Homework Assignment (Data Visualization) - School of AI - Data Lit

by: Fernando Chica

We need to solve a crime. In the crime scene there is no evidence linking the crime directly to the perpetrator, evidence such as DNA. The only hint that we have is broken glass. Near to the crime scene the experts found pieces of glass with fingerprints of a posible suspect. So, we need to determinate if the glass found in the crime scene is the same as the one with the fingerprint.

We are going to use the dataset of Glass Identification avalible in "http://archive.ics.uci.edu/ml/datasets/glass%20identification". This data consists of 214 observations with 10 features of the composition of the glass and 6 different types of glass.

The columns in this dataset are:

1. RI: refractive index
2. Na: Sodium
3. Mg: Magnesium
4. Al: Aluminum
5. Si: Silica
6. K: Potassium
7. Ca: Calcium
8. Ba: Barium
9. Fe: Iron
10. Type of glass (Target label)

The Target label, Type of Glass has 6 classes:

- 1. building_windows_float_processed
- 2. building_windows_non_float_processed
- 3. vehicle_windows_float_processed
- 5. containers
- 6. tableware
- 7. headlamps

In this case we are use the information of only 3 different types of glases (named in the dataset as 1,2 and 7 each type) this are the one with most information in the dataset

The question

Is the glass in the crime scene the same as the found with the fingerprint? *So, let's do it :D.*

Import the necesary libraries

```
import pandas as pd      # Process dataset
import numpy as np       # mathematical tool
import seaborn as sns    # plot tools
import matplotlib.pyplot as plt  # plot tools
from sklearn import decomposition
```

We are going to extract the information of five observations of a type of glass "1" in the dataset that is going to be used as the glass found in the crime scene, and compare its characteristic with other five observations of other type of glass (number 7 in the dataset). The main idea is to compare these observations and determine if its the same glass.

This first section describes the preprocesing process to adapt the dataset from the url to simulate the crime scene. From the begin we know that both glass are different but we have to pretend that we don't know anything xD.

Import the Dataset from the url and adapt to the emule the crime scene

We are going to import directrly from the web the dataset, so we use a condition; Import the dataset from the URL, opposite case, print an error.

After, we are going to drop the information of glass type different of 1,2 or 7. Then, extract two different observationts, the first one (glass type 1) saved in the variable "crime" to be the glass in the crime scene and other (glass type 7) saved in the variable "evidence" to be the glass with the fingerprint. It can be noted that this two observations, one are extracted from dataset also is drop out.

```
try:
    data = pd.read_csv("http://archive.ics.uci.edu/ml/machine-learning-databases/glass/glass.
                       ,names=['RI','Na','Mg','Al','Si','k','Ca','Ba','Fe','Type'])
    data = data.drop(data[data['Type']== 3 ].index)
    data = data.drop(data[data['Type']== 5].index)
    data = data.drop(data[data['Type']== 6 ].index)
    data = data.reset_index(drop=True)

    print('Glass Identification Data Set has {} rows with {} features each'.format(*data.shap


except:
    print('Error! Getting data')
```

```
Glass Identification Data Set has 175 rows with 10 features each
```

List of information about the crime

1. Broken Glass => variable "crime"
2. Broken Glass with fingerprint => variable "eviende"
3. Database to compare => variable "Info"

