THE BILLIONAIRES 2018

NILESH PATEL

NYC DATA SCIENCE ACADEMY: COHORT #15

- OUTLINES:

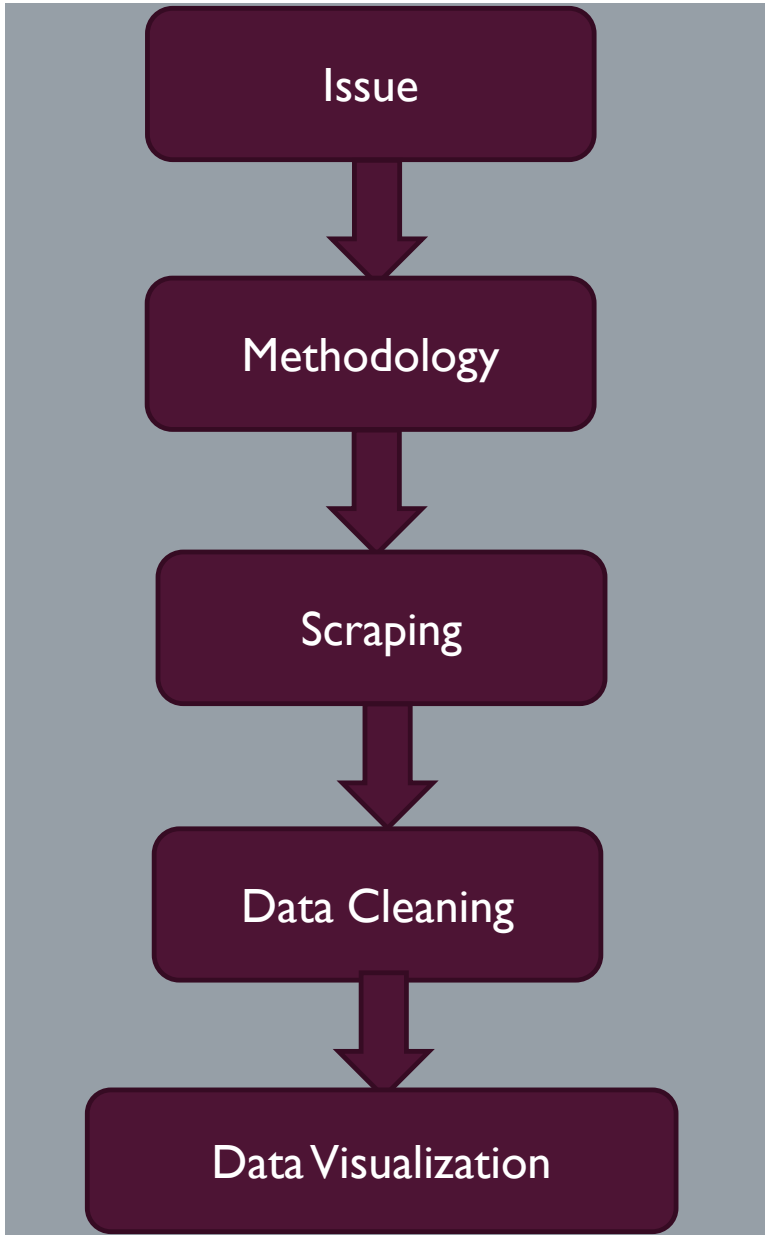
  INTRODUCTION

  SCRAPING & FILTERING

  DATA CLEANING

  DATA VISUALIZATION

  CONCLUSION



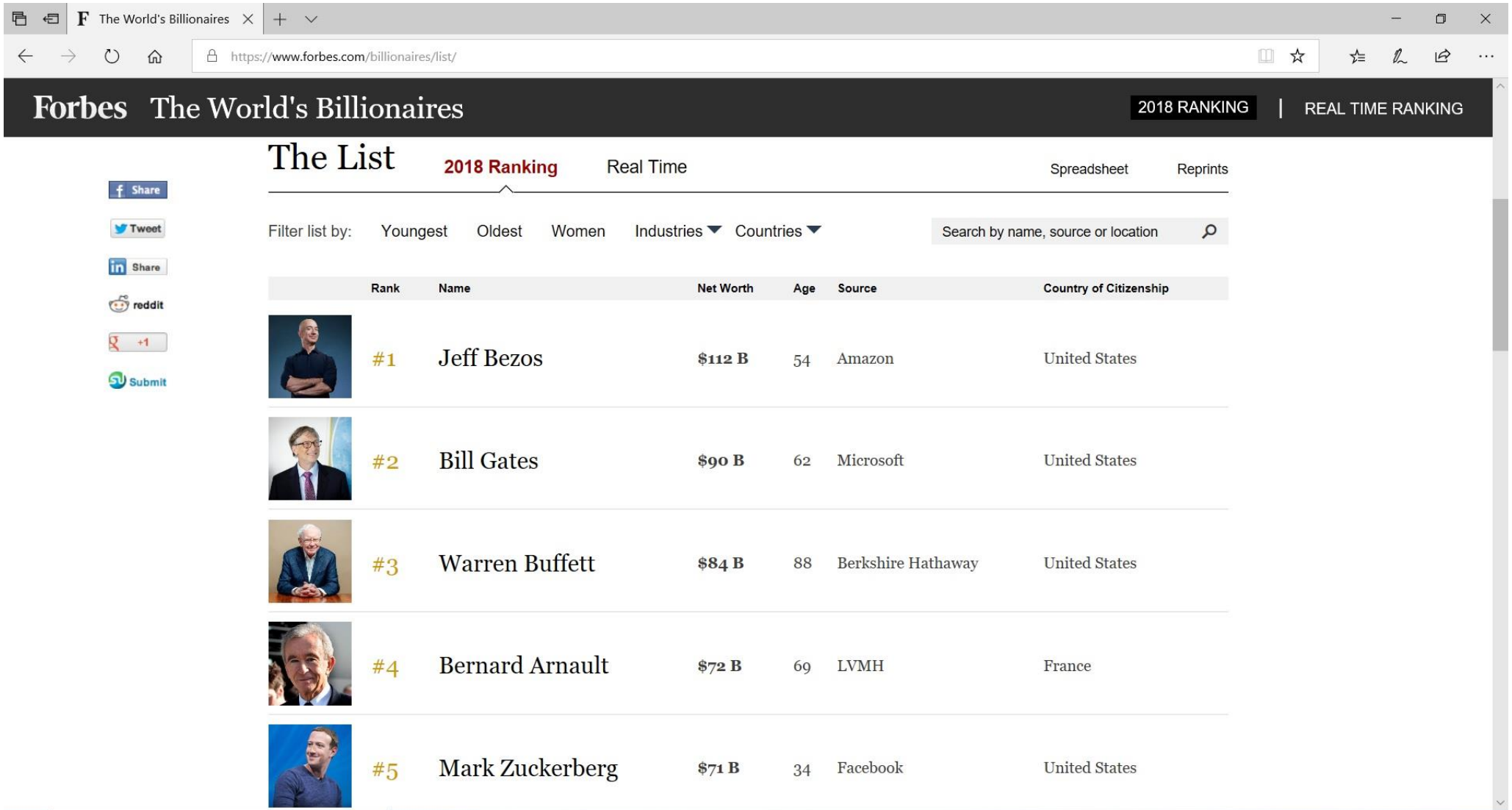
INTRODUCTION

- In order to perform data analysis on the list of all billionaires in the world, I need to get the data information first. By using web scraping tool Selenium, I start to scrape the Forbes.com. The whole process includes:
- go to the Forbes World's Billionaires main page;
- find the URL for page.
- scrape details for each person (Rank, Name, Net Worth 2018, Age, Current Net Worth, Source of Income, Country of Citizenship and Gender.)
- As for me, the challenge is mainly about finding the right X path for each content I want to scrape. Sometimes it is because there is no unique X path when I inspect the page.
- For Gender and Current Net Worth columns I scrape them separately, because all there information was on different pages.

ABOUT FORBES

- Forbes Media is a global media, branding and technology company, with a focus on news and information about business, investing, technology, entrepreneurship, leadership and affluent lifestyles.
- Forbes.com is part of Forbes Digital, a division of Forbes Media LLC.
- 100+ years in business.
- 38 Millions Social Followers
- CEO : Mike Federle

Forbes


The World's Billionaires
https://www.forbes.com/billionaires/list/
Forbes The World's Billionaires
2018 RANKING | REAL TIME RANKING
The List
2018 Ranking
Real Time
Spreadsheet
Reprints
Share
Tweet
Share
reddit
+1
Submit
Filter list by:
Youngest
Oldest
Women
Industries
Countries
Search by name, source or location
Rank
Name
Net Worth
Age
Source
Country of Citizenship
#1
Jeff Bezos
$112 B
54
Amazon
United States
#2
Bill Gates
$90 B
62
Microsoft
United States
#3
Warren Buffett
$84 B
88
Berkshire Hathaway
United States
#4
Bernard Arnault
$72 B
69
LVMH
France
#5
Mark Zuckerberg
$71 B
34
Facebook
United States

SCRAPING

C:\Users\patel\OneDrive\Desktop\forbes static\forbes.py • - Sublime Text (UNREGISTERED)

