PIRI REIS UNIVERSITY

Faculty of Economics and Administrative Sciences

Management Information Systems

Data Mining

Exploring Dataset and Implementing a Data Mining Task (Python)

Kadirhan GÖZÜKOCA and Melih AKBAŞ

201735020-201635501

Volkan USLAN

CONTENTS

1. INTRODUCTION 3
1.1. PROJECT TOPIC AND PURPOSE 3

2. REQUIRED TOOLS 3
2.1. THE PLATFORMS USED TO DEVELOP THE PROJECT 3

2.2. DOWNLOAD LINKS 3

3. SOLUTION APPROACH 3
3.1. FLOWCHART 3

4. SYSTEM DEVELOPMENT PROCESS 4
4.1. DEVELOPMENT ENVIRONMENTS AND PROCESS 4

4.2. DATASET'S KAGGLE LINK AND EXPLAIN: 4

5. BUSINESS CASE 5
5.1. DEVELOPED PROJECT, STEP BY STEP 5

5.1.1. STEP 1 5

5.1.2. STEP 2 6

5.1.3. STEP 3 6

5.1.4. STEP 4 7

5.1.5. STEP 5 8

5.2. USE CASE DIAGRAM 14

5.3. FLOWCHART OF THE TEST BUSINESS SCENARIO 14

5.4. USER'S GUIDE 14

6. CONCLUSION 14
6.1. THE PROJECT CAN BE IMPROVED BY ADDING WHAT? 14

6.2. ADVANTAGES AND DISADVANTAGES OF PROJECT 14

6.3. SIMILAR PROJECTS 15

6.4. WHAT HAVE I GAİNED FROM THIS PROJECT? 15

7. REFERENCES 16

8. IMAGES 16

1. INTRODUCTION

1.1. PROJECT TOPIC AND PURPOSE

In this assignment, We will explore dataset and implement a data mining data mining task using Kaggle. Firstly, We choosed a dataset from kaggle.com. It is include earthquakes in Turkey between 1910-2017. We will analyze some earthquakes like the biggest one, “where was the most earthquake” and also We will visualise them.

2. REQUIRED TOOLS

2.1. THE PLATFORMS USED TO DEVELOP THE PROJECT

The platforms We used to use and make training the Dataset: Jupyter Notebook (Python).

2.2. DOWNLOAD LINKS

- https://www.anaconda.com/products/individual (launch the Jupyter Notebook)

3. SOLUTION APPROACH

3.1. FLOWCHART

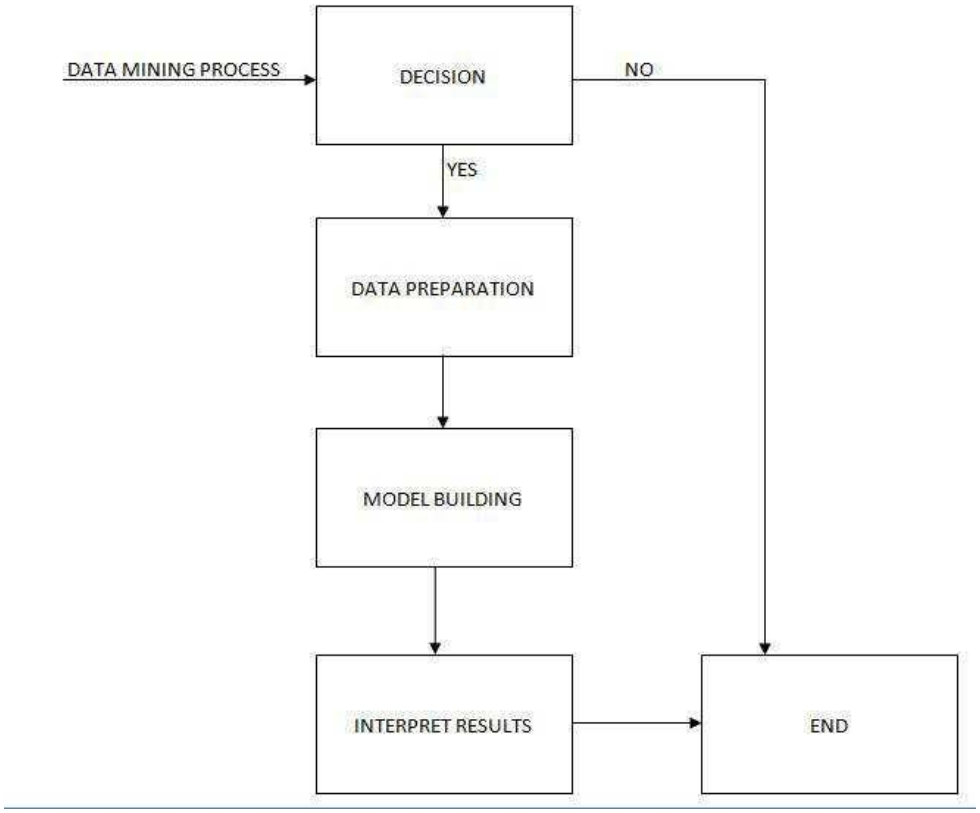

Image 1 : Flowchart for Data Mining

4. SYSTEM DEVELOPMENT PROCESS

4.1. DEVELOPMENT ENVIRONMENTS AND PROCESS

We did our Project using python (Jupyter Notebook). We started to research a data and We found as a result of research in python. We wanted to create an interface access user over Qt-py, but we could not connect code and dataset with any interface. Therefore, there is no user interface in our Project.

