CS910/CS430: Foundations of Data Analytics

Project Briefing

About the Project…

This lecture:

- Requirements for the project
- Admin issues (timescale, format, submission)
- Motivation: why is the project required?
- Suggestions for the project
- Questions

CS910 Project

- Objectives:
  - To let you put the techniques you see in the module into practice
  - To give experience of real data analysis, and all the challenges
  - To do something you find interesting and stimulating
- Module Project (35%): be a data analyst!
  - Select a data set or sets
  - Start investigating it
  - Apply some methods from the module to the data
  - Make some conclusions
  - Write them up as a short report

Select a Dataset

- Many data sets are freely available on the web
  - Some require you to sign an agreement before you can access
- You can bring your own data (e.g. if you have a company connection) – make sure you have permission from owner!
- Try to pick a domain where you have some expertise/insight
  - E.g. soccer fan? Can you get statistics on all teams and all players for last 40 years?
  - E.g. finance expert? Can you get prices for all shares for every 5 minutes for last 5 years?
  - E.g. web guru? Can you get data on web traffic for a big site every second for last 1 year?

Dataset Guidelines

- Data should be “rich” enough to allow study
  - Should not be very small: not much interest in few KB of data
  - Should not be very large: may be difficult to work with many GB
- Data should be varied enough to allow study
  - Should have many attributes of different types
    - A list of Name, Age, Sex will not be yield interesting results
  - Ideally, allow combination with different data sets
    - E.g. use postcode to look up average income in an area
    - E.g. look up number of employees in a company

Dataset Challenges

- Some data sets are quite large
  - May be too big for shared storage in DCS
  - Can use your own resources (computer, storage), at your own risk
  - Can extract necessary information (aggregate) and throw away rest
  - May be slow to download, process
- Be aware of limitations of the data
  - May have missing values, gaps (see “data basics”)
  - May have errors: sanity check the numbers
    - Always think: are my conclusions valid and meaningful?

Recommended Datasets

- Based on previous experience, here are a few data sets and questions that are known to be worth exploring
  - Followed by list of other data sources for the more adventurous
  - If keen, can get your own data (crawl/scrape)
- Course web page includes examples of previous good projects
  - Obviously, you cannot just duplicate these studies!

1. London DataStore

Police Force Strength

Down 0.4% since last month

More

Housing

New Affordable Homes

Up 26% since last year

More

Fire

Primary Fires in Dwellings

Down 24% on last month

Health

Sports Participation Rate

Down 0.2% on last year

Tourism

International Visitors

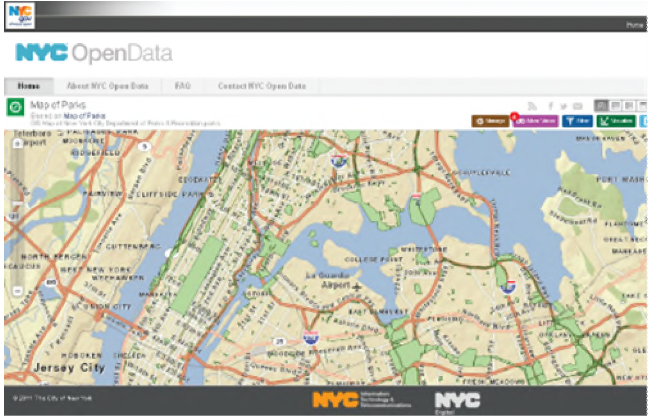
Up 3% since last year

Communities

Census 2011 Population

- Wide variety of data from London http://data.london.gov.uk/
  - Transport statistics: station usage, airport journeys
  - Crime/safety: ambulance trips, incident reports, arrests
  - Transparency: expenses claims, register of interests
- Limitations: many data sets coarsely aggregated
- Possible questions:
  - Compare pairs of indicators (crime, health, wealth) by borough
    - Are there correlations / patterns?
  - Identify how travel patterns change, based on weather/events
  - Find common causes of accidents (across fire, police, ambulance)
  - Categorize expenses: which representatives are best value for money?

