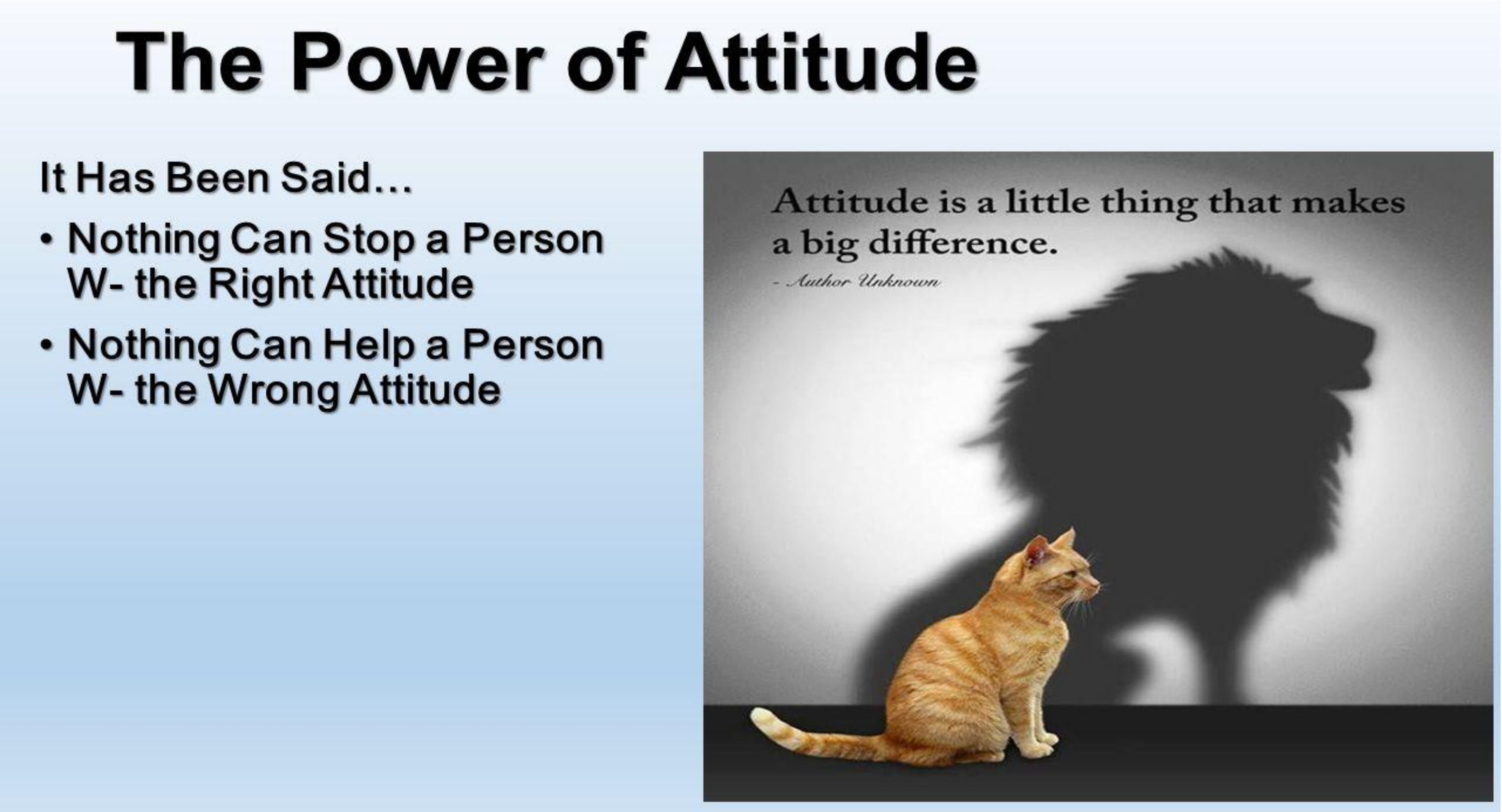
بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

In the name of Allah the most Beneficial ever merciful



The Power of Attitude

It Has Been Said…

- Nothing Can Stop a Person W- the Right Attitude
- Nothing Can Help a Person W- the Wrong Attitude

Attitude is a little thing that makes a big difference.

- *Author Unknown*

Artificial Intelligence (AI) in Software Engineering

Regression

Department of Computer Science , Univeristy of Karachi (DCS-UBIT)
4th May 2021

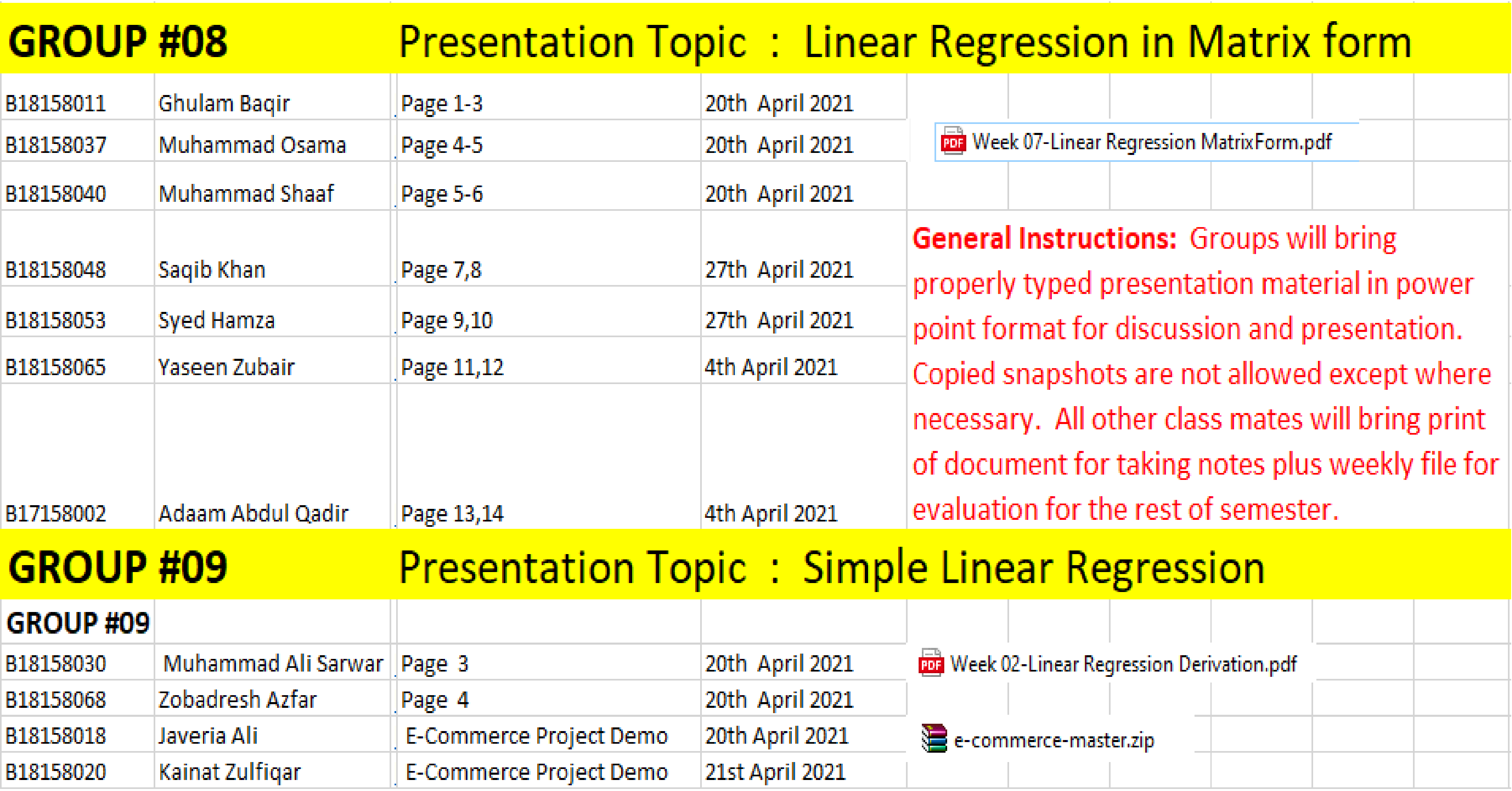
Agenda

1- Mid-Term Lab Help and Support

2- Presentations Group 9 – session 2

3- Presentation Group 8 - session 2

Groups



GROUP #08 Presentation Topic : Linear Regression in Matrix form

B18158011	Ghulam Baqir	Page 1-3	20th April 2021	
B18158037	Muhammad Osama	Page 4-5	20th April 2021	Week 07-Linear Regression MatrixForm.pdf
B18158040	Muhammad Shaaf	Page 5-6	20th April 2021	
B18158048	Saqib Khan	Page 7,8	27th April 2021	
B18158053	Syed Hamza	Page 9,10	27th April 2021	
B18158065	Yaseen Zubair	Page 11,12	4th April 2021	
B17158002	Adaam Abdul Qadir	Page 13,14	4th April 2021	

General Instructions: Groups will bring properly typed presentation material in power point format for discussion and presentation. Copied snapshots are not allowed except where necessary. All other class mates will bring print of document for taking notes plus weekly file for evaluation for the rest of semester.

GROUP #09 Presentation Topic : Simple Linear Regression

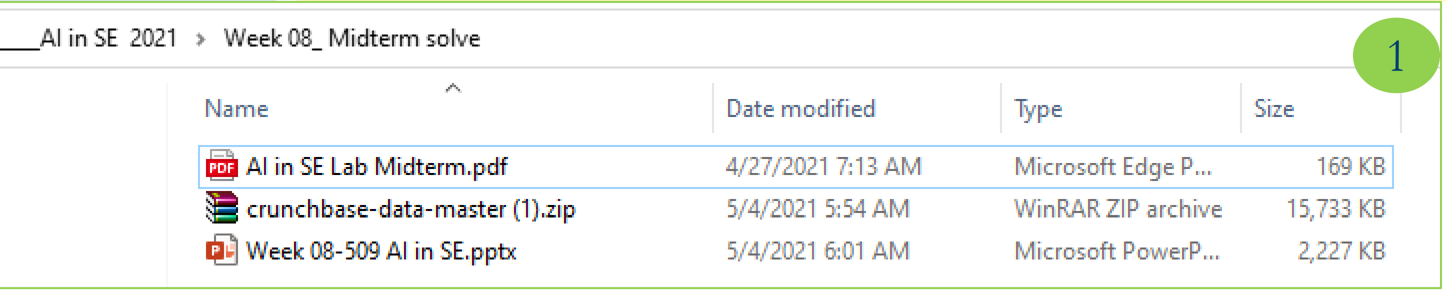
GROUP #09				
B18158030	Muhammad Ali Sarwar	Page 3	20th April 2021	Week 02-Linear Regression Derivation.pdf
B18158068	Zobadresh Azfar	Page 4	20th April 2021	
B18158018	Javeria Ali	E-Commerce Project Demo	20th April 2021	e-commerce-master.zip
B18158020	Kainat Zulfiqar	E-Commerce Project Demo	21st April 2021	

UNIVERSITY OF KARACHI

Step I: Download and Study about following Data set .

3

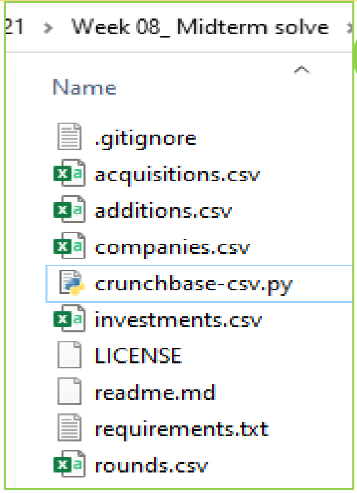
Step I: Download and Study about following Data set .