4.2. DATASET'S KAGGLE LINK AND EXPLAIN:

https://www.kaggle.com/caganseval/earthquake

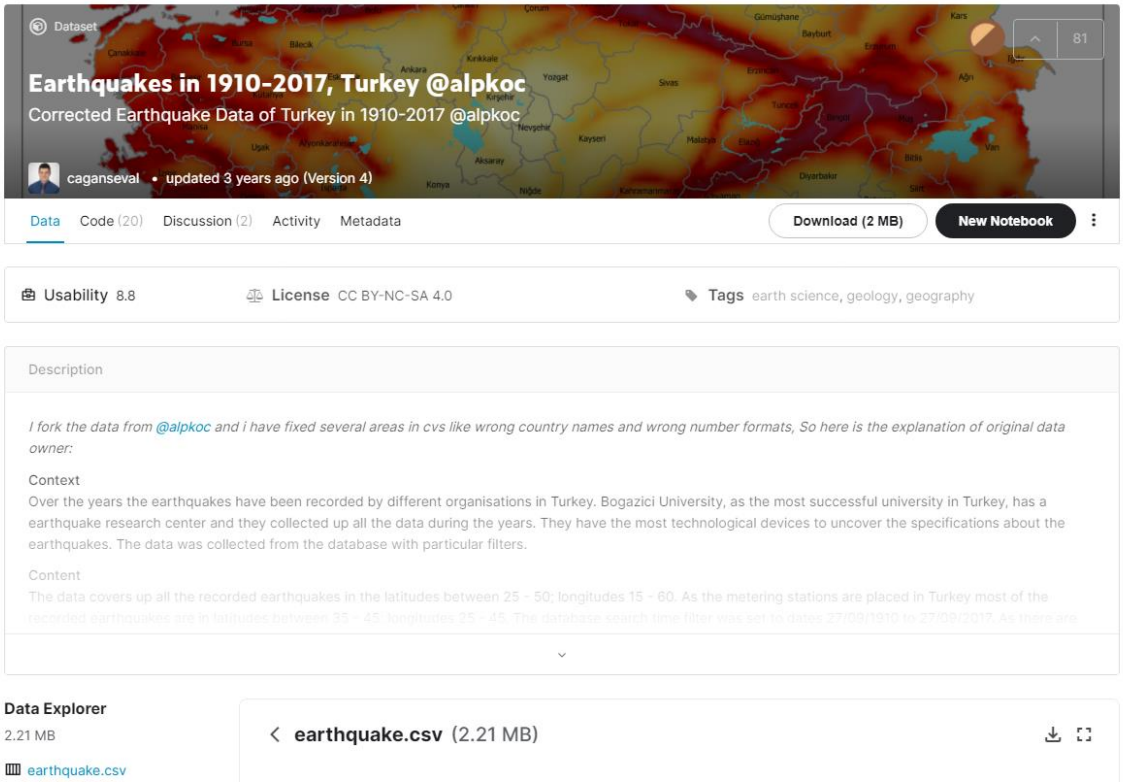

Image 2 : Kaggle Link

5. BUSINESS CASE

5.1. DEVELOPED PROJECT, STEP BY STEP

5.1.1. STEP 1

First of all, in order to use the Jupyter Notebook, We enter the link below and download the anaconda and install it.,
(https://www.anaconda.com/products/individual)