File Edit Selection Find View Goto Tools Project Preferences Help

forbes.py

```python
# Import webdriver and other package
from selenium import webdriver
from selenium.webdriver.support.ui import WebDriverWait
import csv
import re

#creating webdriver
driver = webdriver.Chrome(r'C:\Users\patel\OneDrive\Desktop\chromedriver.exe')
driver.get("https://www.forbes.com/billionaires/list/#version:static")

#open csv file in write mode
csv_file = open('reviews.csv', 'w', encoding='utf-8')
writer = csv.writer(csv_file)

#finding Rank, Name and Net worth by xpath
Rank = list(map(lambda x: x.text, driver.find_elements_by_xpath('//tbody/*[@class="data"]/td[@class = "rank"]')))
Name =list(map(lambda x: x.text, driver.find_elements_by_xpath('//tbody/*[@class="data"]/td[@class = "name"]')))
Net_Worth = list(map(lambda x: x.text, driver.find_elements_by_xpath('//tbody/*[@class="data"]/td[@class = "networth"]')))
wait_button = WebDriverWait(driver, 10)#Wait

#finding Age, Source and Country of citizenship
Age = list(map(lambda x: x.text, driver.find_elements_by_xpath('//tbody/*[@class="data"]//td[5]')))
Source = list(map(lambda x: x.text, driver.find_elements_by_xpath('//tbody/*[@class="data"]//td[6]')))
Country = list(map(lambda x: x.text, driver.find_elements_by_xpath('//tbody/*[@class="data"]//td[7]')))
wait_button = WebDriverWait(driver, 10)#wait

#Using regular expression to clean data
Rank = [(re.sub('[\n,. - #]+',"", i)) for i in Rank]
Name = [(re.sub('[\n,.-]+',"", i)) for i in Name]
Net_Worth = [(re.sub('[\n, - # B $]+',"", i)) for i in Net_Worth]
Age = [(re.sub('[\n,.-]', '', i)) for i in Age]
Source = [(re.sub('[\n,.-]', '', i)) for i in Source]
Country = [(re.sub('[\n,.-]', '', i)) for i in Country]

#using zip function to arrange data
data_dict = zip(Rank, Name, Net_Worth, Age, Source, Country)
```

SELENIUM SCRIPT

jupyter forbes_data Last Checkpoint: 8 hours ago (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

World's Top Billionaires 2018

2,212 billionaires, collectively worth $9.1 trillion.

capitalism's global conquest continues as entrepreneurs around the globe mint fortunes in everything from cryptocurrencies to telecom to bridal dresses. Forbes has pinned down a record 2,212 billionaires from 72 countries and territories including the first ever from Hungary and Zimbabwe. This elite group is worth $9.1 trillion, up 18% since last year. Their average net worth is a record $4.1 billion. Americans lead the way with a record 587 billionaires, followed by mainland China with 373. Centi-billionaire Jeff Bezos secures the list's top spot for the first time, becoming the only person to appear in the Forbes ranks with a 12-figure fortune. Bezos's fortune leapt more than $39 billion, the list's biggest one-year gain ever.

In [172]:
```
1 forbes_data.head(10)
```

Out[172]:

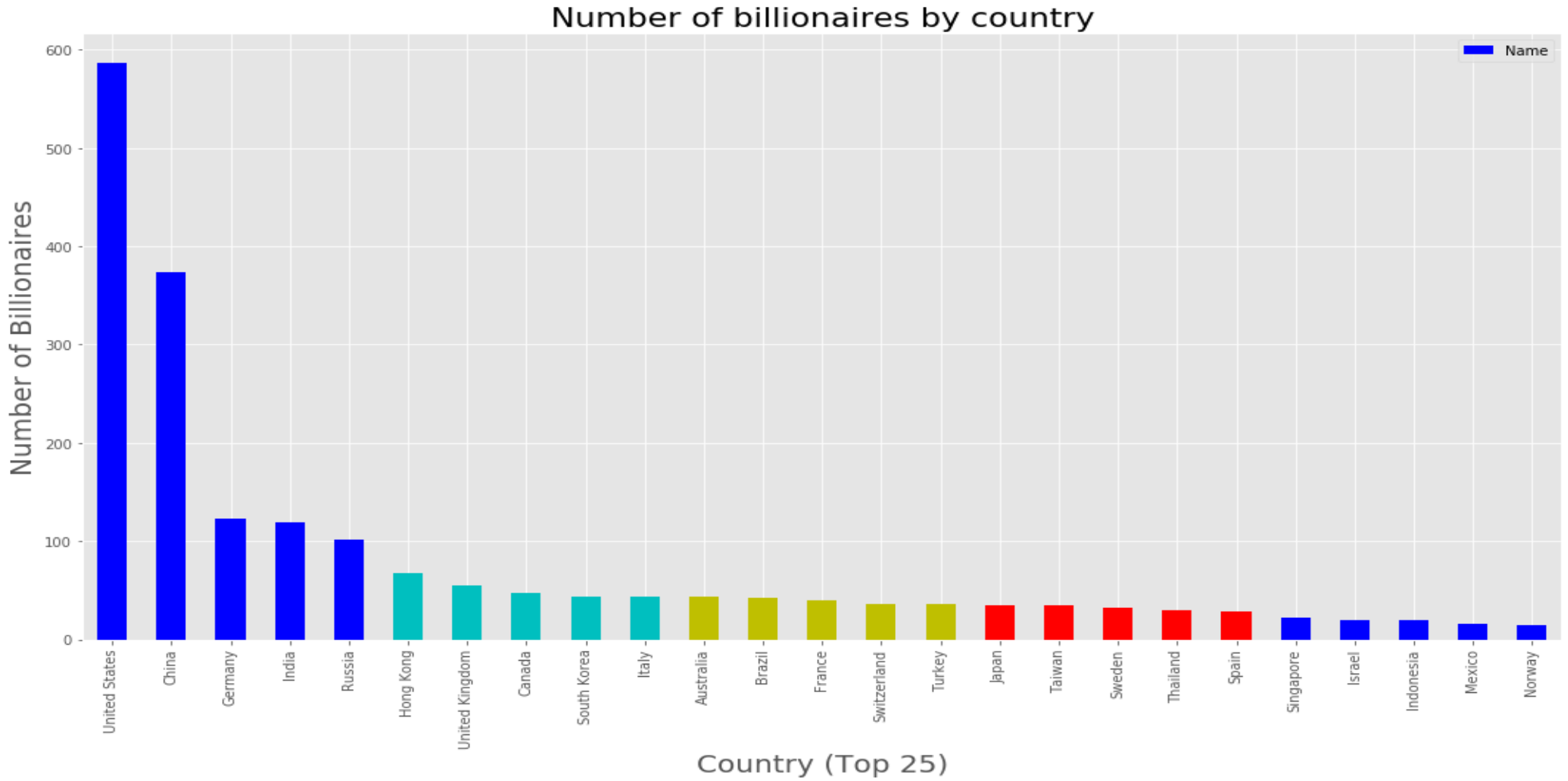
	Rank	Name	Net Worth_x	Age	Source	Country	Net Worth_y	Change	Gender
0	1	Jeff Bezos	112.0	54	Amazon	United States	147.3	35.3	M
1	2	Bill Gates	90.0	62	Microsoft	United States	95.9	5.9	M
2	3	Warren Buffett	84.0	88	Berkshire Hathaway	United States	85.1	1.1	M
3	4	Bernard Arnault	72.0	69	LVMH	France	71.2	-0.8	M
4	5	Mark Zuckerberg	71.0	34	Facebook	United States	57.6	-13.4	M
5	6	Amancio Ortega	70.0	82	Zara	Spain	62.5	-7.5	M
6	7	Carlos Slim Helu	67.1	78	telecom	Mexico	60.8	-6.3	M
7	8	Charles Koch	60.0	82	Koch Industries	United States	51.6	-8.4	M
8	8	David Koch	60.0	78	Koch Industries	United States	51.6	-8.4	M
9	10	Larry Ellison	58.5	74	software	United States	57.9	-0.6	M