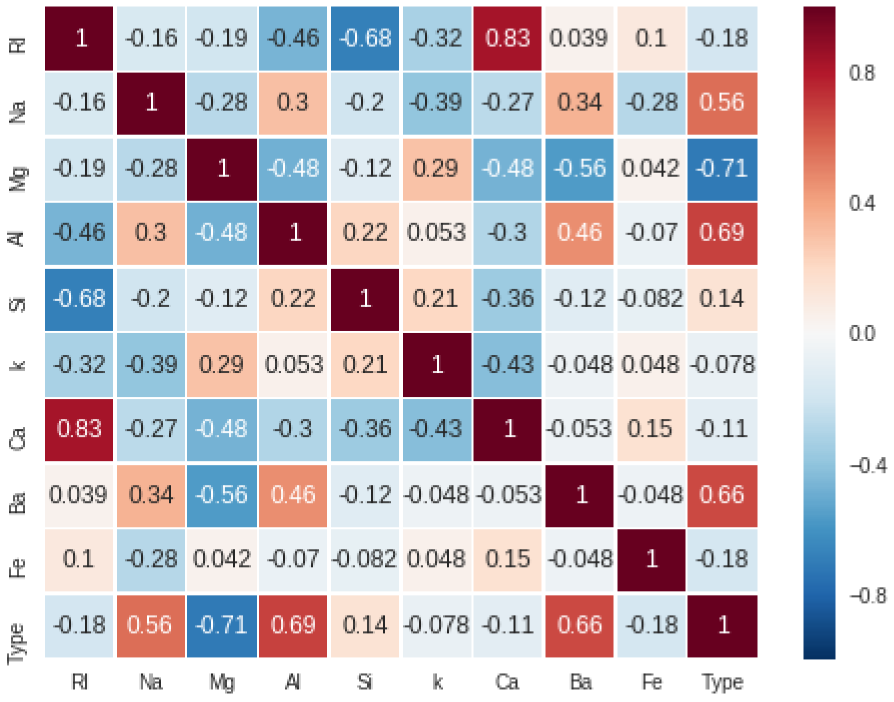
We need to analize the characteristics of the three different type of glass to be able to distinguish one from the other. The problems is that there is lot of characteristics and we are humans, only can see three dimensions (yet). So, first we are going to plot a diagram of the correlations between all the characteristics of the dataset.

Visualization

Heatmap of correlations

```
plt.subplots(figsize=(8,6))
sns.heatmap(Info.corr(),annot=True, linewidth=.5,)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6a79026710>
```



	RI	Na	Mg	Al	Si	k	Ca	Ba	Fe	Type
RI	1	-0.16	-0.19	-0.46	-0.68	-0.32	0.83	0.039	0.1	-0.18
Na	-0.16	1	-0.28	0.3	-0.2	-0.39	-0.27	0.34	-0.28	0.56
Mg	-0.19	-0.28	1	-0.48	-0.12	0.29	-0.48	-0.56	0.042	-0.71
Al	-0.46	0.3	-0.48	1	0.22	0.053	-0.3	0.46	-0.07	0.69
Si	-0.68	-0.2	-0.12	0.22	1	0.21	-0.36	-0.12	-0.082	0.14
k	-0.32	-0.39	0.29	0.053	0.21	1	-0.43	-0.048	0.048	-0.078
Ca	0.83	-0.27	-0.48	-0.3	-0.36	-0.43	1	-0.053	0.15	-0.11
Ba	0.039	0.34	-0.56	0.46	-0.12	-0.048	-0.053	1	-0.048	0.66
Fe	0.1	-0.28	0.042	-0.07	-0.082	0.048	0.15	-0.048	1	-0.18
Type	-0.18	0.56	-0.71	0.69	0.14	-0.078	-0.11	0.66	-0.18	1

It can be noticed that there is some variables that is correlated like Ca and RI or Ba and Al, but in the end of the correlation table there is the feature "Type" that indicates the type of the glass. In this case, the information shows that the features **Na, Al** and **Ba** are the most correlated with the "Type", in other words this features are the one who contribute most information about the kind of glass. So, we are going to use this three features to diferenciate the characteristics of a particular glass and determine if the glass with the fingerprint is the same that was found in the crime scene.

Before to visualize we need to explore this features, for this, we use "describe" from the library pandas and we get the general information of this features in all the dataset.