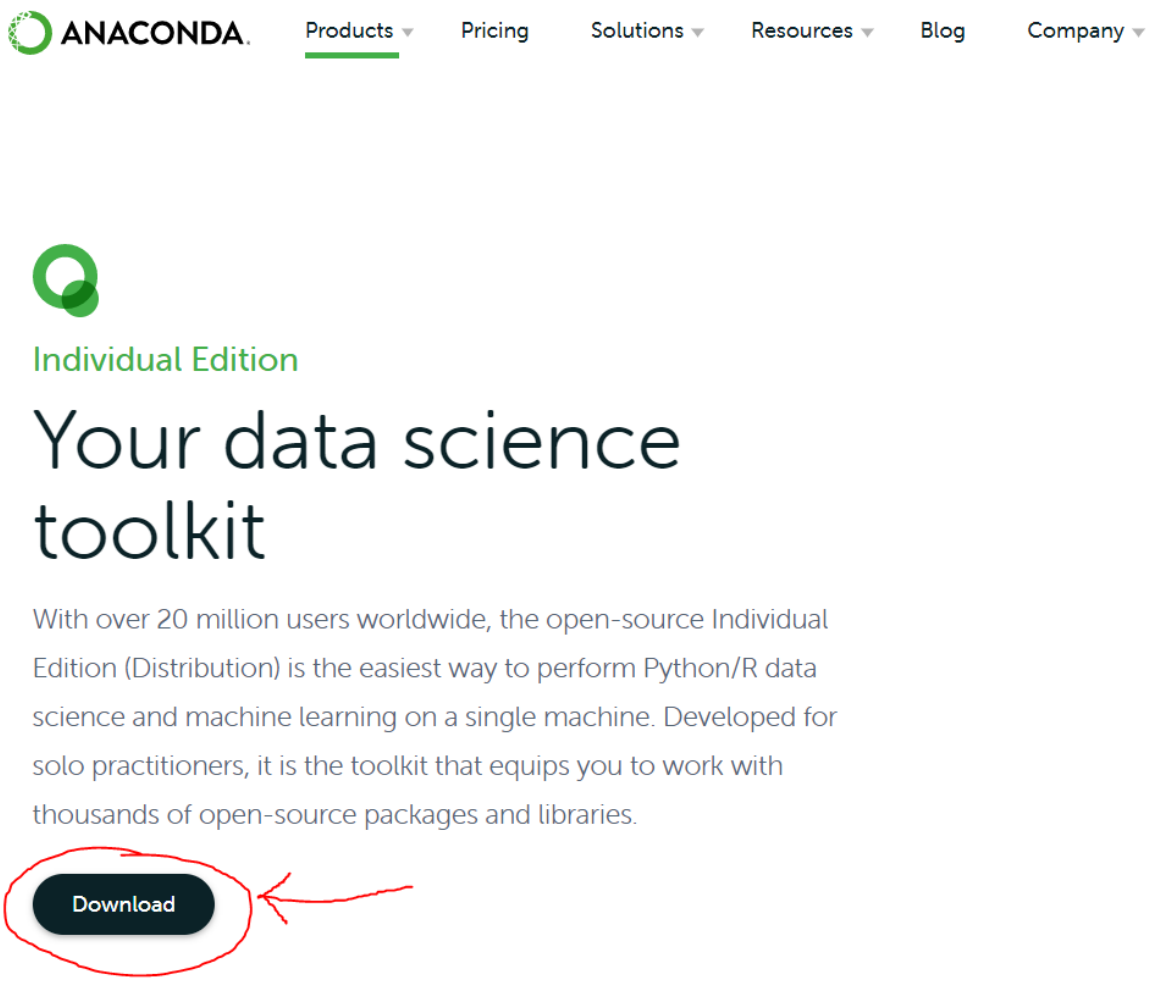

Image 3 : Anaconda Download

5.1.2. STEP 2

We login to the anaconda application we have downloaded and start the Jupyter Notebook.

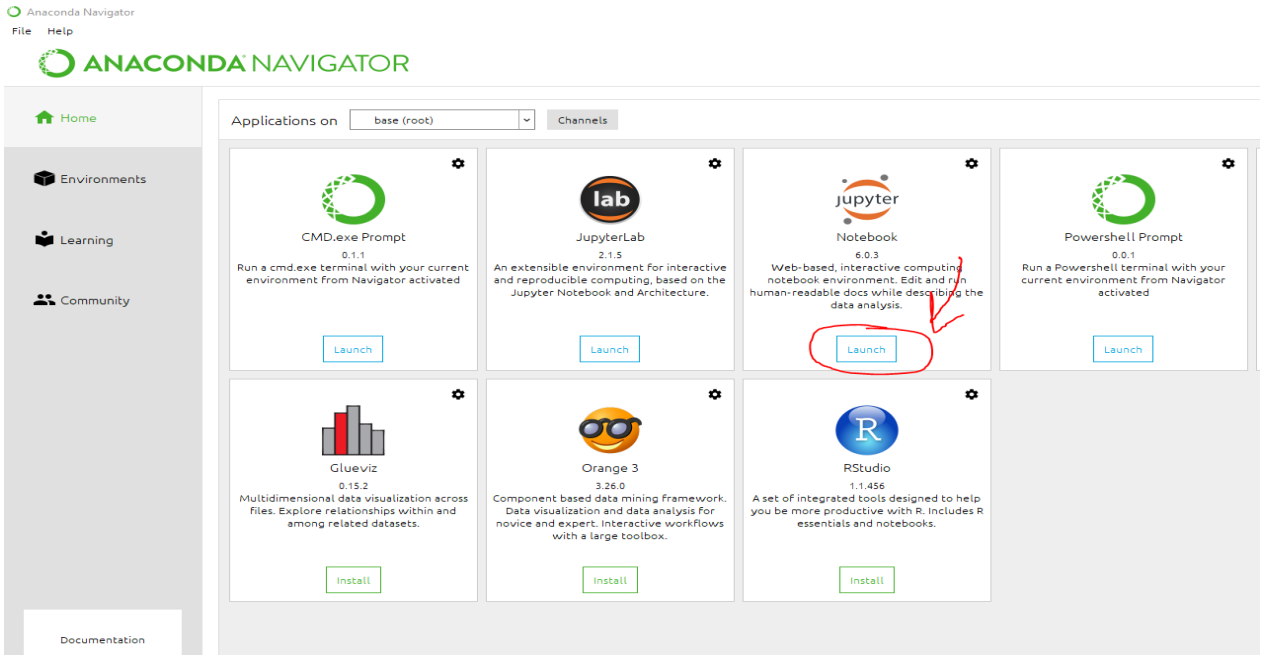

Image 4 : Jupyter Notebook Launch

5.1.3. STEP 3

On the Kaggle.com, you can choice a dataset. This one is ours.

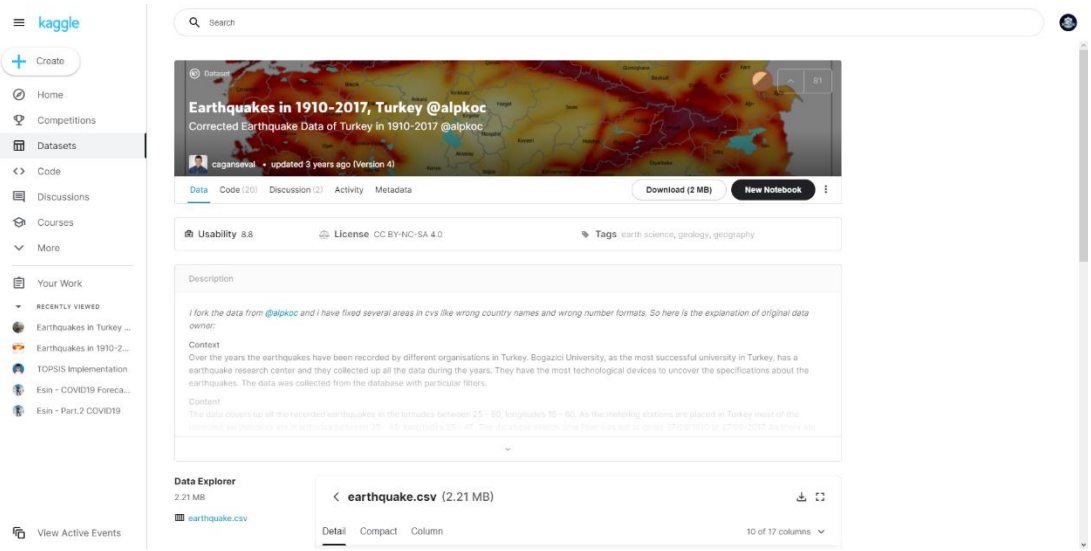

Image 5 : Dataset Page

5.1.4. STEP 4

On the opened web page, on the far right, under the new tab, there is python 3, and We click on it, and now We have access to the page where We will write the code.

