Inf2 – Foundations of Data Science
Week 1: Welcome and logistics

THE UNIVERSITY of EDINBURGH
informatics

FOUNDATIONS OF DATA SCIENCE

The Team

Kobi Gal

David Sterratt

Anna Hadjitofi

History of data science

- Long time ago (thousands of years) science was empirical and people counted stars or crops

Some slides taken from CS109A at Harvard University

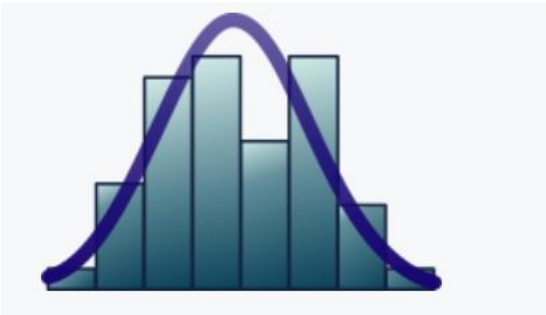
A few hundred years ago…

- Theoretical approaches, try to derive equations to describe general phenomena.

$$\mathbf{F} = \frac{d}{dt}(m\mathbf{v})$$

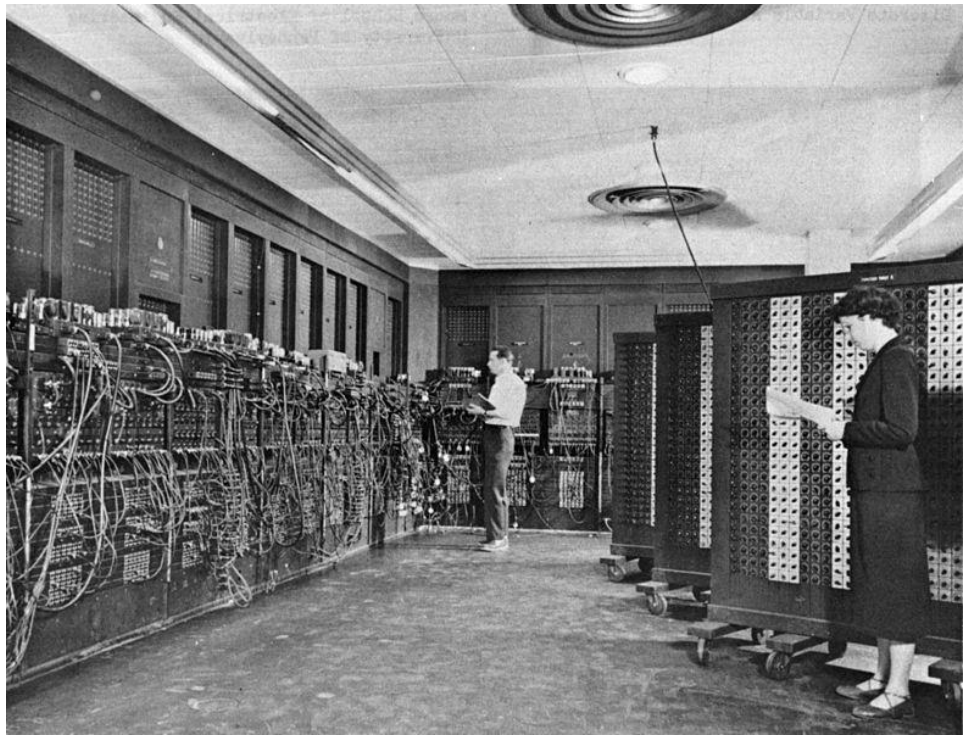
Second law of motion