PYTHON SCRIPT

DATA CLEANING

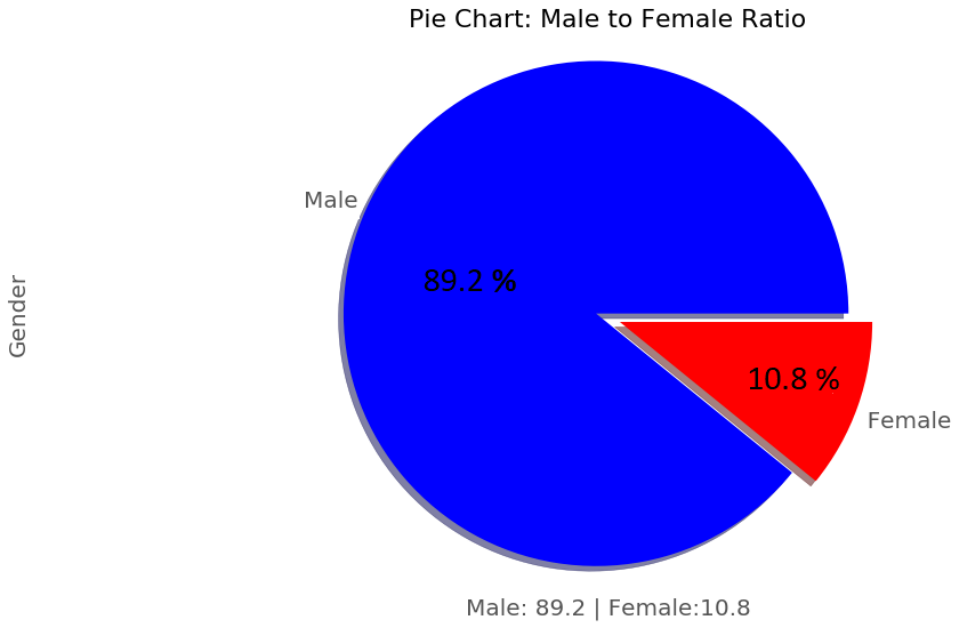
Tasks	Actions
Apply Pandas functions, Methods(read_csv, merge, groupby, fillna, dropna, isnull() etc.)	To read csv files, join, group by values in columns, checking null values, fill NA values, drop NA values from data frame.
NUMPY package	Mathematical operations (i.e sum, log, etc.)
Matplotlib library	Plotting Graph(bar, histogram, scatter, box etc.)
Regular Expressions	Clean data

