Finding Gemstones

# Gemstones

By Noam Bassat

Can we find gemstones by a given location?

01/2021

Data Science project

# GEMSTONES

## *A little background*

*Gemstones are naturally formed mineral crystals which when cut and polished reveal their beauty.*

They are created inside the earth when the right physical conditions occur.
They are brought to the surface by various geological forces and can be carried away by rivers before they are deposited.
They are defined by their chemical formula and trace minerals as well as their crystal structure.
Of the more than 2,000 identified natural minerals, fewer than 100 are used as gemstones and only 16 have achieved importance.

# The importamce of gemstones

*Gemstones have been chosen for their beauty and durability, then cut and polished mainly for use as human adornment.*

01
Jewelry and ornamental purposes

02
Health, religion and spiritual benefits

03
Status symbol and economical value

04
Scientific and industrial uses

# Main Steps

*Main process steps of the project*

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| **Obtaining Data** | **Scrubbing Data** | **Exploring Data** | **Modeling Data** | **interpreting Data** |
| Web crawling for collecting and obtaining the data | Cleaning, formatting, and filtering the data | Visualizing and understanding the data | Clustering the data into groups, modeling and the algorithm | Presentation of data, understanding and delivering the results + Final predict model |

# Web Crawling

*Finding main gemstones to explore, and crawling their data*

---

### Gemstones list

From "Geology.com" I took the main gemstones list that I wanted to explore.

### Gemstones and Locations

From "mindat.org" I took every relevant information about the gemstone. After I had their locations references, I took the relevant details of each location, and put it into a different dataframe.



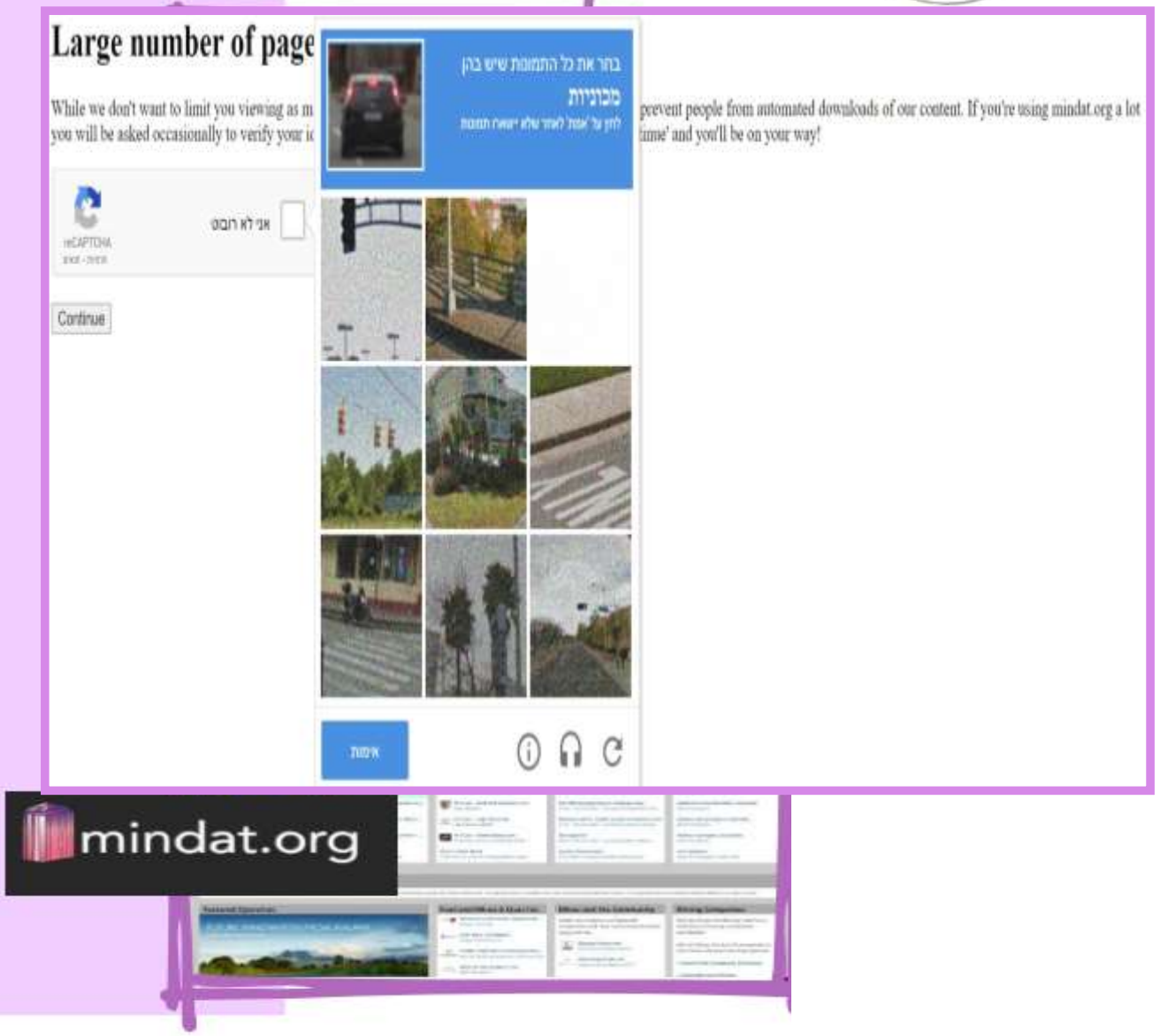
## Main tools:

**BeautifulSoup and Selenium**

# Web Crawling

## But, not every thing was so easy…

After a number of pages visits, the "Midnat" Site wanted to check if I am a bot.
of course I am not, but my program is.
So, I made my calculations and found out that this bot-checker appears every 23 minutes. (including the sleeping and searching time). So, I had to stay close, and pass the test every 23 minutes.
Finally, after 5 and a half hours, I had a dataset that contains 2,621 locations records, and another dataset that contains information about 77 gemstones.



Gemstones project
by Noam Bassat

## Locations Dataframe

| | Location URL | Stone Refernced | Minerals Recorded | Location Name | Rock Types | Climate | Decimal Coordinates | Minerals |
|---|---|---|---|---|---|---|---|---|
| 0 | https://www.mindat.org//loc-4047.html | Amazonite Apatite Aquamarine Beryl Opal Smoky ... | None | , Preitenegg, Wolfsberg District, Carinthia, A... | [] | Dfb : Warm-summer humid continental climate | 46.96177,14.95181 | [Valid Species, Al, C, Ca, Ce, F, Fe, H, K, La... |
| 1 | https://www.mindat.org//loc-250161.html | Amazonite Aquamarine Beryl Kyanite Smoky Quartz | None | , Spruce Pine, Spruce Pine Mining District, Mi... | [] | Cfb : Temperate oceanic climate | 35.91667,-82.06667 | [Al, B, Be, C, Ca, Cd, Ce, Cl, Cu, F, Fe, H, K... |
| 2 | https://www.mindat.org//loc-8991.html | Amethyst Apatite Spinel | None | , Thanh Son District, Phú Thọ Province, Vietna... | [] | Cwa : Monsoon-influenced humid subtropical cli... | 21.18306,105.25000 | [Al, B, Be, F, Fe, H, K, Mg, O, Si, Ti] |
| 3 | https://www.mindat.org//loc-56146.html | Amethyst | None | , Eggenburg, Waldviertel, Lower Austria, Austr... | [] | Cfb : Temperate oceanic climate | 48.64833,15.82083 | [Valid Species, O, Si] |
| 4 | https://www.mindat.org//loc-232658.html | Amethyst | None | , Krimml, Zell am See District, Salzburg, Aust... | [] | Dfb : Warm-summer humid continental climate | 47.15161,12.13835 | [Valid Species, O, Si] |

## Gemstones Dataframe

| | Name | Formula | Elements | Colours | Image_url | Locations | Locations_url |
|---|---|---|---|---|---|---|---|
| 0 | Amazonite | K(AlSi3O8) | Al, K, O, Si | NaN | ['/imagecache/11/7a/047660800149462460164 26.jp... | ['Argentina', '\xa0', 'Australia', '\xa0', 'Au... | ['https://www.mindat.org//loc-269722.html', 'h... |
| 1 | Amethyst | SiO2 | O, Si | Violet - purpleC | ['/imagecache/8a/ca/030397500149462461632 78.jp... | ['Afghanistan', '\xa0', 'Angola', '\xa0', 'Arg... | ['https://www.mindat.org//loc-404929.html', 'h... |
| | | | O, Si | NaN | ['/imagecache/d2/c6/045327600149463921965 42.jp... | ['Bolivia', '\xa0', 'Brazil', '\xa0', 'Canada'... | ['https://www.mindat.org//loc-11804.html', 'ht... |
| | | | NaN | NaN | ['/imagecache/b4/a9/002072300149466039882 22.jp... | ['Austria', '\xa0', 'Canada', '\xa0', 'Madagas... | ['https://www.mindat.org//loc-205904.html', 'h... |
| | | | Bonfim W | Pink to red brown, occasionally yellow, green,... | ['/imagecache/ff/30/013988700149463306149 53.jp... | ['Andorra', '\xa0', 'Antarctica', 'AmMin 36:36... | ['https://www.mindat.org//loc-31239.html', 'ht... |
| | | | ... | ... | ... | ... | ... |
| | | | Cu, H, O, P | bright blue, sky-blue, pale green, blue-green,... | ['/imagecache/6d/d1/053307900149565739551 92.jp... | ['Afghanistan', 'Gübelin, E., Wolgensinger, M.... | ['https://www.mindat.org//loc-14319.html', 'htt... |
| | | | NaN | NaN | ['/imagecache/5a/7d/056483900156019819743 23.jp... | [] | [] |
| 75 | Variscite | AlPO4 · 2H2OIMA Formula:Al(PO4) · 2H2O | Sn–W–Pb–Ag–Zn Huanuni Deposit, Bolivia. Minera... | Pale to emerald-green, bluish green, colourles... | ['/imagecache/77/23/081230900149771923762 73.jp... | ['Antigua and Barbuda', '\xa0', 'Argentina', '... | ['https://www.mindat.org//loc-147083.html', 'h... |
| 76 | Zircon | Zr(SiO4)May contain minor U, Th, Pb, Hf, Y/REE... | , Pr, Nd, and Sm) from a Carbonatite Deposit: ... | Colourless, yellow, grey, reddish-brown, green... | ['/imagecache/40/37/055760600149463205732 25.jp... | ['Afghanistan', '\xa0', 'Algeria', '\xa0', 'An... | ['https://www.mindat.org//loc-3.html', 'https:... |
| 77 | Zoisite | Ca2Al3[Si2O7][SiO4]O(OH) | Bonfim W | Colourless, purple, greyish-white, grey, yello... | ['/imagecache/ce/e7/040961500149462731040 80.jp... | ['Afghanistan', '\xa0', 'Antarctica', 'AmMin 3... | ['https://www.mindat.org//loc-8.html', 'https:... |

# Scrubbing Data

## Dealing with missing values:

Dropping rows and columns with too many missing values.