1

___AI in SE 2021 › Week 08_ Midterm solve

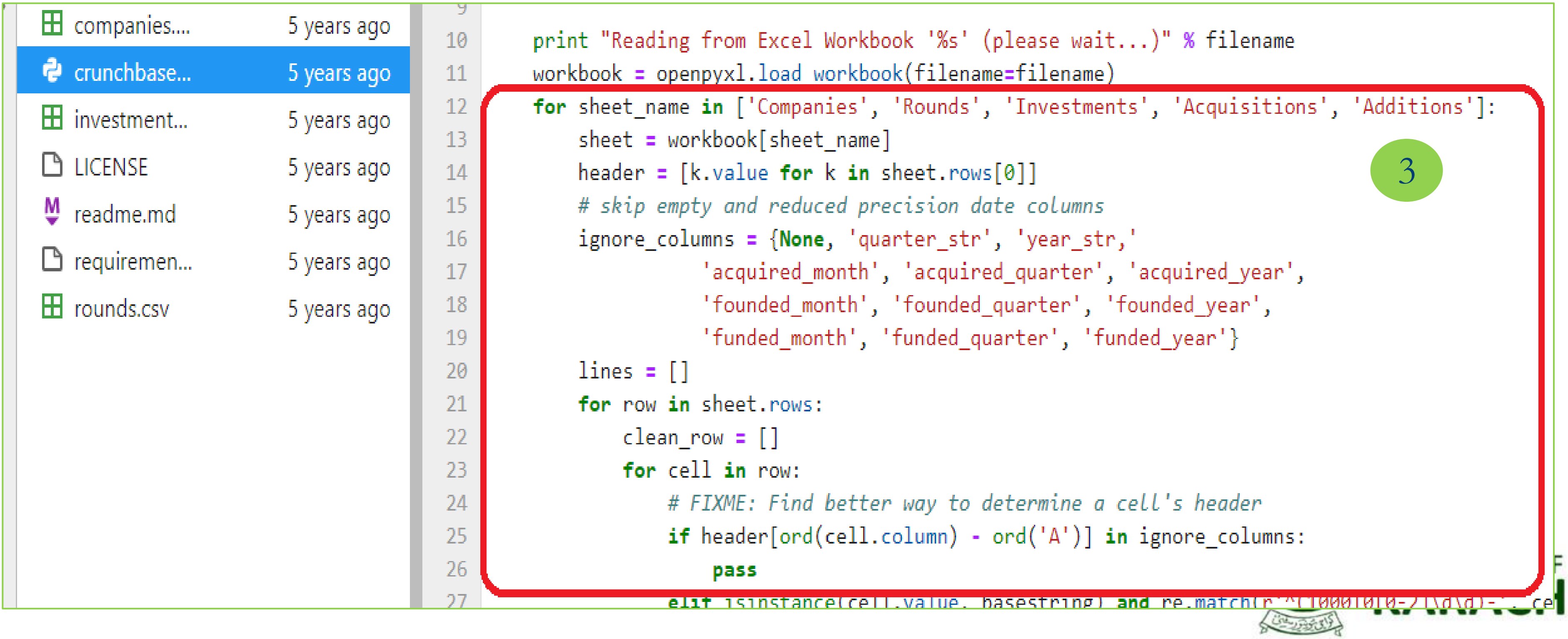
Name	Date modified	Type	Size
AI in SE Lab Midterm.pdf	4/27/2021 7:13 AM	Microsoft Edge P...	169 KB
crunchbase-data-master (1).zip	5/4/2021 5:54 AM	WinRAR ZIP archive	15,733 KB
Week 08-509 AI in SE.pptx	5/4/2021 6:01 AM	Microsoft PowerP...	2,227 KB



2

21 › Week 08_ Midterm solve ›

Name
.gitignore
acquisitions.csv
additions.csv
companies.csv
crunchbase-csv.py
investments.csv
LICENSE
readme.md
requirements.txt
rounds.csv

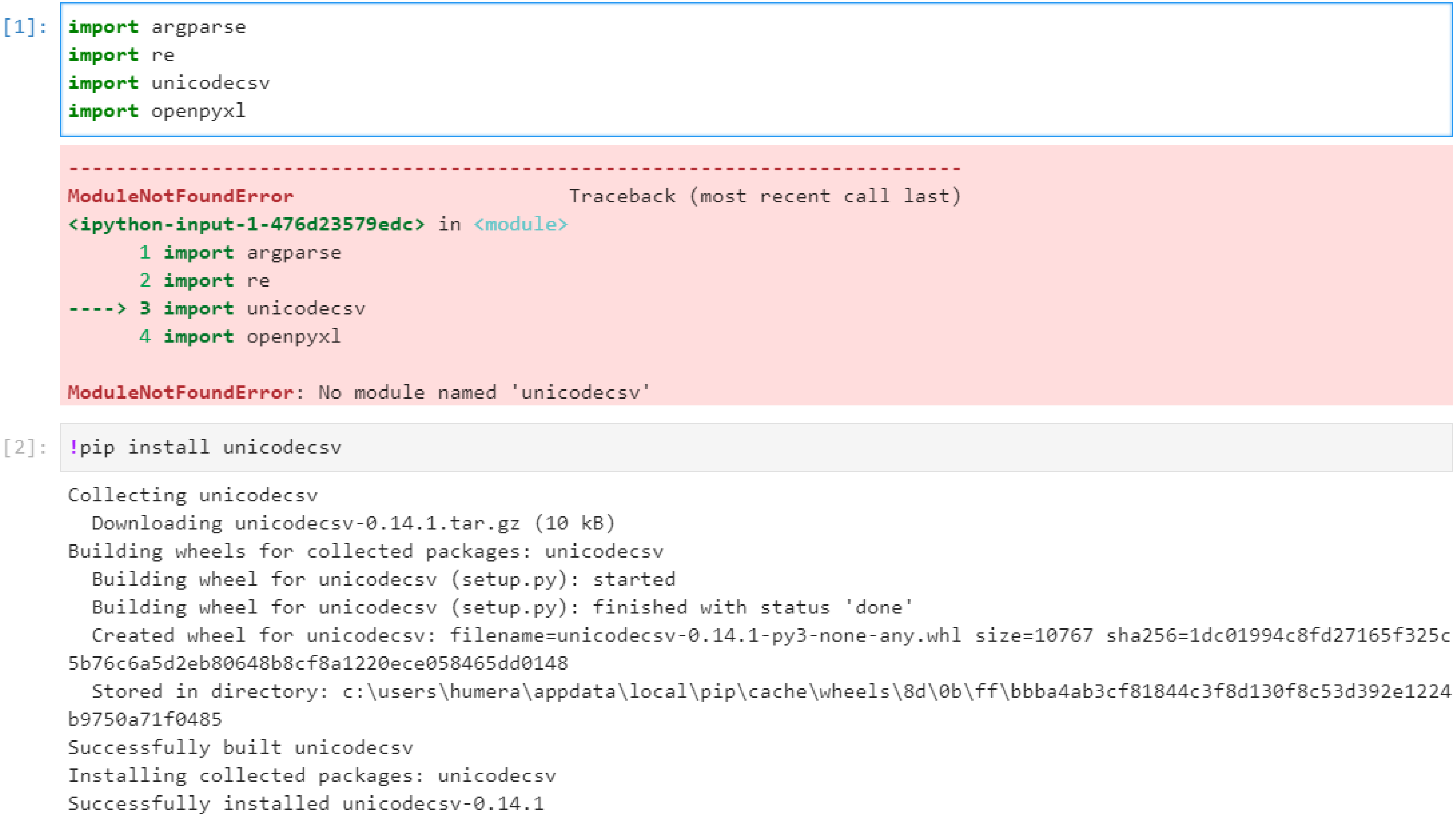


3

companies....	5 years ago
crunchbase...	5 years ago
investment...	5 years ago
LICENSE	5 years ago
readme.md	5 years ago
requiremen...	5 years ago
rounds.csv	5 years ago

```
print "Reading from Excel Workbook '%s' (please wait...)" % filename
workbook = openpyxl.load_workbook(filename=filename)
for sheet_name in ['Companies', 'Rounds', 'Investments', 'Acquisitions', 'Additions']:
    sheet = workbook[sheet_name]
    header = [k.value for k in sheet.rows[0]]
    # skip empty and reduced precision date columns
    ignore_columns = {None, 'quarter_str', 'year_str,'
               'acquired_month', 'acquired_quarter', 'acquired_year',
               'founded_month', 'founded_quarter', 'founded_year',
               'funded_month', 'funded_quarter', 'funded_year'}
    lines = []
    for row in sheet.rows:
        clean_row = []
        for cell in row:
            # FIXME: Find better way to determine a cell's header
            if header[ord(cell.column) - ord('A')] in ignore_columns:
                pass
            elif isinstance(cell.value, basestring) and re.match(r'^(000|0[0-2]\d\d)-', ce
```

Step I (A): Open and Attempt to run .py File



```
[1]: import argparse
     import re
     import unicodecsv
     import openpyxl
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
<ipython-input-1-476d23579edc> in <module>
      1 import argparse
      2 import re
----> 3 import unicodecsv
      4 import openpyxl

ModuleNotFoundError: No module named 'unicodecsv'
```

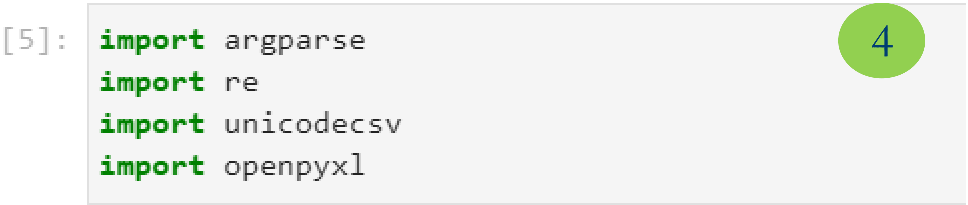
```
[2]: !pip install unicodecsv
```

```
Collecting unicodecsv
  Downloading unicodecsv-0.14.1.tar.gz (10 kB)
Building wheels for collected packages: unicodecsv
  Building wheel for unicodecsv (setup.py): started
  Building wheel for unicodecsv (setup.py): finished with status 'done'
  Created wheel for unicodecsv: filename=unicodecsv-0.14.1-py3-none-any.whl size=10767 sha256=1dc01994c8fd27165f325c
5b76c6a5d2eb80648b8cf8a1220ece058465dd0148
  Stored in directory: c:\users\humera\appdata\local\pip\cache\wheels\8d\0b\ff\bbba4ab3cf81844c3f8d130f8c53d392e1224
b9750a71f0485
Successfully built unicodecsv
Installing collected packages: unicodecsv
Successfully installed unicodecsv-0.14.1
```

UNIVERSITY OF KARACHI

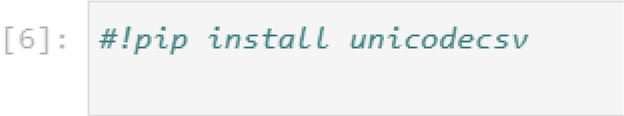
5

Tale of 1000 Crunchbase Startups. Introduction | by Susan Li | Towards Data Science

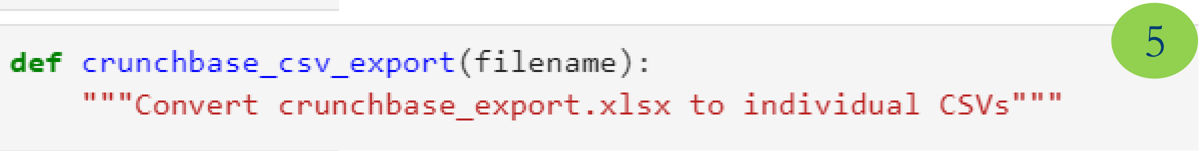


```
[5]: import argparse
     import re
     import unicodecsv
     import openpyxl
```

4

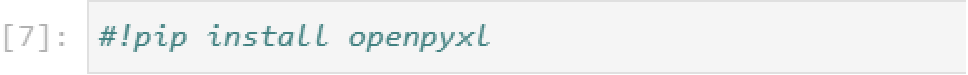


```
[6]: #!pip install unicodecsv
```



```
def crunchbase_csv_export(filename):
    """Convert crunchbase_export.xlsx to individual CSVs"""
```

5



```
[7]: #!pip install openpyxl
```

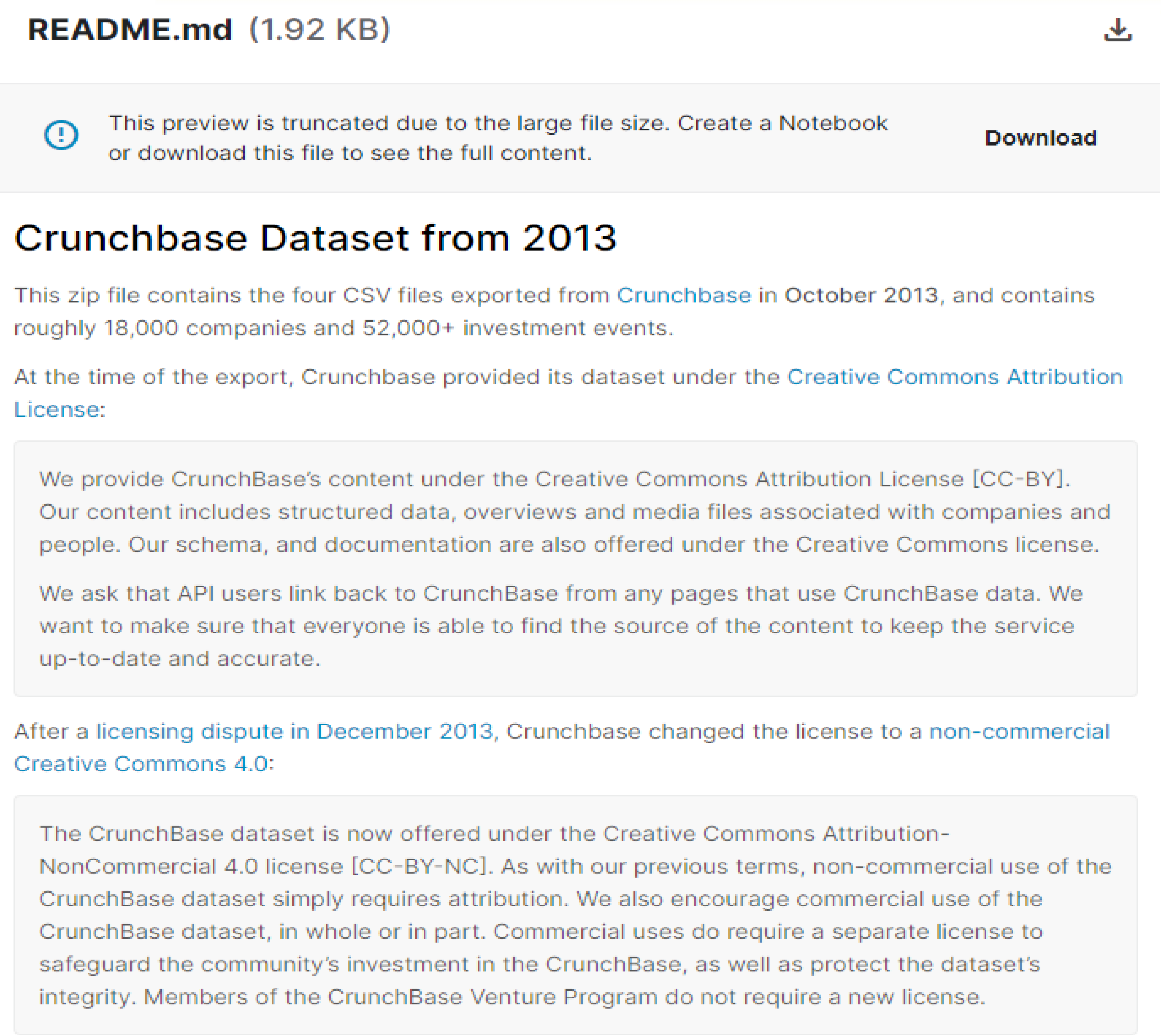
6

Crunchbase is a website that crowd sources information about the fundraising of many startups. It is an excellent resource for discovering innovative companies and learning about the people behind them.