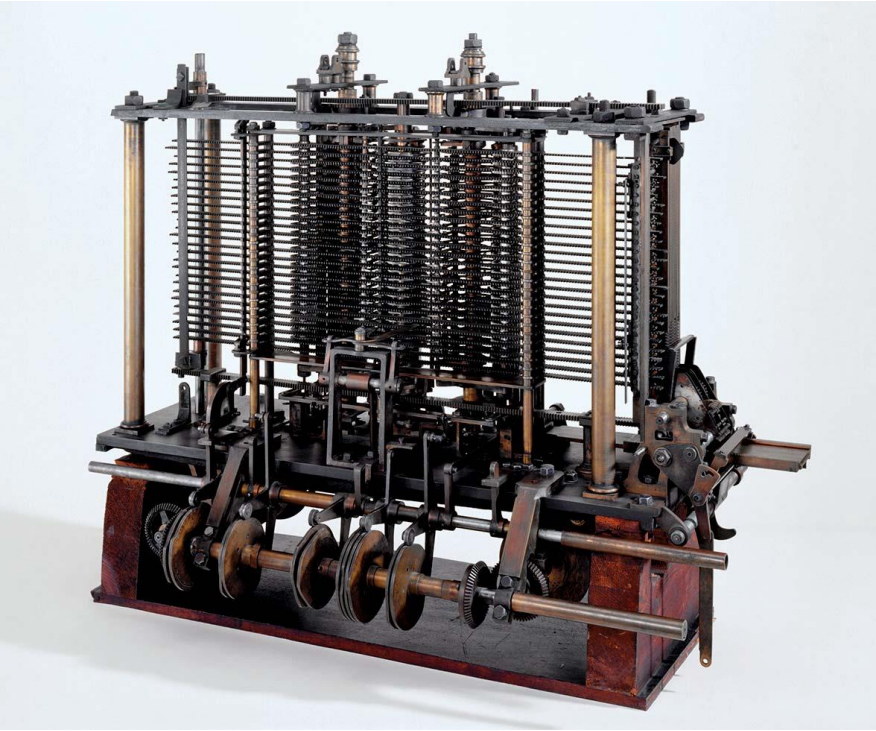
- Statistics and probability.



A hundred years ago…

- Computational approaches





And now, Data Science

Gaining insights into data through computation, statistics, and visualization

Week 1 Plan

- Introduction to Foundation to Data Science
  - **What?**
  - Why?
  - How?
- Course logistics

A Data Scientist Is…

“Data Scientist = statistician + programmer + storyteller + artist”

- Shlomo Aragmon

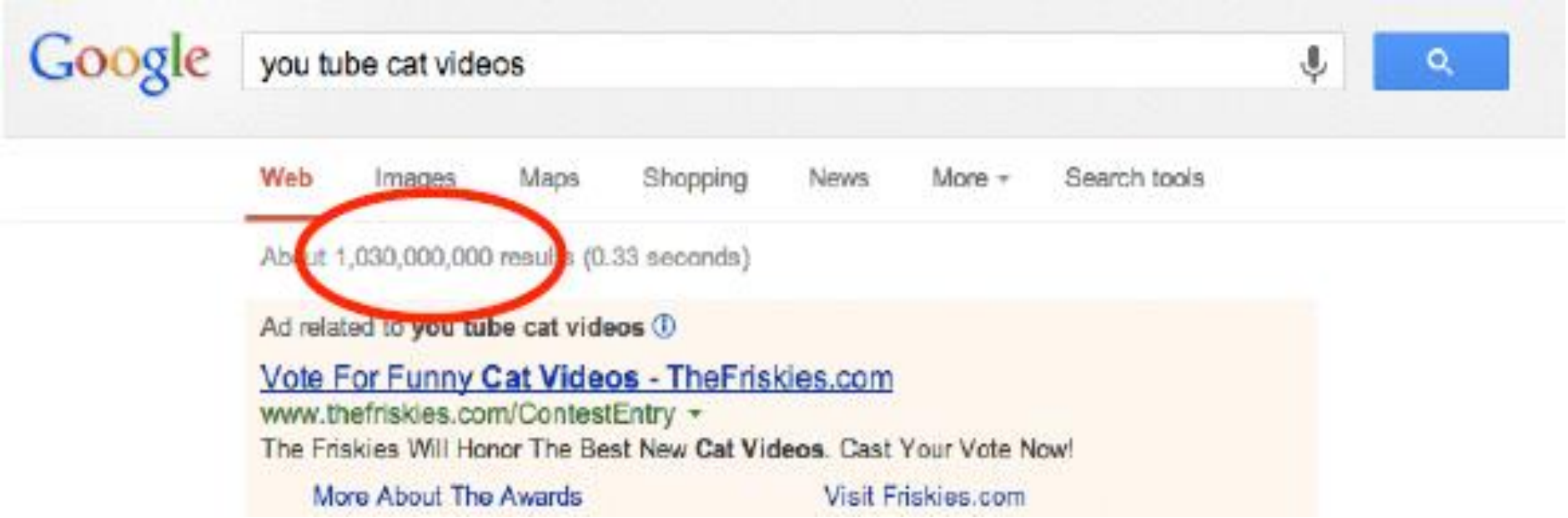
Plan for today

- Introduction to data analysis and data science
  - What?
  - **Why?**
  - How?
- Course logistics