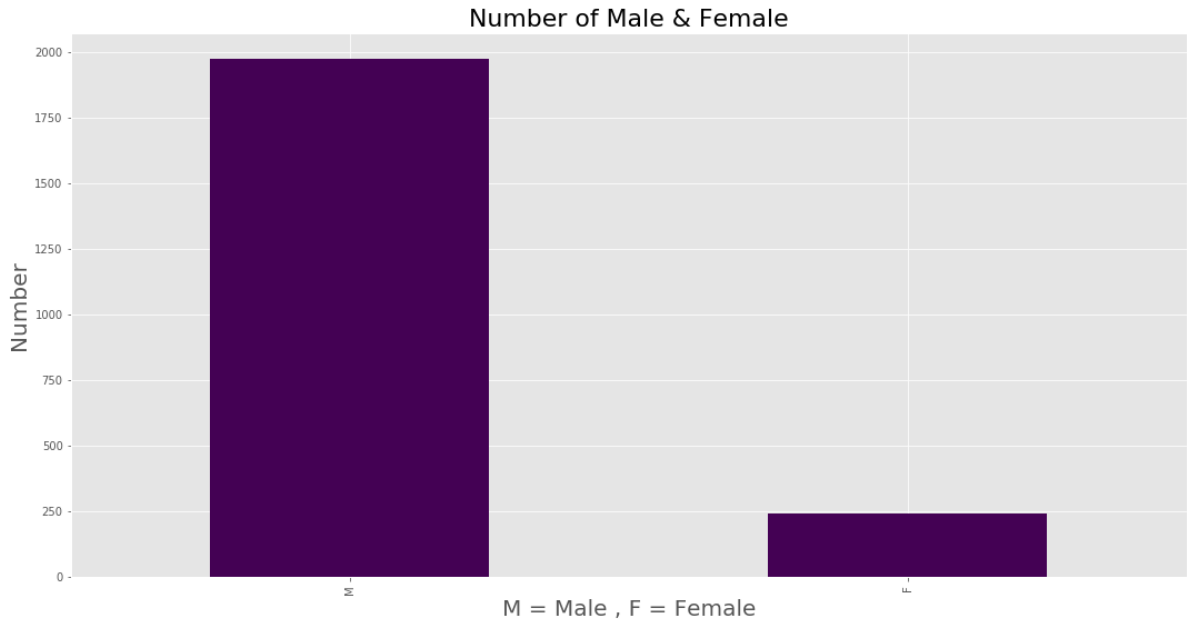
GRAPH 1: BAR PLOT



- In the list of 2212 Billionaires of the world, compare to all other 71 countries United States has maximum number or people in the list.