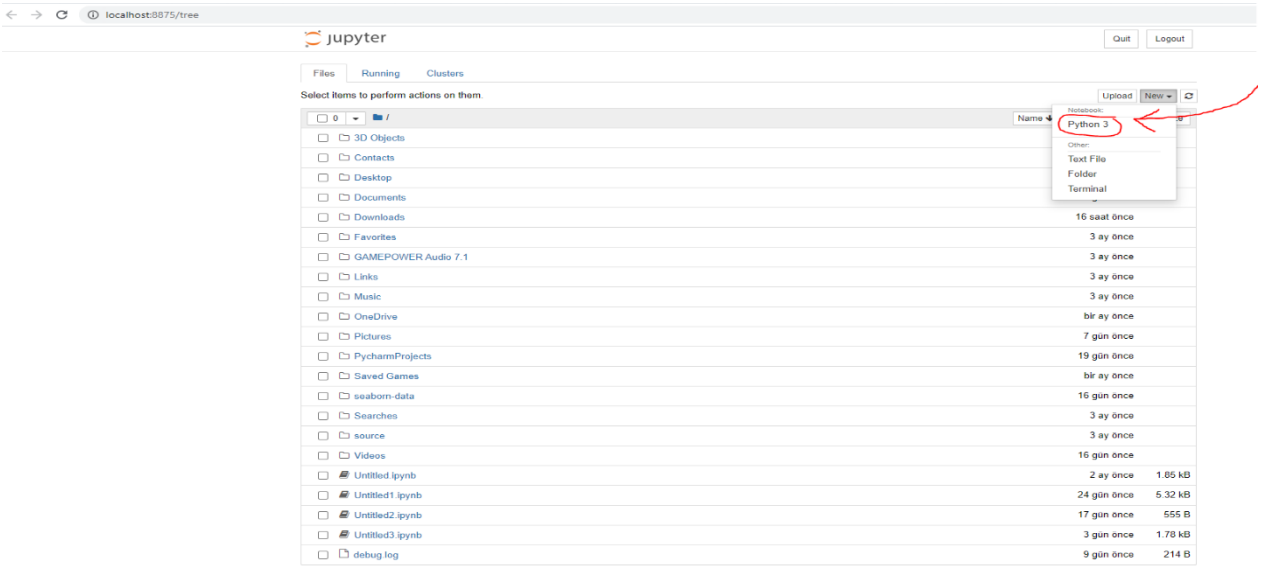

Image 2 : New Page in Jupyter Notebook

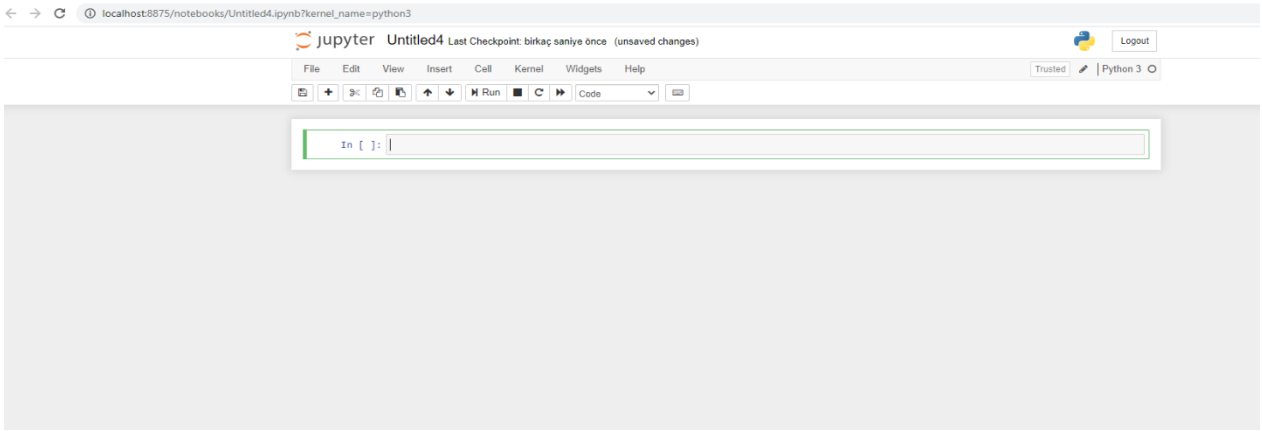

Image 3 : New Page

5.1.5. STEP 5

Now We write the codes on the opened page.

