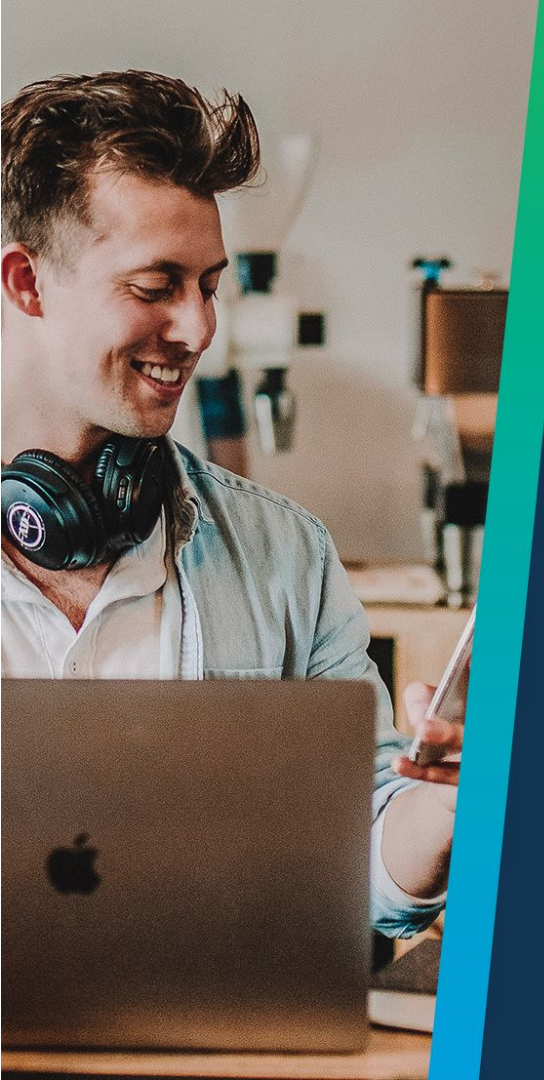


DS Workshop

Hyperiondev

Data Visualisation and Analysis

Welcome

Your Lecturer for this session

Sanana Mwanawina

Workshop – Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment - please engage accordingly.
- No question is daft or silly - **ask them!**
- There are Q/A sessions midway and at the end of the session, should you wish to ask any follow-up questions.
- You can also submit questions here: www.hyperiondev.com/sbc4-ds-questions
- For all non-academic questions, please submit a query: www.hyperiondev.com/support
- Report a safeguarding incident: hyperiondev.com/safeguardreporting
- We would love your feedback on lectures and workshops: https://hyperionde.wufoo.com/forms/zsgv4m40ui4i0g/

GitHub repo

Go to: github.com/HyperionDevBootcamps

Then click on the "**C4_DS_lecture_examples**" repository, do view or download the code.

Objectives

1. Provide an overview of what we have covered
2. Understand the difference between standardisation and normalisation

Types of variables

1. Categorical variables: cannot be quantified

   - Nominal: no natural order e.g., mode of transport to work
   - Ordinal: there is some relation e.g., food ratings can be “good” or “average” we know good is better than average, but you cannot quantify by how much

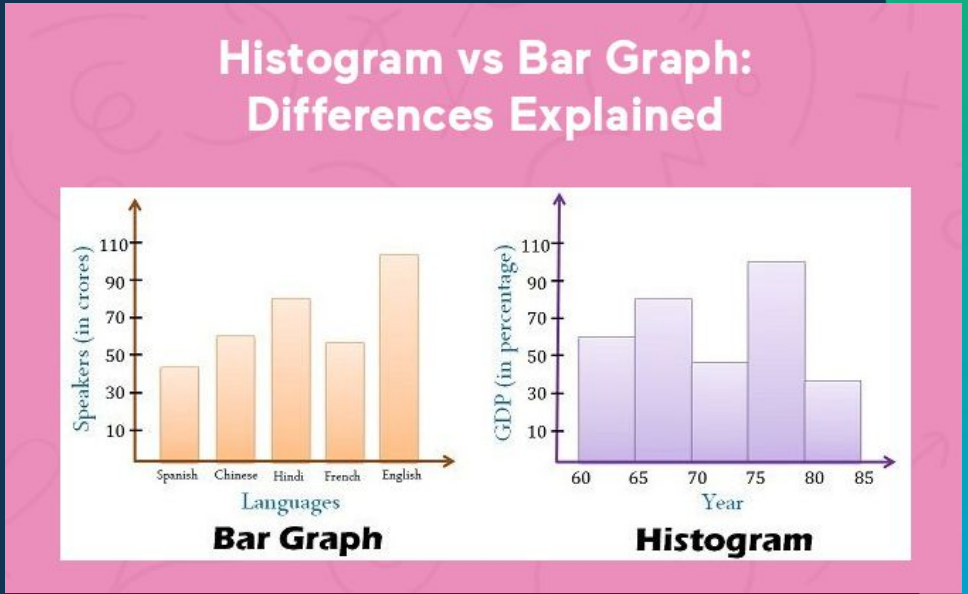
2. Numeric variables: can be quantified

   - Continuous: can assume an infinite number of real values within a given interval e.g., height
   - Discrete: takes on a distinct, countable values e.g., number of cars that drive into a car wash on a given day

Types of visualisations

Bar graphs vs Histograms

1. Histograms have bars that show the number of observations that fall within a certain range.
2. In a bar graph, the bars represent a category and the height shows the number of observations in that category
3. The x-axis in a histogram is always numeric



Types of visualisations

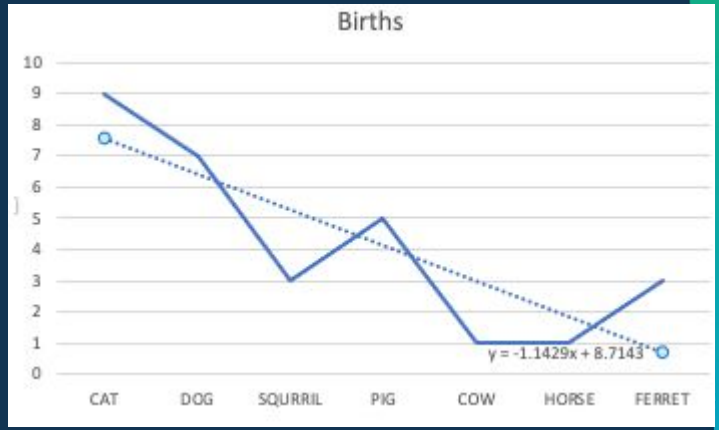
Line graphs

- Great for identifying trends, especially in a time series analysis
- A typical line graph will have continuous data along both the x-axis and y-axis

Types of visualisations

Line graphs

- Variables need to make sense
- On the right, we have a categorical x-axis. We can produce a trend line quite alright, but it does not make any sense we trying to interpret it.



Working with datasets

1. Importing datasets

   Using the pandas module to open .txt and .csv files

```
import pandas as pd

df = pd.read_csv(r'balance.txt', sep=' ')
```

2. We saw that there are many ways to manipulate datasets:
   - .head() and .tail() show us the first and last 5 observations in our datasets
   - .columns gives us the column names
   - sort_values() enables us to arrange observations in a well ordered manner
   - select a range of column and rows for viewing

```
df.iloc[5:10,[1,7]]
```

Data cleaning

1. Dropping columns. Necessary when our dataset contains column we do not need
2. Replacing values
3. Grouping data e.g., by ethnicity
4. Inconsistent data entries: fuzzywuzzy

Working with missing data

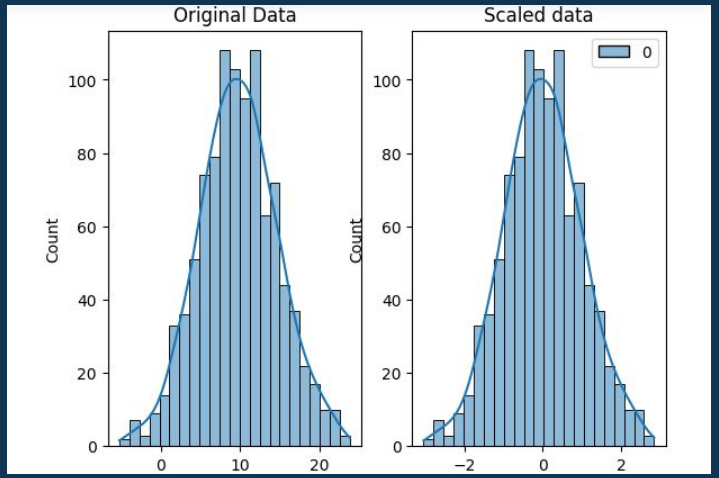
Missing values can be dealt with in many ways:

1. Drop the missing values
2. Replace with a zero
3. Replace with the value of the observation right before or right after - works well when observations are sorted or follow some kind of logical order
4. Imputation: mean, median, mode, linear regression and k-nearest meighbour

Standardisation and Normalisation

The goal is to change the values of numeric columns in our dataset to a common scale.

1. Standardisation: makes our features center around zero and have a standard deviation of one.



Standardisation and Normalisation

Scenario: you are working with a dataset that contains information about customer transactions for an e-commerce business. The dataset has the following variables:

1. Customer age in years
2. Amount spent on purchase in pounds
3. Number of items purchased
4. Time spent on website in minutes
5. Customer satisfaction rating on a scale of 1 to 5

This dataset contains variables measured in different units or sclaes. To ensure meaningful comparisons, we will standardise this dataset.

