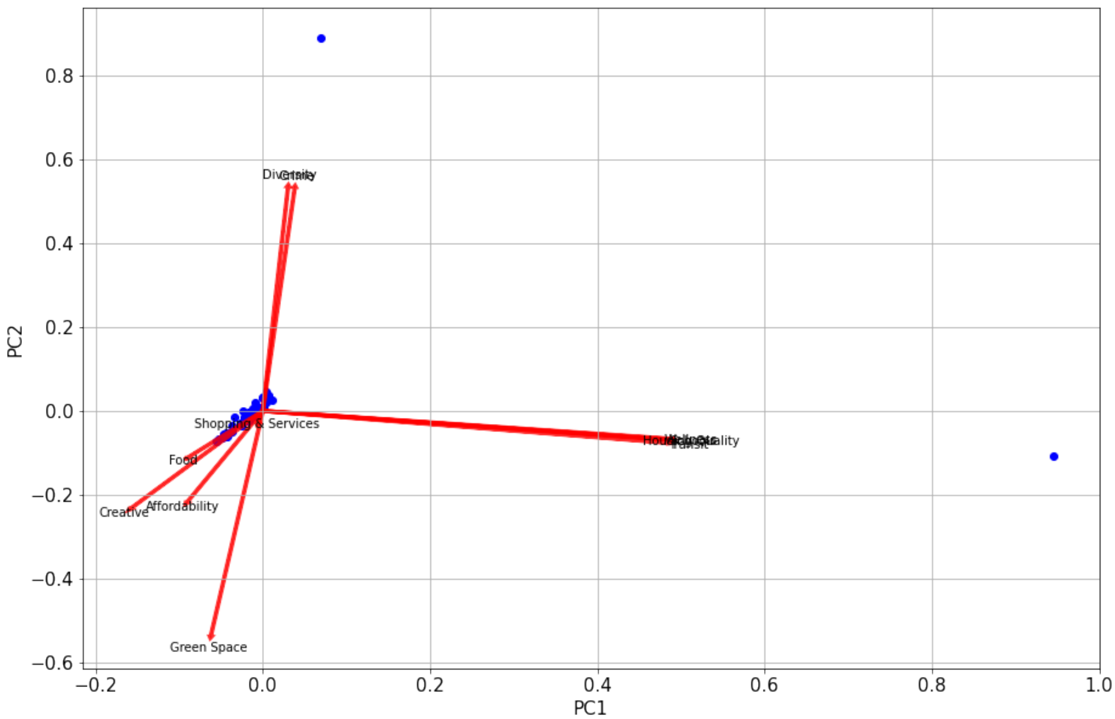

```
[338]: loading_plot(np.transpose(pca_new.components_[0:2,:]),list(df_new.columns))
```



[332]:
```
biplot(principal_components_new[:,0:2],np.transpose(pca_new.components_[0:2,:
 ↪]),list(df_new.columns))
plt.show()
```



```
[348]: biplot(principal_components_new[:,0:2],np.transpose(pca_new.components_[0:2,:
 ↪]),list(df_new.columns))
```



In the above plots, the variables have been scaled. Not scaling the variables hinders good visualization because the length of the eigenvectors is really small as compared to the scale of the non-scaled data.

Introducing outliers affects the calculation of PCs in the sense that they are now biased. They cluster all the remaining data points into one cluster which does not let us observe what happens to the majority of the data. In this case, the values of PC3 and PC4 are also significant enough to not be ignored. Hence, it's best to remove outliers and then perform PCA.

7 Conclusion

The eigenvectors that do not have a larger angle between them are more closely correlated than others. The length of the eigenvector describes the variance of that particular component with respect to PC1 and PC2. Hence, PCA proves to be a very efficient technique to visualize data which has larger dimensions since it reduces the dimensions (through SVD).