GRAPH II:
PIE CHART & BAR PLOT





Pie Chart:
Gender:
Male: 89.2 %
Female: 10.8 %

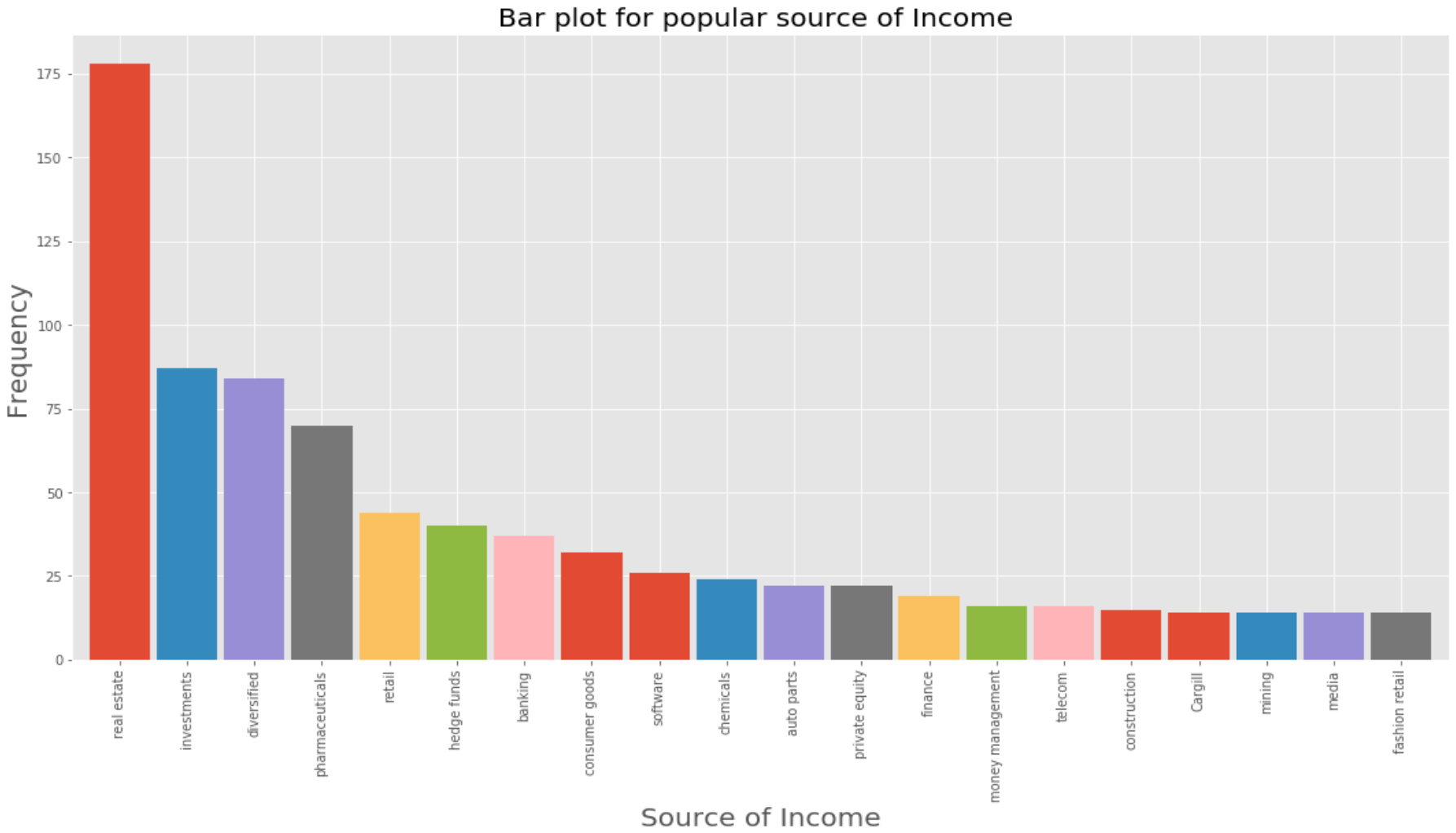
X = M: Male
F: Female

Y =Number of Male/Female

RESULT: In the list of 2212 billionaires from 72 different countries.

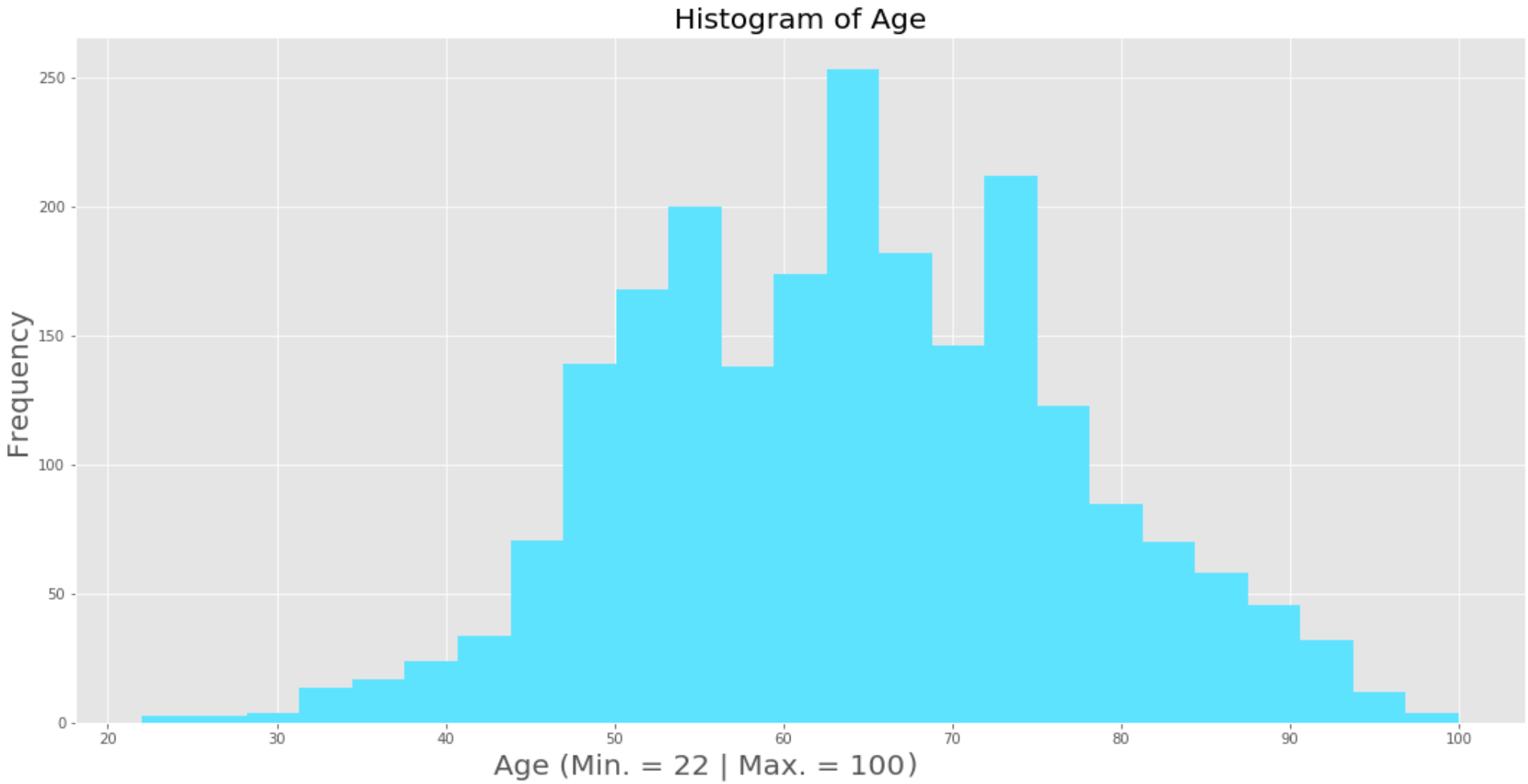
Male = 1972 Female 240

GRAPH III: BAR PLOT



- Graph shows similarity in their source of income.
- Real estate is most popular among all of them.
- Investments is on number two.
- Other then that Pharmaceutical, retail, hedge funds, banking, consumer goods, software are also popular.