Standardisation and Normalisation

Simple example:

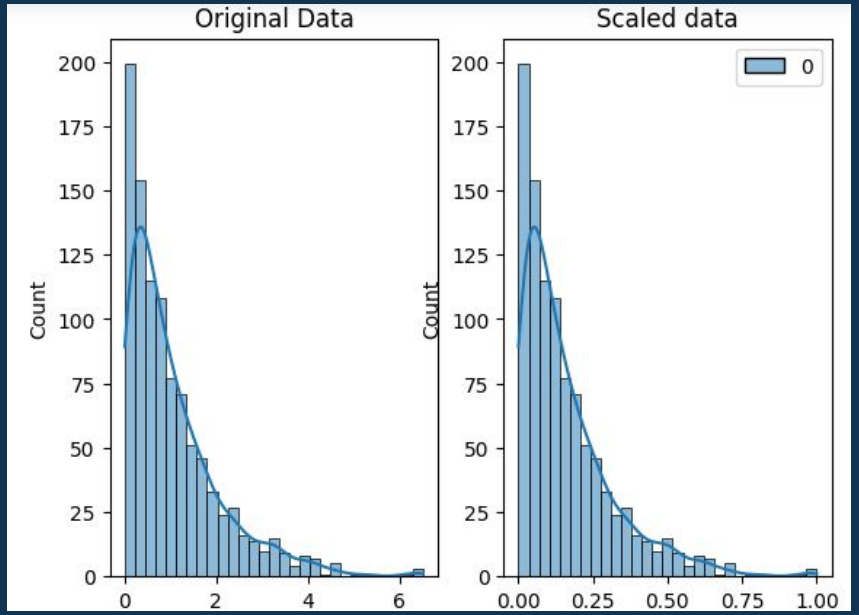
We want to see how time spent on the website measured in minutes (x_1) and the age of the customer measured in years (x_2) affect how much they spend on their purchase (y).

We develop two linear models to try to predict the amount spent given these variables.

Model 1: $y = 3\ x_1$
Model 2: $y = 5\ x_2$

Standardisation and Normalisation

2. Normalisation ensures our observations all lie between 0 and 1



Standardisation and Normalisation

Scenario: you are analysing data for a company that provides electricity. The dataset contains the following variables:

1. Household size ranging from 1 to 10 people
2. Monthly electricity consumption in kilowatt-hours ranging
3. Annual income ranging from 20,000 to 100,000
4. Distance from the city center in kilometers

Standardisation and Normalisation

When to use standardisation?

- Standardisation assumes observations follow a Guassian distribution (normal distribution).
- This does not strictly have to be true in order to use standardisation, but it is a more effective technique if your attribute distribution is Gaussian.

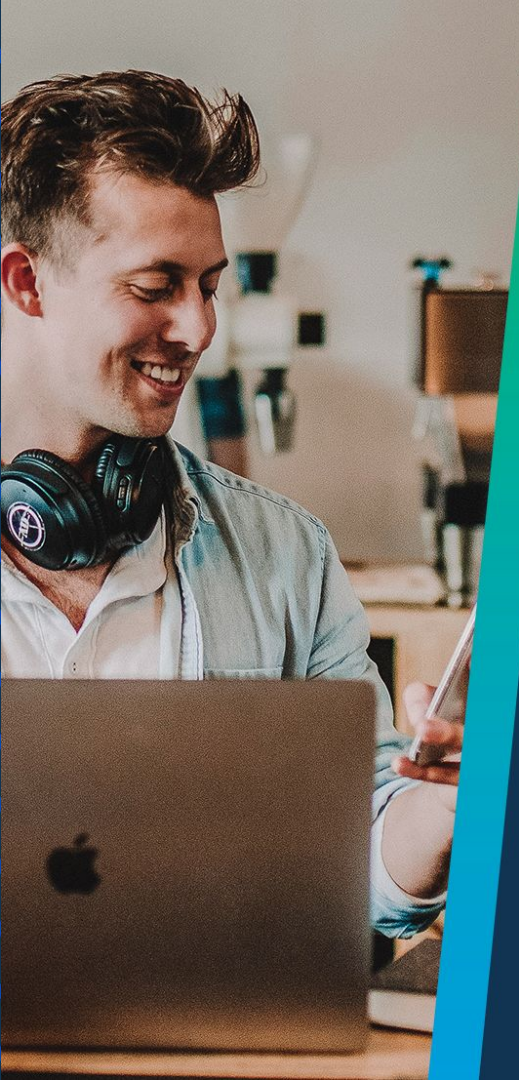
When to use normalisation?

- When you do not know the distribution of your data or when your observations are clearly not Guassian distributed

Hyperiondev

Q & A Section

Please use this time to ask any questions relating to the topic explained, should you have any



Hyperiondev

Thank you for joining us