Practical Data Science

Dr. Yongli Ren

(yongli.ren@rmit.edu.au)

Computer Science & IT

School of Science

Outline

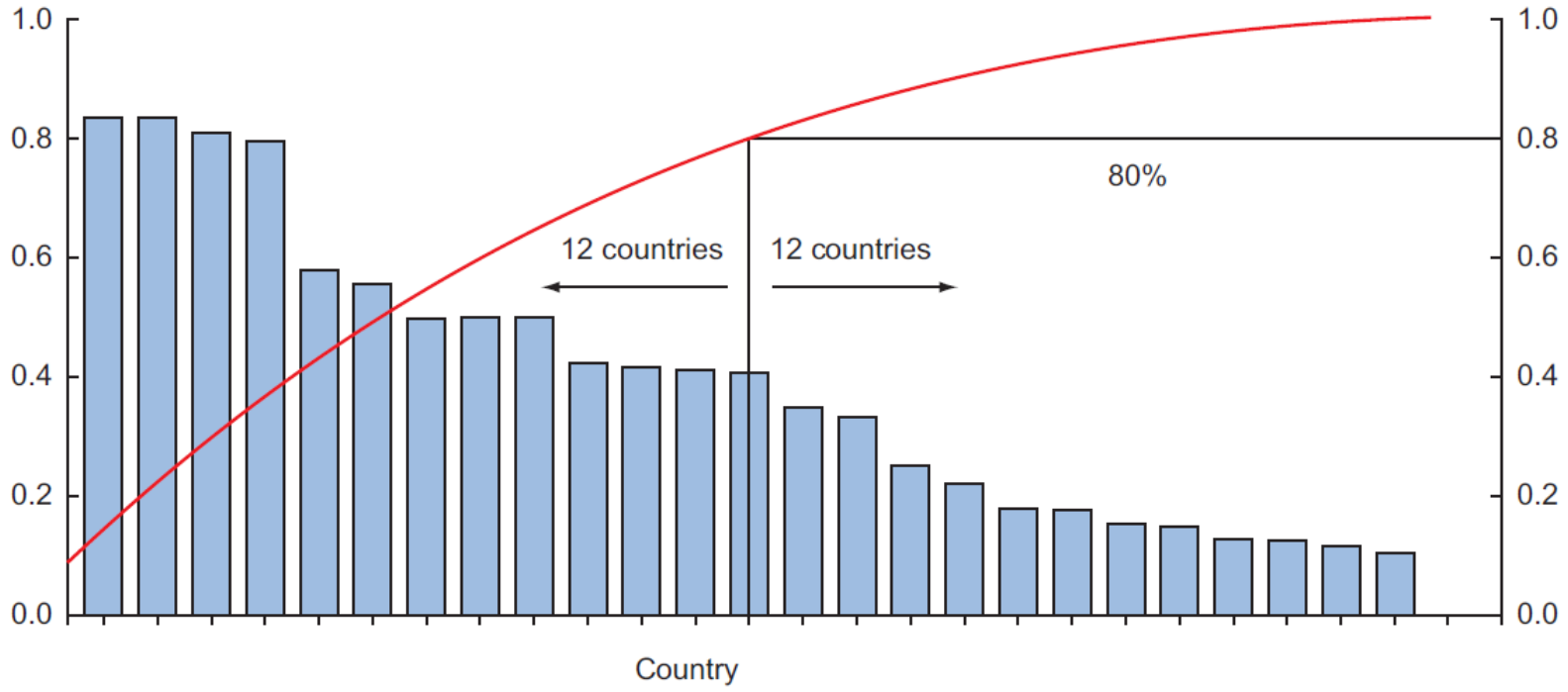
- Highlights
- Assignment 1
- Q/A

Data Summarisation

- During exploratory data analysis,
  - You take a deep dive into the data
  - Information becomes much easier to grasp when shown in a picture,
    - therefore you mainly use graphical techniques
      - to gain an understanding of
        - each feature and
        - the interactions between features.
  - This phase is about exploring data, so
    - keeping your *mind open* and *your eyes peeled*.
- The goal isn't to cleanse the data,
  - but it's common that you'll still discover anomalies you missed before, forcing you to take a step back and fix them.

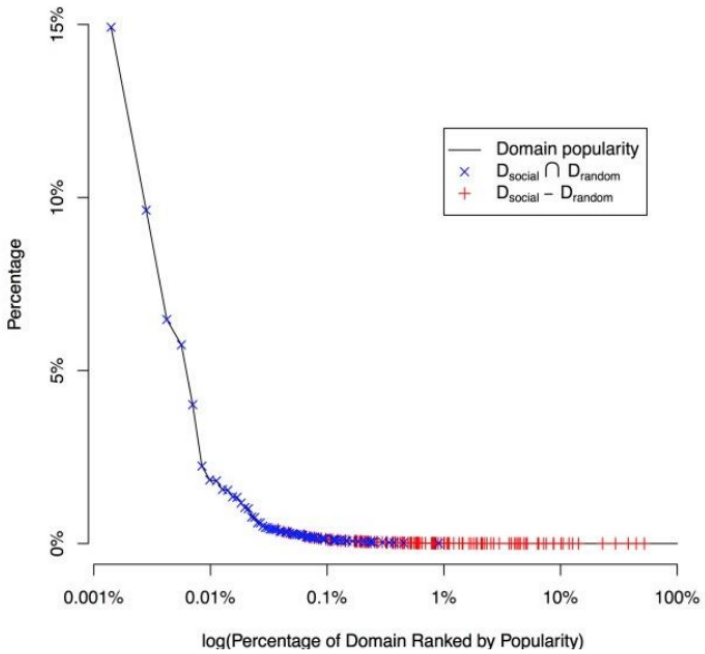
Data Summarisation

- Pareto Diagram
  - The values (Bar Chart) + Cumulative Distribution

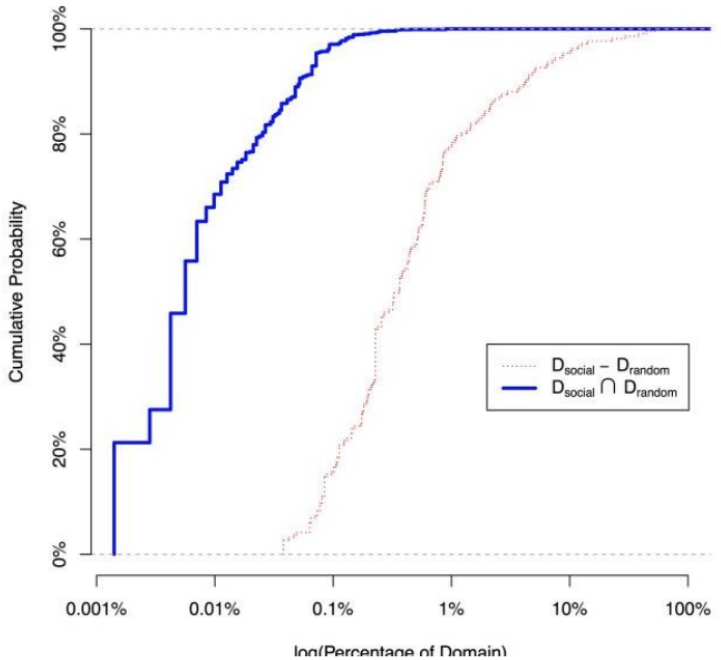


A Pareto diagram is a combination of the values and a cumulative distribution. It's easy to see from this diagram that the first 50% of the countries contain slightly less than 80% of the total amount. If this graph represented customer buying power and we sell expensive products, we probably don't need to spend our marketing budget in every country; we could start with the first 50%.

A Shopping Mall Project: Visualization (Summarisation)



(a) the log plot of domain popularity and the distribution of $D_{social} \cap D_{random}$ and $D_{social} - D_{random}$ over it.