Unfortunately, unlike other public data sources, one had to pay a Pro membership in order to download the data from Crunchbase. Therefore, I decided to download data from here, which is not the most up to date; however, it is fine for my purposes.

UNIVERSITY OF KARACHI

README



README.md (1.92 KB)

This preview is truncated due to the large file size. Create a Notebook or download this file to see the full content.

Crunchbase Dataset from 2013

This zip file contains the four CSV files exported from Crunchbase in October 2013, and contains roughly 18,000 companies and 52,000+ investment events.

At the time of the export, Crunchbase provided its dataset under the Creative Commons Attribution License:

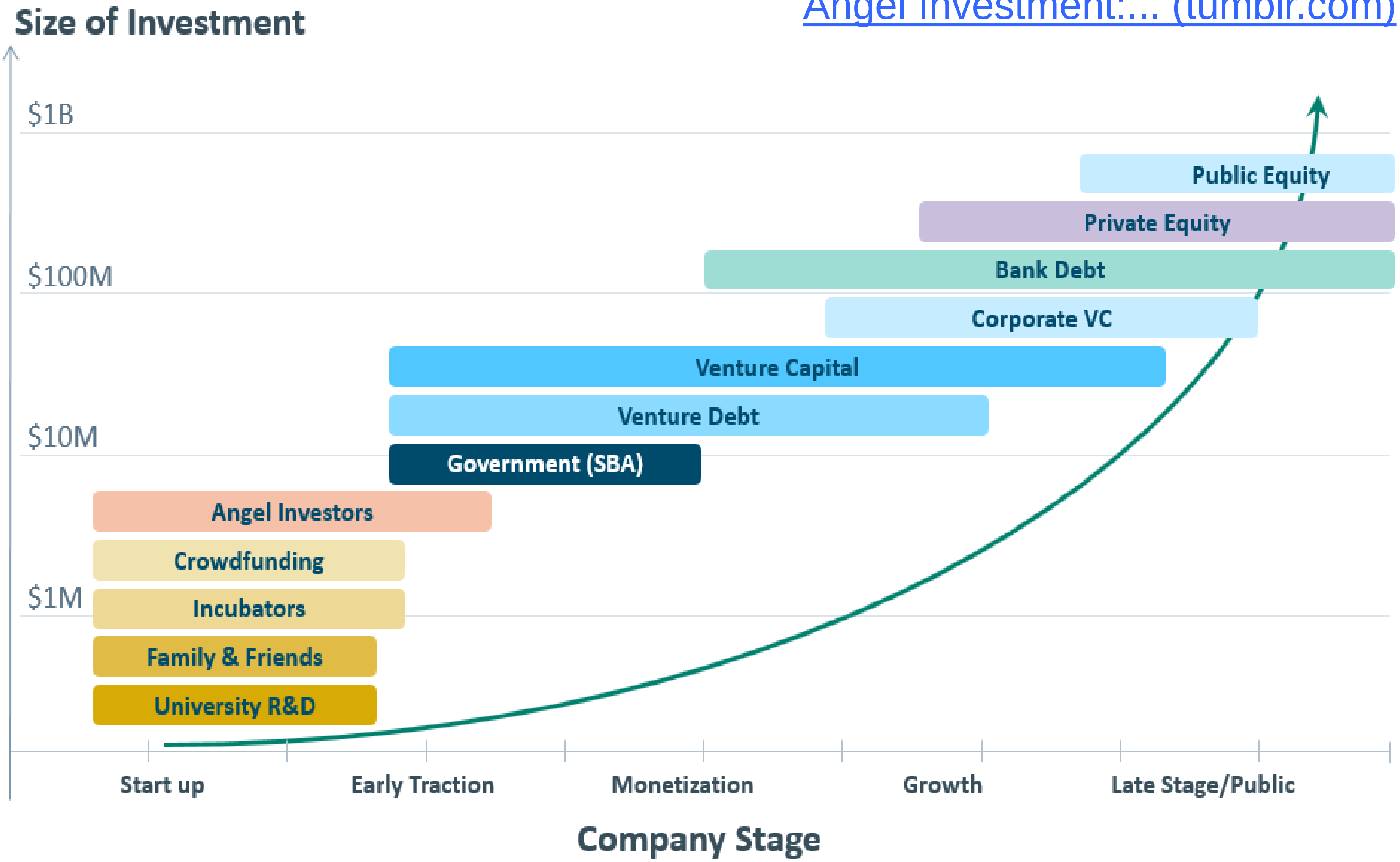
> We provide CrunchBase's content under the Creative Commons Attribution License [CC-BY]. Our content includes structured data, overviews and media files associated with companies and people. Our schema, and documentation are also offered under the Creative Commons license.
>
> We ask that API users link back to CrunchBase from any pages that use CrunchBase data. We want to make sure that everyone is able to find the source of the content to keep the service up-to-date and accurate.

After a licensing dispute in December 2013, Crunchbase changed the license to a non-commercial Creative Commons 4.0:

> The CrunchBase dataset is now offered under the Creative Commons Attribution-NonCommercial 4.0 license [CC-BY-NC]. As with our previous terms, non-commercial use of the CrunchBase dataset simply requires attribution. We also encourage commercial use of the CrunchBase dataset, in whole or in part. Commercial uses do require a separate license to safeguard the community's investment in the CrunchBase, as well as protect the dataset's integrity. Members of the CrunchBase Venture Program do not require a new license.

Seeking startup funding

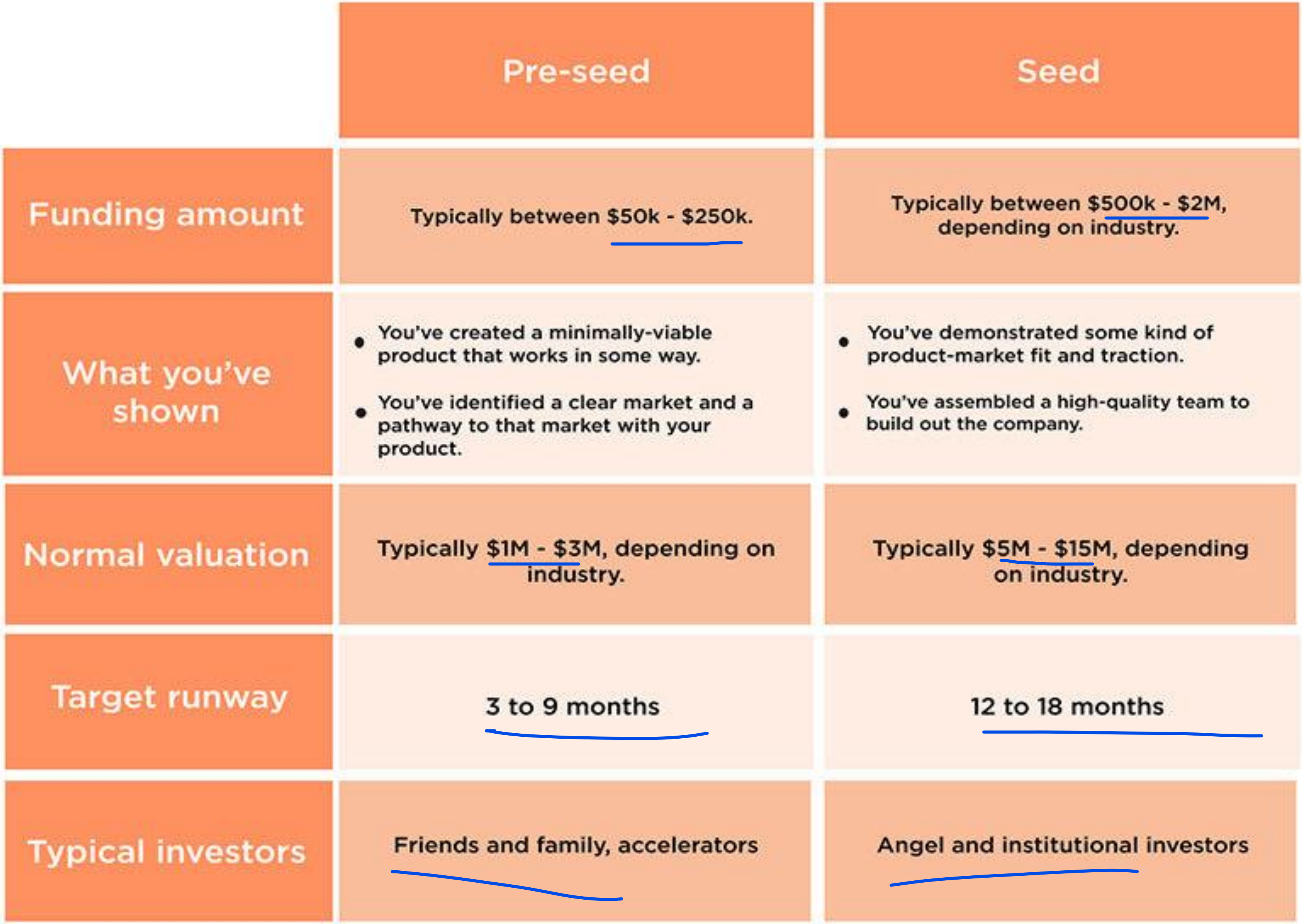
[VC — How to attract Venture Capital & Angel Investment:... (tumblr.com)]



UNIVERSITY OF KARACHI

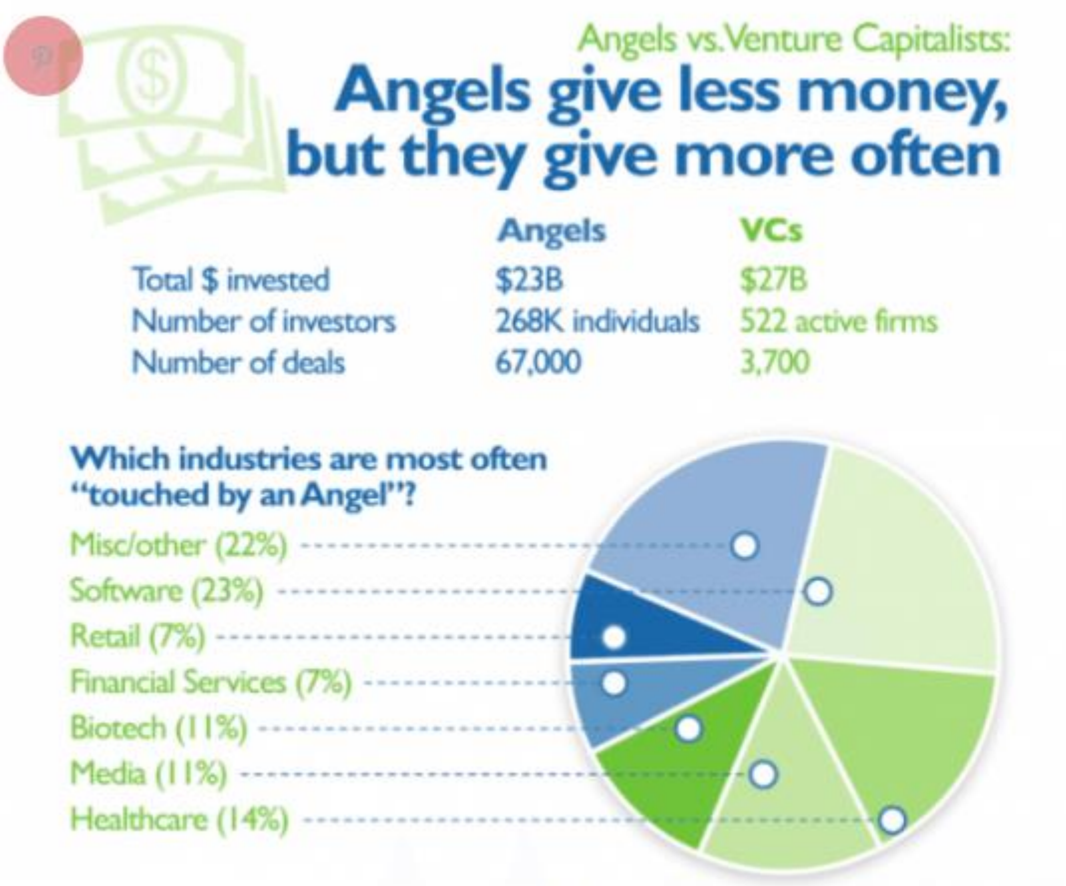
VC — How to attract Venture Capital & Angel Investment:... (tumblr.com)

Stages of Funding



	Pre-seed	Seed
Funding amount	Typically between $50k - $250k.	Typically between $500k - $2M, depending on industry.
What you've shown	• You've created a minimally-viable product that works in some way. • You've identified a clear market and a pathway to that market with your product.	• You've demonstrated some kind of product-market fit and traction. • You've assembled a high-quality team to build out the company.
Normal valuation	Typically $1M - $3M, depending on industry.	Typically $5M - $15M, depending on industry.
Target runway	3 to 9 months	12 to 18 months
Typical investors	Friends and family, accelerators	Angel and institutional investors

Angel vs. VC's



Angels vs. Venture Capitalists:

Angels give less money, but they give more often

	Angels	VCs
Total $ invested	$23B	$27B
Number of investors	268K individuals	522 active firms
Number of deals	67,000	3,700

Which industries are most often "touched by an Angel"?

- Misc/other (22%)
- Software (23%)
- Retail (7%)
- Financial Services (7%)
- Biotech (11%)
- Media (11%)
- Healthcare (14%)

Source: Score.org

Data Exploration

The dataset contains three tables: investments, companies, and acquisitions.

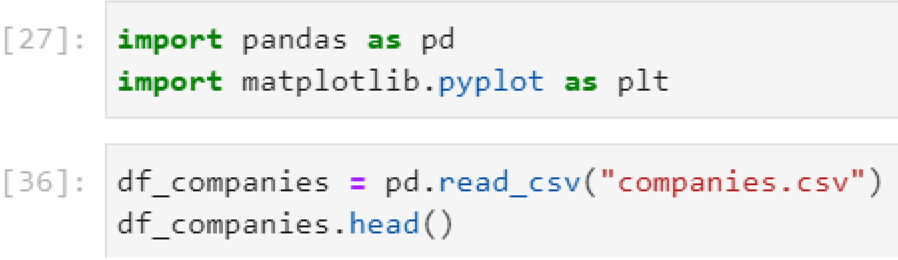
acquisitions.csv
additions.csv
companies.csv

It includes more than 66,000 companies that were founded between 1977 and 2015.

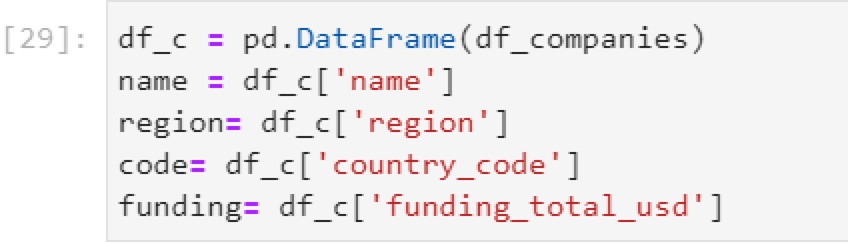
Among these 66,000 companies, there were approximately 18,000 companies that were subsequently acquired.

Step II: Loading Data Set

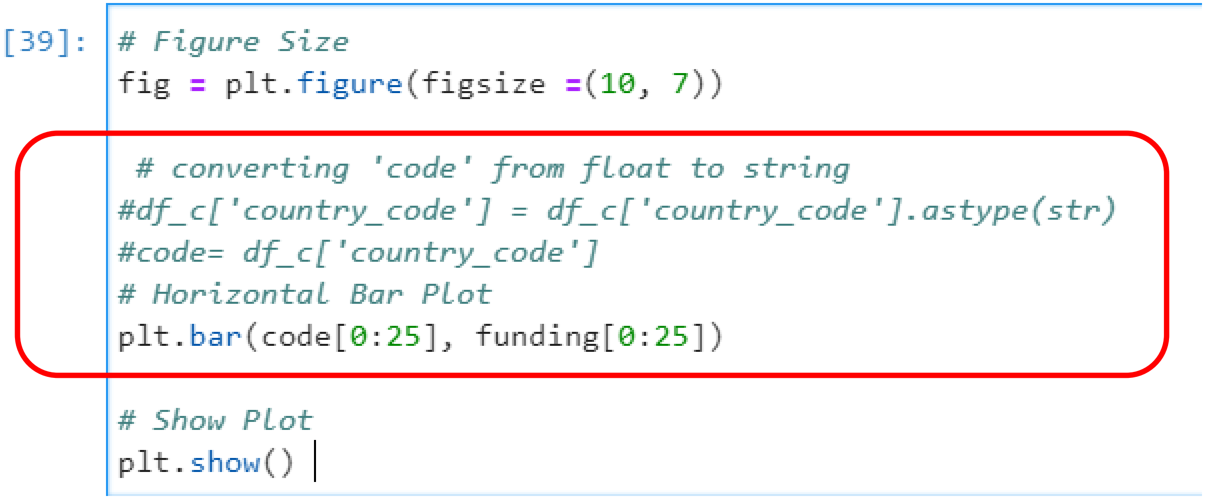
Pandas and data frames



```
[27]: import pandas as pd
      import matplotlib.pyplot as plt

[36]: df_companies = pd.read_csv("companies.csv")
      df_companies.head()
```

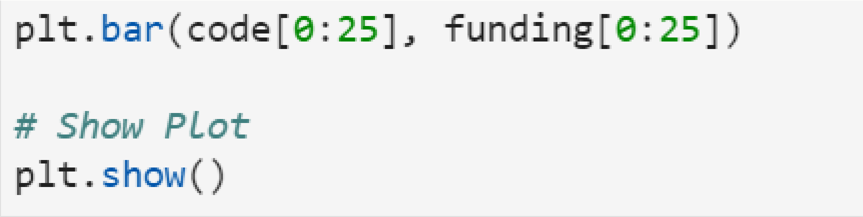


```
[29]: df_c = pd.DataFrame(df_companies)
      name = df_c['name']
      region= df_c['region']
      code= df_c['country_code']
      funding= df_c['funding_total_usd']
```

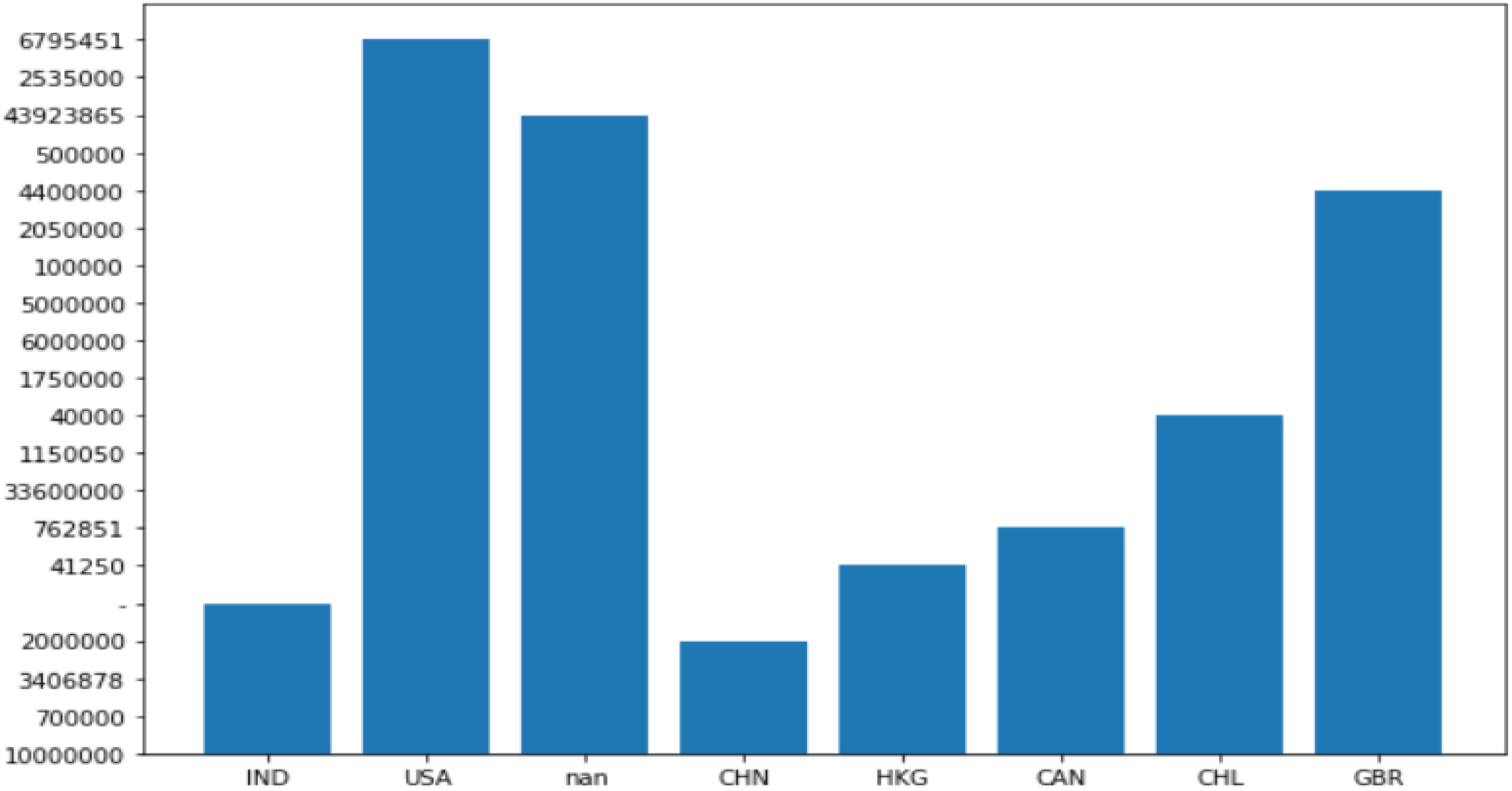


```
[39]: # Figure Size
      fig = plt.figure(figsize =(10, 7))

       # converting 'code' from float to string
      #df_c['country_code'] = df_c['country_code'].astype(str)
      #code= df_c['country_code']
      # Horizontal Bar Plot
      plt.bar(code[0:25], funding[0:25])

      # Show Plot
      plt.show()
```

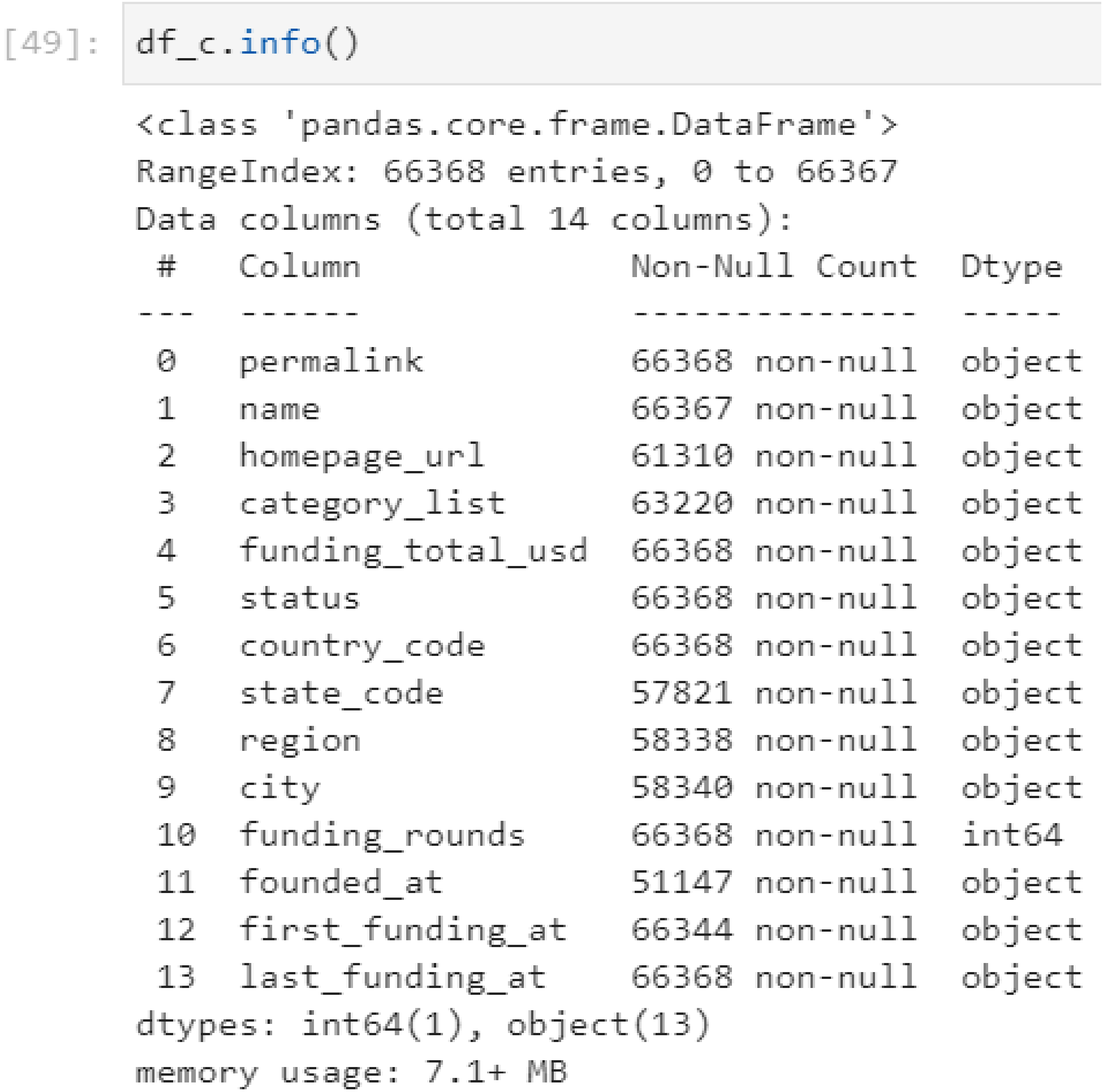
Bar plot between Country code and funding



```
plt.bar(code[0:25], funding[0:25])

# Show Plot
plt.show()
```