DATA, Lot’s of data

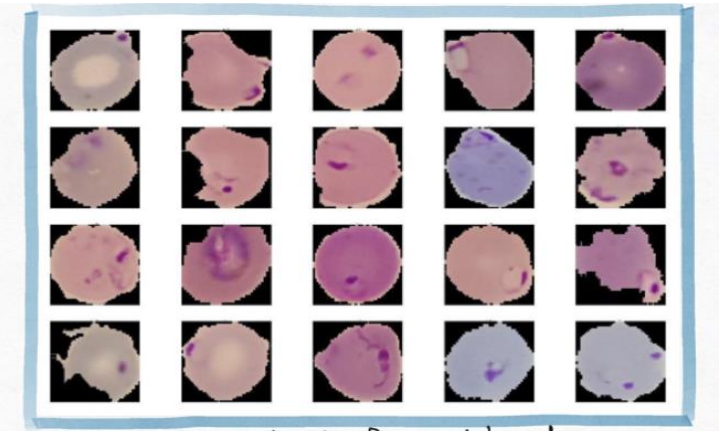
“Between the dawn of civilization and 2003, we only created five exabytes of information; now we’re creating that amount every two days.”

---Eric Schmidt, Google (and others)



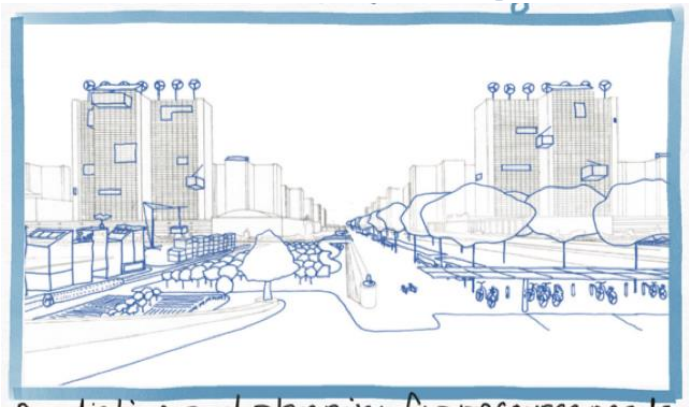
The Potential of Data Science

Disease Diagnosis



Detecting malaria from blood smears

Urban Planning



Predicting and planning for resource needs

Drug Discovery

Quickly discovering new drugs for COVID

Agriculture

Precision agriculture

The Dangers of Data Science

Gender Bias

Racial Bias

50 Best Jobs in America for 2020

	Job Title	Median Base Salary	Job Satisfaction	Job Openings	
#1	Front End Engineer	$105,240	3.9/5	13,122	View Jobs
#2	Java Developer	$83,589	3.9/5	16,136	View Jobs
#3	Data Scientist	$107,801	4.0/5	6,542	View Jobs
#4	Product Manager	$117,713	3.8/5	12,173	View Jobs
#5	DevOps Engineer	$107,310	3.9/5	6,603	View Jobs
#6	Data Engineer	$102,472	3.9/5	6,941	View Jobs
#7	Software Engineer	$105,563	3.6/5	50,438	View Jobs

The ability to take **data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate.** It is going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and **ubiquitous data**."