2. NYC city data



- https://data.cityofnewyork.us/
  - Transportation: parking tickets issued, parking facilities, MTA data
  - Business: movie locations, electronics stores, building permits
  - Weather: hurricane shelters, areas affected by flooding
- Possible questions:
  - Analyze parking in NYC: which areas are ‘best’ and ‘worst’?
    - Which drivers (states) have worst record…
    - Which vehicles are most ticketed?
  - Which buildings have appeared in most movies/TV shows?
  - How many people are affected by different weather events?
  - Compare differences between London and NY (tricky…)

3. US Healthcare (Medicare)

- Spending breakdowns by claim: https://data.medicare.gov/dataset/Spending-Breakdown-by-Claim/b3t5-5kfi
  - For each hospital, type of care, period of care, shows avg. costs
- Many possible questions:
  - How do costs vary across states?
  - Which are most expensive facilities?
  - When is most cost incurred (before, during, after hospitalization?)
  - Which facilities are best value for money?
  - Are there clusters of similarly behaving facilities?

4. EMI Music data

- EMI Million Interview Dataset
  https://www.kaggle.com/c/MusicHackathon/data
  - Feedback from music fans on their music preferences
  - Ratings and keywords for (anonymized) artists and songs
  - Fans' opinions on their own music habits
- Many possible questions:
  - Predict how users will rate new tracks (holdout data)
  - Find correlations between artists and keywords
  - Cluster to find similar artists

5. Machine Learning Repository

- UCI Machine learning Repository
  http://archive.ics.uci.edu/ml/
- Many (often well-studied) data sets from different domains
  - 1987 National Indonesia Contraceptive Prevalence Survey
    https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice
    - How do various demographic factors affect use of contraception?
  - Objects taken from the web that may or may not be ads
  - Social networks, road networks, location networks
- Often there is a ‘target’ attribute to predict the value of
  - Can you do as well as or better than published results?

National open data

DATA.GOV
EMPOWERING PEOPLE

DATA.GOV.UK (beta)
Opening up Government

- UK: http://data.gov.uk/ USA: http://www.data.gov/
  - Healthcare: obesity, diet, drug use, dental health…
  - Education: exam results, demographics
  - Geography: postcodes, natural resources, lakes, buildings
- Possible questions:
  - What factors most affect health outcomes?
  - What factors most correlate with good exam results?
  - How do natural resources correlate with wealth/health?
  - How do these compare between US and UK?
    - May need to carefully correct for different measurements

Kaggle

- Kaggle (kaggle.com) hosts competitions to analyze data
  - Provide data, and evaluation criteria
- A selection of past competitions:
  - Rank hotels to maximize customer satisfaction
  - Analyze questions on forums to predict tags
  - Decide how a list of tweets relate to the weather
  - Decide whether a review will be rated ‘useful’
- You don’t have to tackle the posed question
  - You can use the data to answer related questions
  - You are welcome to compete, but it won’t affect your grade

Academic data

- Academics write papers, which cite other papers
  - Digital Bibliography & Library Project (DBLP), http://dblp.org
    - Large collection of papers and authors
  - ArnetMiner: citations among papers, referenced against DBLP http://arnetminer.org/citation
- Academics seem to love analysis of citation data:
  - Identify most “influential” people, most “significant” works
  - Identify changes in publishing/citing patterns across time
  - Find individuals who are ‘similar’ in their career trajectory

News Data

- Reuters Text Corpus
  http://www.daviddlewis.com/resources/testcollections/
- Documents that appeared on Reuters news wire in 1987
  - Mostly the text content
  - Some metadata about the TOPIC, PLACE, and PEOPLE involved
- Possible questions:
  - Identify most important (not necessarily most frequent) people
  - Identify significant associations between people and places
  - Identify significant associations between pairs of people

Collect your own data (advanced)

- Perhaps you know another data source of interest on the web
- You can write a ‘crawler’ and ‘scraper’ to download web-pages
  - Parse the downloaded HTML to extract data
  - Or use an API to extract information
- Caution is needed!
  - Some websites do not want their data to be collected
    - E.g. Facebook will detect and block data collection efforts
  - Be respectful: do not ‘hammer’ web servers
    - Google limits to 1000 queries per day, more through API

Estimated level of effort

- ♦ 15 CATS is approximately 150 hours of effort
  - 35% of 150 is ~50 hours: at most 1 (long!) working week of effort
- ♦ Due date will be Jan 11th, 2023 (12 noon)
- ♦ Suggestion: start thinking about projects and data sets NOW
  - Make steady progress to find and analyze data
- ♦ Most important advice: make it fun!
  - Drawing interesting conclusions is very rewarding
  - Working on a topic you find interesting will help