## Dealing with duplicates;

I have a tiny number of duplicates rows, which contained the exact same values, so I kept only the first ones.

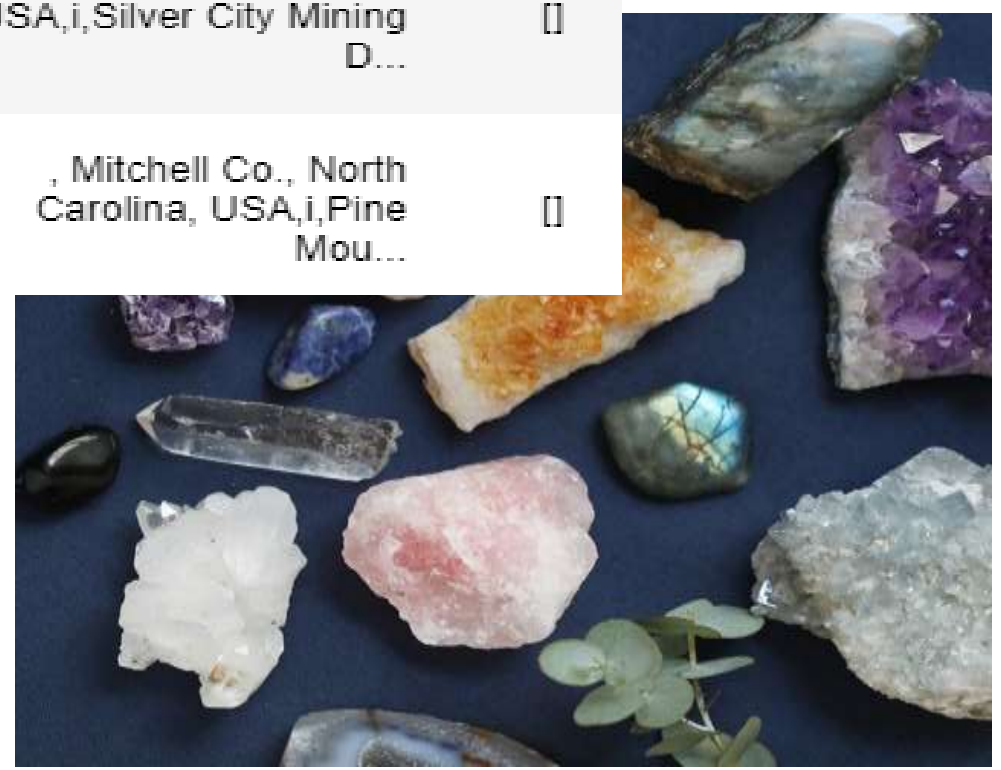
## Cleaning the noise:

Many strings contained unnecessary signs, duplicated words and so on, so I built some “cleaning” methods.

## Formatting:

Importing every non–numeric variable into a fitted categorical data. Also, splitting columns such as coordinates into two columns, one for the latitude value and one for the longitude.

| | | |
|---|---|---|
| Sand and Gravel | , Palm Springs, Riverside Co., California, USA... | [] |
| None | , Lyon Co., Nevada, USA,i,Silver City Mining D... | [] |
| Feldspar | , Mitchell Co., North Carolina, USA,i,Pine Mou... | [] |



# Full - Locations Dataframe

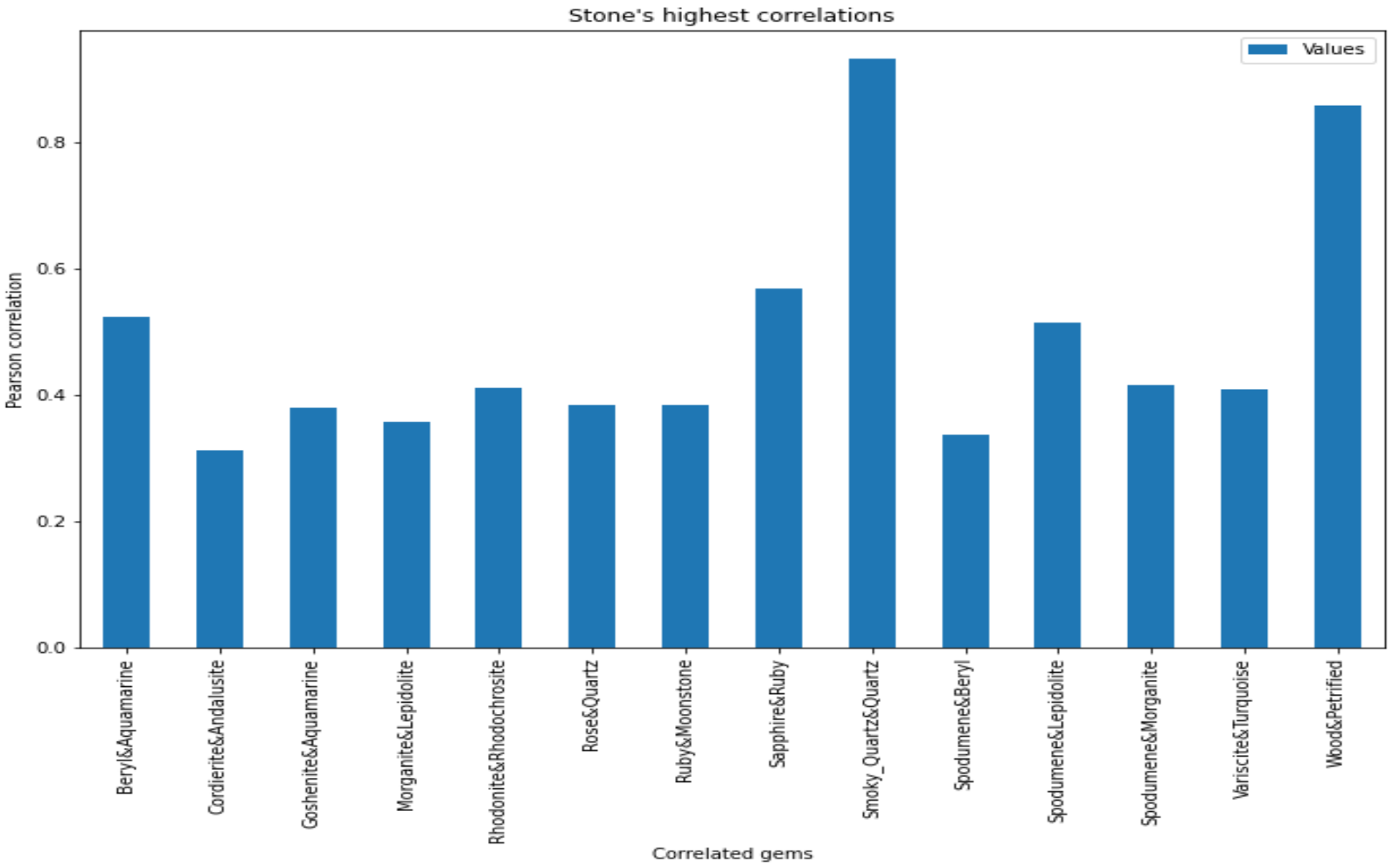
| | Location URL | Stone Refernced | Location Name | Climate | Minerals | Latitude | Longitude | Ag | Al | As | ... | Ti | Tl | U | V | W | Y | Yb | Zn | Zr | Gro |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | https://www.mindat.org/loc-14762.html | Lapis Lazuli Unakite | ,i,Israel,Country, | 15 | Ag Al B Ba C Ca Ce Cl Cr Cu F Fe Gd H K La Mg ... | 31.261220 | 35.214581 | 1 | 1 | 0 | ... | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | |
| 1 | https://www.mindat.org//loc-145486.html | Lapis Lazuli Unakite | , Israel,i,Central District (HaMerkaz District... | 15 | NaN | 32.620121 | 35.014771 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 2 | https://www.mindat.org//loc-145489.html | Lapis Lazuli Unakite | , Israel,i,Haifa District,District, Israel,Cou... | 15 | Al B C Ca Cl Cr F Fe H K Mg N Na O P Si Sn Ti ... | 31.795924 | 35.211981 | 0 | 1 | 0 | ... | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | |
| 3 | https://www.mindat.org//loc-145488.html | Lapis Lazuli Unakite | , Israel,i,Jerusalem District,District, Israel... | 15 | Al C Ca H O Si | 32.607559 | 35.289 | | | | | | | | | | | | | | |
| 4 | https://www.mindat.org//loc-205351.html | Lapis Lazuli Unakite | , Israel,i,Northern District (HaZafon District... | 7 | Al C Ca Cl Fe H K Mg Mn Na Ni O Si Ti Zn | 31.261220 | 35.214 | | | | | | | | | | | | | | |
| ... | ... | ... | ... | ... | ... | ... | | | | | | | | | | | | | | | |
| 1603 | https://www.mindat.org//loc-193574.html | Diopside Lapis Lazuli Unakite Zoisite | , Cochise Mining District, Little Dragoon Moun... | 13 | Valid Species Al Ca Cu Fe H K Mo O S Si Zn | 32.065560 | -109.969 | | | | | | | | | | | | | | |
| | | Diopside | | | Al B C | | | | | | | | | | | | | | | | |