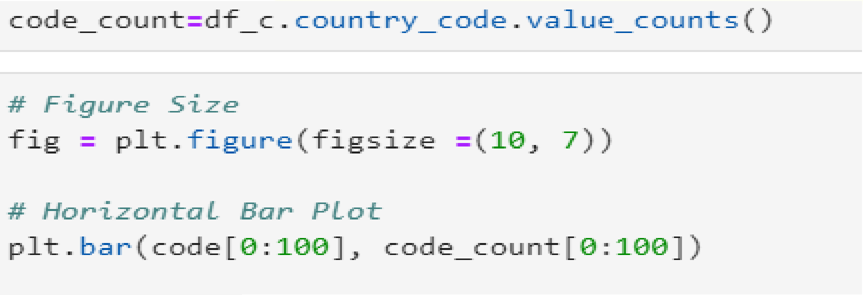
Observations, data type and count



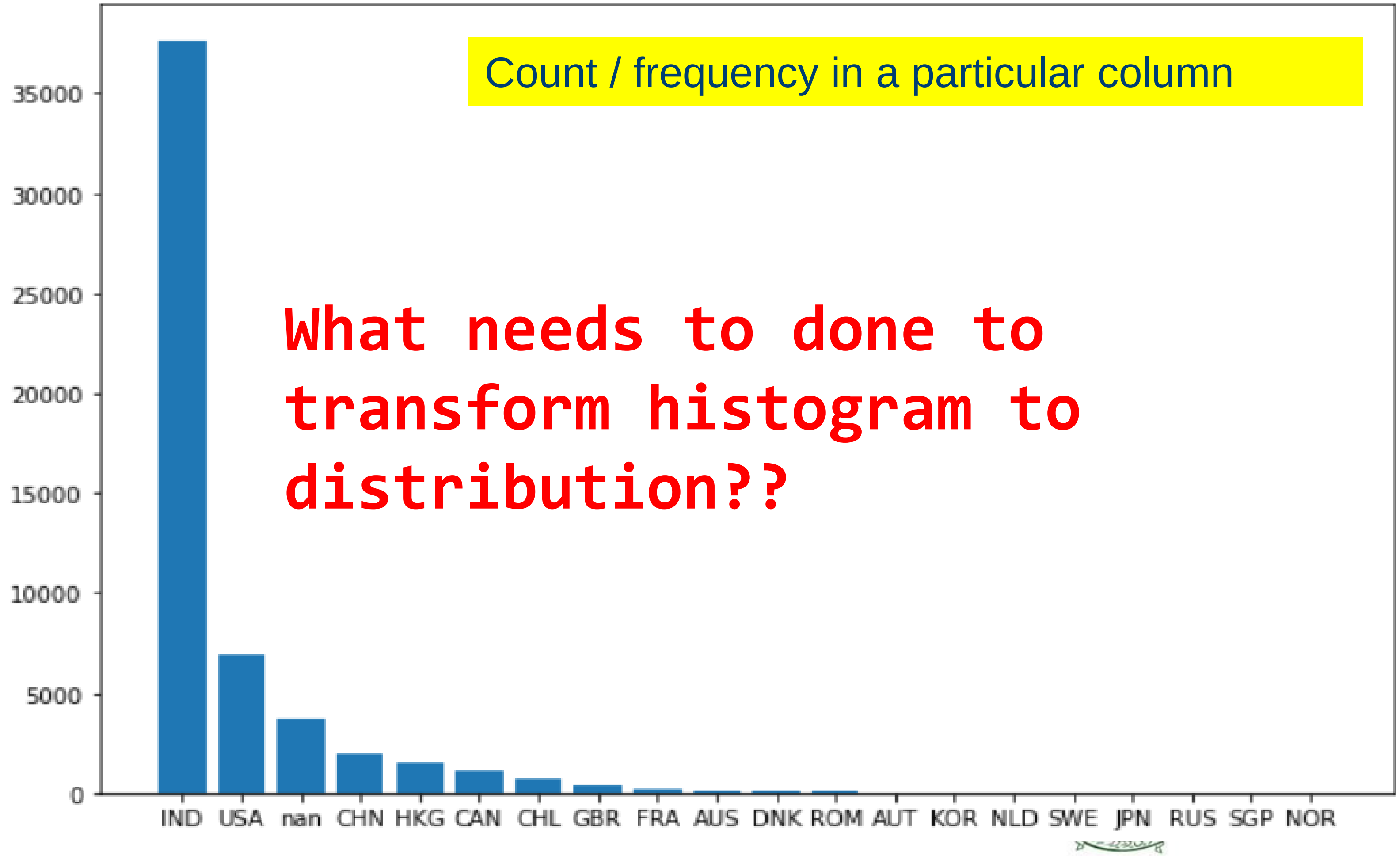
```
[49]: df_c.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 66368 entries, 0 to 66367
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype 
---  ------             --------------  ----- 
 0   permalink          66368 non-null  object
 1   name               66367 non-null  object
 2   homepage_url       61310 non-null  object
 3   category_list      63220 non-null  object
 4   funding_total_usd  66368 non-null  object
 5   status             66368 non-null  object
 6   country_code       66368 non-null  object
 7   state_code         57821 non-null  object
 8   region             58338 non-null  object
 9   city               58340 non-null  object
 10  funding_rounds     66368 non-null  int64 
 11  founded_at         51147 non-null  object
 12  first_funding_at   66344 non-null  object
 13  last_funding_at    66368 non-null  object
dtypes: int64(1), object(13)
memory usage: 7.1+ MB
```

Name the pattern hidden inside histogram ?



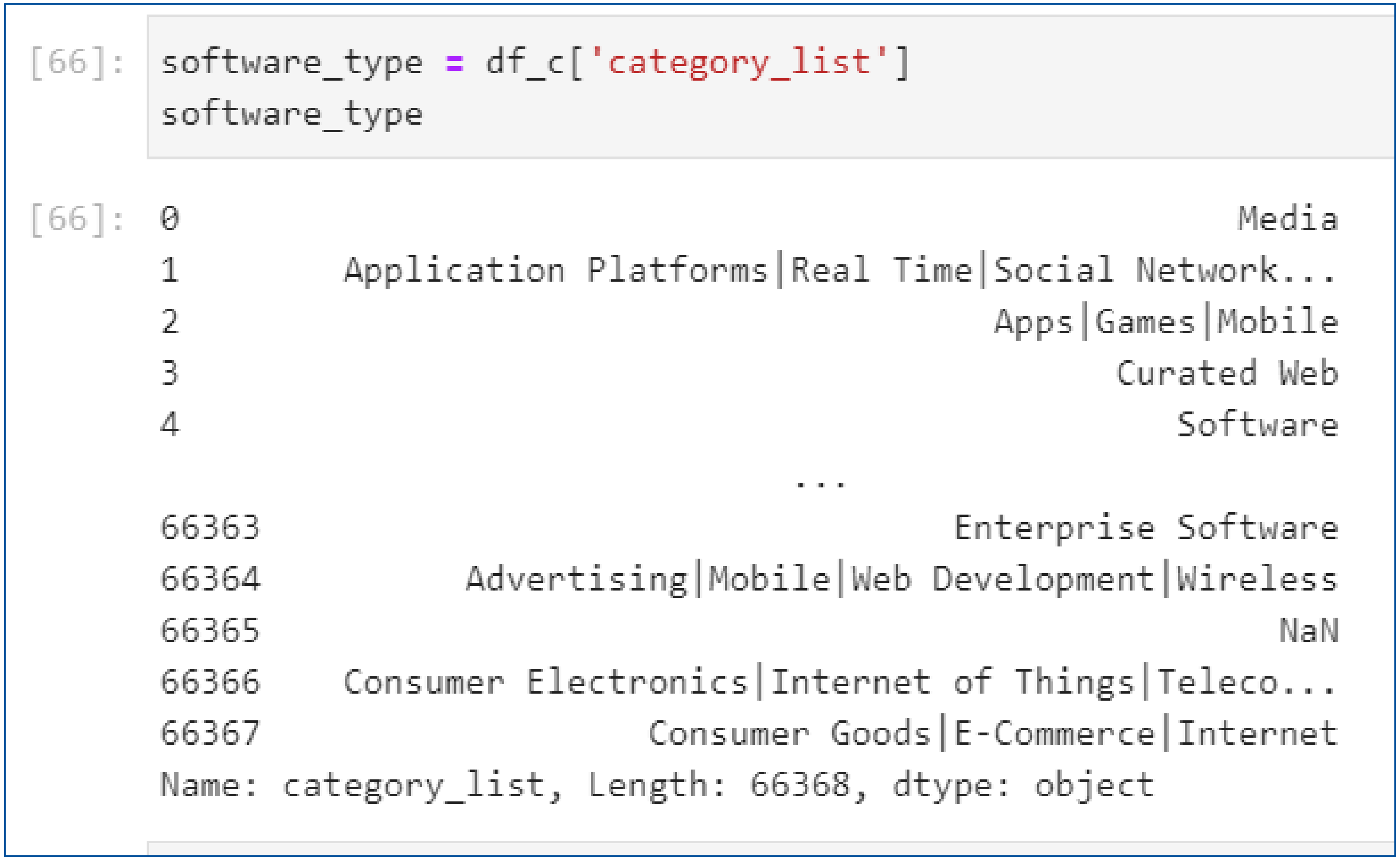
```
code_count=df_c.country_code.value_counts()

# Figure Size
fig = plt.figure(figsize =(10, 7))

# Horizontal Bar Plot
plt.bar(code[0:100], code_count[0:100])

# Show Plot
plt.show()
```



Count / frequency in a particular column

What needs to done to transform histogram to distribution??

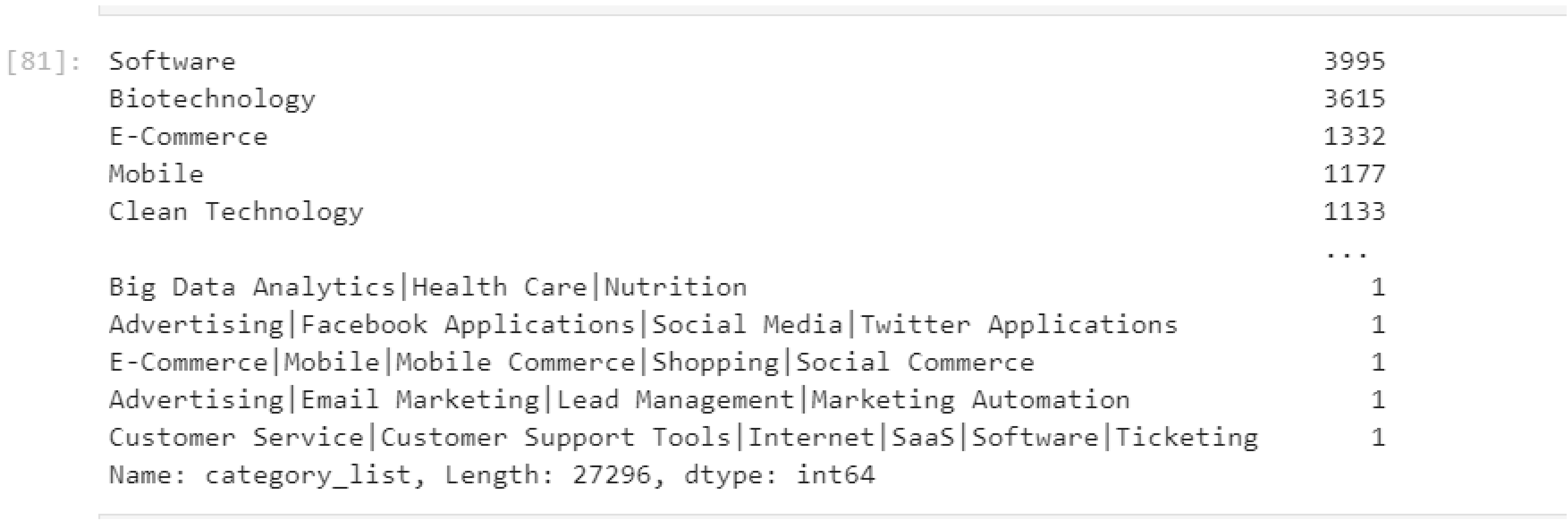
Product Types



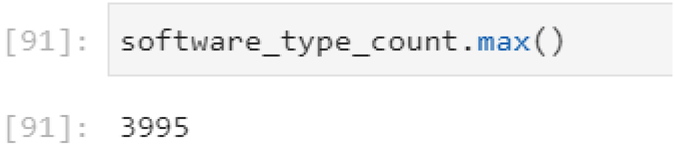
```
[66]: software_type = df_c['category_list']
      software_type
```

```
[66]: 0                                                    Media
      1        Application Platforms|Real Time|Social Network...
      2                                         Apps|Games|Mobile
      3                                               Curated Web
      4                                                  Software
                                     ...                         
      66363                                   Enterprise Software
      66364            Advertising|Mobile|Web Development|Wireless
      66365                                                   NaN
      66366     Consumer Electronics|Internet of Things|Teleco...
      66367                    Consumer Goods|E-Commerce|Internet
      Name: category_list, Length: 66368, dtype: object
```

Histogram of category vs its count/peak

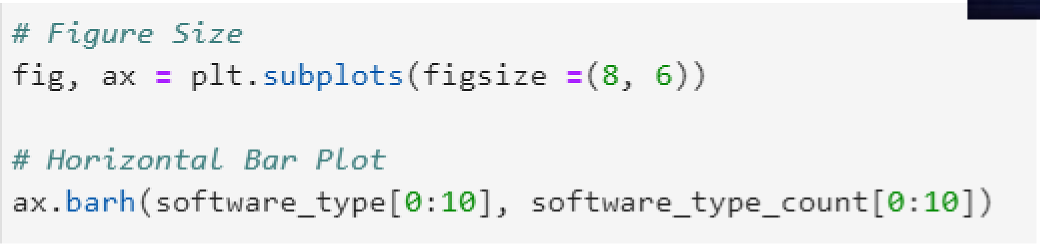


```
[81]: Software                                                               3995
      Biotechnology                                                          3615
      E-Commerce                                                             1332
      Mobile                                                                 1177
      Clean Technology                                                       1133
                                                                             ...
      Big Data Analytics|Health Care|Nutrition                                  1
      Advertising|Facebook Applications|Social Media|Twitter Applications       1
      E-Commerce|Mobile|Mobile Commerce|Shopping|Social Commerce                1
      Advertising|Email Marketing|Lead Management|Marketing Automation          1
      Customer Service|Customer Support Tools|Internet|SaaS|Software|Ticketing  1
      Name: category_list, Length: 27296, dtype: int64
```

