Practical Session 9

Foundations Spatial Data Science

Today Goals & Aims

Practical goals

- Introduction to transformation and dimensionality
- Work with census data using methods of transformation and dimensionality reduction
- Build the foundation for the topics of **Classification** and **Clustering**

Why are we doing this

- Understand the concepts and methods of transformation and dimensionality reduction is essential to analyse complex datasets.
- This is an **introduction to the topic** and a demonstration of how to work with these topics in python. *We don't expect you to know all the maths behind it and fully understand these topics straight away.*
- Considerate how you might incorporate these areas in your final assessment.

Today Goals & Aims

Term Calendar

	Weekly Topic		WORKSHOP Lead	WORKSHOP Date (Monday)	PRACTICAL Date Groups 1,2,3 (Tuesday)	PRACTICAL Date Groups 4,5,6 (Wednesday)
1	Getting Oriented	initiate	David, Nicolas	4 Oct	4 Oct	5 Oct
2	Foundations (Part 1)	initiate	Nicolas	11 Oct	11 Oct	12 Oct
3	Foundations (Part 2)	initiate	Nicolas	18 Oct	18 Oct	19 Oct
4	Objects & Classes	initiate	David	25 Oct	25 Oct	26 Oct
5	Numeric Data	engage	David	1 Nov	1 Nov	2 Nov
	Reading Week					
6	Spatial Data	engage	Nicolas	15 Nov	15 Nov	16 Nov
7	Textual Data	engage	Nicolas	22 Nov	22 Nov	23 Nov
8	Visualising Data	solve	David	29 Nov	29 Nov	30 Nov
9	Classifying Data	solve	David	6 Dec	6 Dec	7 Dec
10	Clustering Data	solve	Nicolas	13 Dec	13 Dec	14 Dec

Transformation & Dimensionality Reduction

Image source: https://realpython.com/python-statistics/

Today's Agenda

1. Work with the MSOA Atlas Excel (.xls) file in python to clean it and transform it into a .csv file.
   This will be useful when you need to work with geo-data (merge the MSOA Atlas with the MSOA geometries)
2. Train & Test Data Set - Why is it important to split your data to develop a model (machine learning algorithm)?
3. Data Normalisation & Standardisation - How to choose between methods?
4. Non-Linear Transformation (e.g Log-Normal, Exponential, Poisson)
5. PCA and t-sne

Read .xls File

- How to read multiple sheets or select one sheet of the .xls file?
- How to deal with multiple column headers - merge 3 into 1.
- How to remove specific character (e.g £,$,&,%) at the beginning, end of a string?
- How to drop specific columns, rows in a dataframe - think of axis.

Importance of Data Splitting

Given an Airbnb listings can I predict its price?

What is this problem in Machine Learning called?Is it a **supervised or unsupervised** learning problem?

Supervised: Pre-assigned labels are given to train the model e.g decision trees, linear regression, classification.
Unsupervised: No pre-assigned labels are provided. Model first self-discover any naturally occurring patterns in the data e.g clustering.

Importance of Data Splitting

Given an Airbnb listings can I predict its price?

Supervised Machine Learning - Model need to map the inputs (independent variables) to the given outputs (dependent variables).

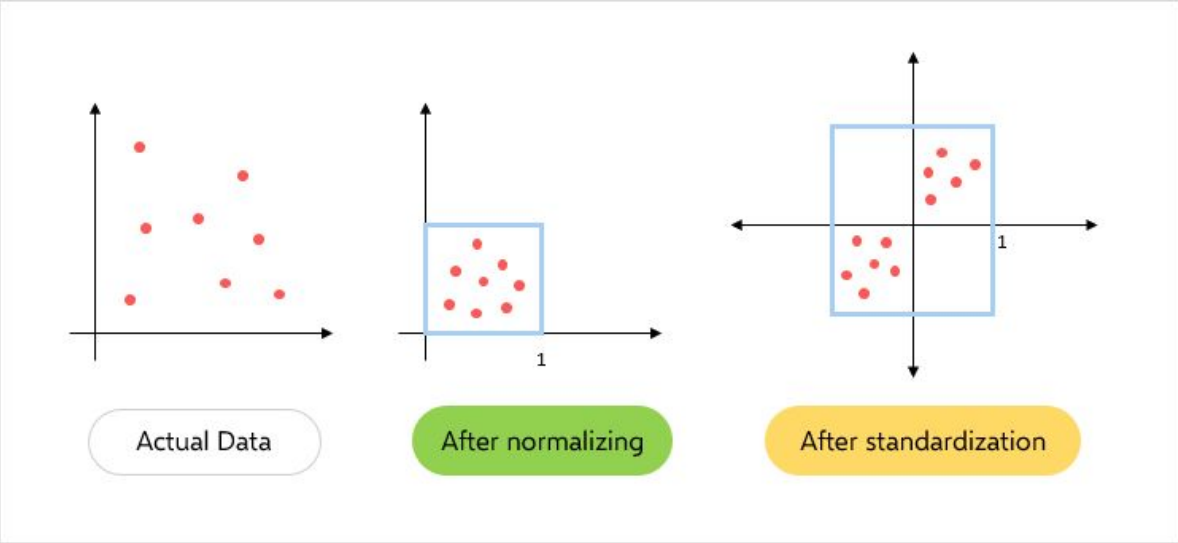
Unbiased Evaluation of the model. Training, Validation, Test Data

Training Set: Use to train “fit” the model. Find optimal parameters.
Validation: Use to evaluate hyperparameter.
Test Set: Use to test the model. Can I use it on unseen data?

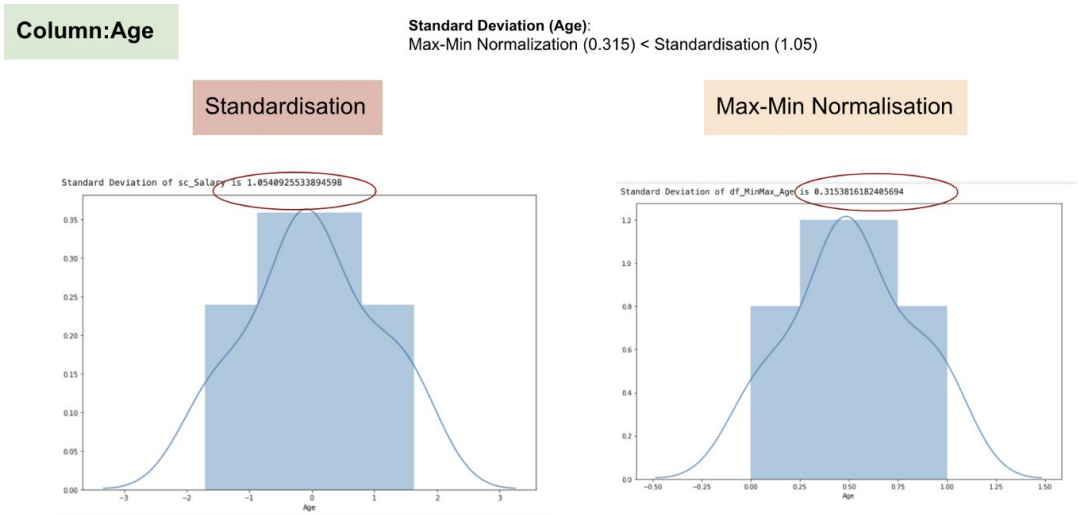
Normalisation & Standardisation

Normalisation: Rescale the values into a range of [0,1]

Standardisation: Rescale the data to have a mean of 0 and a standard deviation of 1.



Source: https://becominghuman.ai/what-does-feature-scaling-mean-when-to-normalize-data-and-when-to-standardize-data-c3de654405ed



Source: https://www.kdnuggets.com/2020/04/data-transformation-standardization-normalization.html

Next Week

Build on the work of this week - transforming data and dimensionality reduction - to explore clustering and classification techniques.

Week 10 READINGS

- Shapiro, W. and Yavuz, M. (2017) *Rethinking 'distance' in New York City*, Medium Article, URL
- Wolf, L. *et al.* (2020) *Quantitative geography III: Future challenges and challenging futures*, *Progress in Human Geography* DOI
- Arribas-Bel, D. and Singleton, A. (2019) *Geographic Data Science*, *Geographical Analysis* DOI

Time to practice !

Good Reads

- **12 Useful Things to know about Machine Learning** by James Le, Link: https://jameskle.com/writes/12-useful-things-about-ml
- **An end-to-end comprehensive guide for PCA** on Analytics Vidhya, *This article will give you a solid understanding of the mathematics of PCA.* Link: https://www.analyticsvidhya.com/blog/2020/12/an-end-to-end-comprehensive-guide-for-pca/
- **PCA using Python (scikit-learn)** by Michael Galarnyk on Medium. Link: https://medium.com/apprentice-journal/pca-application-in-machine-learning-4827c07a61db
- **Guide on t-sne. Implementation in R and Python** on Analytics Vidyha, Link:https://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/