# Locations Dataset

| | Climate | Latitude | Longitude | Ag | Al | As | Au | B | Ba | Be | ... | Th | Ti | Tl | U | V | W | Y | Yb | Zn | Zr |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 15 | 31.261220 | 35.214581 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | ... | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 1 | 15 | 32.620121 | 35.014771 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 15 | 31.795924 | 35.211981 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | ... | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 3 | 15 | 32.607559 | 35.289086 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 7 | 31.261220 | 35.214581 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 1603 | 13 | 32.065560 | -109.959440 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 1604 | 15 | 33.897470 | -115.462220 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | ... | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1605 | 12 | 39.263610 | -118.359440 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 1606 | 2 | 35.943050 | -81.910000 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | ... | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1607 | 8 | 40.139200 | -112.174950 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

# Exploring & visualizing the Data

| | Names | Values |
|---|---|---|
| 0 | Beryl&Aquamarine | 0.521820 |
| 1 | Cordierite&Andalusite | 0.312613 |
| 2 | Goshenite&Aquamarine | 0.377900 |
| 3 | Morganite&Lepidolite | 0.356887 |
| 4 | Rhodonite&Rhodochrosite | 0.410045 |
| 5 | Rose&Quartz | 0.383578 |
| 6 | Ruby&Moonstone | 0.384450 |
| 7 | Sapphire&Ruby | 0.567988 |
| 8 | Smoky_Quartz&Quartz | 0.931150 |
| 9 | Spodumene&Beryl | 0.335434 |
| 10 | Spodumene&Lepidolite | 0.513158 |
| 11 | Spodumene&Morganite | 0.415219 |
| 12 | Variscite&Turquoise | 0.407828 |
| 13 | Wood&Petrified | 0.858764 |



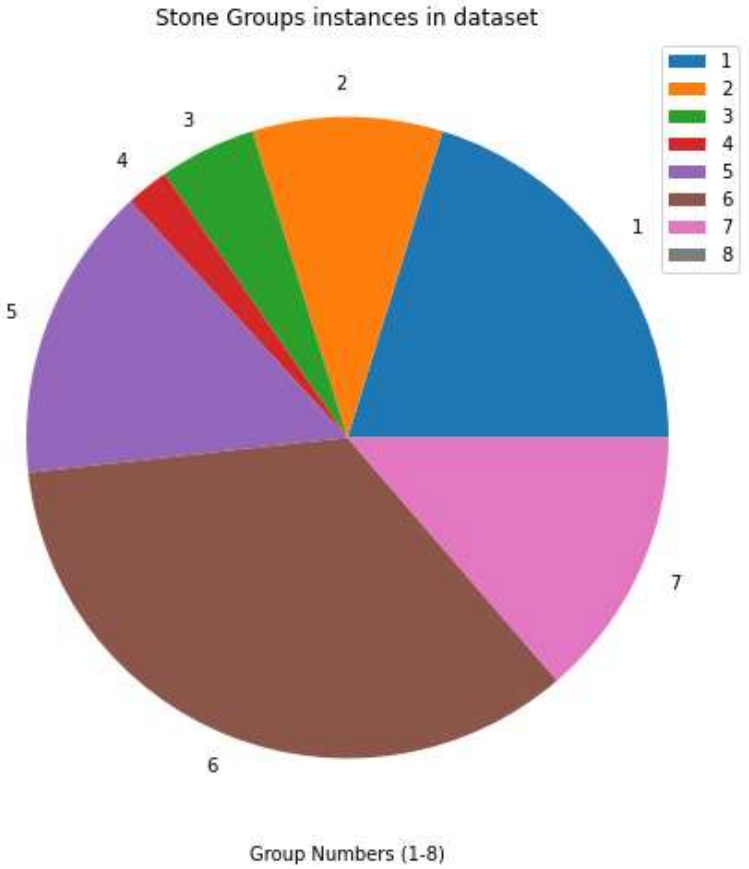

After seeing the correlations I dropped one of two stones that appeared to be the exact same stone.
Stones and minerals such as: “Lapis” and “Lasuli”, “Titanite” and “Sphene” and more..

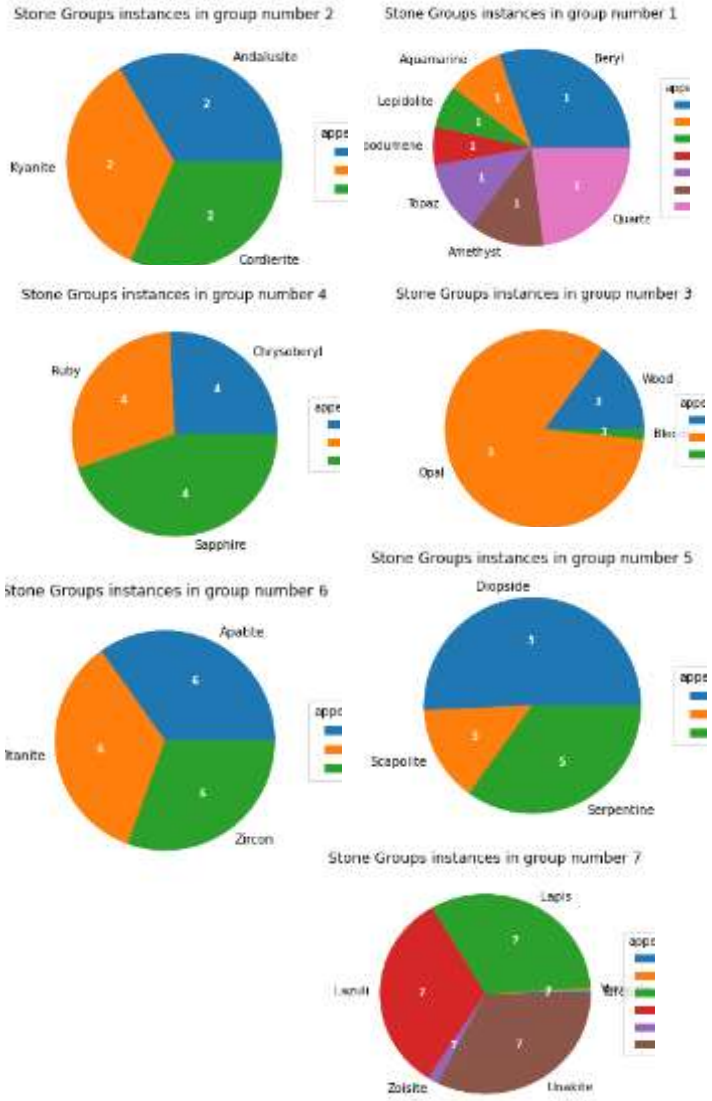
Gemstones project
by Noam Bassat

# Exploring & visualizing the Data

## Splitting the Stones into correlated stone groups

| | Stones in Group |
|---|---|
| 1 | ['Beryl', 'Aquamarine', 'Lepidolite', 'Spodume... |
| 2 | ['Andalusite', 'Kyanite', 'Cordierite'] |
| 3 | ['Wood', 'Opal', 'Opalize', 'Bloodstone'] |
| 4 | ['Chrysoberyl', 'Ruby', 'Sapphire', 'spinel'] |
| 5 | ['Diopside', 'Scapolite', 'Serpentine'] |
| 6 | ['Apatite', 'Titanite', 'Zircon'] |
| 7 | ['Turquoise', 'Variscite', 'Lapis', 'Lazuli', ... |
| 8 | ['Opal', 'Opalize', 'Bloodstone', 'wood'] |