Density Plots

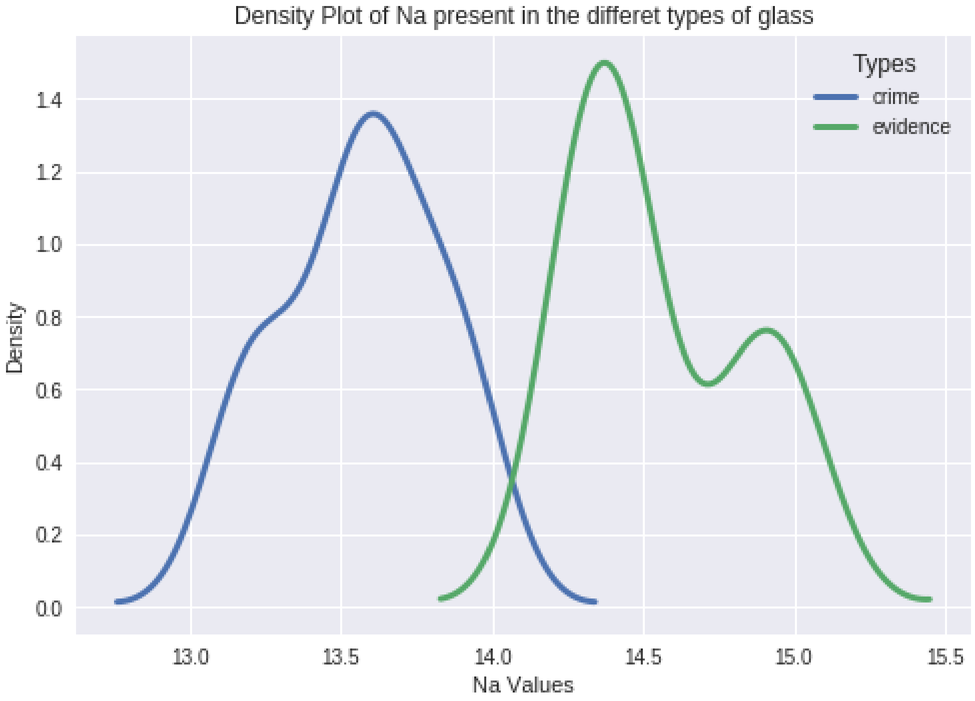
We can notice that in the feature Ba there is a lot of 0 values in the glass type 1 and 2. Thus, we use the other two features to density plots and visualize the characteristic of the different kind of glass. A Density Plot visualises the distribution of data over a continuous interval or time period, an advantage Density Plots have over Histograms is that they're better at determining the distribution shape because they're not affected by the number of bins used (which benefits us by the amount of observations we have of the evidence of the crime).

Other important fact is that the data are in numerical order regarding to type, so the from the position 0 to 64 are type 1, from 65 to 141 are type 2 and from 142 to 165 are type 7, this can be useful if we use pca.

First we compare between the glass found in the crime scene and the other with the fingerprints.

```
sns.distplot(crime['Na'], hist = False, kde = True, kde_kws = {'linewidth': 3},label='crime')
sns.distplot(evidence['Na'], hist = False, kde = True, kde_kws = {'linewidth': 3},label='evid

plt.legend(prop={'size': 10}, title = 'Types',loc= 'best')
plt.title('Density Plot of Na present in the differet types of glass')
plt.xlabel('Na Values')
plt.ylabel('Density')
```

