Data Science Institute | Lancaster University

Available Group Projects for SCC460

Project Number	Project Name	Host
1	Exploratory Data Analysis and Inference of Organisational Database	Relative Insight
2	Fly-Tipping	Lancaster City Council
3	Digging grantmaking data	360Giving
4	Exploring COVID response grants	360Giving
5	Understanding government grantmaking	360Giving
6	Predicting the Likelihood of a Patient becoming Stranded	NHS Wrightington, Wigan and Leigh
7	Substance Regulations – Volume and Frequency	Yordas Group
8	Player - On Ball Value Model	Burnley FC
9	Does previous load effect match day physical performance?	Burnley FC
10	Weather trends against railway on-time performance	Avanti West Coast
11	Estimating current value of domestic property in England and Wales	GBG Plc
12	NHS Covid Predictions	NHS Lancashire Teaching
13	Electric Vehicle Charging Dashboard	Miralis Data
14	Factors affecting length of stay in Mental Health wards	NHS Lancashire and South Cumbria

Project 1: Exploratory Data Analysis and Inference of Organisational Database

Project name	Exploratory Data Analysis and Inference of Organisational Database
Project owner / sponsor	Relative Insight – Josh Bailey
Project description	This project is to undertake a full exploratory data analysis (EDA) and inference models of a real-world example of a table from one of our databases. Students should look to the specific questions get started but this project is not limited to those. Aims & Valuable insights: - We would like a full understanding of the distributions and correlations of the variables - A breakdown of the summary statistics of the table as well as a detailed investigation into the nature of the data. - Summary statistics tables of all the strongest inference models built - An analysis of those models and a summary of the insights gleaned from them. Specific Questions: - What are the data types of each variable? - How are the variables distributed? What does this look like? - Which variables correlate strongly (Positively/negatively) with deal length or account manager id? - Can any of the variables you've explored be used for feature engineering – if so, how? How do they affect the models built? - What inferences can we take from the models built? - Which dependent variables affect the health of an account?
Project Data	The data will be one of the organisation's database tables pulled from our SQL database. The column names of this table are: 1. account_manager_id 2. active_users 3. combines

	4. contract_length 5. created_at 6. deleted_at 7. enabled 8. health 9. health_score 10. id 11. insight_style 12. insights 13. language_sets 14. last_language_set 15. last_login 16. last_project 17. logins 18. metadata 19. project_limit 20. project_open_time 21. projects 22. prospect 23. questions 24. refresh 25. restored_at 26. saml_config_id 27. splits 28. start_date 29. total_active_users 30. total_combines 31. total_insights 32. total_language_sets 33. total_logins 34. total_projects 35. total_questions 36. total_splits 37. total_word_count 38. type 39. updated_at 40. word_count This will be provided as a .csv file

Organisation Details

Organisation name	Relative Insight
Main contact name	Josh Bailey
Main contact position	Data Scientist
Main contact email	josh@relativeinsight.com

Project 2: Lancaster City Council: Fly-Tipping

Project name	Lancaster City Council: Fly-Tipping
Project owner / sponsor	Jez Bebbington, Executive Support Manager Thomas Kelly, Public Realm BI & Data Developer
Project description	What are the aims of the project? The illegal dumping of waste- known as fly-tipping- is a high-priority and high-visibility challenge for local authorities, due to the very obvious detrimental effects to the local environment. For Councils to make the most of their scarce resources, they must invest in preventing and minimising fly-tipping in their area, rather than collecting waste which has been illegally dumped. Given the large expanses of open space in the Lancaster district, which includes Carnforth and the surrounding rural area, an evidence- and data-informed approach is required so the Council can predict and target fly-tipping hot spots. What research question(s) should be addressed? 1) What are the fly-tipping hot spots within the Lancaster district? 2) What correlations, if any, exist between the location, type, and size of waste dumped? What insights would be valuable? Any information that can assist the Council with a preventative rather than reactive approach- in particular the prediction and targeting of hot spots. An analysis of the overall location within the district, i.e. the level of fly-tipping that is occurring in Morecambe, Heysham, Lancaster or Carnforth- or within the different Council Wards- would also be helpful in understanding the challenge.
Project Data	What data will be analysed? The Council holds a full record (including type and amount of waste) of each fly-tipping incident from the start of 2018, along with a location record of each incident from September 2015. In what format is it currently held?

	Two Excel sheets contain the full records and the location records, with a corresponding reference number for each incident so full records can be linked to location records.

Organisation Details

Please give the following details about your organisation

Organisation name	Lancaster City Council
Main contact name	Jez Bebbington
Main contact position	Executive Support Manager
Main contact email	jbebbington@lancaster.gov.uk

Project 3: Digging grantmaking data

Project name	Digging grantmaking data
Project owner / sponsor	360Giving – we're a charity that helps funders publish data about the grants they make
Project description	Many of the UK's grant making funders have published data about the grants they have made using the 360Giving data standard: http://standard.threesixtygiving.org/ The aim of this project is to use data published using the 360Giving Data Standard to better understand the landscape of UK grantmaking. Research questions - How does grantmaking vary by theme or topic? How have these topics changed over time? The title and description fields provide a text description of each grant which might be useful for this. - Do grantmakers fund the same organisations? You can use organisation identifiers included to track which organisations are funded by two or more grantmakers. Insights from this dataset will help us to understand grantmaking in the UK better and how to improve it.
Project Data	A CSV of the 360Giving corpus, extracted from the GrantNav website [https://grantnav.threesixtygiving.org/].
Other comments	You might want to look at variation by funders, or by award year, or by geography. The 360Giving standard website describes the fields found in the data: http://standard.threesixtygiving.org/en/latest/#. Bear in mind when using the data that not all UK grantmakers publish data, and not all data publishers have covered the same time period, so you'll need to consider how to account for incomplete or low-quality data. It's also important to note that there are large outliers in the data – for example the National Lottery Community Fund accounts for a large proportion of all the grants available. Please note that the organisation identifiers (used to uniquely identify organisations that receive grants) are not always complete or correct, please contact us if you run through difficulties with them.

Organisation Details

Organisation name	360Giving
Main contact name	David Kane
Main contact position	Product Lead
Main contact email	david@threesixtygiving.org

Project 4: Exploring COVID response grants

Project name	Exploring COVID response grants
Project owner / sponsor	360Giving – we're a charity that helps funders publish data about the grants they make
Project description	During the covid-19 pandemic UK grantmaking foundations provided support to charities and other organisations to allow them to keep services running and provide additional support in a very difficult financial climate. Many of these funders have published data about the grants they have made using the 360Giving data standard: http://standard.threesixtygiving.org/ Potential research questions - What kinds of activities have been funded by grantmakers in during covid? What kinds of organisations - How does the funding compare to previous years funding? Insights from this dataset will help us to understand the response by grantmakers to the Covid pandemic.
Project Data	The grants that they have made are available through the 360Giving Covid Grants Tracker: https://covidtracker.threesixtygiving.org/. A CSV download of the grants is available. The Covid tracker itself pulls data from the 360Giving Datastore: https://www.threesixtygiving.org/data/360giving-datastore/ - if you are interested in the wider corpus of data published under the 360Giving standard we can provide access to the datastore, or you can explore it through the GrantNav tool: https://grantnav.threesixtygiving.org/
Other comments	You might want to look at variation by funders, or by award year, or by geography. The 360Giving standard website describes the fields found in the data: http://standard.threesixtygiving.org/en/latest/#. Bear in mind when using the data that not all UK grantmakers publish data, and not all data publishers have covered the same time period. It's also important to consider that a large number of the grants in the dataset are made by one funder - the National Lottery Community Fund. Please notice that the organisation identifiers (used to uniquely identify organisations that receive grants) are not always complete or correct, please contact us if you run through difficulties with them.

Organisation Details

Organisation name	360Giving
Main contact name	David Kane
Main contact position	Product Lead
Main contact email	david@threesixtygiving.org

Project 5: Understanding government grantmaking

Project name	Understanding government grantmaking
Project owner / sponsor	360Giving – we're a charity that helps funders publish data about the grants they make
Project description	In November 2020 the UK Government published data about the grants that it makes, following a cross-government effort to collect and standardise this information, using the 360Giving data standard: http://standard.threesixtygiving.org/ This dataset, which now contains data for 2018/19 and 2019/20, plus some supplementary data on covid-specific grants and incomplete data from earlier years, can be used to explore how government spends money on grants, and which organisations have received grants. