University of Lancashire

Data Science: Fundamentals

Karen O’Shea

Where opportunity creates success

Learning Objectives

University of Lancashire

Discuss data analysis design

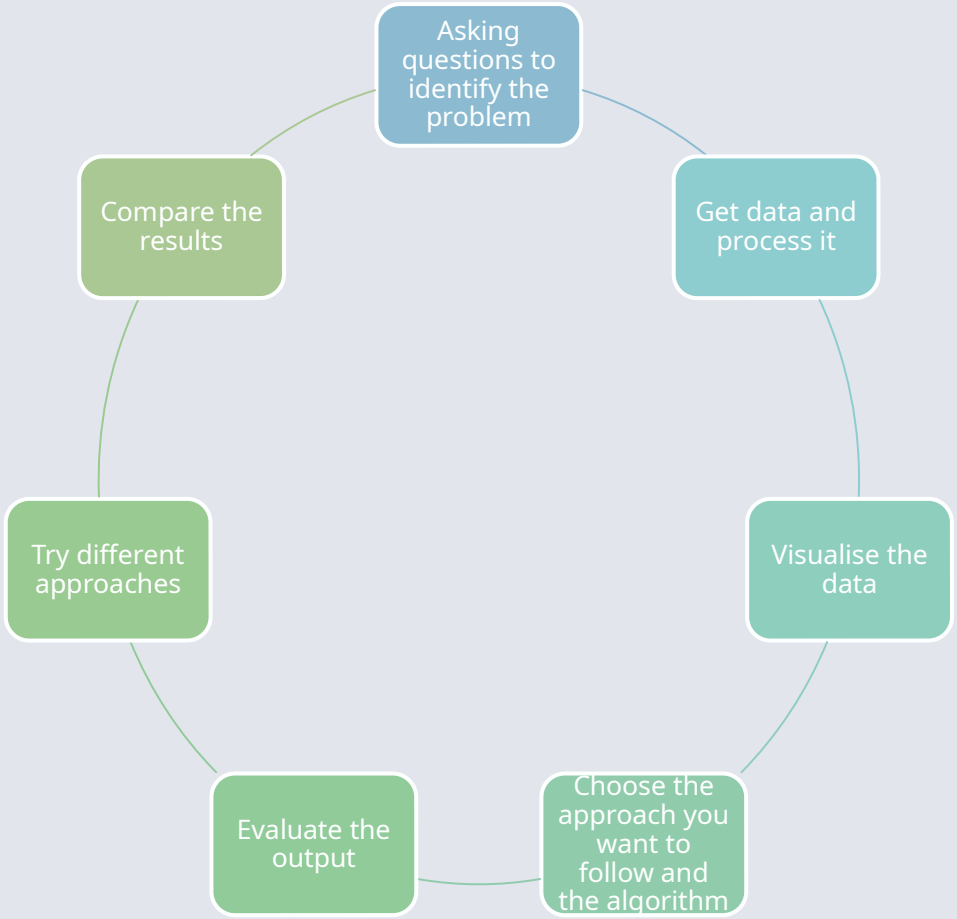
Evaluate tools for Data Science

Data Analysis

University of Lancashire

- Terms used interchangeably
- Asking intelligent questions
- Finding patterns
- Deriving insight
- Decision making
- Machine Learning...

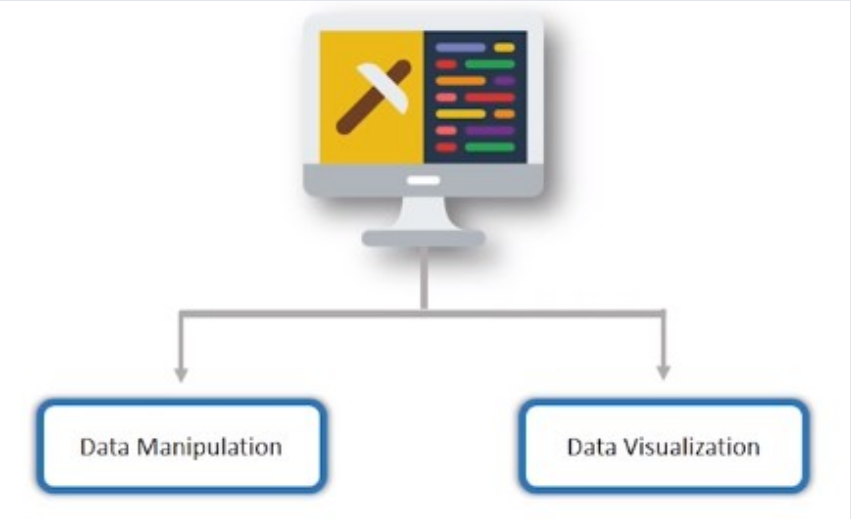
University of Lancashire



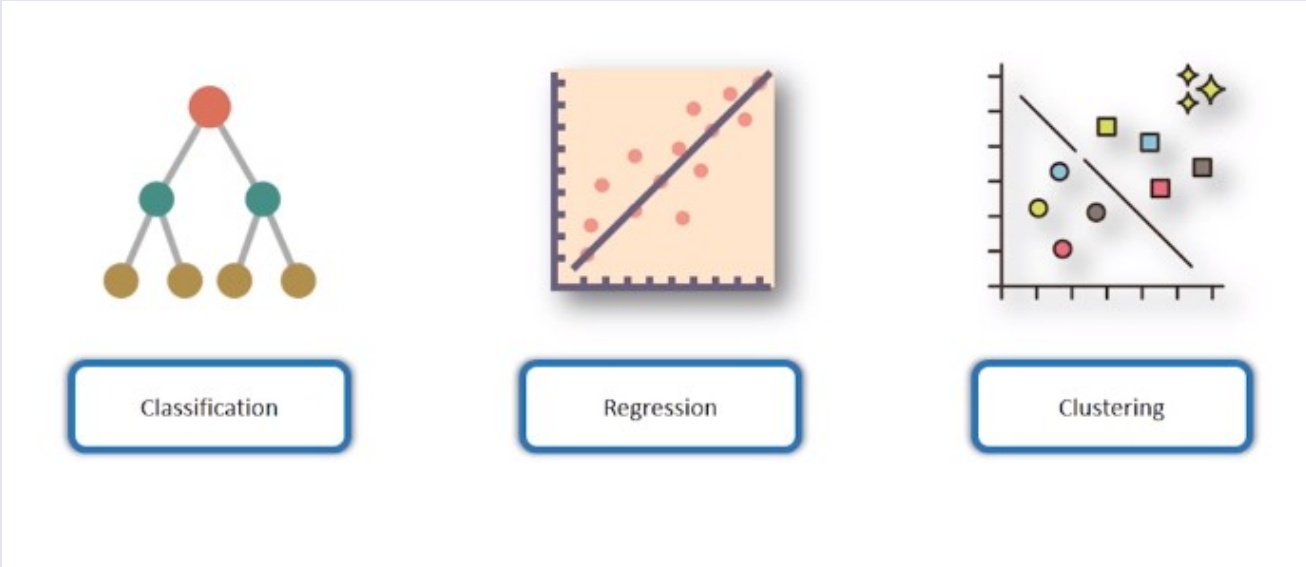
Data Science: Workflow Model

Data Science Workflow (1)

Data Science Workflow (2)



Data Science Workflow (3)



University of Lancashire

Python for Data Science

- Versatile platform
- Easy to learn
- Open-source
- Widely used language

Dataset examples (last week's lab)

- Dataset available on Blackboard (download from 'Kaggle')

https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset

- Examine dataset first. Task:
  - What questions could be asked of the data?

- We will use a Jupyter Notebook or Google Colab. Set up a folder to store dataset and notebook

University of Lancashire

Data Analysis Python Libraries

Many libraries:

- NumPy
- Matplotlib
- SciPy
- Scikit Learn
- Statmodels
- Seaborn
- Blaze
- Scrapy
- SymPy
- Bokeh
- **PANDAS**

Data Science using PANDAS

University of Lancashire

Data Analysis and Manipulation Tool

One of the most useful libraries

Extensive means of data analysis

Methods for data filtering

Fast, flexible, and user friendly

Data Science algorithms provide benefit to increase company profit, gain insight into business intelligence and enhance processing ability.

PANDAS – Data Structures

University of Lancashire

DataFrame and Series form the basic data model in Pandas:

- Series – one-dimensional indexed array. Easy to access elements of array
- DataFrame – similar to an Excel workbook. Column names (indexes) and row numbers.

Import Dataset and Libraries

University of Lancashire

```
import pandas as pd

df=pd.read_csv("dataset.csv")
```

*Reads dataset into DataFrame ready for viewing

Data Exploration

University of Lancashire

df.head(15)

* Will display 15 rows

df.describe()

* Will provide you with mean, quartile, count, min, max and standard deviation as well as their outputs.

df[‘Property_Area’].value_counts()

*Accessing a particular column in the data frame

```
df = pd.read_csv("LoanPredictionDataset.csv")
df.head()
```

	Loan_ID	Gender	Married	Dependents	Education	Sel
0	LP001002	Male	No	0	Graduate	
1	LP001003	Male	Yes	1	Graduate	
2	LP001005	Male	Yes	0	Graduate	
3	LP001006	Male	Yes	0	Not Graduate	
4	LP001008	Male	No	0	Graduate	

```
df.describe()
```

	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	614.000000	614.000000	592.000000	600.00000	564.000000
mean	5403.459283	1621.245798	146.412162	342.00000	0.842199
std	6109.041673	2926.248369	85.587325	65.12041	0.364878
min	150.000000	0.000000	9.000000	12.00000	0.000000
25%	2877.500000	0.000000	100.000000	360.00000	1.000000
50%	3812.500000	1188.500000	128.000000	360.00000	1.000000
75%	5795.000000	2297.250000	168.000000	360.00000	1.000000
max	81000.000000	41667.000000	700.000000	480.00000	1.000000

```
df['Property_Area'].value_counts()
```

Property_Area	count
Semiurban	233
Urban	202
Rural	179

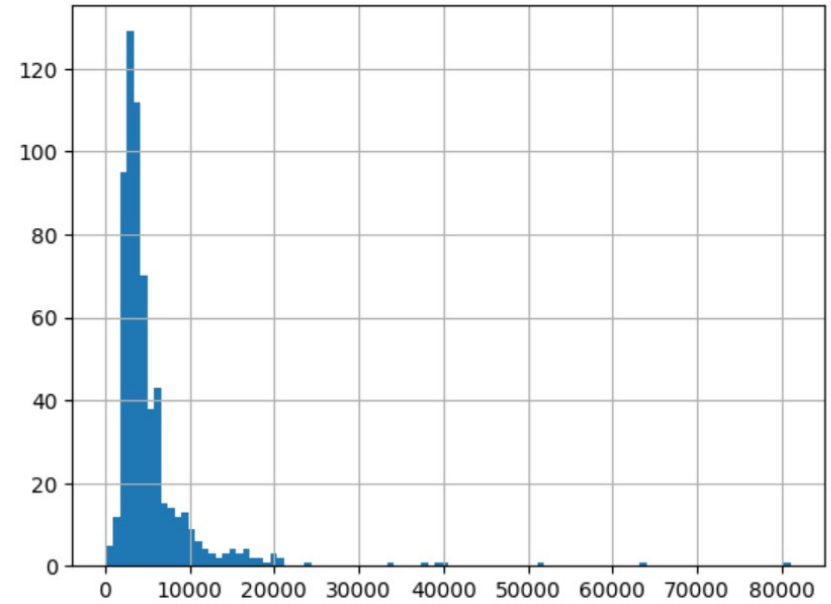
Distribution Analysis

df[‘ApplicantIncome’].hist(bins=100)

* Using a histogram to observe extreme values

```
df['ApplicantIncome'].hist(bins=100)
```

```
<Axes: >
```



Learning Objectives

University of Lancashire

Discuss data analysis design

Evaluate tools for Data Science