- Hal Varian,
chief economist, google

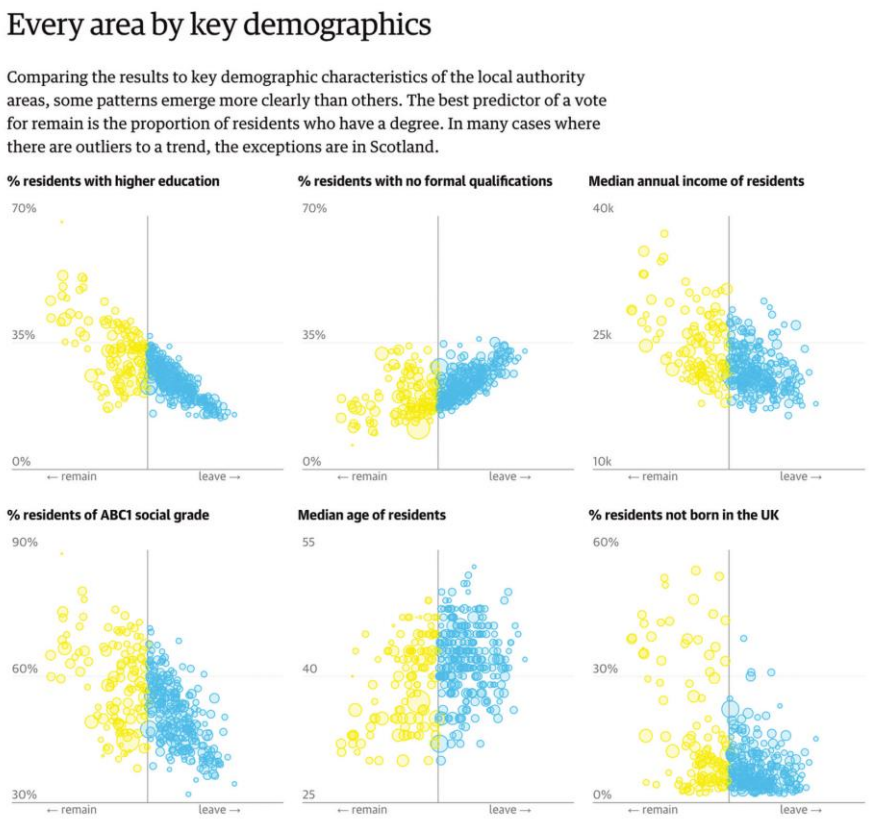
Learning outcomes

By the end of this course you should be able to...

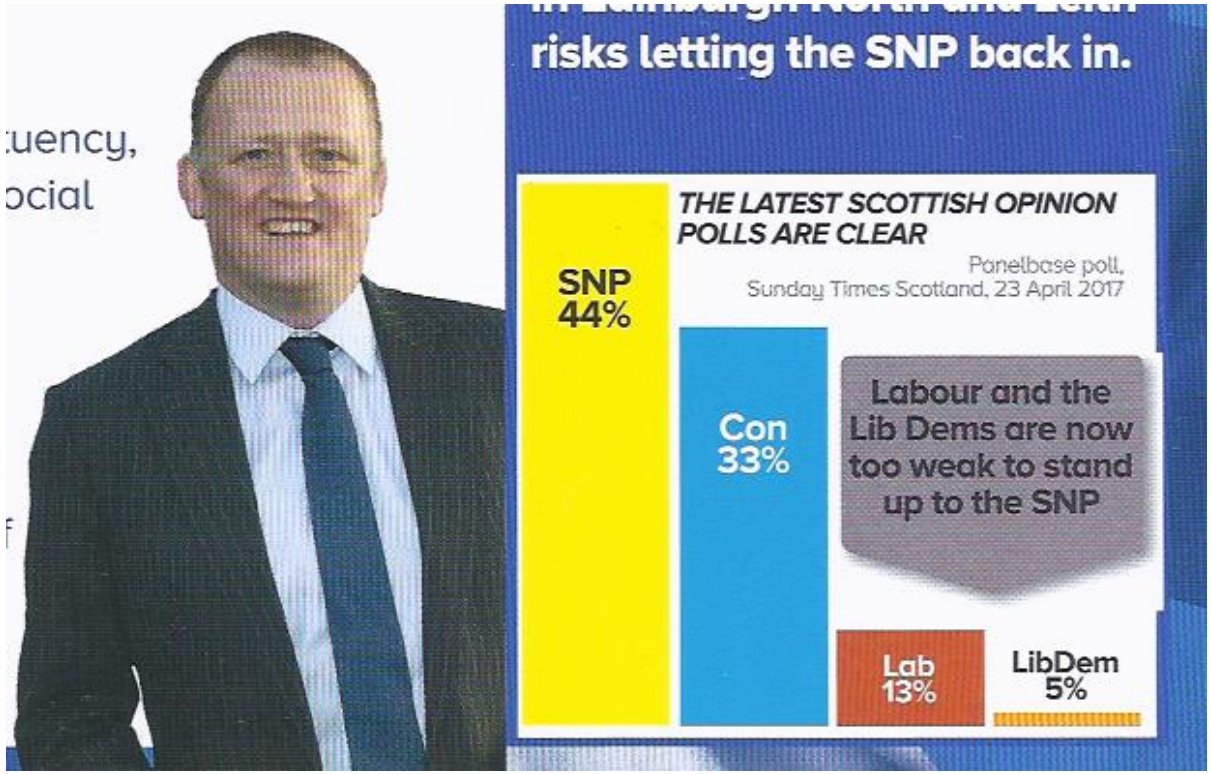
Intended Learning Outcome:

1. Describe and apply good practices for storing, manipulating, summarising, and visualising data.
2. Use standard packages and tools for data analysis and describing this analysis, such as Python and LaTeX.
3. Apply basic techniques from descriptive and inferential statistics and machine learning; interpret and describe the output from such analyses.
4. Critically evaluate data-driven methods and claims from case studies, in order to identify and discuss a) potential ethical issues and b) the extent to which stated conclusions are warranted given evidence provided.
5. Complete a data science project and write a report describing the question, methods, and results.

Describe and apply good practices for storing, manipulating, summarising, and visualising data.



The Guardian



Election leaflet

Use standard packages and tools for data analysis and describing this analysis, such as Python and LaTeX

seaborn

statsmodels

The LaTeX Project

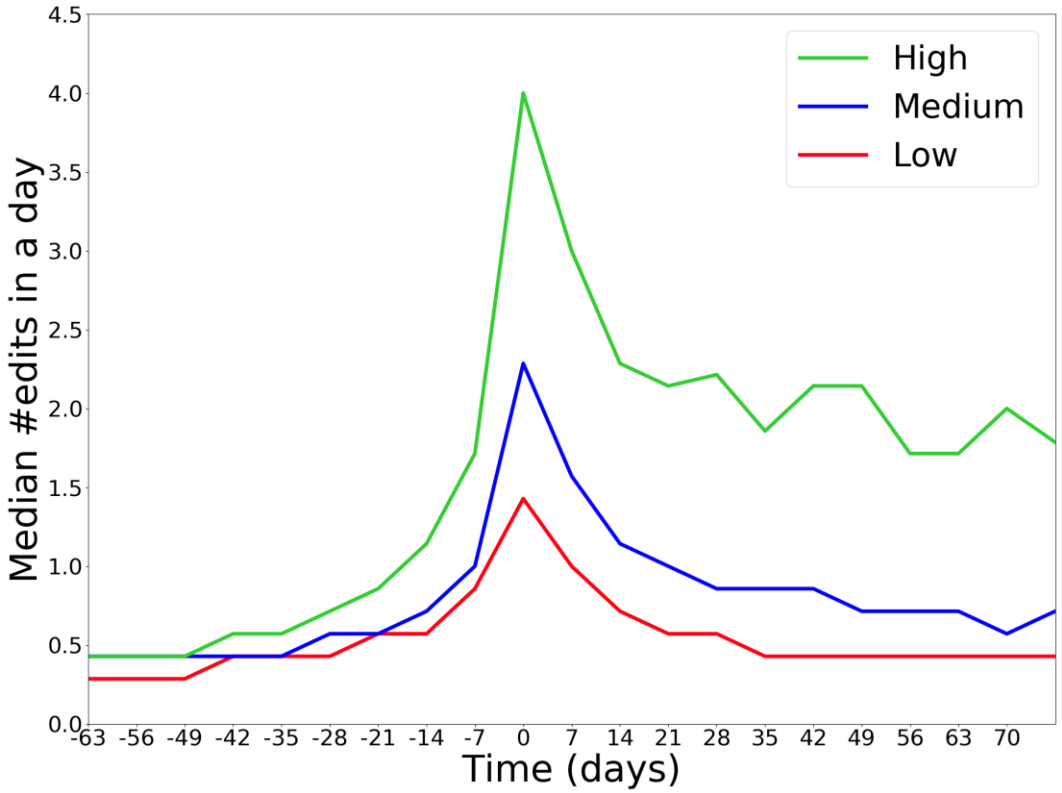
matplotlib

pandas

python

jupyter

Apply basic techniques from descriptive and inferential statistics and machine learning; interpret and describe the output from such analyses.



Badge Behavior in SO

Critically evaluate data-driven methods and claims from case studies, in order to identify and discuss a) potential ethical issues and b) the extent to which stated conclusions are warranted given evidence provided.