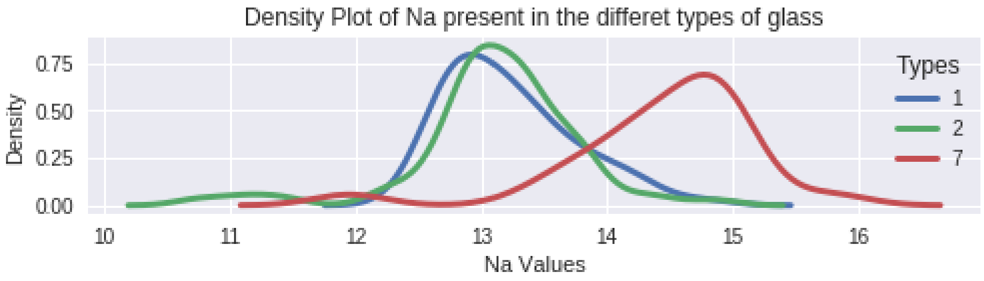
```
Text(0, 0.5, 'Density')
```

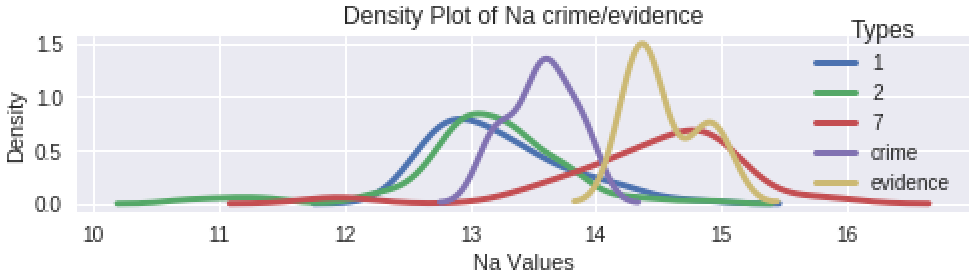


Observing the graph, it's easy to distinguish that they have a different distribution of data between both, this fact point to that there are different types of glass. Now, let's check with the density plots of the all dataset.

```
types = [1,2,7]

plt.subplot(311)
# Iterate through the types
for typ in types :
    # Subset to the Info
    subset = Info[Info['Type'] == typ]

    # Draw the density plot
    sns.distplot(subset['Na'], hist = False, kde = True,
                 kde_kws = {'linewidth': 3},label=typ,
                 )

# Plot formatting
plt.legend(prop={'size': 10}, title = 'Types',loc= 'best')
plt.title('Density Plot of Na present in the differet types of glass')
plt.xlabel('Na Values')
plt.ylabel('Density')



plt.subplot(313)
# Iterate through the types
for typ in types :
    # Subset to the Info
    subset = Info[Info['Type'] == typ]

    # Draw the density plot
    sns.distplot(subset['Na'], hist = False, kde = True,
                 kde_kws = {'linewidth': 3},label=typ,
                 )


sns.distplot(crime['Na'], hist = False, kde = True, kde_kws = {'linewidth': 3},label='crime')
sns.distplot(evidence['Na'], hist = False, kde = True, kde_kws = {'linewidth': 3},label='evid


# Plot formatting
plt.legend(prop={'size': 10}, title = 'Types',loc= 'best')
plt.title('Density Plot of Na crime/evidence')
plt.xlabel('Na Values')
plt.ylabel('Density')
```

```
Text(0, 0.5, 'Density')
```

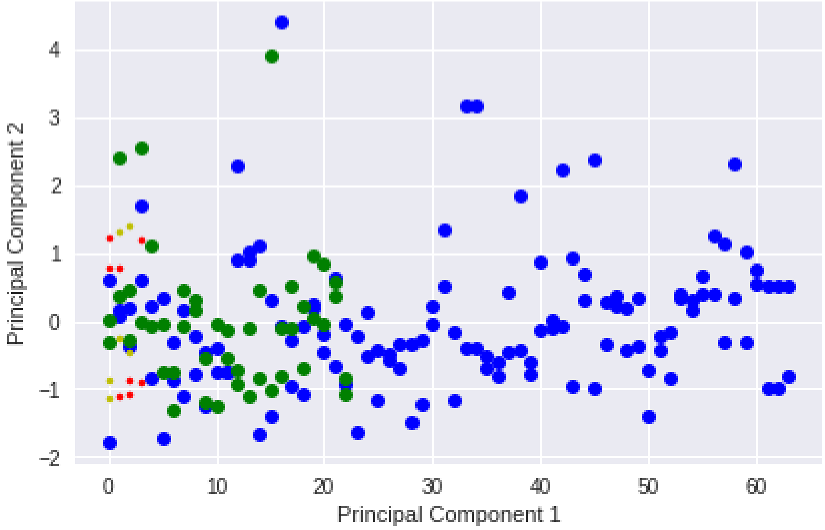




0 to 64 are type 1, from 142 to 165 are type 7

```
#Visualize PCA

plt.figure(figsize=(6, 4))
plt.plot(t1, 'bo')
plt.plot(t3, 'go')
plt.plot(c1,'r.')
plt.plot(e1,'y.')

plt.xlabel('Principal Component 1')
plt.ylabel('Principal Component 2')
plt.legend(loc='lower center')
plt.tight_layout()
plt.show()
```

```
No handles with labels found to put in legend.
```



Using PCA we can't find information about if the two glass are different, PCA is a linear transformation technique and the results shows that there is no difference in the different types of glass even in all the dataset. Thus, because of the information in the density plots we conclude that glass found in the crime scene and the other with the fingerprint are different, we need to will back to the crime scene to find more evidence.