(b) the empirical CDFs of $D_{social} \cap D_{random}$ and $D_{social} - D_{random}$

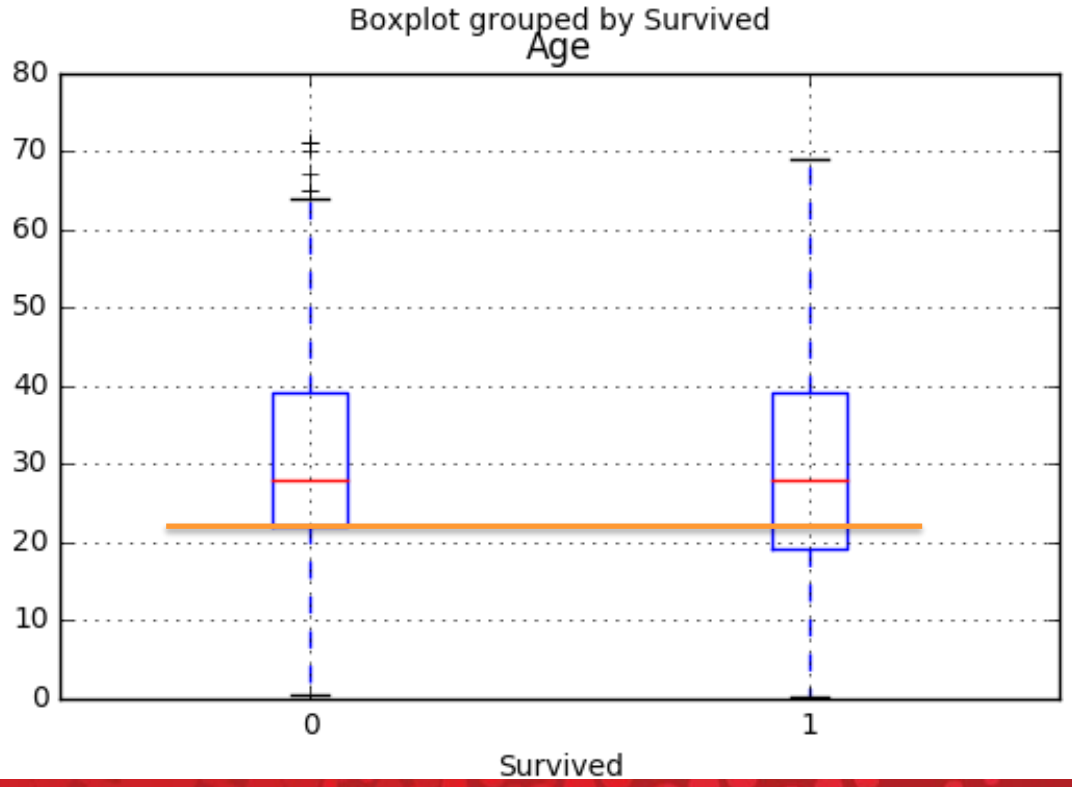
Figure 6. The domain popularity and the relationship between D_{social} and D_{random}

Thus, $D_{social} \cap D_{random}$ reflects the domains that are commonly accessed by an indoor user regardless of whether they are accompanied or not, and $D_{social} - D_{random}$ reflects the domains that are shared among accompanying users but not non-accompanying users. Finally, we obtain $|D_{social}| = 208$, $|D_{random}| = 88$, $|D_{social} \cap D_{random}| = 70$ and $|D_{social} - D_{random}| = 138$.

boxPlot
- by groups

- If there are groups in the data (from categorical variables),
  - just point out the variable for which you need the boxplot and specify that you need to have the data separated by the groups:

```
titanic.dropna().boxplot(column='Age', by = 'Survived')
```



Titanic Survival Rates by Sex
(two categorical variables: Sex vs Survival or not?)

Load packages and dataset

```
import pandas as pd
import matplotlib.pyplot as plt

titanic_filename = 'Titanic.csv'
titanic = pd.read_csv(titanic_filename, sep=',', decimal='.', index_col=0)

sex_counts = titanic.dropna()['Sex'].value_counts()

mask_sex_f = titanic['Sex'] == 'female'
mask_sex_m = titanic['Sex'] == 'male'

f_survive = titanic.dropna().loc[mask_sex_f, 'Survived'].value_counts()
m_survive = titanic.dropna().loc[mask_sex_m, 'Survived'].value_counts()

rate = [f_survive[1] / float(sex_counts['female']), m_survive[1] / float(sex_counts['male'])]
plt.bar(list(range(2)), rate, color='b', align='center')
plt.xticks(list(range(2)), ['female', 'male'])
plt.xlabel('Sex')
plt.ylabel('Propotion Survived')
plt.title('Titanic Survival Rates by Sex')
```