STARTING WITH LIBRARIY IMPORTS

In [28]:
```
import numpy as np # linear algebra
import pandas as pd # data processing
import matplotlib.pyplot as plt
import seaborn as sns #visualization
import os
##author @Kadirhan Gözükoca
```

LET'S CHECK OUR DATASET

In [25]:
```
data = pd.read_csv (r'C:\Users\trapn\Desktop\earthquake.csv')
```

In [26]:
```
data.head(10)
```

Out[26]:

	id	date	time	lat	long	country	city	area	direction	dist	depth	xm	md	richter	mw	ms	mb
0	2.000000e+13	2003.05.20	12:17:44 AM	39.04	40.38	turkey	bingol	baliklicay	west	0.1	10.0	4.1	4.1	0.0	NaN	0.0	0.0
1	2.010000e+13	2007.08.01	12:03:08 AM	40.79	30.09	turkey	kocaeli	bayraktar_izmit	west	0.1	5.2	4.0	3.8	4.0	NaN	0.0	0.0
2	1.980000e+13	1978.05.07	12:41:37 AM	38.58	27.61	turkey	manisa	hamzabeyli	south_west	0.1	0.0	3.7	0.0	0.0	NaN	0.0	3.7
3	2.000000e+13	1997.03.22	12:31:45 AM	39.47	36.44	turkey	sivas	kahvepinar_sarkisla	south_west	0.1	10.0	3.5	3.5	0.0	NaN	0.0	0.0
4	2.000000e+13	2000.04.02	12:57:38 AM	40.80	30.24	turkey	sakarya	meseli_serdivan	south_west	0.1	7.0	4.3	4.3	0.0	NaN	0.0	0.0
5	2.010000e+13	2005.01.21	12:04:03 AM	37.11	27.75	turkey	mugla	demirciler_milas	south_west	0.1	32.8	3.5	3.5	0.0	NaN	0.0	0.0
6	2.010000e+13	2012.06.24	12:07:22 AM	38.75	43.61	turkey	van	ilikaynak	south_west	0.1	9.4	4.5	0.0	4.5	NaN	0.0	0.0
7	1.990000e+13	1987.12.31	12:49:54 AM	39.43	27.98	turkey	balikesir	dikkonak_bigadic	south_east	0.1	26.0	3.8	3.8	0.0	NaN	0.0	0.0
8	2.000000e+13	2000.02.07	12:11:45 AM	40.05	34.07	turkey	kirikkale	kocabas_delice	south_east	0.1	1.0	3.8	3.8	0.0	NaN	0.0	0.0
9	2.010000e+13	2011.10.28	12:47:56 AM	38.76	43.54	turkey	van	degirmenozu	south_east	0.1	3.1	4.3	0.0	4.2	NaN	0.0	4.3

In [27]:
```
data.columns
```

Out[27]:
```
Index(['id', 'date', 'time', 'lat', 'long', 'country', 'city', 'area',
       'direction', 'dist', 'depth', 'xm', 'md', 'richter', 'mw', 'ms', 'mb'],
      dtype='object')
```

Image 4 : ASSIGNMENT 1

CREATING A NEW COLUMN FOR YEARS

In [29]:
```
def yeardate(x):
    return x[0:4]
data["yeardate"] = data.date.apply(yeardate)
#We have to change the object to int.
data['yeardate'] = data.yeardate.astype(int)
print(data.yeardate.dtypes)
data.head(3)
```

```
int32
```

Out[29]:

	id	date	time	lat	long	country	city	area	direction	dist	depth	xm	md	richter	mw	ms	mb	yeardate
0	2.000000e+13	2003.05.20	12:17:44 AM	39.04	40.38	turkey	bingol	baliklicay	west	0.1	10.0	4.1	4.1	0.0	NaN	0.0	0.0	2003
1	2.010000e+13	2007.08.01	12:03:08 AM	40.79	30.09	turkey	kocaeli	bayraktar_izmit	west	0.1	5.2	4.0	3.8	4.0	NaN	0.0	0.0	2007
2	1.980000e+13	1978.05.07	12:41:37 AM	38.58	27.61	turkey	manisa	hamzabeyli	south_west	0.1	0.0	3.7	0.0	0.0	NaN	0.0	3.7	1978

Correlation

In [30]:
```
plt.figure(figsize=(15,15))
sns.heatmap(data.corr(), annot = True, fmt= ".1f", linewidths = .3)
plt.show()
```

