ST662 Topics in Data Analytics
2020-21 Semester 2
List of projects and students in each group

1. Airline Safety (click here for details)
   Students: Abhishek Pandey; Kaimo Zhang; Lara Butler; Patrick Gorry

2. Portuguese Banking Product (click here for details)
   Students: Adam Mills; Annamalai Meenakshisundaram; Aveek Das; Niall Martin; Sujit Krishnankutty

3. Sudoku puzzles (click here for details)
   Students: Chenlin Liu; Gavin Simpson; Joseph Ogun; Meadhbh Healy; Victor San Vicente

4. Fire Arms Background Checks (click here for details)
   Students: Brian Webb; Dean Rickaby; Jake Meehan; Manasi Tondulkar

5. Irish Residential Property Price Register (click here for details)
   Students: Akankha Raut; Madhusudan Panwar; Niall Ryan; Snehal Deshmukh

6. US Consumer Complaints (click here for details)
   Students: Anita Donaldson; Atikant Negi; Eoghan Quinn-Nealon; Stephen Foley; Susan Mooney

7. Speed Dating (click here for details)
   Students: Alisha Ratigan; Andrew Clarke; Daniel Tracey; James Slevin

8. NYC Flights (click here for details)
   Students: Aaron Doyle; Jack Hickey; James Ferris; Pengyu Yang; Susan Edgeworth

9. Breast Cancer (click here for details)
   Students: Avantika Singh; Gavin Keane; Kevin Horan; Mathews Venattu; Orla Marnell

10. Mite Abundance Response to Climate Change (click here for details)
    Students: Ben Guilfoyle; Caeleen Richardson; Niamh Harford; Shreya Agarwal; Siobhain Topping

Project 1: Airline Safety

A published article titled "Should Travelers Avoid Flying Airlines That Have Had Crashes in the Past?" is available here.

The dataset behind this article can be downloaded and further information on it is available here.

The variables in the dataset are:

Header	Definition
airline	Airline (asterisk indicates that regional subsidiaries are included)
avail_seat_km_per_week	Available seat kilometers flown every week
incidents_85_99	Total number of incidents, 1985-1999
fatal_accidents_85_99	Total number of fatal accidents, 1985-1999
fatalities_85_99	Total number of fatalities, 1985-1999
incidents_00_14	Total number of incidents, 2000-2014
fatal_accidents_00_14	Total number of fatal accidents, 2000-2014
fatalities_00_14	Total number of fatalities, 2000-2014

Carry out your own assessment of the article title.

Project 2: Portuguese Banking Product

A dataset from a Portuguese banking institution marketing campaign recorded many variables and a binary response (y) indicating whether or not the client subscribed to a term deposit.

The dataset can be downloaded and further information on it is available here.

The variables in the dataset are:

Variable	Details
age	(numeric)
job	type of job (categorical: ‘admin’, ’blue-collar’, ’entrepreneur’, ’housemaid’, ’management’, ’retired’, ’self-employed’, ’services’, ’student’, ’technician’, ’unemployed’, ’unknown’)
marital	marital status (categorical: ’divorced’, ’married’, ’single’, ’unknown’)
education	(categorical: ’basic.4y’, ’basic.6y’, ’basic.9y’, ’high.school’, ’illiterate’, ’professional.course’, ’university.degree’, ’unknown’)
default	has credit in default? (categorical: ’no’, ’yes’, ’unknown’)
housing	has housing loan? (categorical: ’no’, ’yes’, ’unknown’)
loan	has personal loan? (categorical: ’no’, ’yes’, ’unknown’)
contact	contact communication type (categorical: ’cellular’, ’telephone’)
month	last contact month of year (categorical: ’jan’, ’feb’, ’mar’, ’nov’, ’dec’)
day_of_week	last contact day of the week (categorical: ’mon’, ’tue’, ’wed’, ’thu’, ’fri’)
duration	last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y=’no’). The duration is not known before a call is performed, also, after the end of the call, y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model
campaign	number of contacts performed during this campaign and for this client (numeric, includes last contact)
pdays	number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
previous	number of contacts performed before this campaign and for this client (numeric)
poutcome	outcome of the previous marketing campaign (categorical: ’failure’, ’nonexistent’, ’success’)
emp.var.rate	employment variation rate (numeric)
cons.price.idx	consumer price index (numeric)
cons.conf.idx	consumer confidence index (numeric)
euribor3m	euribor 3 month rate (numeric)
nr.employed	number of employees (numeric)
y	has the client subscribed a term deposit? (binary: ’1’, means ’Yes’, ’0’ means ’No’)