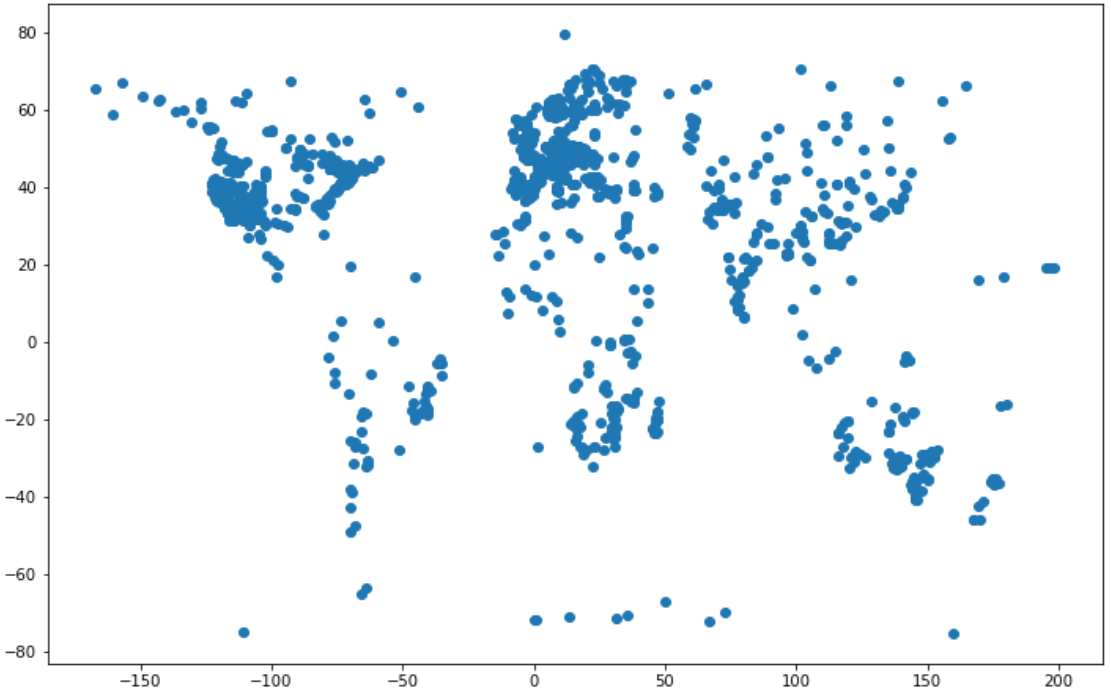


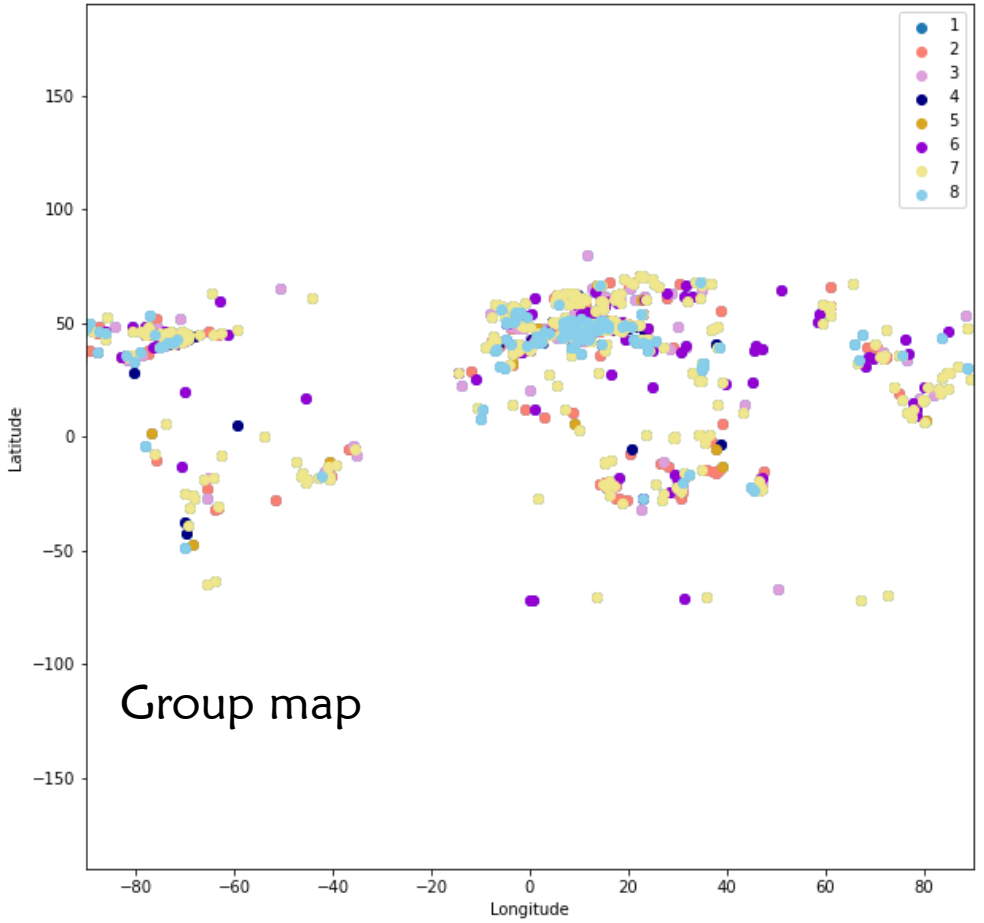



Gemstones project
by Noam Bassat

# Exploring & visualizing the Data

## Visualizing locations with scatter plot




No significant separation can be seen in the group graph…

# Exploring & visualizing the Data

## Mapping scatter plot

With plotly.express


There was a tiny amount of outliers that I didn't want to drop, so I manually fixed their coordinates..


After fixing the outliers

# Exploring & visualizing the Data

Gemstones project by Noam Bassat

*"Gemstone deposits that are carried by rivers are called placers or alluvial deposits. Rivers can transport gem-bearing rock many hundreds of miles. When the force of the current diminishes, the denser gems, such as diamond, zircon, garnet, sapphire, chrysoberyl, topaz, peridot and tourmaline, are deposited before the lighter quartz sand. Thus the gems left behind by the river tend to be concentrated in certain places. This makes mining the deposit much easier and more productive."*
*https://www.gemselect.com/english/other-info/gemstone-deposits.php*