Image 5 : ASSIGNMENT 2

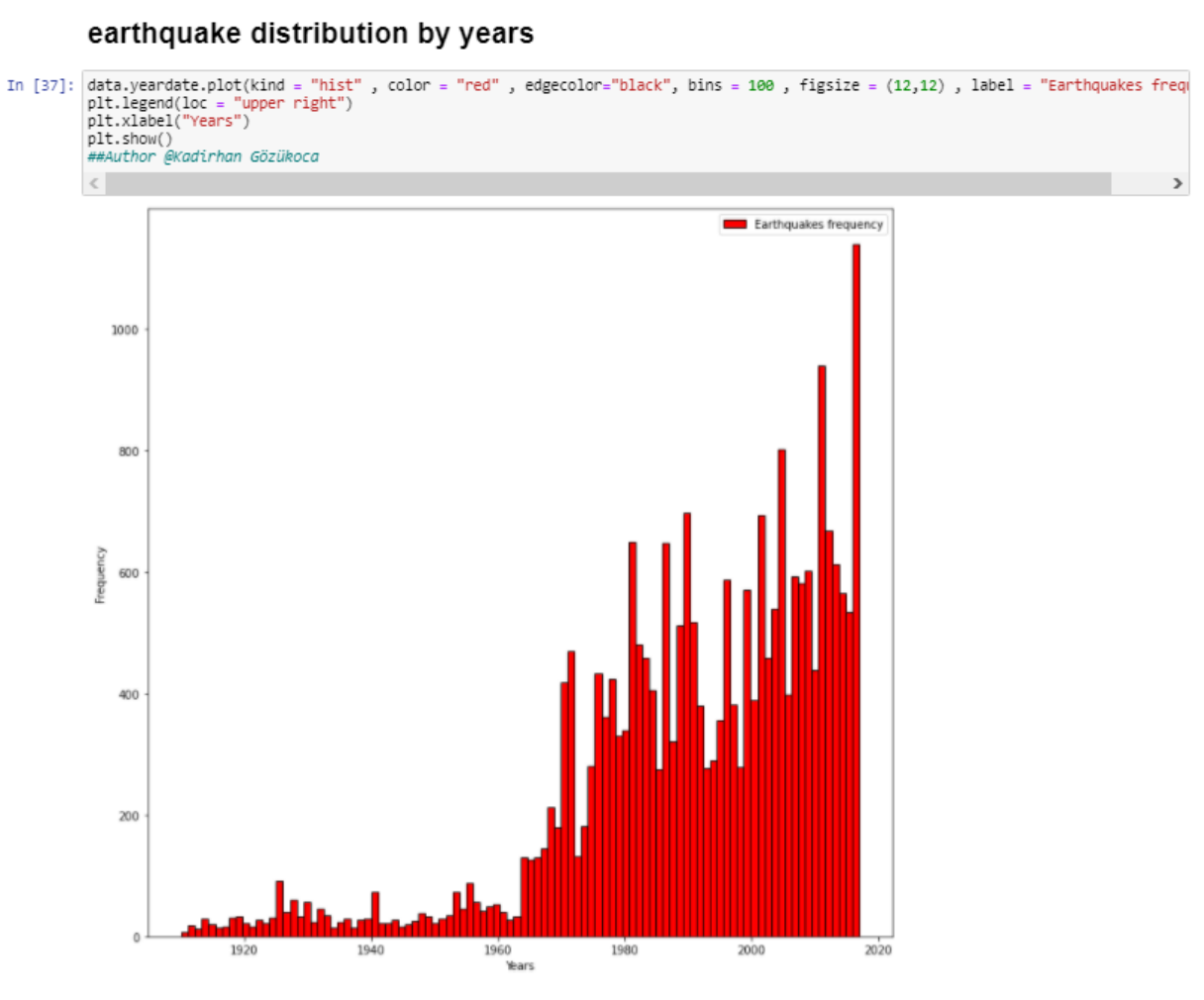

Image 6 : ASSIGNMENT 3

We used four libraries, pandas, NumPy, matplotlib and seaborn.We imported the libriaries. We created the data reader. We created an account on www.kaggle.com and We choosed a dataset about "Earthquakes between 1910-2017 in Turkey". Before the starting, We research some tools and packages which We are planning to use and We imported the libriaries (pandas, Matplotlib and NumPy) and packages which We will use. Also We started the data analysis. Here is our contents:

-Correlation between features.

-Which year had the most earthquakes in turkey?

-Where was the most earthquake?

-How long did the earthquake last?

-Where and when did the most severe earthquake occur?

Columns:

lat: latitude of earthquake

dist: distance of direction in km

richter: intensity of earthquake

md: depending on time magnitude

mw: moment magnitude

ms: surface-wave magnitude

mb: body-wave magnitude

xm: biggest value in specified magnitude values

We created the columns for the filtering years. It is important to understanding better the statistics. Also, We have to change the object to int because years are integer as you know.

We did our first visualization with correlation graph. Correlation is basically used to show the relationship between two variables. The simplest way to learn is to calculate the covariance value, which shows the change of these two variables relative to each other. As you can see we have possitive correlation **xm** between as **mw**, **ms** and **mb.**

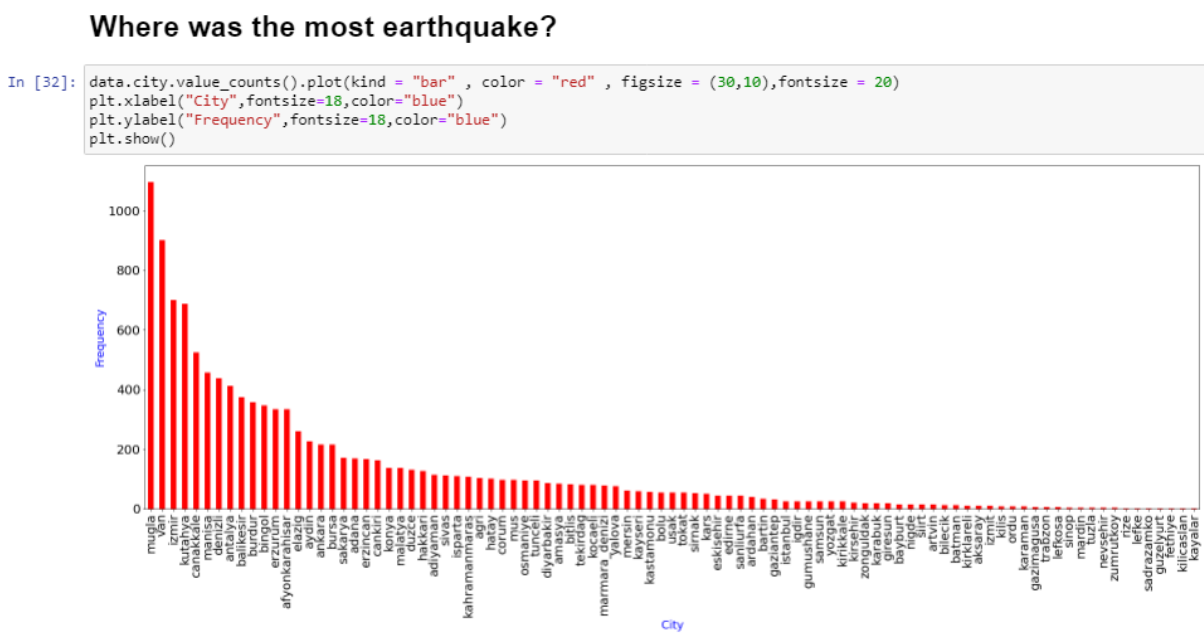

Image 11 : ASSIGNMENT 4

Countries close to Turkey

In [33]:
```
data.country.value_counts().plot(kind = "bar" , color = "red" , figsize = (30,10),fontsize = 20)
plt.xlabel("Country",fontsize=18,color="blue")
plt.ylabel("Frequency",fontsize=18,color="blue")
plt.show()
```