Check the number of female/male passengers

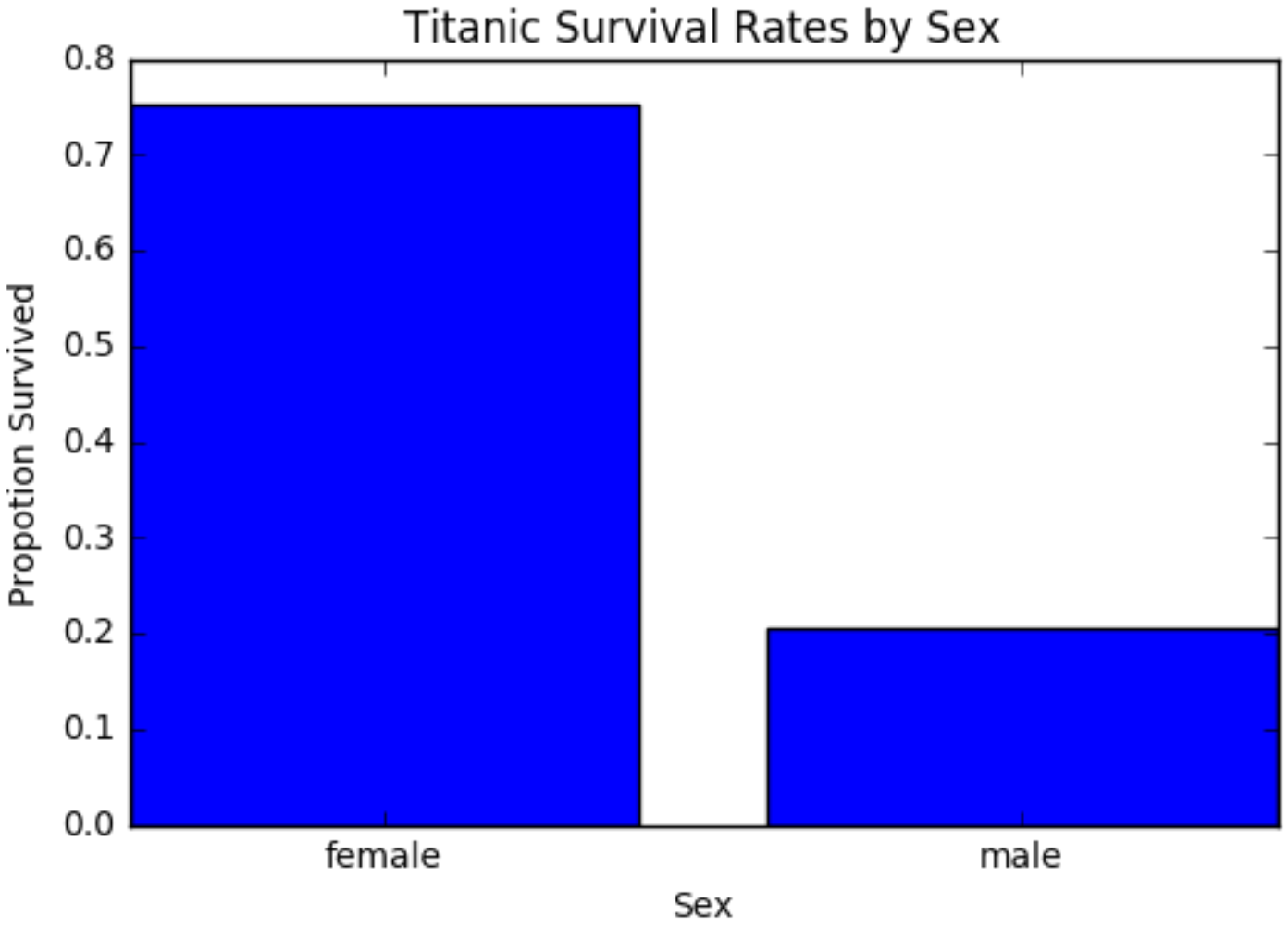
Create two masks for female and male

Use masks to select the survived female/male passengers

Calculate the survival rates for female and male

Plot the survival rates as bar graph
Specify the xticks, x/ylables and title of the graph

Titanic Survival Rates by Sex



Data Science

Thanks!