Build an appropriate logistic regression model and provide a detailed interpretation of the final model. Compare and contrast the model fitting process and capabilities for fitting logistic regression models using SAS, R, Python and Minitab software.

Project 3: Sudoku puzzles

An in-class experiment involving Sudoku puzzles was carried out with eight different classes at Maynooth University between 2009 and 2013. For each experiment, the students in the class were given a Sudoku puzzle to complete. There were four different types of puzzles which were effectively the same puzzle but with different symbols.

Details on the experiment are in Brophy and Hahn 2014 (Journal of Statistics Education, Volume 22, 2014 - Issue 1).

The data from all eight repeats of the experiment are provided in SudokuCombined.csv.

The variables in the dataset are:

Variable	Details
Class	Indicator of class group 1 to 8
Before1	Have played Sudoku before: yes or no
Type	Type of Sudoku puzzle: Numbers, Greek, Letters or Symbols
Correct	Puzzle was correct: yes or no
Time1	Time to completion: Mins:Secs
Mins	Time to completion: minutes component
Seconds	Time to completion: seconds component
Time2	Time to completion: total seconds
Before2	Have played Sudoku before: no, yes in last three months (Within_3 or Inside), yes but not in last three months (Out_3 or Outside)
Logic	Enjoy playing logic puzzles: yes, no, indifferent

Decide on three hypotheses that you will then address with this data. For example: Is the probability of successfully completing the Sudoku related to the type of Sudoku played?

Project 4: Fire Arms Background Checks

In the US, data is recorded on fire arms background checks.

The dataset can be downloaded and further information on it is available here.

The following variables are in the dataset:

month, state, permit, permit_recheck, handgun, long_gun, other, multiple, admin, prepawn_handgun, prepawn_long_gun, prepawn_other, redemption_handgun, redemption_long_gun, redemption_other, returned_handgun, returned_long_gun, returned_other, rentals_handgun, rentals_long_gun, private_sale_handgun, private_sale_long_gun, private_sale_other, return_to_seller_handgun, return_to_seller_long_gun, return_to_seller_other, totals.

The data is available in pdf and csv. While csv will be more useful for analysis, the pdf contains informative footnotes that should be looked at.

Generate some insightful visualisations to display this data. Does the rate of change in total firearms background checks over time vary across states?

Project 5: Irish Residential Property Price Register

In 2010, a residential property house price register was set up in Ireland, listing the addresses of houses sold and the price for which they were sold.

Information on this register and access to the database is available here.

The list of houses sold in a town during a single year can be viewed by putting the town name in the address line and filling in the county, year, and January to December in the relevant boxes.

Compare the rate of change in house prices from 2010 to 2020 for a range of Irish towns. Generate some insightful visualisations to display this data.

Project 6: US Consumer Complaints

The US government records data on consumer complaints.

The dataset can be downloaded and further information on it is available here.

The variables in the dataset are

Field name	Description
Date received	The date the CFPB received the complaint.
Product	The type of product the consumer identified in the complaint.
Sub-product	The type of sub-product the consumer identified in the complaint.
Issue	The issue the consumer identified in the complaint.
Sub-issue	The sub-issue the consumer identified in the complaint.
Consumer complaint narrative	Consumer complaint narrative is the consumer-submitted description of 'what happened' from the complaint. Consumers must opt-in to share their narrative. We will not publish the narrative unless the consumer consents, and consumers can opt-out at any time.
Company public response	The company's optional, public-facing response to a consumer's complaint.
Company	The complaint is about this company.
State	The state of the mailing address provided by the consumer.
ZIP code	The mailing ZIP code provided by the consumer.
Tags	Data that supports easier searching and sorting of complaints submitted by or on behalf of consumers.
Consumer consent provided?	Identifies whether the consumer opted in to publish their complaint narrative.
Submitted via	How the complaint was submitted to the CFPB.
Date sent to company	The date the CFPB sent the complaint to the company.
Company response to consumer	This is how the company responded.
Timely response?	Whether the company gave a timely response.
Consumer disputed?	Whether the consumer disputed the company's response.
Complaint ID	The unique identification number for a complaint.