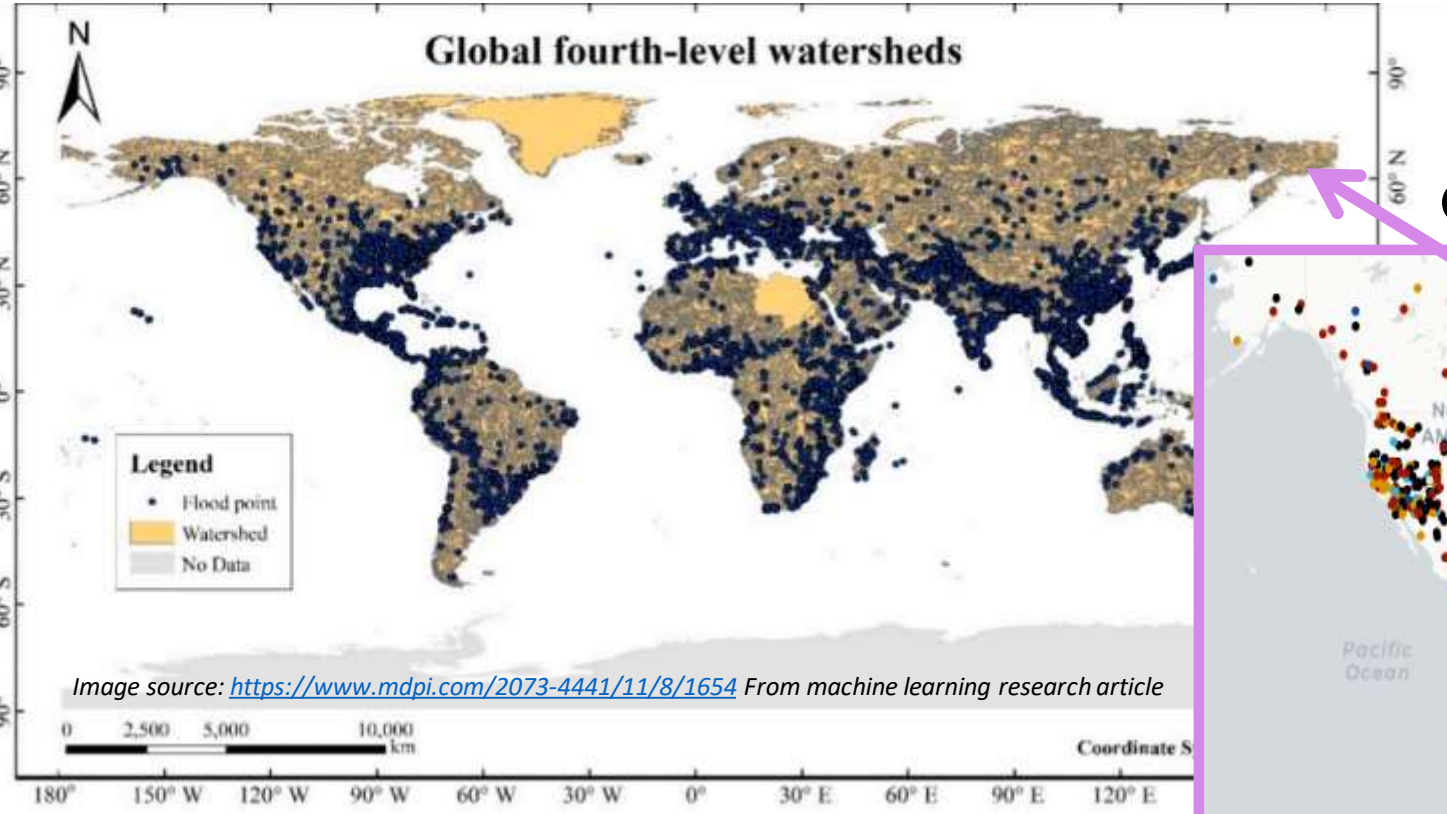


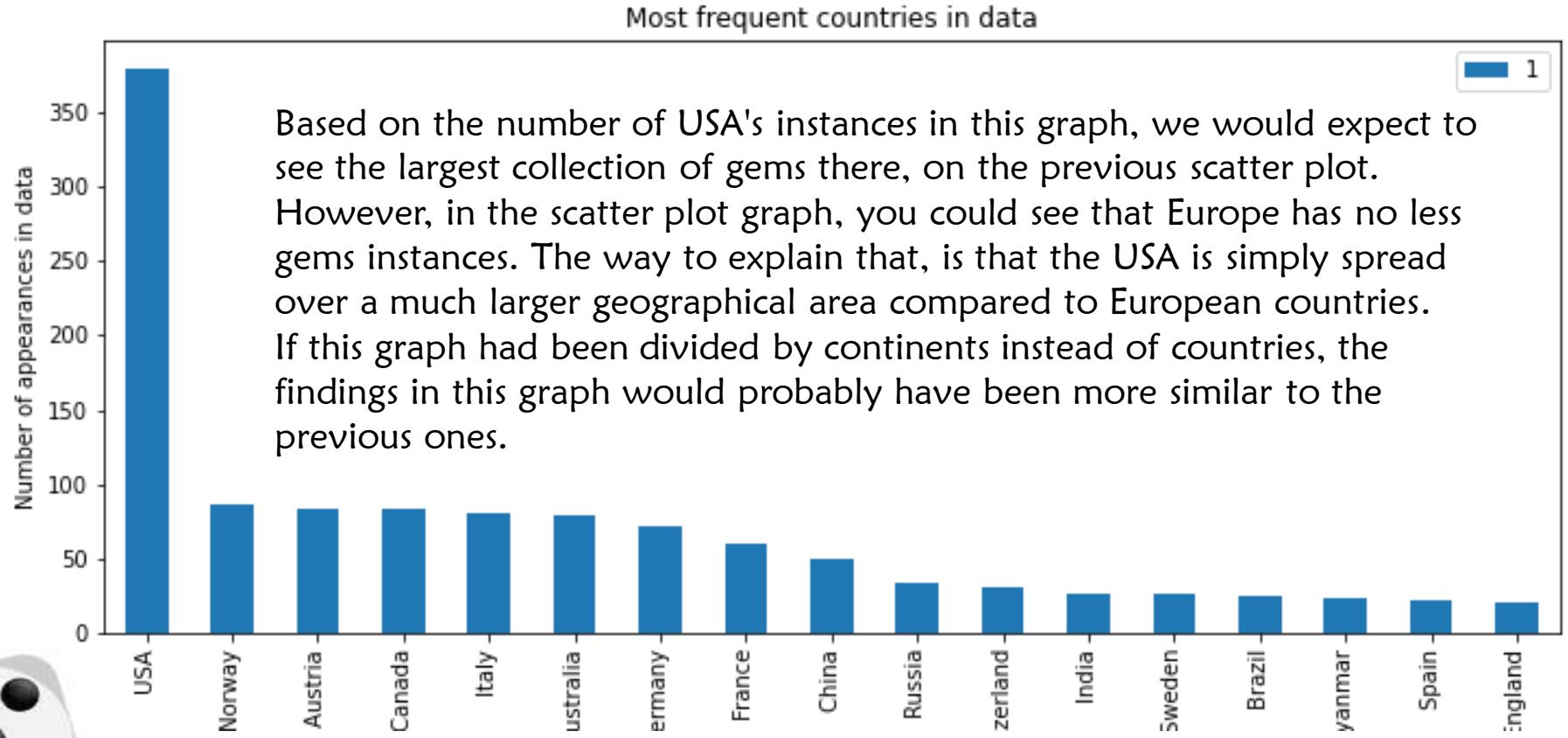

*Image source: https://www.mdpi.com/2073-4441/11/8/1654 From machine learning research article*

**Figure 1.** Global fourth-level watersheds and the location of flood inventori

Can you see the similarity?

# Exploring & visualizing the Data

## Visualizing Countries




Based on the number of USA's instances in this graph, we would expect to see the largest collection of gems there, on the previous scatter plot. However, in the scatter plot graph, you could see that Europe has no less gems instances. The way to explain that, is that the USA is simply spread over a much larger geographical area compared to European countries. If this graph had been divided by continents instead of countries, the findings in this graph would probably have been more similar to the previous ones.

# Modeling Data

## Part 1 – Unsupervised Learning: Clustering by Kmeans

Unfortunately, my division into groups gave really poor results when it came to the machine learning phase.
I tried to apply many algorithm models based on my groups labels, but the best accuracy they gained was lower the 0.1

So, the Kmeans algorithm has created some new clustering for me…
And it did it much better than I did.
Now that I have 8 clusters I tried some supervised models again and the accuracy has increased multiple times.

```
X = dataset_df.copy() #feature matrix

dataset_df["clusters"] = KMeans(n_clusters=8, n_init=500, max_iter=1000).fit_predict(X)
y=dataset_df["clusters"]
full_df["clusters"]=dataset_df["clusters"]
```

# Modeling Data

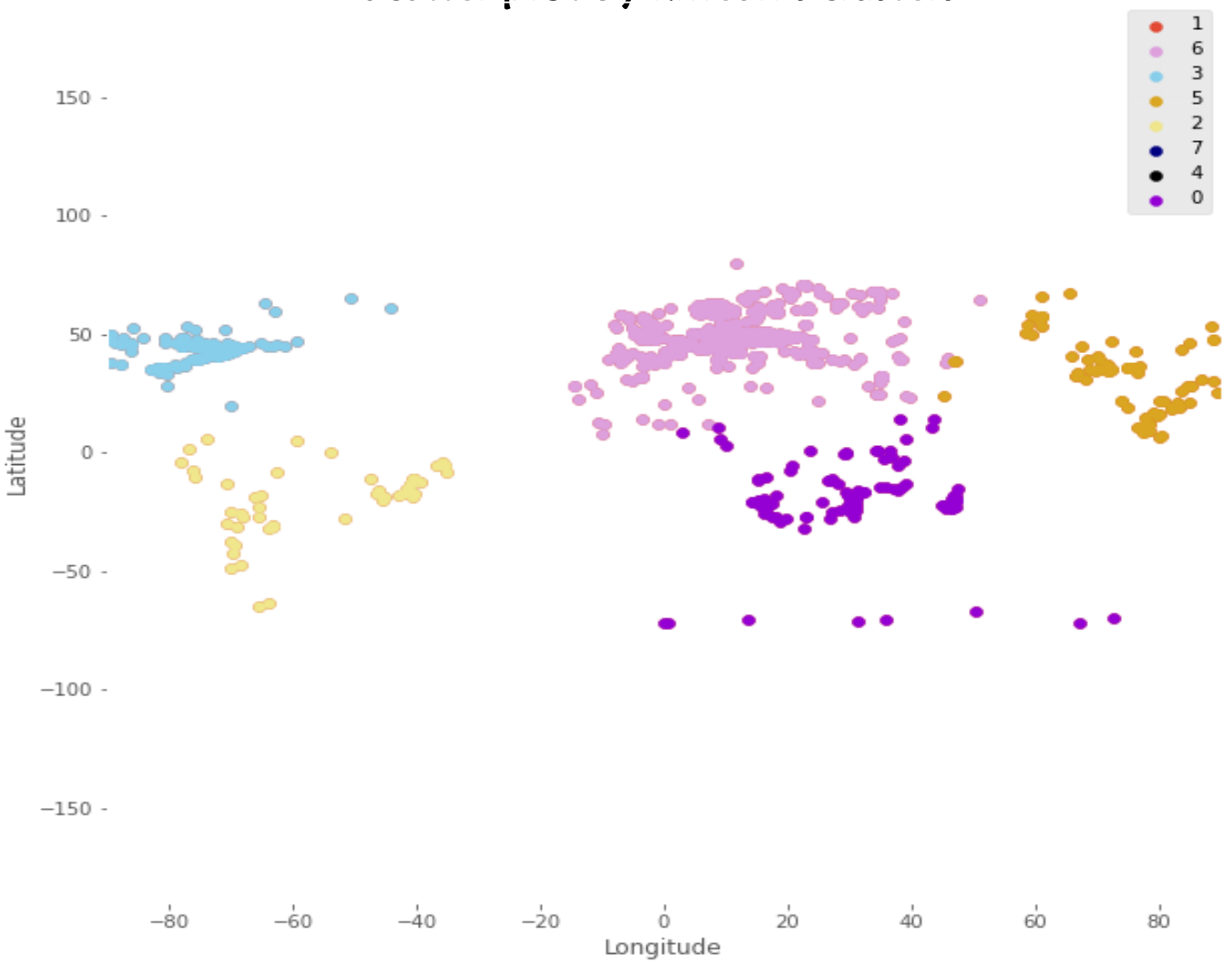
Part 2 – Supervised Learning: Comparison of the performance of 3 supervised algorithms: RandomForestClassifier, DecisionTreeClassifier, and KNN, using Sklearn.

```
DT
======
accuracy on train data 0.9955555555555555
accuracy on test data 0.9751552795031055
[[ 26   0   0   0   0   0   0   0]
 [  0 205   0   0   0   0   0   0]
 [  0   0  79   0   0   0   0   0]
 [  0   2   0  33   0   0   0   3]
 [  0   0   0   0  67   2   0   0]
 [  0   0   1   0   0  13   0   0]
 [  0   3   0   0   0   0  37   0]
 [  0   0   0   1   0   0   0  11]]
RandForest
======
accuracy on train data 0.9857777777777778
accuracy on test data 0.9606625258799172
[[ 26   0   0   0   0   0   0   0]
 [  0 205   0   0   0   0   0   0]
 [  0   0  78   0   1   0   0   0]
 [  0   5   0  32   0   0   0   1]
 [  0   0   2   0  67   0   0   0]
 [  0   0   1   0   5   7   1   0]
 [  0   2   0   0   0   0  38   0]
 [  1   0   0   0   0   0   0  11]]
KNN
======
accuracy on train data 0.7217777777777777
accuracy on test data 0.5548654244306418
[[ 15   5   0   5   0   0   0   1]
 [  0 177  11   0  16   1   0   0]
 [  0  28  36   0  15   0   0   0]
 [  3  14   5  15   1   0   0   0]
 [  0  39  13   1  16   0   0   0]
 [  1   4   4   0   2   1   2   0]
 [  2  23   0   3   5   0   7   0]
 [  0   8   2   1   0   0   0   1]]
```

**Decision Tree Classifier has the best results**

with accuracy on train data 0.9955555555555555 and accuracy on test data 0.9751552795031055