Lessons learned from previous years

- Don’t be underambitious
  - Several projects did simple counts/plots of data
  - Use methods from in the module to show you understand them
- Don’t be overambitious
  - Some projects studied several complex hypotheses over multiple datasets
  - Took a lot of time to find good datasets, clean for use
  - Some successful, but spent too much time
  - Better to do initial analysis, then describe possible extensions
- A few example (good) projects posted on module website
- “Project forum” on module for discussion / questions

Report Format

- Project report: 7 pages (about 4-6000 words) + references
  - The main constraint will be the page limit
  - References can go beyond 7 page limit
  - At least 10pt font, “sensible margins” (2cm on all sides)
  - Some marks will be for presentation: attempts to cram too much in (super narrow fonts, squeezing line spacing) will be penalized
- Why? Practice in communicating your ideas as a data analyst
  - Your aim is to clearly communicate insights into data you found
  - Must be comprehensible to a non-subject matter expert
  - Compare to the papers from the case studies: these are ~10 pages
    - Look at these for guidance as to structure, content
  - Plots may say more than words, so good use of figures will help

Suggested outline format

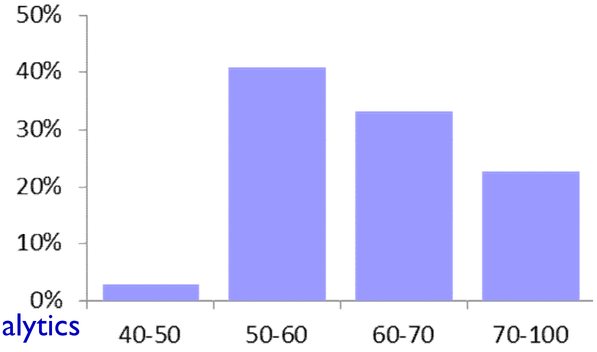
- Introduction and executive summary (abstract)
- Description of data set used (source, size, attributes)
  - Include details of any preparations/reformatting performed
  - Enough detail to allow someone else to repeat your process
- Results (core of the report)
  - Structure into subsections on different aspects to guide the reader
  - Explain methodology used at each step
  - Include plots, reference and explain each plot in the text
- Conclusions and context
  - Describe any related work on similar questions
  - Identify hypotheses / other directions that could be taken in future

Motivation: preparation for dissertation

- The FoDA project also serves as a warm up for MSc dissertations
  - Start thinking about domains where you want to do deeper study
  - Project may lead to questions that need more work to understand
  - Prepares you for independent work and time management
- The FoDA project cannot be re-used for the dissertation!
  - 75 CATS vs 5 CATS: very different in scope
    - Dissertation from January to September
  - FoDA project: apply existing tools to a small case-study
  - Dissertation project: more research oriented, more novelty

Submission Details

- Submission is electronic submission to Tabula
  - Upload a PDF of your report: make sure it is readable!
  - Include your university id (15xxxxx) on first page, not your name
- Due Jan 11th, 2023 (12 noon):
  - Usual late scheme applies: loss of 5% of marks per working day late
  - Usual policy: you can discuss, but **work must be your own**
    - Cannot make use of material submitted for other modules
- Grading will follow the Faculty of Science guidelines: www2.warwick.ac.uk/services/academicoffice/quality/categories/examinations/marking/pgt/sciencecriteria/
  - 70%+: MSc distinction
  - 60-69%: MSc merit
  - 50-59%: MSc pass
  - <49%: inadequate



Evaluation criteria

- Framing material, description of data and cleaning [20%]
  - Did introduction set the scene, and provide overview?
  - Is data set and attributes described, any necessary cleaning done?
- Questions addressed and methods used [20%]
  - Were suitable, novel questions/hypotheses formulated?
  - Did the project make good use of methods from the module?
- Analysis and discussion/interpretation [20%]
  - Were the results explained? Was there a convincing argument?
- Quality of background/conclusion [20%]
  - Did it include appropriate references, and consider further steps?
- Quality of Presentation [20%]
  - Was the report well-presented, easy to read, clear?

Questions?