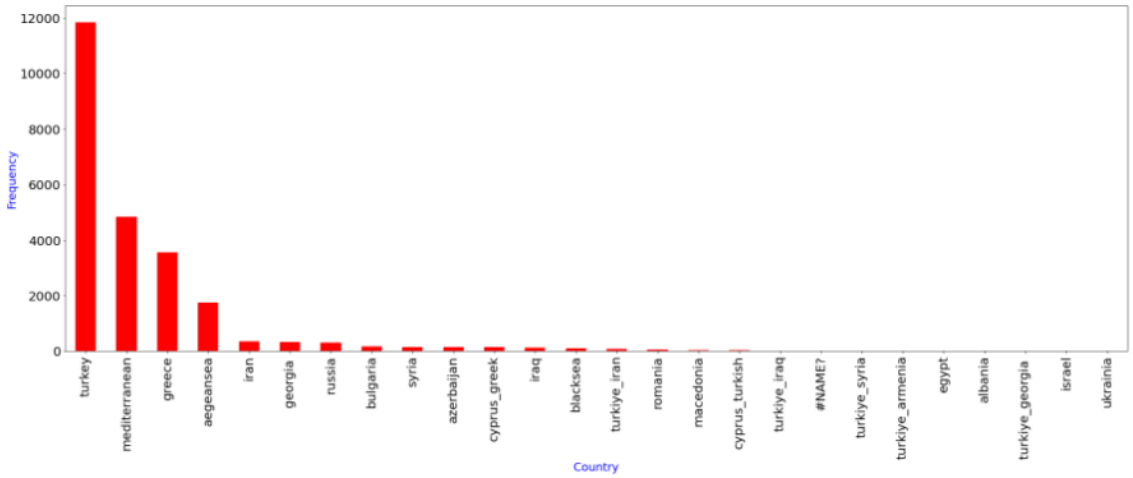

Image 12 : ASSIGNMENT 5

How long did the earthquake last?

In [34]:
```
data.long.max()
filtre = data.long == 48.0
data[filtre]
```

Out[34]:

	id	date	time	lat	long	country	city	area	direction	dist	depth	xm	md	richter	mw	ms	mb	yeardate
10064	1.910000e+13	1910.12.04	12:02:00 AM	39.30	48.0	azerbaijan	NaN	NaN	NaN	NaN	37.0	5.5	5.3	5.3	5.5	5.4	5.3	1910
10068	1.910000e+13	1911.06.23	12:30:02 AM	40.00	48.0	azerbaijan	NaN	NaN	NaN	NaN	18.0	5.3	5.0	4.9	5.3	5.0	5.0	1911
10115	1.920000e+13	1915.10.06	12:59:03 AM	41.00	48.0	russia	NaN	NaN	NaN	NaN	15.0	4.8	4.7	4.6	4.8	4.6	4.7	1915
10481	1.930000e+13	1932.03.15	12:18:06 AM	34.00	48.0	iran	NaN	NaN	NaN	NaN	35.0	5.6	5.4	5.3	5.6	5.5	5.4	1932
10751	1.950000e+13	1948.08.30	12:42:01 AM	41.90	48.0	russia	NaN	NaN	NaN	NaN	31.0	5.5	5.3	5.3	5.5	5.4	5.3	1948
10872	1.950000e+13	1953.12.30	12:09:05 AM	34.00	48.0	iran	NaN	NaN	NaN	NaN	5.0	5.1	5.1	0.0	NaN	0.0	0.0	1953
11332	1.970000e+13	1965.05.15	12:43:01 AM	39.90	48.0	azerbaijan	NaN	NaN	NaN	NaN	10.0	4.4	4.2	4.2	4.4	4.0	4.3	1965
12540	1.980000e+13	1976.04.14	12:25:04 AM	40.10	48.0	azerbaijan	NaN	NaN	NaN	NaN	33.0	4.3	0.0	0.0	NaN	0.0	4.3	1976
12695	1.980000e+13	1977.01.18	12:48:54 AM	33.11	48.0	iran	NaN	NaN	NaN	NaN	49.0	5.2	0.0	0.0	NaN	5.2	5.2	1977

Image 13 : ASSIGNMENT 6

The biggest earthquake in Turkey

In [35]:
```
data.xm.max()
filtering = data.country == "turkey"
filtering2 = data.xm == 7.9
data[filtering & filtering2]
```

Out[35]:

	id	date	time	lat	long	country	city	area	direction	dist	depth	xm	md	richter	mw	ms	mb	yeardate
6717	1.940000e+13	1939.12.26	12:57:21 AM	39.8	39.51	turkey	erzincan	kurutilek	north_east	3.0	20.0	7.9	7.2	7.2	7.7	7.9	7.1	1939

Image 14 : ASSIGNMENT 7

Magnitude level

In [36]:
```
threshold = sum(data.xm) / len(data.xm)
data["magnitude-level"] = ["hight" if i > threshold else "low" for i in data.xm]
data.loc[:10,["magnitude-level","xm","city"]]
```

Out[36]:

	magnitude-level	xm	city
0	hight	4.1	bingol
1	low	4.0	kocaeli
2	low	3.7	manisa
3	low	3.5	sivas
4	hight	4.3	sakarya
5	low	3.5	mugla
6	hight	4.5	van
7	low	3.8	balikesir
8	low	3.8	kirikkale
9	hight	4.3	van
10	low	3.5	kahramanmaras

Image 15 : ASSIGNMENT 8

5.2. USE CASE DIAGRAM

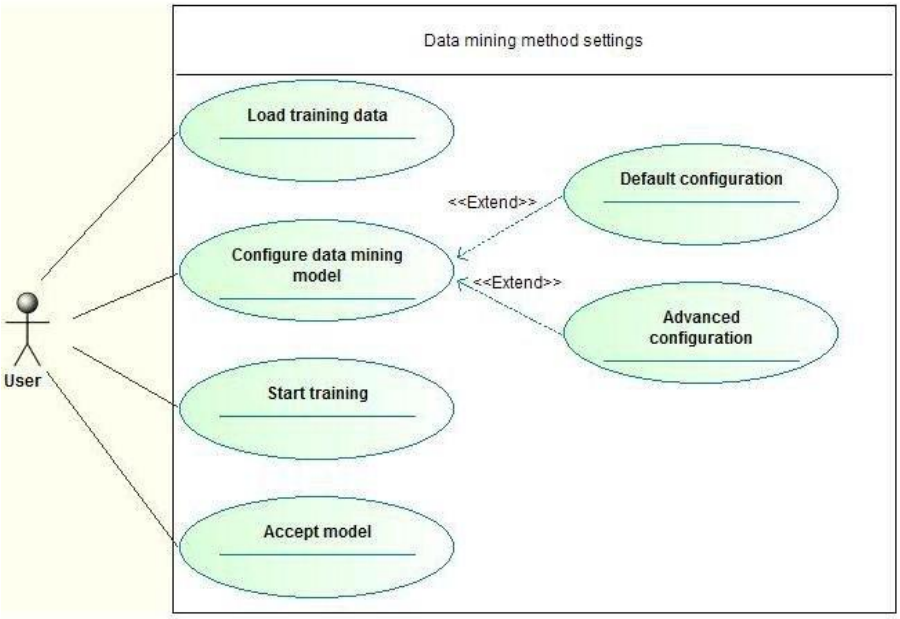

Image 16 : Use Case Diagram for Data Mining

5.3. FLOWCHART OF THE TEST BUSINESS SCENARIO

We don't have a test business scenario as we cannot do our user interface.

5.4. USER'S GUIDE

We don't have a user guide as we cannot do our user interface.

6. CONCLUSION

6.1. THE PROJECT CAN BE IMPROVED BY ADDING WHAT?

An interface is missing in our project. A nice user interface can be created in this project. For example: The user should be able to upload the datasets and excel files, and select the graph and then the user can see the results.

6.2. ADVANTAGES AND DISADVANTAGES OF PROJECT

Advantages:

- It helps companies gather reliable information
- Data mining uses both new and legacy systems
- It enables complex datas to be reduced to simple.
- It helps data scientists easily analyze enormous of data quickly.
- It helps detect businesses make informed decisions.

- It helps data scientists quickly initiate automated predictions of behaviors and trends and discover.

Disadvantages:

- Data mining requires large databases, making the process hard to manage.
- Many data analytics tools are complex and challenging to use. Data Scientists need the right training to use the tools effectively.
- Companies can potentially sell the customer data they have gleaned to other businesses and organizations.

6.3. SIMILAR PROJECTS

Data mining can be written not only in python, but also in other languages, as well as R, SQL, S.A.S.and many other languages.

To give examples of projects similar to mines:

1.Data Mining Techniques on Earthquake Data - (https://www.researchgate.net/publication/297056372_Data_Mining_Techniques_on_Earthquake_Data_Recent_Data_Mining_Approaches)

2.Earthquake Prediction Using Data Mining - (https://www.ijsr.net/archive/v6i11/ART20177914.pdf)

6.4. WHAT HAVE I GAİNED FROM THIS PROJECT?

We learned how important data analysis is in many situations, in business life.

We learned what should be considered while preparing a word document and how a more effective and professional report can be prepared.

7. REFERENCES

Anaconda Individual Download. (2020). (Anaconda) https://www.anaconda.com/products/individual adresinden alındı

International Journal Science and Research (IJSR) ISSN(ONLINE): 2319-7064

Great Learning (2020, July 23). *Youtube*. (Great Learning Channel) https://www.youtube.com/watch?v=4rymD1Hpnho

TheEngineeringWorld (2020, Mart 24). *Youtube*. (Cleaning Data) https://www.youtube.com/watch?v=xcKXmXilaSw

Kaggle.com: https://www.kaggle.com/caganseval/earthquake*Online Output*. (2017).

Wikipedia. (2021, December 17). Wikipedia.org: https://en.wikipedia.org/wiki/List_of_earthquakes_in_Turkey

8. IMAGES

Image 1 : Flowchart for Data Mining 3

Image 3 : Kaggle Link 4

Image 4 : Anaconda Download 5

Image 5 : Jupyter Notebook Launch 6

Image 4 : Dataset Page 6

Image 6 : New Page in Jupyter Notebook 7

Image 7 : New Page 7

Image 8 : Assignment1 8

Image 9 : Assignment2 9

Image 10 : Assignment3 10

Image 11 : Assignment4 11

Image 12 : Assignment5 12

Image 13 : Assignment6 12

Image 14 : Assignment7 13

Image 15 : Assignment8 13

Image 16 : Use Case Diagram for Data Mining 14