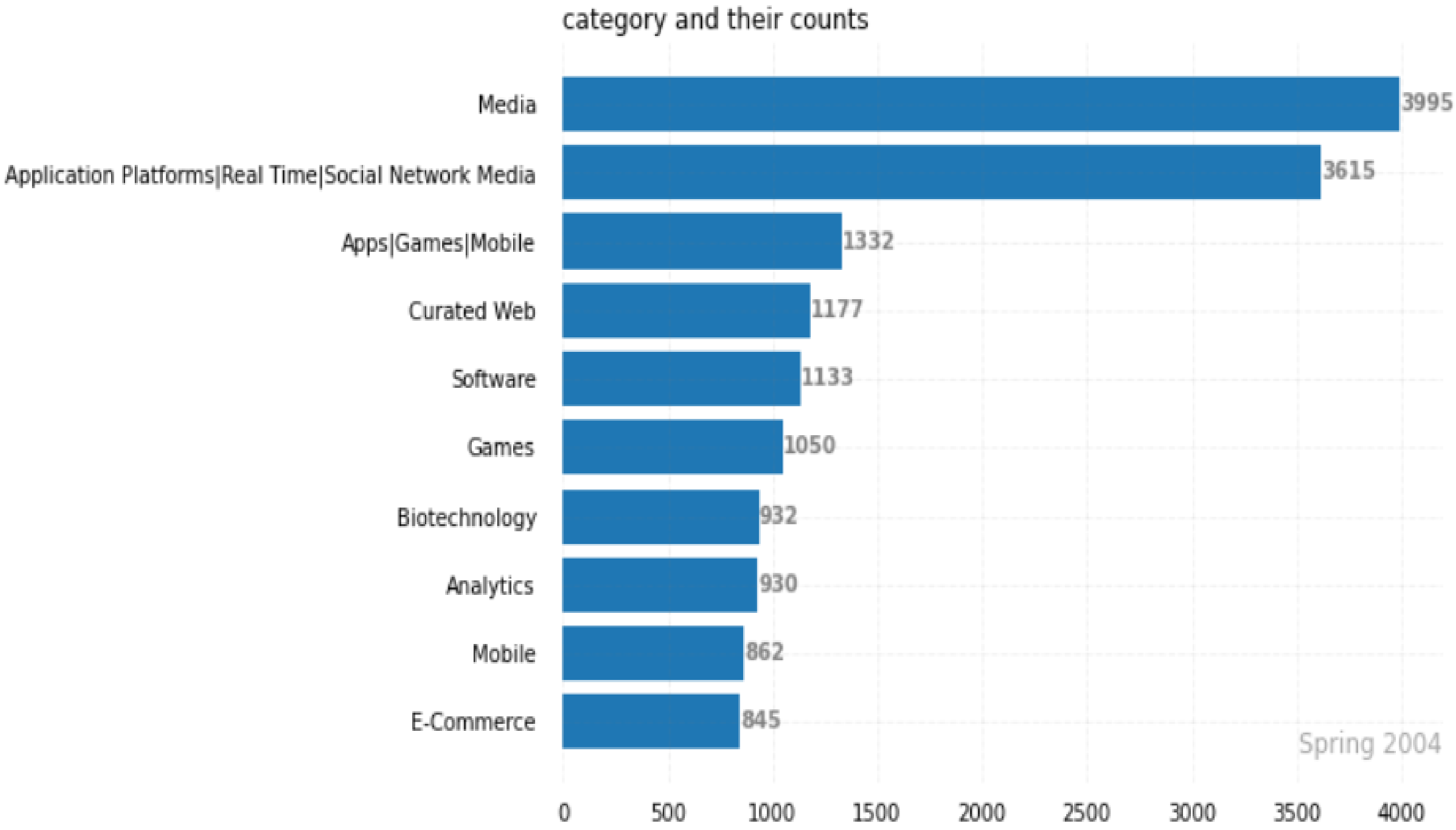


```
[91]: software_type_count.max()

[91]: 3995
```

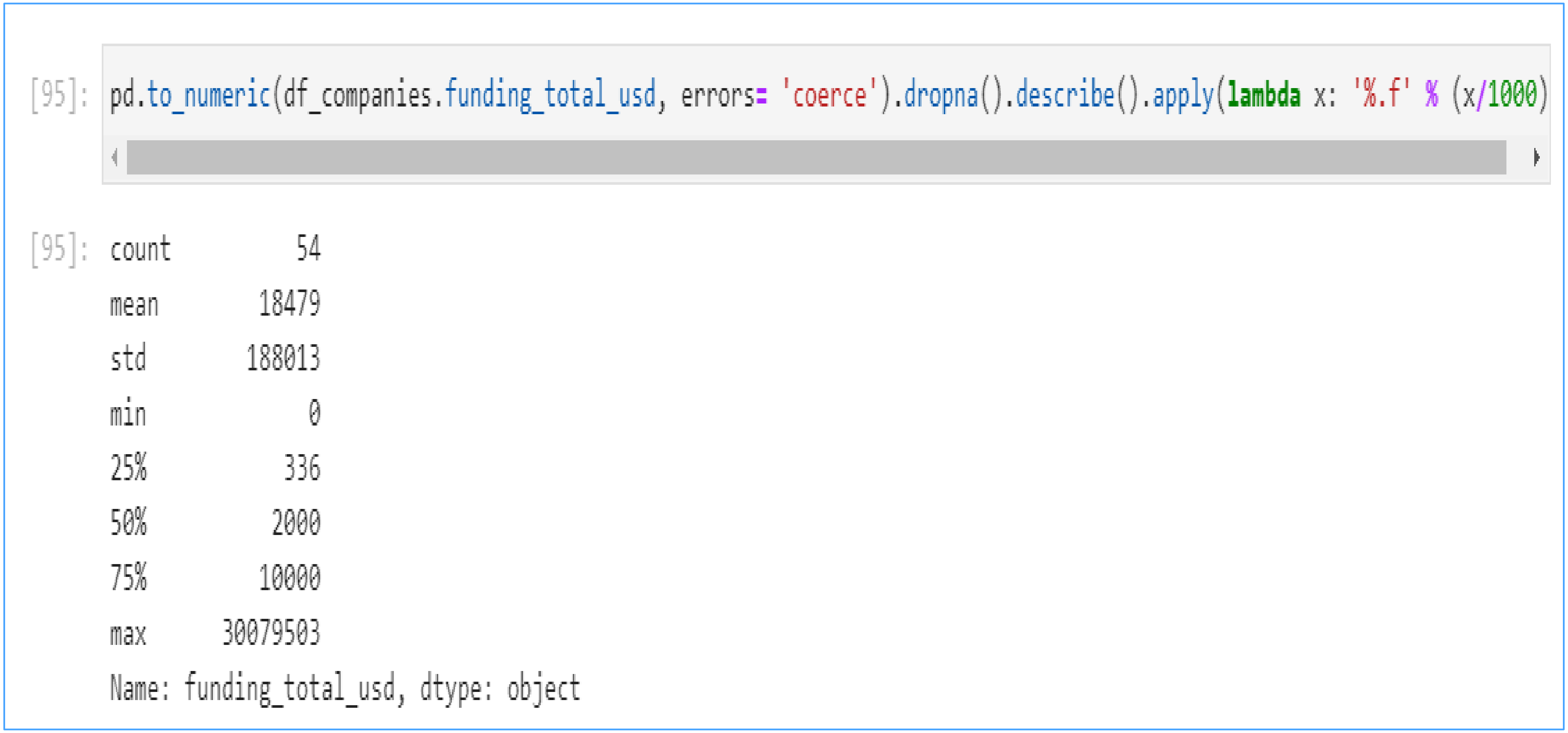
Submission due time: 3:30 pm today



```
# Figure Size
fig, ax = plt.subplots(figsize =(8, 6))

# Horizontal Bar Plot
ax.barh(software_type[0:10], software_type_count[0:10])
```



Mean, Std, min , max of total funding



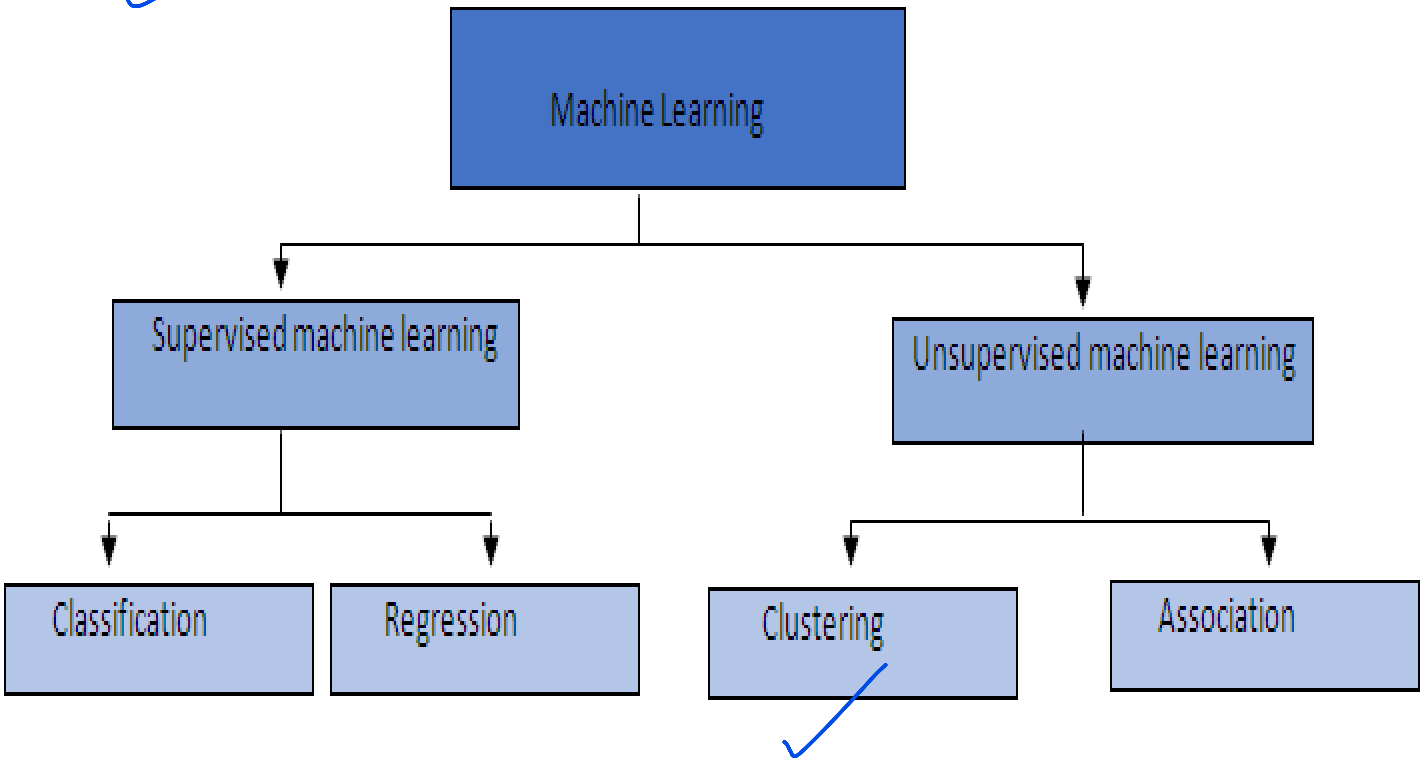
```
[95]: pd.to_numeric(df_companies.funding_total_usd, errors= 'coerce').dropna().describe().apply(lambda x: '%.f' % (x/1000)
```

```
[95]: count          54
      mean        18479
      std        188013
      min             0
      25%           336
      50%          2000
      75%         10000
      max      30079503
      Name: funding_total_usd, dtype: object
```