# Scatter plot by kmean's clusters



Gemstones project by Noam Bassat

# Results

## Attributes of Kmean's Clusters

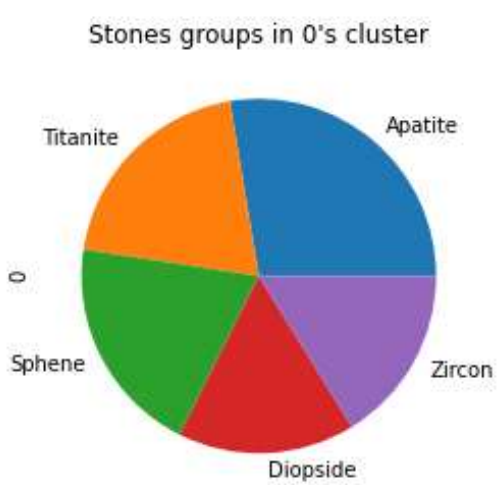
| | Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 |
|---|---|---|---|---|---|---|---|---|
| Number of locations | 679 | 68 | 216 | 104 | 102 | 261 | 125 | 53 |
| Average coordinates | -20.895355 , -56.725037 | , 48.719148 11.227510 | 39.069659, 112.783420 - | -29.315328 , 143.127625 | 39.803252, 123.978676 | 30.223338, 82.134699 | -19.690474 , 29.756632 | 42.901839 -75.451651 |
| Min Coordinates | -74.700000, 110.666670 - | 7.495260 -14.492780 , | 16.883330, 166.854170 - | -75.283330 , 107.731640 | 16.262780, 101.270000 | -4.650000, 45.083330 | -72.000000 , 0.083330 | 19.639860, -94.661670 |
| Max Coordinates | 5.632360, -35.270000 | 79.799840, 51.061020 | ,67.061940 -93.736110 | -2.330250, 179.958330 | 70.880000, 164.733330 | ,66.916670 107.200000 | 14.000000 72.500000 | 67.399170, -44.152500 |
| Most frequent Climate | 6 **Aw : Tropical savanna, wet** | 2 **Cfb : Temperate oceanic climate** | 13 **BSk : Cold semi-arid (steppe) climate** | 2 **Cfb : Temperate oceanic climate** | 5 **Cfa : Humid subtropical climate** | 3 **Cwa : Monsoon-influenced humid subtropical climate** | 7 **BSh : Hot semi-arid (steppe) climate** | 1 **Dfb : Warm-summer humid continental climate** |
| Most significant minerals | Ag Al As Au B Ba Be Bi | Ag Al As Au B Ba Be Bi Br C Ca Cd Ce | Ag Al As Au B Ba Be | Ag Al As Au B Ba Be Bi | Ag Al As Au B Ba Be | Ag Al As Au B | Ag Al As Au B Ba Be | Ag Al As Au B Ba |

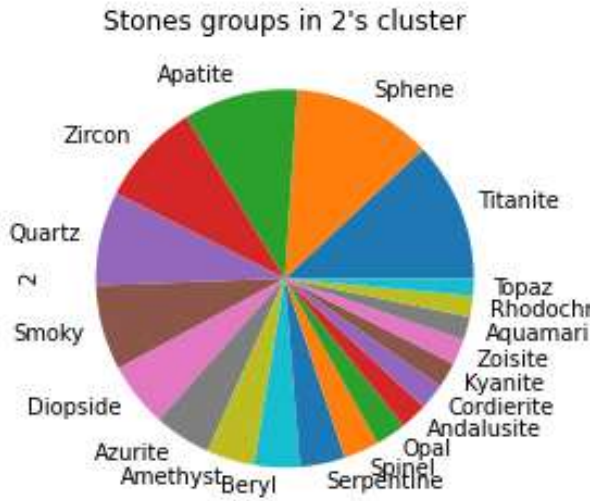
Gemstones project
by Noam Bassat

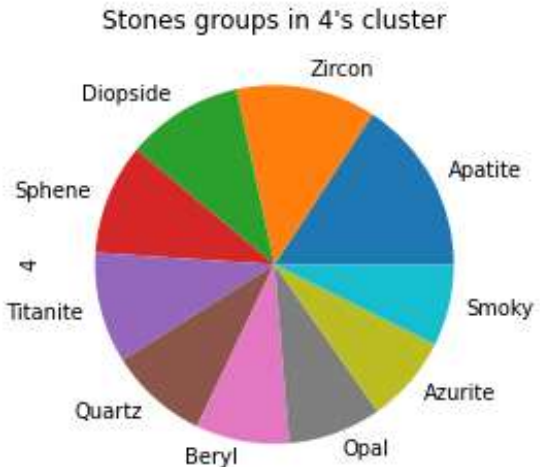
# Stones in each of kmean's clusters, sorted by their appearance in the data

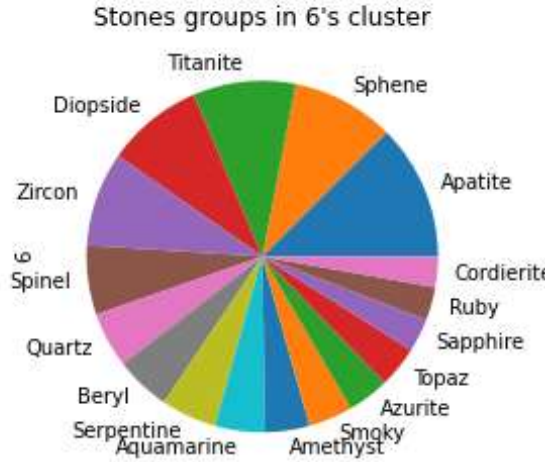
| Cluster 0 | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 | Cluster 7 |
|---|---|---|---|---|---|---|---|
| Total stones 319 Zircon 24 Diopside 20 Sphene 19 Quartz 17 Beryl 16 Opal 16 Azurite 15 Smoky Quartz 14 Spinel 10 Serpentine 10 Aquamarine 8 Sapphire 8 Topaz 8 Scapolite 7 Zoisite 6 Emerald 6 Variscite 5 Diamond 5 Lepidolite 4 Kyanite 4 Heliodor 4 Amethyst 4 Rhodochrosite 4 Andalusite 4 Ruby 3 Turquoise 3 Cordierite 3 Gaspeite 2 Goshenite 2 Rhodonite 2 Rose Quartz 2 Sodalite 2 Moonstone 2 Wood 2 Spodumene 2 Chrysoprase 1 Jade 1 Amazonite 1 Malachite 1 Sphalerite 1 Petrified 1 Opalized 1 | Total stones 2348 Sphene 252 Zircon 186 Quartz 169 Smoky Quartz 154 Diopside 116 Azurite 103 Amethyst 86 Beryl 82 Serpentine 81 Spinel 64 Opal 55 Andalusite 49 Cordierite 47 Kyanite 46 Zoisite 46 Aquamarine 42 Rhodochrosite 38 Topaz 31 Scapolite 28 Sodalite 28 Labradorite 18 Rhodonite 16 Sapphire 14 Rose 12 Lepidolite 11 Diamond 11 Variscite 10 Chrysoberyl 9 Turquoise 9 Ruby 9 Citrine 8 Wood 7 Petrified 5 Spodumene 5 Heliodor 4 Emerald 4 Jadeite 3 Goshenite 2 Iolite 2 Gaspeite 2 Garnet 2 Opalized 2 Jade 2 Moonstone 2 Bloodstone 2 Nephrite 2 Sit 2 Amazonite 2 Fluorite 2 Chrysoprase 1 Maw 1 Fire 1 Malachite 1 Sunstone 1 Hematite 1 Moldavite 1 Agate 1 | Total stones 766 Quartz 56 Azurite 51 Zircon 44 Smoky Quartz 40 Diopside 39 Opal 37 Sphene 36 Beryl 36 Amethyst 34 Serpentine 32 Rhodochrosite 31 Aquamarine 20 Rose 14 Rhodonite 14 Spinel 12 Wood 12 Morganite 12 Heliodor 12 Andalusite 12 Zoisite 10 Topaz 10 Petrified 10 Citrine 10 Spodumene 9 Lepidolite 9 Scapolite 8 Cordierite 8 Goshenite 6 Turquoise 6 Kyanite 6 Sodalite 4 Chrysoberyl 3 Variscite 3 Sapphire 3 Diamond 3 Nephrite 3 Jade 3 Labradorite 3 Emerald 2 Ruby 2 Moonstone 2 Opalized 2 Amazonite 1 Benitoite 1 Fluorite 1 Bloodstone 1 Agate 1 Opalite 1 Hematite 1 Tourmaline 1 Iris 1 Sunstone 1 Peridot 1 | Total stones 504 Sphene 37 Diopside 35 Zircon 34 Spinel 25 Quartz 20 Serpentine 20 Beryl 20 Aquamarine 18 Smoky Quartz 16 Amethyst 16 Topaz 15 Azurite 15 Sapphire 12 Ruby 12 Cordierite 11 Andalusite 9 Scapolite 9 Lepidolite 7 Spodumene 7 Sodalite 7 Rhodochrosite 6 Heliodor 6 Emerald 6 Chrysoberyl 5 Moonstone 5 Rhodonite 4 Zoisite 4 Goshenite 4 Morganite 4 Labradorite 4 Rose 4 Kyanite 3 Opal 2 Citrine 2 Turquoise 2 Aventurine 2 Jadeite 2 Iolite 1 Fire 1 Agate 1 Petrified 1 Wood 1 Tourmaline 1 Sunstone 1 Peridot 1 | Total stones 211 Zircon 22 Beryl 14 Quartz 12 Aquamarine 12 Sphene 11 Smoky Quartz 10 Emerald 8 Spinel 7 Diopside 7 Amethyst 6 Opal 6 Spodumene 5 Lepidolite 5 Kyanite 4 Cordierite 4 Serpentine 4 Chrysoberyl 3 Zoisite 3 Azurite 3 Wood 3 Rose 2 Sapphire 2 Goshenite 2 Citrine 2 Diamond 2 Topaz 2 Opalized 2 Andalusite 2 Jadeite 2 Morganite 2 Labradorite 1 Rhodochrosite 1 Petrified 1 Rhodonite 1 Sodalite 1 | 397 Zircon Total stones 27 Beryl 27 Quartz 26 Sphene 24 Aquamarine 22 Smoky Quartz 18 Diopside 15 Topaz 15 Spinel 15 Sapphire 11 Serpentine 11 Amethyst 10 Lepidolite 9 Ruby 9 Opal 8 Emerald 8 Rose 8 Goshenite 8 Azurite 7 Kyanite 7 Sodalite 7 Morganite 6 Diamond 6 Spodumene 6 Andalusite 5 Chrysoberyl 5 Cordierite 4 Zoisite 3 Scapolite 3 Citrine 2 Labradorite 1 Tourmaline 1 Garnet 1 Iolite 1 Amazonite 1 Variscite 1 Rhodonite 1 Chrysoprase 1 Rhodochrosite 1 Peridot 1 | 861 Total stones Quartz 92 Smoky Quartz 70 Sphene 68 Zircon 58 Beryl 44 Diopside 44 Amethyst 38 Serpentine 37 Aquamarine 26 Scapolite 25 Opal 23 Kyanite 22 Rose 18 Spinel 18 Spodumene 15 Azurite 14 Sodalite 9 Lepidolite 9 Andalusite 8 Morganite 8 Heliodor 8 Amazonite 7 Rhodochrosite 7 Topaz 7 Citrine 6 Cordierite 6 Zoisite 5 Sapphire 5 Labradorite 5 Sunstone 5 Variscite 4 Turquoise 4 Chrysoberyl 3 Rhodonite 3 Wood 3 Goshenite 2 Garnet 2 Jadeite 2 Petrified 2 Red 2 Tourmaline 2 Jade 1 Charoite 1 Hematite 1 Nephrite 1 Opalized 1 Ruby 1 Diamond 1 | Total stones 191 Sphene 21 Zircon 17 Diopside 17 Serpentine 7 Spinel 5 Azurite 5 Chrysoberyl 5 Beryl 5 Rhodonite 5 Topaz 5 Zoisite 5 Sodalite 4 Amethyst 4 Opal 3 Quartz 3 Kyanite 3 Cordierite 3 Rhodochrosite 3 Scapolite 3 Ruby 2 Smoky Quartz 2 Andalusite 2 Emerald 2 Lepidolite 2 Sapphire 2 Jade 1 Spodumene 1 Nephrite 1 Labradorite 1 Diamond 1 Variscite 1 |

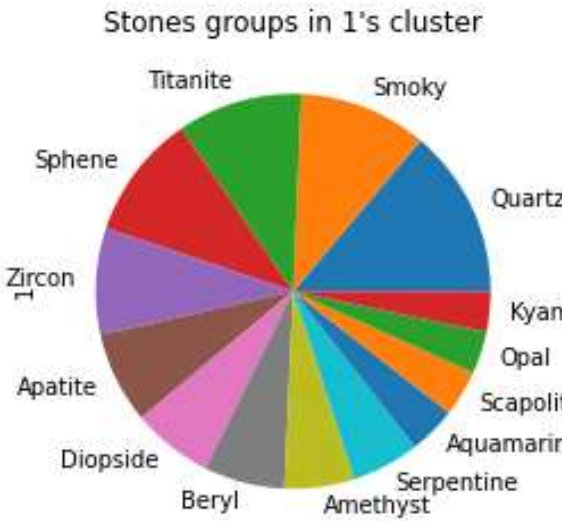
# Stones in each of kmean's clusters, sorted by their appearance in the data

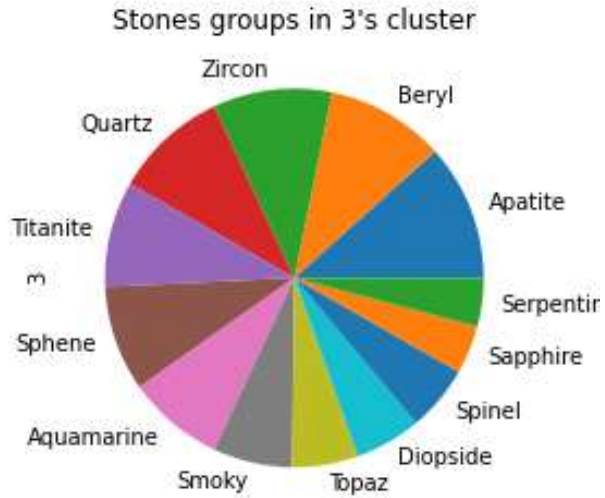


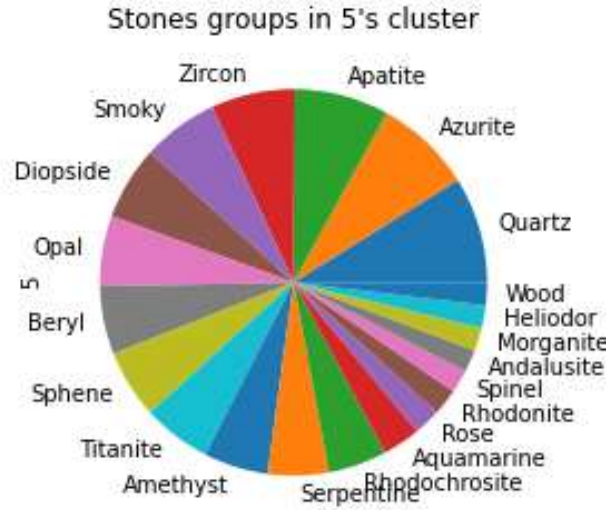



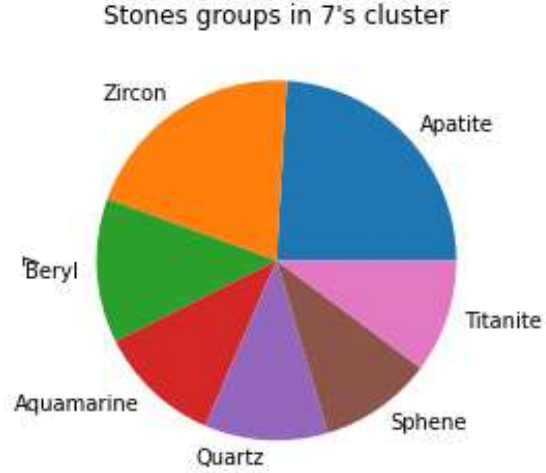



# Accuracy change based on specific features

```
X_locations=X.loc[:,'Latitude':'Longitude']
```

```
=== Locations ===
   DT
   ======
    accuracy on train data 0.9946666666666667
    accuracy on test data 0.9834368530020704
   RandForest
   ======
    accuracy on train data 0.9991111111111111
    accuracy on test data 0.9917184265010351
   KNN
   ======
    accuracy on train data 0.992
    accuracy on test data 0.989648033126294
```

```
X_Minerals = X.loc[:,'Ag':'Zr']
```

```
=== Minerals ===
   DT
   ======
    accuracy on train data 0.5955555555555555
    accuracy on test data 0.35403726708074534
   RandForest
   ======
    accuracy on train data 0.5902222222222222
    accuracy on test data 0.4451345755693582
   KNN
   ======
    accuracy on train data 0.56
    accuracy on test data 0.38095238095238093
```

```
X_climate = pd.DataFrame(X['Climate'])
```

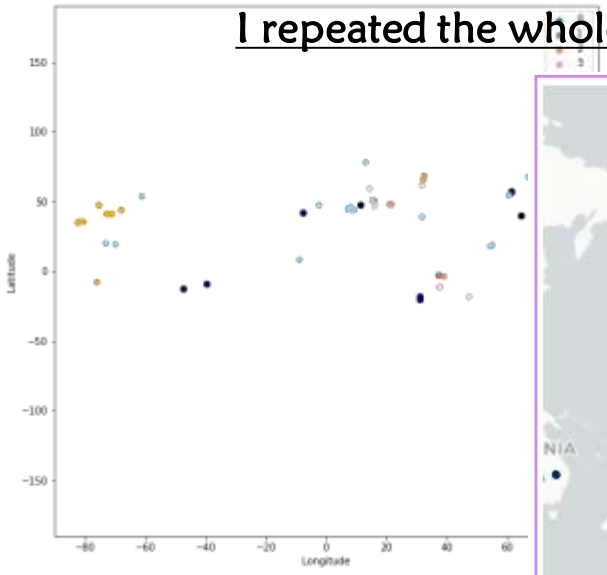
```
=== Climate ===
   DT
   ======
    accuracy on train data 0.6133333333333333
    accuracy on test data 0.6024844720496895
   RandForest
   ======
    accuracy on train data 0.6088888888888889
    accuracy on test data 0.5962732919254659
   KNN
   ======
    accuracy on train data 0.5235555555555556
    accuracy on test data 0.494824016563147
```

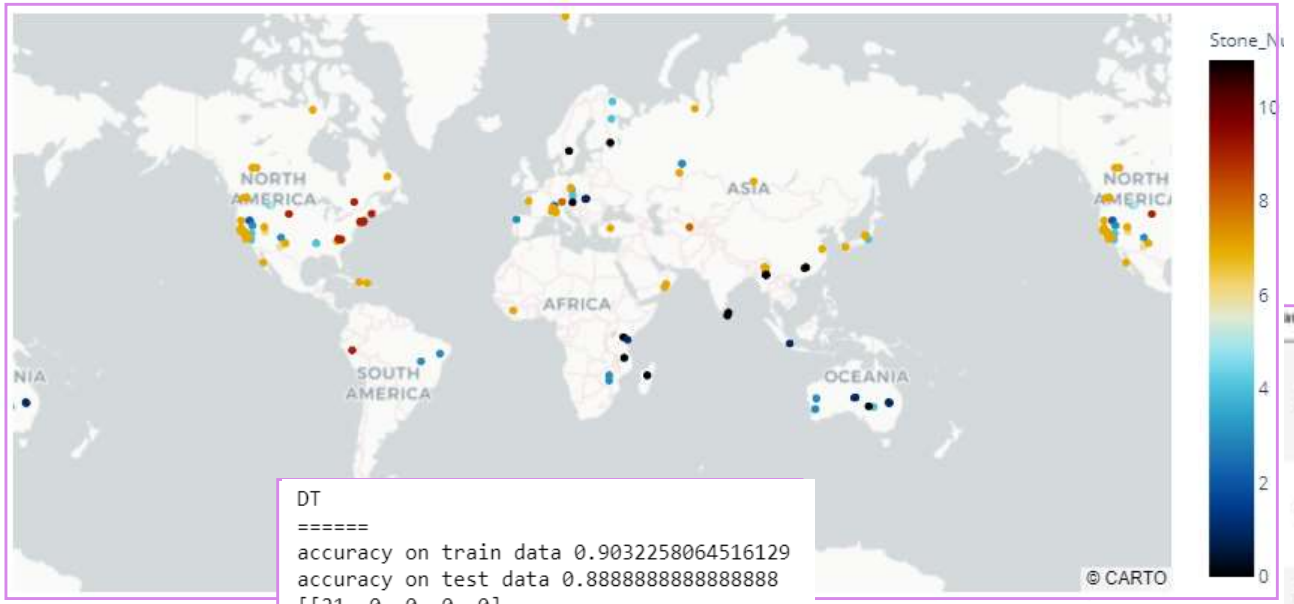
# +Bonus:

## The rarest stones in the world

Gemstones project
by Noam Bassat

I repeated the whole process to another dataframe, that contains 11 stones that are the rarest in the world.




| | Stone Refernced | Climate | Latitude | Longitud | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Tanzanite | 6.0 | -2.41667 | 37.00833 | 0 | 1 | 1 | 0 | 1 | 0 | ... | | | | | | | | | | | | | | | |
| 1 | Black_Opal | 18.0 | -3.47167 | 38.73139 | 0 | 1 | 0 | 0 | 1 | 0 | ... | | | | | | | | | | | | | | | |
| 2 | Black_Opal | 7.0 | -28.56854 | 147.98010 | 0 | 1 | 0 | 0 | 0 | 0 | ... | | | | | | | | | | | | | | | |
| 3 | Black_Opal | 7.0 | -28.53667 | 147.76083 | 0 | 0 | 0 | 0 | 0 | 0 | ... | | | | | | | | | | | | | | | |
| 4 | Black_Opal | 7.0 | -28.55167 | 147.78167 | 0 | 0 | 0 | 0 | 0 | 0 | ... | | | | | | | | | | | | | | | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | | | | | | | | | | | | | | | |
| 153 | Taaffeite | 4.0 | 61.55000 | 31.50000 | 1 | 1 | 1 | 1 | 1 | 1 | ... | | | | | | | | | | | | | | | |
| 154 | Taaffeite | 19.0 | 7.78060 | 80.81840 | 0 | 1 | 0 | 0 | 0 | 0 | ... | | | | | | | | | | | | | | | |
| 155 | Taaffeite | 19.0 | 6.63702 | 80.36177 | 0 | 1 | 0 | 0 | 0 | 0 | ... | | | | | | | | | | | | | | | |
| 156 | Taaffeite | 4.0 | 59.85538 | 14.26480 | 1 | 1 | 1 | 0 | 1 | 1 | ... | | | | | | | | | | | | | | | |
| 157 | Taaffeite | 6.0 | -10.95944 | 37.32917 | 0 | 1 | 0 | 0 | 1 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | | | | | | 11 |

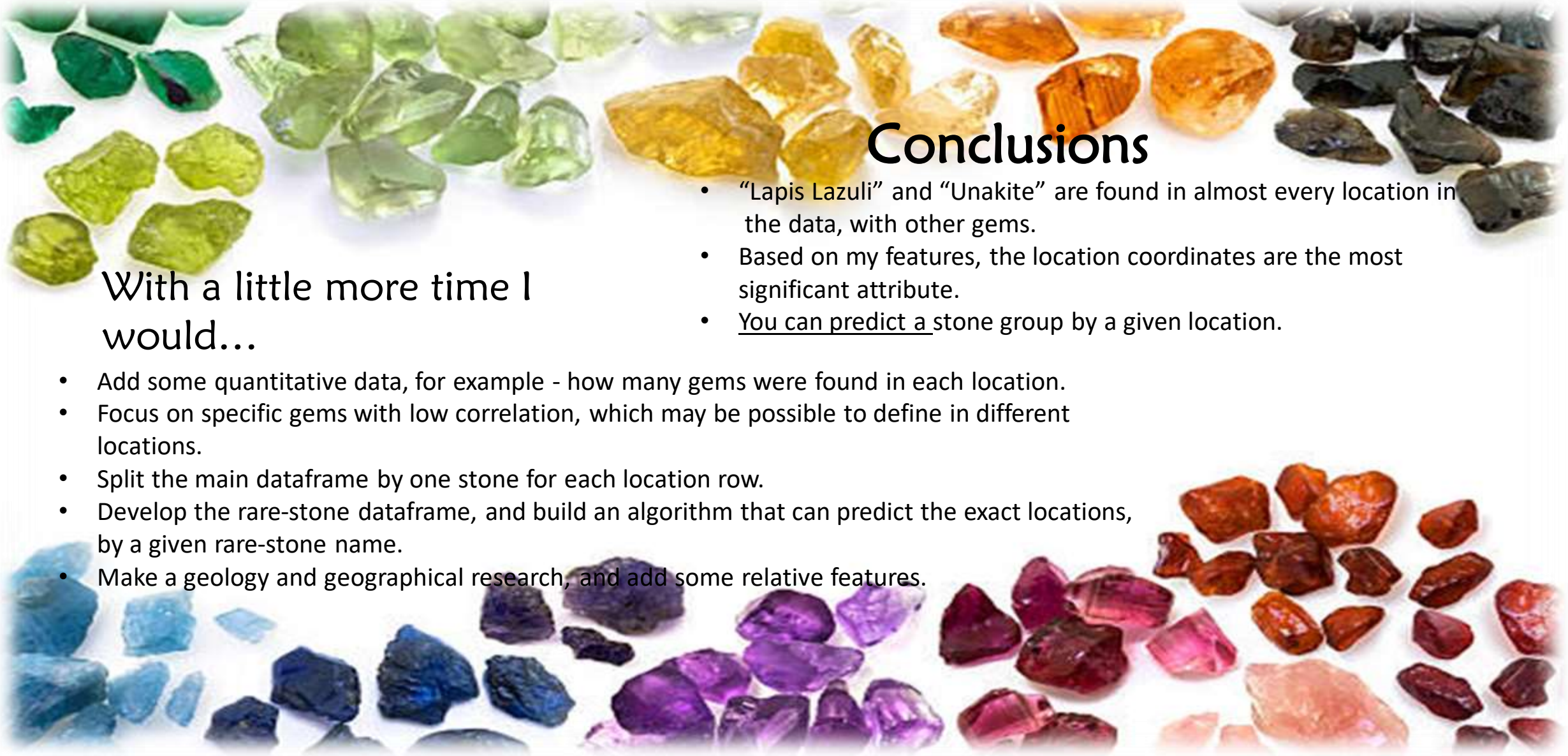
156 rows × 69 columns

```
DT
======
accuracy on train data 0.9032258064516129
accuracy on test data 0.8888888888888888
[[21  0  0  0  0]
 [ 0 16  0  0  0]
 [ 0  0 12  0  0]
 [ 0  2  0  7  0]
 [ 5  0  0  0  0]]
RandForest
======
accuracy on train data 0.967741935483871
accuracy on test data 0.8253968253968254
[[16  0  3  2  0]
 [ 0 14  2  0  0]
 [ 0  0 12  0  0]
 [ 0  1  2  6  0]
 [ 0  0  1  0  4]]
```

| | anate | Minerals | Latitude | Longitude | Ag | Al | As | ... | Th | Ti | Tl | U | V | W | Y | Zn | Zr | Stor |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 6.0 | Al As B C Ca Ce Cl Cr Cu F Fe Ge Ge H K Mg Mn ... | -2.41667 | 37.00833 | 0 | 1 | 1 | ... | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | |
| | 18.0 | Al B C Ca Ce Cl Cr Cu F Fe Gd H La Li Mg Na Nd ... | -3.47167 | 38.73139 | 0 | 1 | 0 | ... | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | |
| | 7.0 | Al Ca H Mg Na O Si | -28.56854 | 147.98010 | 0 | 1 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | | | | Australia,Lighting... | | | | | | | | | | | | | | |
| 3 | https://www.mindat.org/loc-235813.html | Black_Opal | | Lightning Ridge, Finch Co., New South Wales,... | 7.0 | H O Si | | -28.53667 | 147.76083 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 4 | https://www.mindat.org/loc-146633.html | Black_Opal | | Coocoran Opal Fields, Lightning Ridge, Finch... | 7.0 | Valid Species H O Si | | -28.55167 | 147.78167 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 153 | https://www.mindat.org/loc-3142.html | Taaffeite | | Republic of Karelia, Russia,Pitkyarantami... | 4.0 | Ag Al As Au B Ba Be Bi C Ca Ce Cl Co Cu F Fe O... | | 61.55000 | 31.50000 | 1 | 1 | 1 | ... | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | |
| 154 | https://www.mindat.org/loc-123009.html | Taaffeite | | Polonnaruwa District, North Central Province... | 19.0 | Al Be Ca H K Mg Na O Si W | | 7.78060 | 80.81840 | 0 | 1 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | |
| 155 | https://www.mindat.org/loc-3167.html | Taaffeite | | Ratnapura, Ratnapura District, Sabaragamuwa... | 19.0 | Al Be Mg O | | 6.63702 | 80.36177 | 0 | 1 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| | | | | Långban Ore | | Ag Al As B Ba Be Bi C Ca | | | | | | | | | | | | |

Gemstones project
by Noam Bassat

# My final machine



# Conclusions

- “Lapis Lazuli” and “Unakite” are found in almost every location in the data, with other gems.
- Based on my features, the location coordinates are the most significant attribute.
- You can predict a stone group by a given location.

## With a little more time I would…

- Add some quantitative data, for example - how many gems were found in each location.
- Focus on specific gems with low correlation, which may be possible to define in different locations.
- Split the main dataframe by one stone for each location row.
- Develop the rare-stone dataframe, and build an algorithm that can predict the exact locations, by a given rare-stone name.
- Make a geology and geographical research, and add some relative features.

# THANKS FOR WATCHING

NoamBassat92@gmail.com