GRAPH IV: HISTOGRAM



X = Age

Y = Frequency

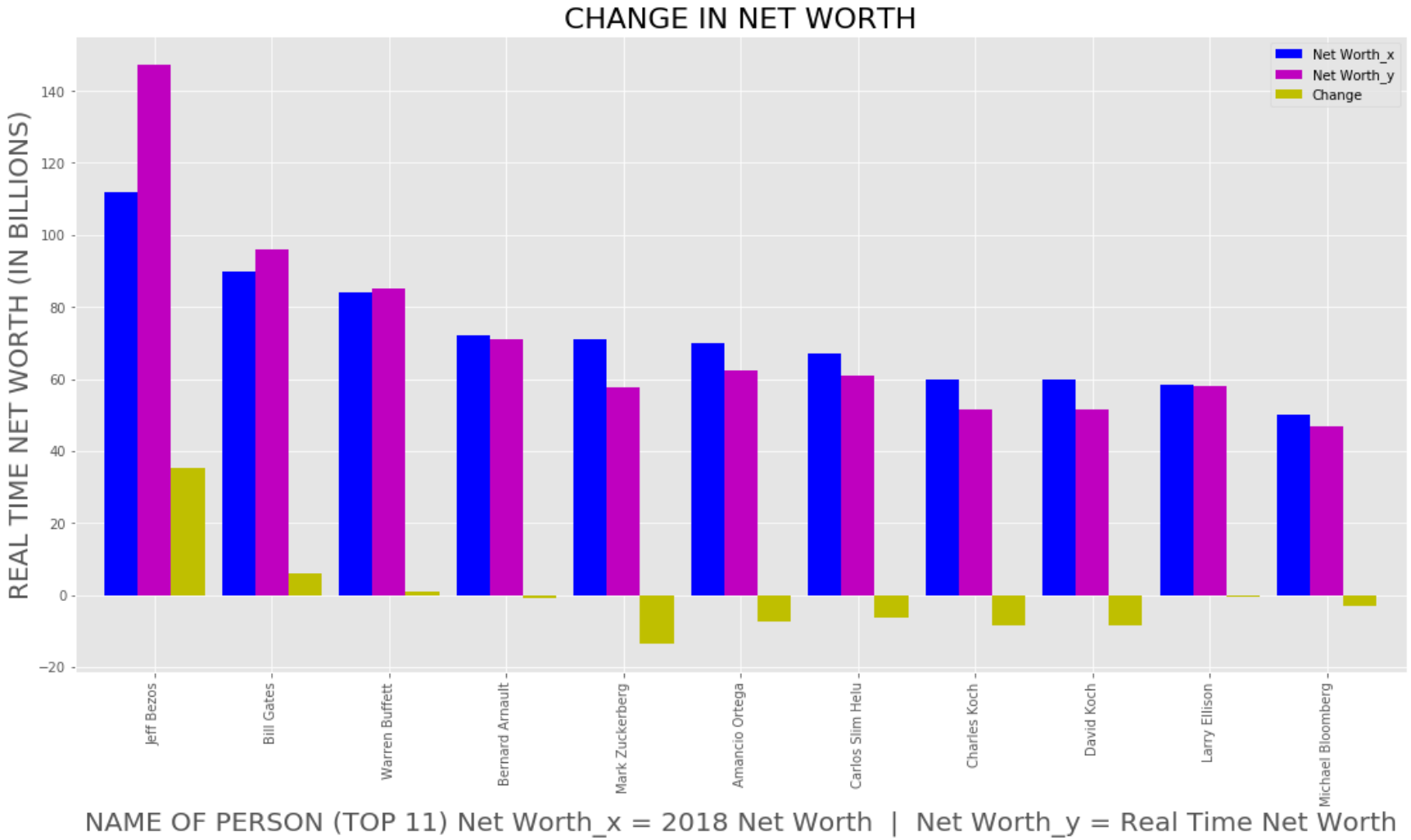
Min. Age = 22 years
Country = 'Norway

Max. Age = 100 years
Country
'Singapore'

Mean.Age = 62.828103

Median.Age = 64

GRAPH V:

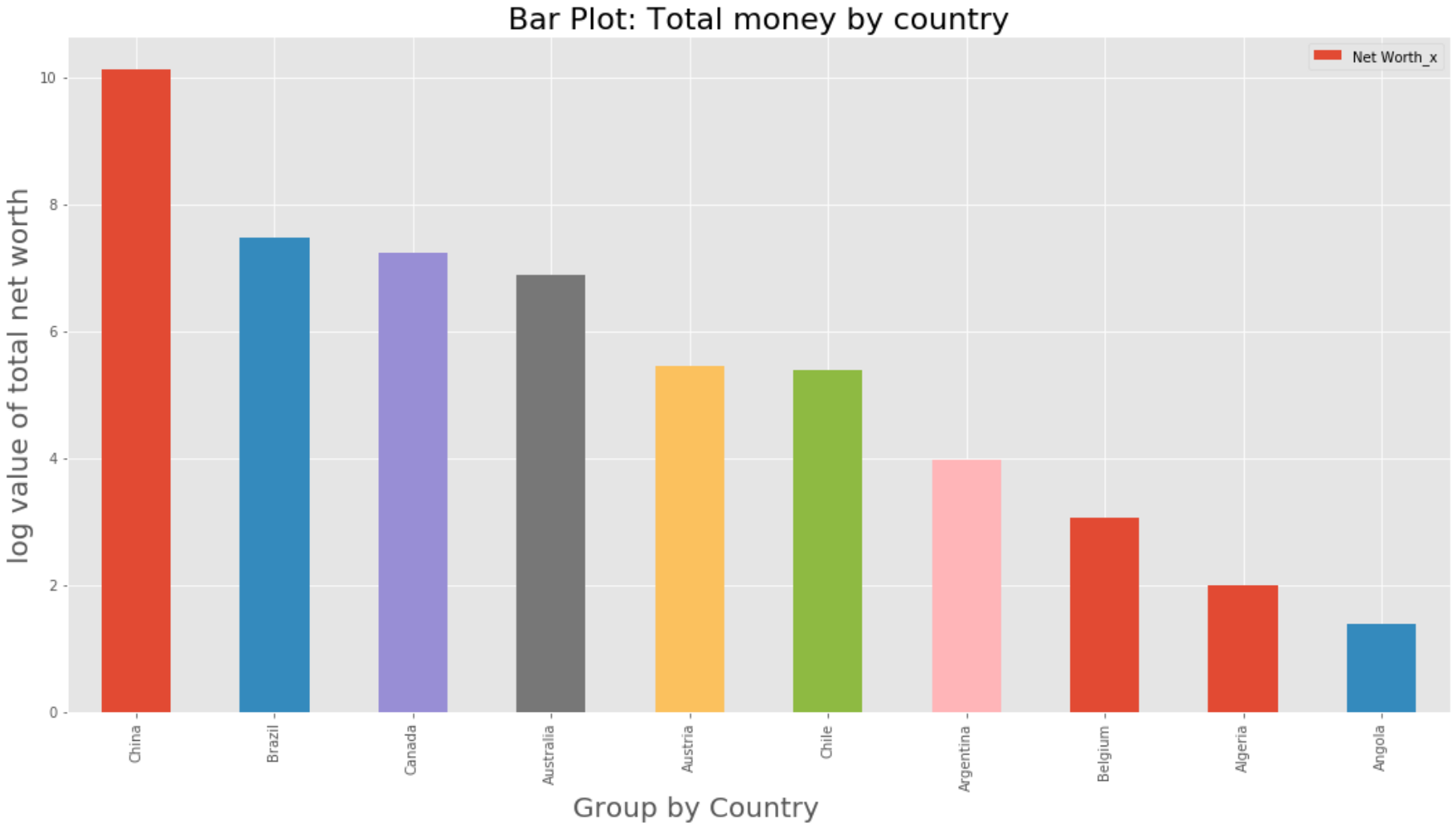


Net Worth_x:
Net worth according to 2018.

Net Worth_y:
Real time Net worth.

Change:
Difference in Net worth and real time Net worth.
(positive or Negative)

GRAPH V1: BAR PLOT



Country	Net Worth_x
United States	3104.8
China	1120.4
Germany	578.7
India	440.1
Russia	409.3
Hong Kong	334.7
France	320.4
United Kingdom	207.4
Brazil	176.4
Italy	170.4

X = Top 10 Countries

Y =Log values of total net worth group by country(In Billions)

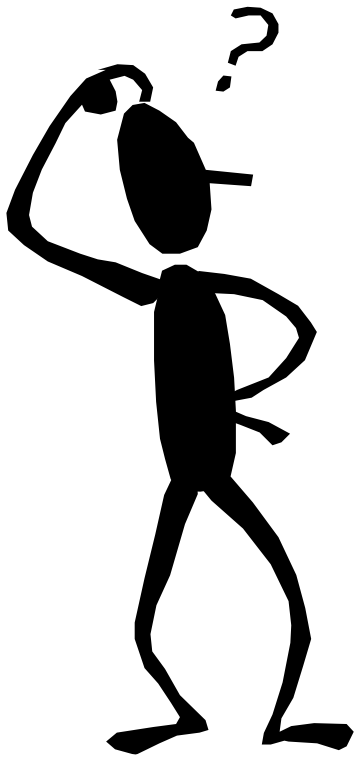
GRAPH VII: PLOT

	Name	Net Worth_y	Country
0	Jeff Bezos	147.3	United States
1	Bernard Arnault	71.2	France
2	Amancio Ortega	62.5	Spain
3	Carlos Slim Helu	60.8	Mexico
4	Mukesh Ambani	43.1	India
5	Jack Ma	34.6	China
6	Li Kashing	31.2	Hong Kong
7	Beate Heister & Karl Albrecht Jr	25.7	Germany
8	Jorge Paulo Lemann	23.0	Brazil
9	David Thomson	27.8	Canada
10	Dietrich Mateschitz	22.6	Austria

- Graph showing one person from each country having maximum amount of money,

Conclusion:

- These 2,212 billionaires, collectively worth $9.1 trillion.
- United States having $3.1 trillion contribution in that, is around 33% of the total value of billionaires around the world.\
- Main Source of income is Real estate.
- Men to women ratio in this Billionaires list is approximately 9:1.
- Youngest age is 22 years and oldest is 100 years. Mean of age is 62 and median is around 64 years.



Any Questions !!

THANK YOU