Mail Online — Science & Tech

Home | News | U.S. | Sport | TV&Showbiz | Australia | Femail | Health | Science | Money | Video | Travel | Best Buys | Discounts

Latest Headlines | NASA | Apple | Twitter — Login

Site Web Enter your search Search

Teenagers who dislike their own body are THREE TIMES more likely to be depressed as adults, study finds

- UK researchers measured body dissatisfaction at 14 and depression scores at 18
- Their sample was of boys and girls born in the west of England between 1991-92
- Boys were found to be more likely to experience severe depression than the girls
- Social media could account for young males experiencing body dissatisfaction

Original research

Body dissatisfaction predicts the onset of depression among adolescent females and males: a prospective study

Anna Bornioli, Helena Lewis-Smith, Amy Slater, Isabelle Bray

► Supplemental material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/jech-2019-213033).

University of the West of England Bristol, Bristol, UK

Correspondence to Erasmus University, Rotterdam, The Netherlands; bornioli@ese.eur.nl

ABSTRACT

Rationale Body dissatisfaction is prevalent in mid-adolescence and may be associated with the onset of depression.

Objective The study assessed the influence of body dissatisfaction on the occurrence of later depressive episodes in a population-based sample of British adolescents.

Method Participants were 2078 females and 1675 males from the Avon Longitudinal Study of Parents and Children (ALSPAC) cohort. Logistic regression was used to test if body dissatisfaction at 14 years old predicted the

extent to which a person cognitively 'buys into' socially determined ideals of beauty) and appearance comparisons (ie, the extent to which a person compares their own appearance with that of others). This model has received substantial support[7 8]; and scholars have also highlighted the prominent impact of the media in body dissatisfaction processes.[9 10] With regard to changes in body image across appearance; there is no clear consensus on how body image changes within adolescence. Wertheim and Paxton[11] state that among female adolescents, once body dissatisfaction is established, it 'does not

J Epidemiol Community Health: first published as 10.1136/jech-201

Complete a data science project and write a report describing the question, methods, and results

The Impact of Lockdown on Bicycle Usage in Edinburgh

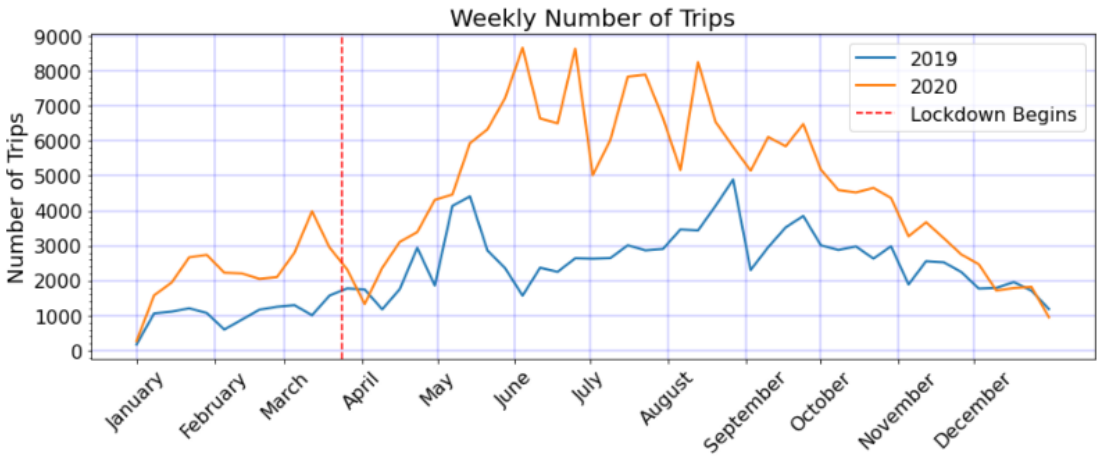
6th April 2021

1 Overview

Governments across the world have introduced restrictive measures, such as enforcing "lockdown" periods, to reduce social interaction and halt the spread of the COVID-19 virus. To analyse the impact of these lockdown periods on bicycle usage in Edinburgh, data from the shared bicycle hire service Just Eats Cycles was used. A predictive model was used to measure the change in bicycle use and test whether a lockdown period, where home confinement was required, implied a decrease in the number of trips throughout 2020. The use of descriptive and inferential statistics, such as bootstrap, allowed evaluation

4 Exploration and analysis

4.1 Effect of Covid-19 and Lockdown on Edinburgh Bicycle Usage



Timeline

- S1, weeks 1-8: Ethics, data wrangling, visualisation, linear models
- S1, week 9: Intro to supervised machine learning
- S1, week 10, 11: Statistical inference (after DMP has introduced probability)
- S2, weeks 1, 2: Statistical inference (continued)
- S2, week 3: Intro to unsupervised learning
- S2, weeks 4-6: Ethical issues, software engineering
- S2, weeks 7-11: Project
- See "Schedule" in Learn for detailed schedule
    - We are a bit behind due to Monday's holiday, but hope to catch up