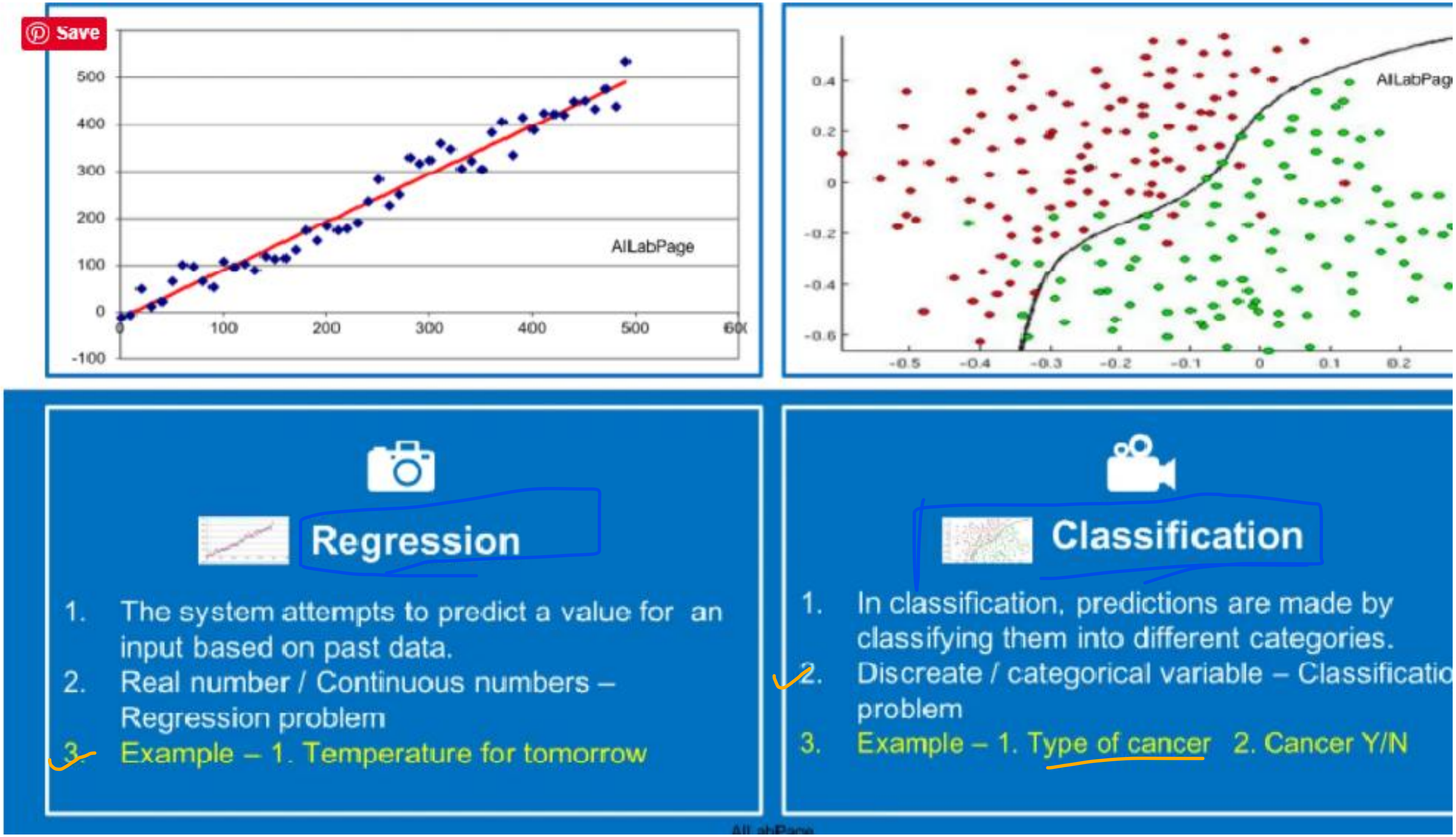
Step III: Which task you prefer to perform this dataset:

Regression and Classification

Nature of Problem/ Learning



Classification vs. Regression



Regression

1. The system attempts to predict a value for an input based on past data.
2. Real number / Continuous numbers – Regression problem
3. Example – 1. Temperature for tomorrow

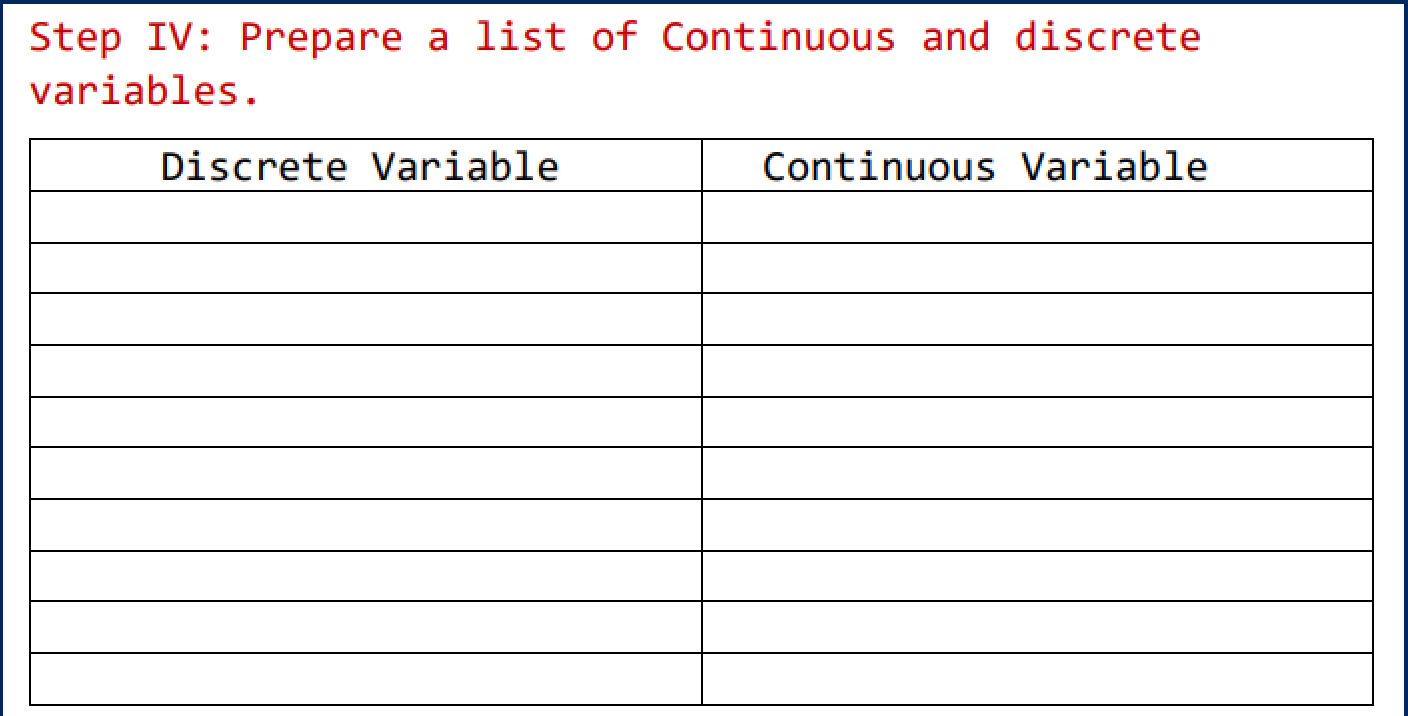
Classification

1. In classification, predictions are made by classifying them into different categories.
2. Discreate / categorical variable – Classificatio problem
3. Example – 1. Type of cancer 2. Cancer Y/N

AILabPage

UNIVERSITY OF KARACHI

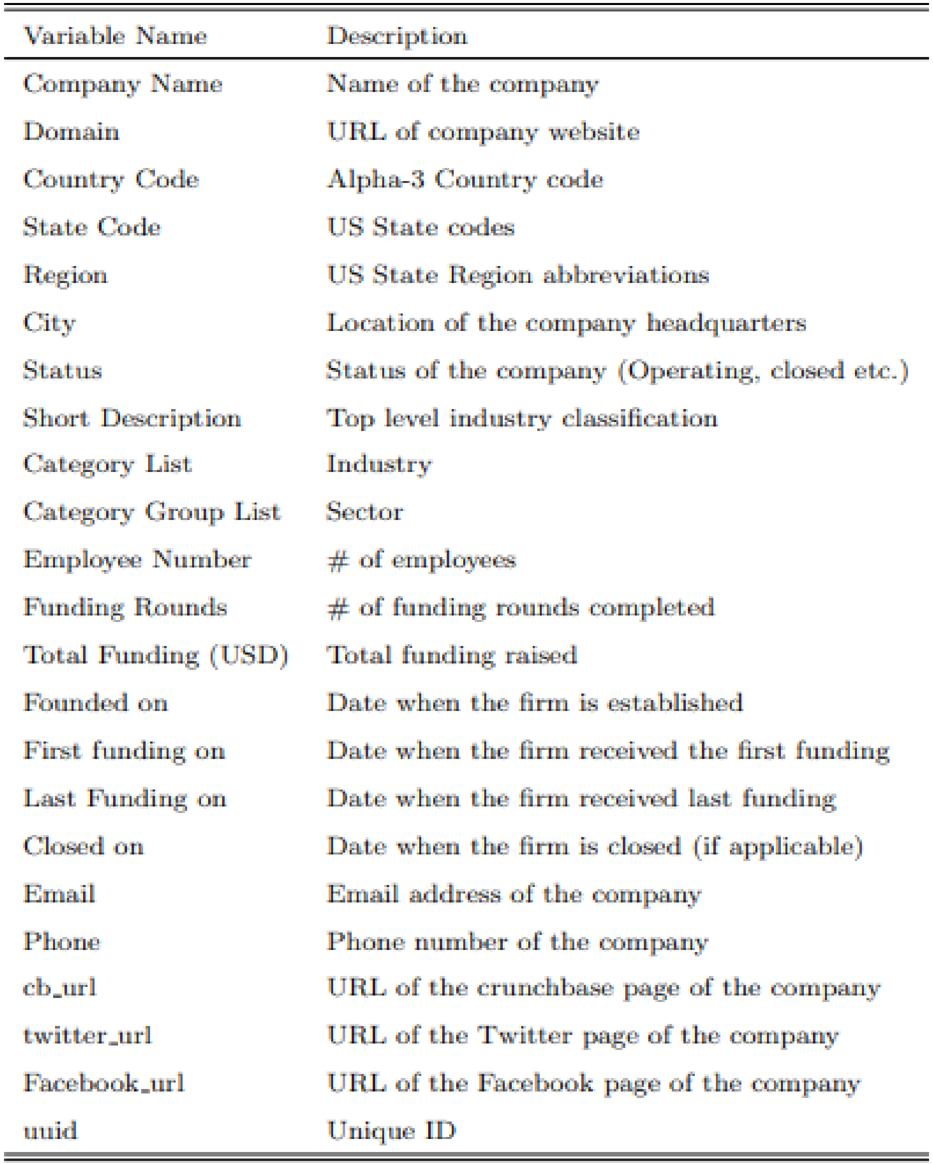
Step IV



Step IV: Prepare a list of Continuous and discrete variables.

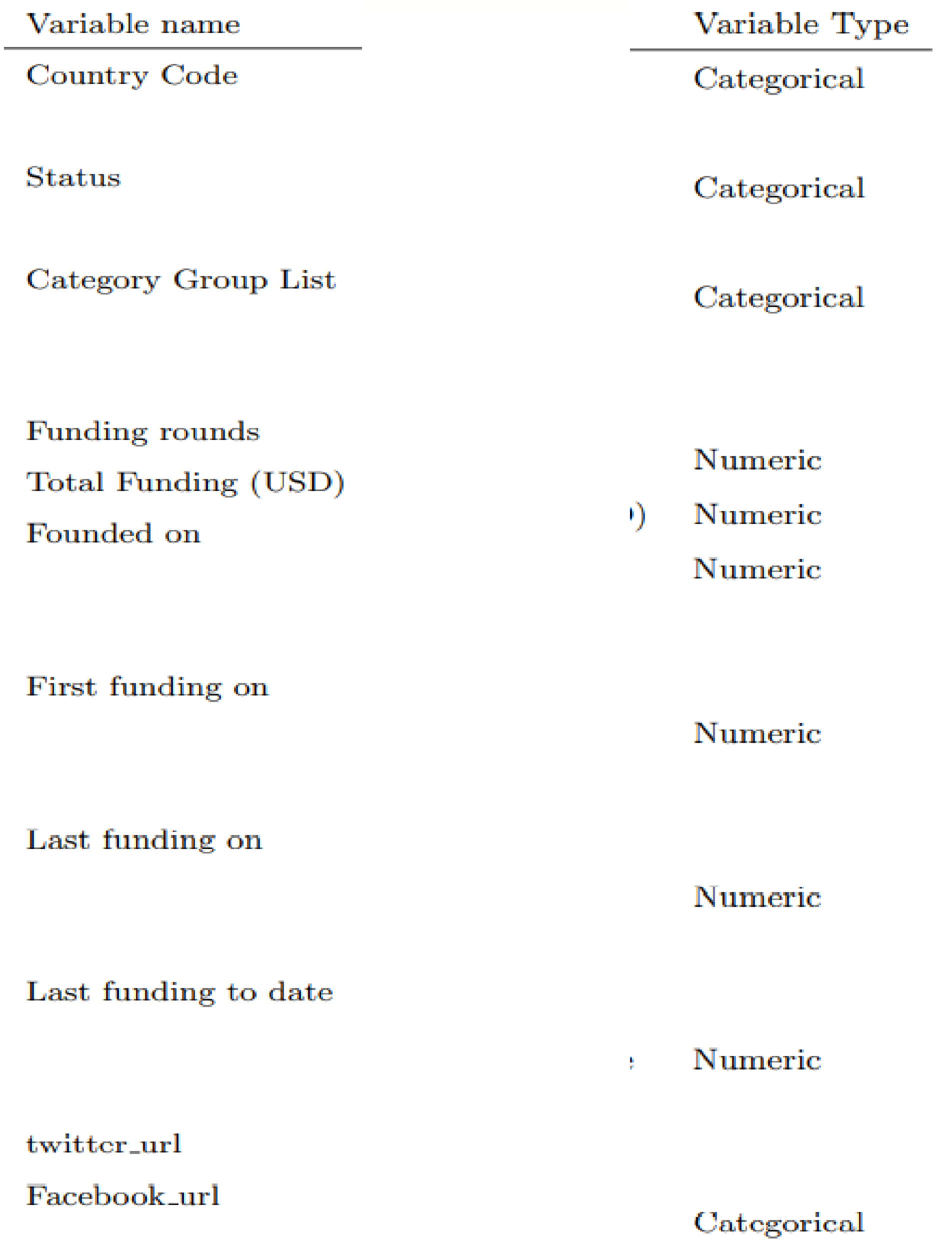
Discrete Variable	Continuous Variable

Description of raw variable



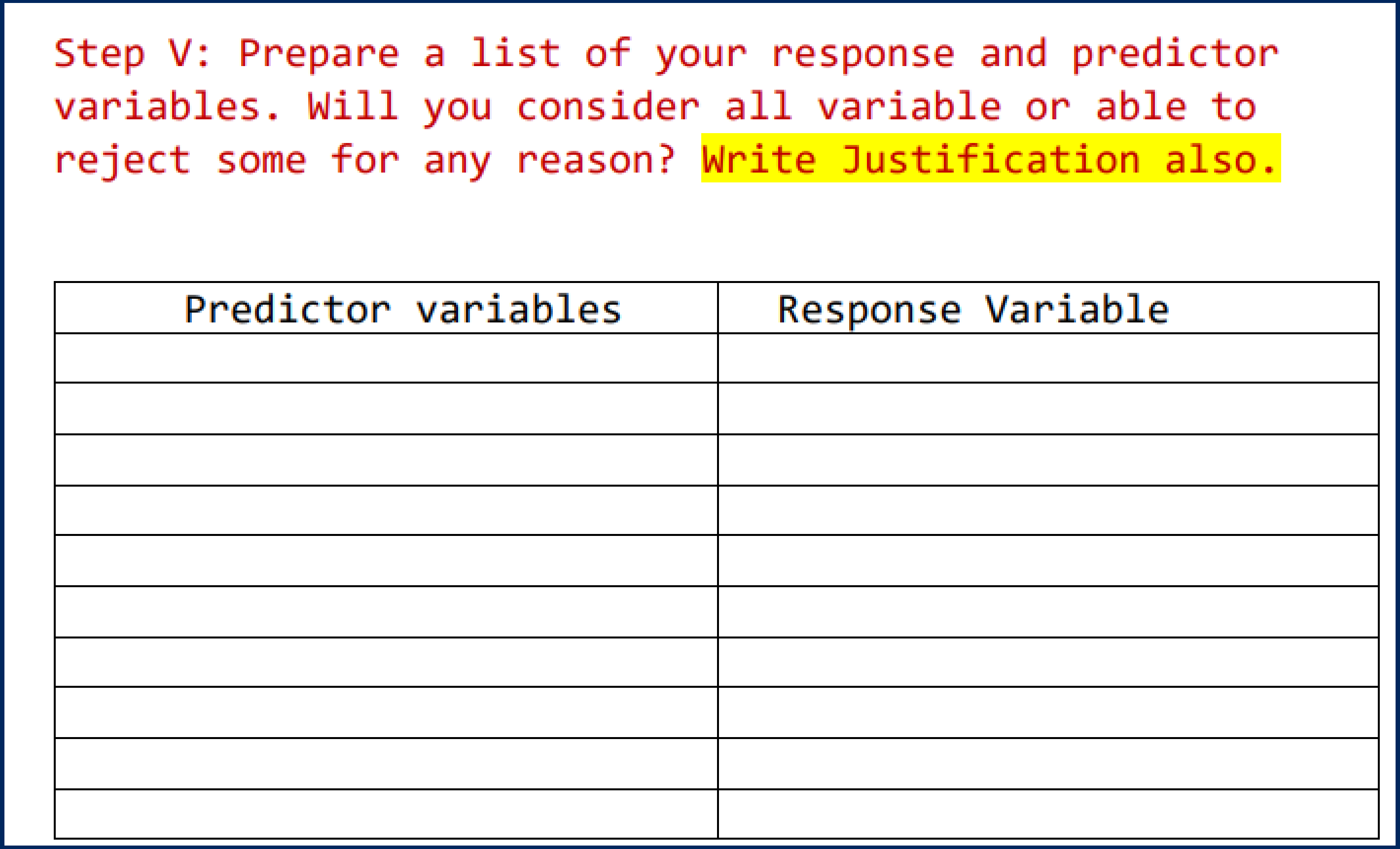
Variable Name	Description
Company Name	Name of the company
Domain	URL of company website
Country Code	Alpha-3 Country code
State Code	US State codes
Region	US State Region abbreviations
City	Location of the company headquarters
Status	Status of the company (Operating, closed etc.)
Short Description	Top level industry classification
Category List	Industry
Category Group List	Sector
Employee Number	# of employees
Funding Rounds	# of funding rounds completed
Total Funding (USD)	Total funding raised
Founded on	Date when the firm is established
First funding on	Date when the firm received the first funding
Last Funding on	Date when the firm received last funding
Closed on	Date when the firm is closed (if applicable)
Email	Email address of the company
Phone	Phone number of the company
cb_url	URL of the crunchbase page of the company
twitter_url	URL of the Twitter page of the company
Facebook_url	URL of the Facebook page of the company
uuid	Unique ID

Discrete vs. Categorical



Variable name	Variable Type
Country Code	Categorical
Status	Categorical
Category Group List	Categorical
Funding rounds	Numeric
Total Funding (USD)	Numeric
Founded on	Numeric
First funding on	Numeric
Last funding on	Numeric
Last funding to date	Numeric
twittcr_url Facebook_url	Catcgorical

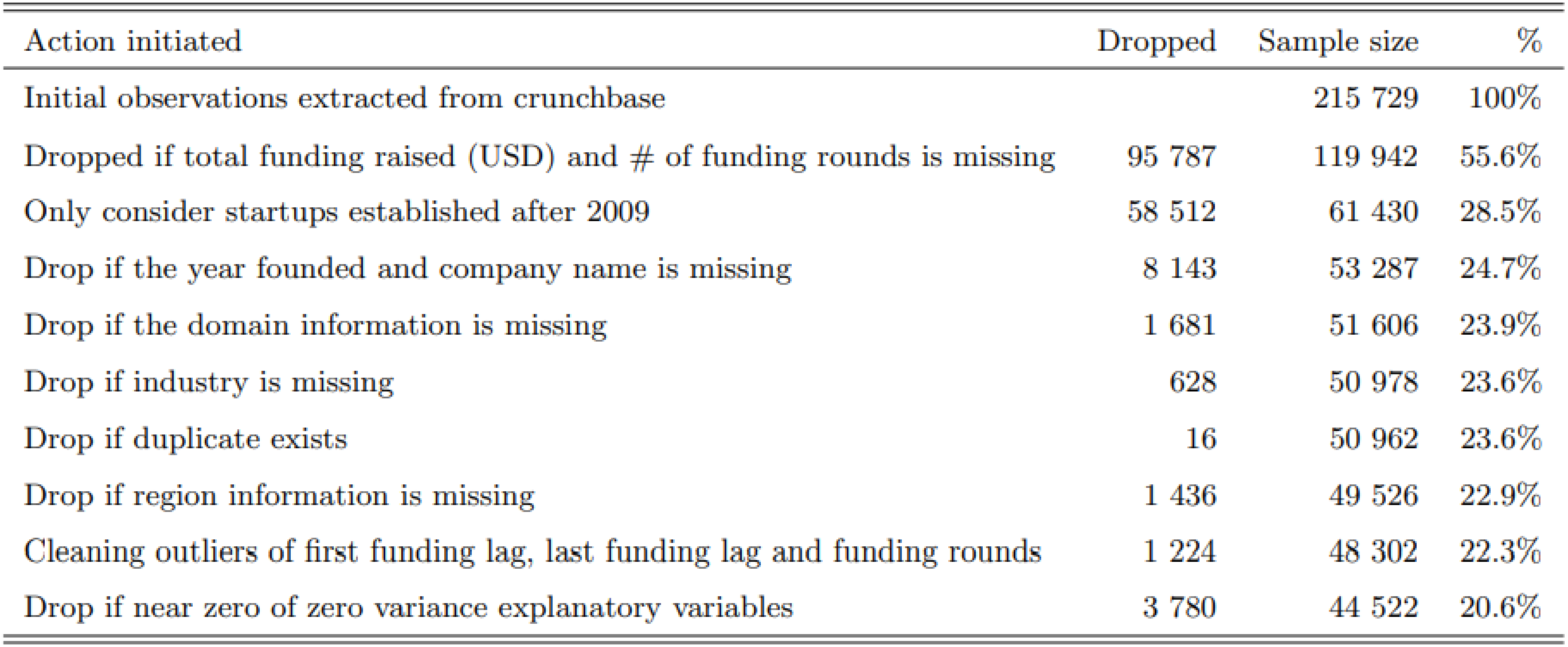
Step V



Step V: Prepare a list of your response and predictor variables. Will you consider all variable or able to reject some for any reason? Write Justification also.

Predictor variables	Response Variable

Example of Data Cleaning Steps



Action initiated	Dropped	Sample size	%
Initial observations extracted from crunchbase		215 729	100%
Dropped if total funding raised (USD) and # of funding rounds is missing	95 787	119 942	55.6%
Only consider startups established after 2009	58 512	61 430	28.5%
Drop if the year founded and company name is missing	8 143	53 287	24.7%
Drop if the domain information is missing	1 681	51 606	23.9%
Drop if industry is missing	628	50 978	23.6%
Drop if duplicate exists	16	50 962	23.6%
Drop if region information is missing	1 436	49 526	22.9%
Cleaning outliers of first funding lag, last funding lag and funding rounds	1 224	48 302	22.3%
Drop if near zero of zero variance explanatory variables	3 780	44 522	20.6%

UNIVERSITY OF KARACHI

Submission due time: 12:00 noon Friday

Tale of 1000 Crunchbase Startups. Introduction | by Susan Li | Towards Data Science

https://hackersandslackers.com/compare-rows-pandas-dataframes/

Let's build a function called dataframe_difference() which answers any of 4 questions

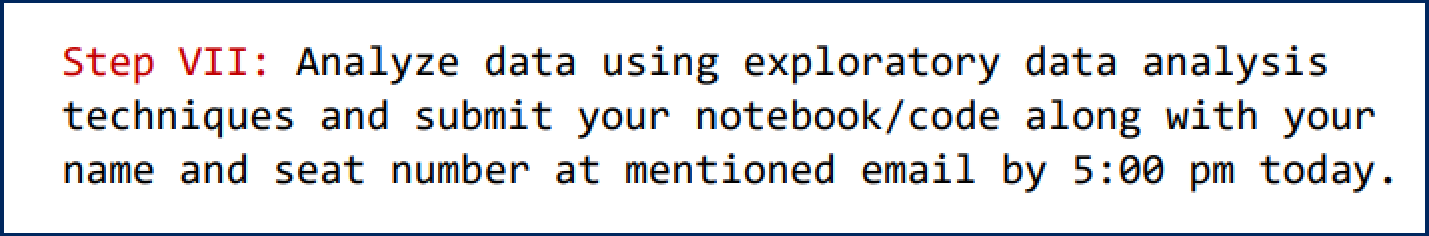
Which rows were only present in the first DataFrame?

Which rows were only present in the second DataFrame?

Which rows were present in both DataFrames?

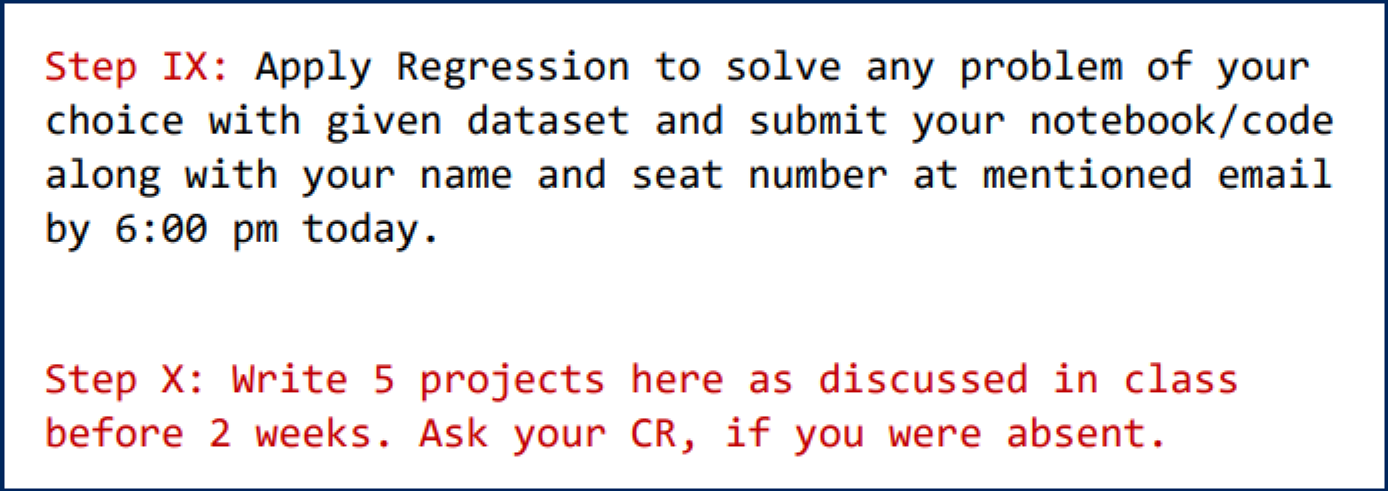
Which rows were not present in both DataFrames, but present in one of them?

Step VII



Step VII: Analyze data using exploratory data analysis techniques and submit your notebook/code along with your name and seat number at mentioned email by 5:00 pm today.

Step VIII, IX, X



Step IX: Apply Regression to solve any problem of your choice with given dataset and submit your notebook/code along with your name and seat number at mentioned email by 6:00 pm today.

Step X: Write 5 projects here as discussed in class before 2 weeks. Ask your CR, if you were absent.