Are some products more likely to be "Closed with monetary relief" than others? Are there other factors influencing this? Explore some other interesting questions possible with this data. (The variable "Company response to consumer" can be recoded to be "Closed with monetary relief" or "Other" prior to any analysis.)

Project 7: Speed Dating

What influences love at first sight? (Or, at least, love in the first four minutes?) This dataset was compiled by Columbia Business School professors Ray Fisman and Sheena Iyengar for their paper Gender Differences in Mate Selection: Evidence From a Speed Dating Experiment.

Data was gathered from participants in experimental speed dating events from 2002-2004. During the events, the attendees would have a four minute ”first date” with every other participant of the opposite sex. At the end of their four minutes, participants were asked if they would like to see their date again. They were also asked to rate their date on six attributes: Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests.

The dataset also includes questionnaire data gathered from participants at different points in the process. These fields include: demographics, dating habits, self-perception across key attributes, beliefs on what others find valuable in a mate, and lifestyle information. The dataset and related info on the variables can be downloaded here.

What is the attribute that influences the most how people decide whether they want to see someone again romantically? Explore some other interesting questions with this data.

Project 8: NYC Flights

This dataset includes information on all flights that departed New York City in 2013.

It is available through the `nycflights13` in R, see here for details.

The variables are

Variable	Details
year, month, day	Date of departure
dep_time, arr_time	Actual departure and arrival times (format HHMM or HMM), local time zone
sched_dep_time, sched_arr_time	Scheduled departure and arrival times (format HHMM or HMM), local time zone
dep_delay, arr_delay	Departure and arrival delays, in minutes. Negative times represent early departures/arrivals
carrier	Two letter carrier abbreviation. See the `airlines` dataset to get name.
flight	Flight number
tailnum	Plane tail number. See the `planes` dataset for additional metadata.
origin, dest	Origin and destination. See the `airports` dataset for additional metadata.
air_time	Amount of time spent in the air, in minutes.
distance	Distance between airports, in miles.
hour, minute	Time of scheduled departure broken into hour and minutes.
time_hour	Scheduled date and hour of the flight as a `POSIXct` date. Along with origin, can be used to join flights data to weather data.

Generate some insightful visualisations to display this data. Is there a particular airline that is more on time than others? What other questions could you explore with this data?

Project 9: Breast Cancer

Clinical features were observed or measured for 64 patients with breast cancer and 52 healthy controls.

There are 10 predictors, all quantitative, and a binary dependent variable, indicating the presence or absence of breast cancer (`Classification` variable, 1 means healthy and 2 means patient).

The predictors are anthropometric data and parameters which can be gathered in routine blood analysis.

Prediction models based on these predictors, if accurate, can potentially be used as a biomarker of breast cancer.

You can download the dataset here.

How are the predictors related to the occurrence of breast cancer? Is it possible to build a predictive model from this data to help doctors identify breast cancer?

Project 10: Mite Abundance Response to Climate Change

An experiment was set up to simulate the greenhouse effect and to record the response of Mesostigmata mites in the soil during a year, in different ecosystems.

During a period of one year, six samplings were carried out in the first week of every two months: February, April, June, August, October and November 2010.

The description of each recorded variable is in the file MiteResponse_Info.docx, and the dataset is in the file MiteResponse.csv

How is the abundance of Mesostigmata mites affected by the simulated greenhouse effect? Also, create predictive models to assess which predictors are important to estimate the abundance of these mites.