Course Logistics

Each week…

- 2 Lectures
  - 3 out of 4 lectures will contain new information. The 4th lecture should be Q&A and demos.
  - Lecture notes available.
- Comprehension questions – do discuss in Piazza
- Practical labs on campus
- Workshops on campus

Labs

- (Almost) every week
- 3 on-campus labs, each 2 hours long
- Assigned to groups, but if you need to, go drop-in to another lab
    - (see "Course Information" Learn for times)
- We recommend **pair programming** to get the most out of the labs
- No software installation required – you can use the **Noteable** service

Foundation Data Sciences

Week 02: Introduction to Jupyter Notebooks and Pandas

Learning outcomes: In this lab you will learn the very basics of the python library pandas, which is used for data management. By the end of the lab you should be able to:

- use jupyter notebook,
- load different data file types,
- display data,
- filter your data for specific values, and
- apply basic statistical computations on the data.

Prerequisites

- Basic knowledge of `python` is assumed for this course. If you haven't used Python before or need a refresher, we can recommend the following python tutorial as a starting point.
- Basic knowledge of `numpy` is assumed for this course. If you haven't used numpy before or need a refresher, we can recommend the following numpy tutorial.

We will try to cover a different research question every week. This week we will take the position of a historian and try to answer the following question.

Research question: Which passenger group had the worst survival rate on the *Titanic*?

Tasks and workshops

- "Conversation makes you smart" - Jon Oberlander
- Discussion and problem solving on topics such as
    - Ethics of data
    - Visualisation
- Designed to link to coursework
- You will be assigned to groups in Week 1 or Week 2
- Change groups using the Group Change Request Form

Summative assessment

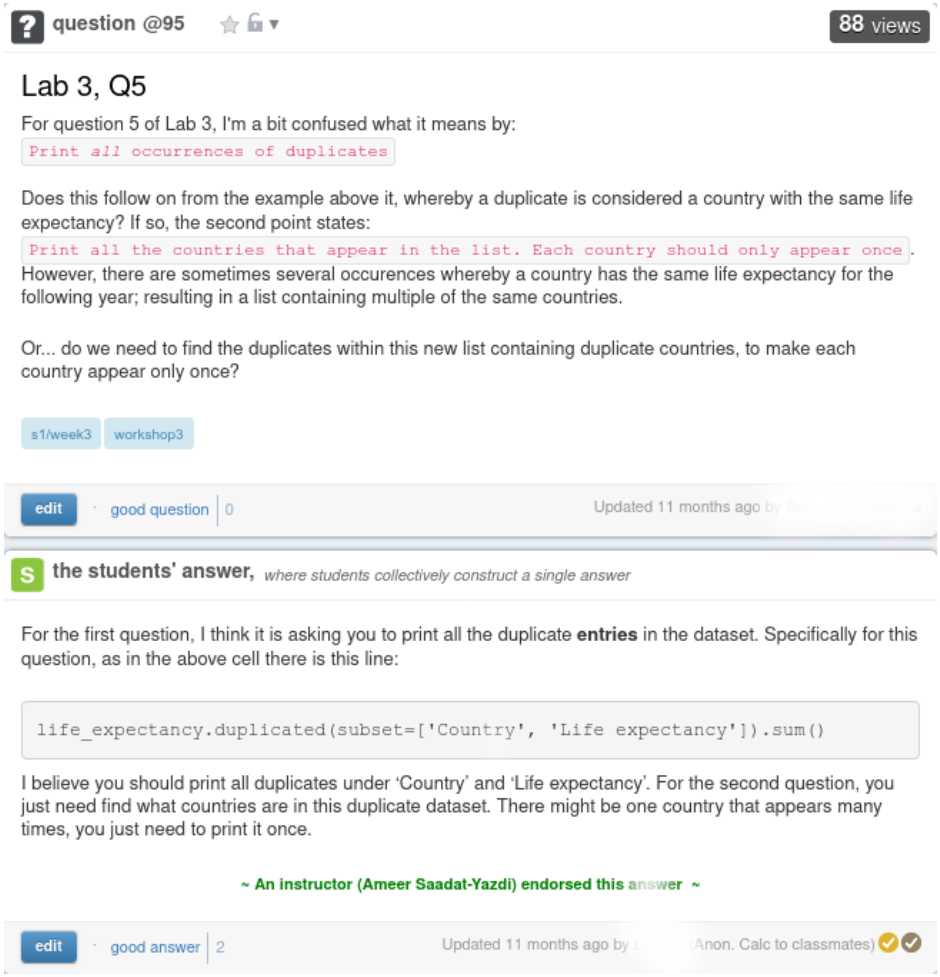
- CW1: Data-wrangling and visualization Exercise (20%), released S1 Week 5.
- CW2: Critical evaluation (20%) S2 Week 1
- Group Project (40%), released S2 Week 7
- CT: Class test (20%), S2 revision period
- All dates and late rules in "Assessment" in Learn
- We have tried to coordinate deadlines with other Year 2 courses

Feedback on your progress

- Comprehension questions on lecture videos
- Piazza
- Mock class test
- Feedback on coursework
- Solutions to workshop exercises
- Workshop session during project period

Piazza

- "Conversation makes you smart"…
- … and let's make it pleasant as well as helpful
    - Please check out the "standards of conduct" and "posting guidelines" in the Welcome post
- Please **do** try to answer each others' questions – you should all benefit
- Anna, David and Kobi will be watching.
- Depending on the question we may give you a bit of time to try answering before we jump in.



Getting help

- Ask your peers on Piazza
- Ask private questions via Piazza
- Ask questions in the lectures

Student advice from 2020/21

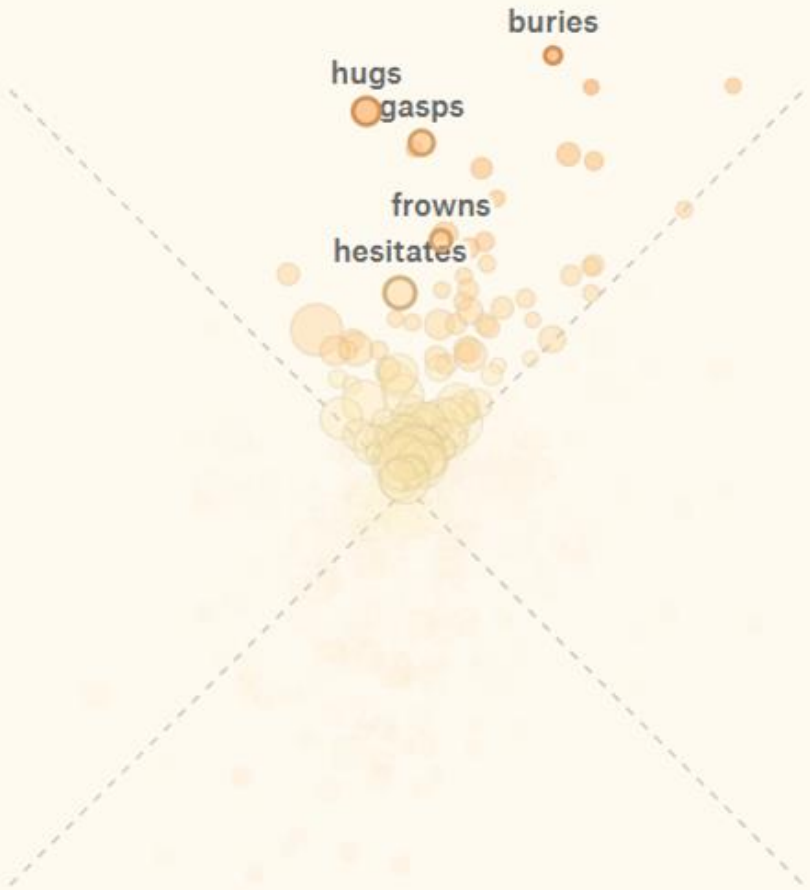
- "Do quizes at the end of each week and go over all of them before class tests"
- Do the labs and type them yourselves. I often referred to the solutions, because I didn't really know how to handle things, but I always typed things myself and thus I am learning a lot. This really showed in the final project!"
- "Don't get too hung up on understanding the formal mathematical motivation behind each topic. As long as you understand the methods you are taught and can apply them at a fairly detailed level".

https://pudding.cool/2017/08/screen-direction/

WRITTEN FOR

FEMALES	MALES	OPPOSITE GENDER	SAME GENDER

Words more likely used to direct **female** characters by **female & male** writers.



What was required to do the data analysis we just looked at?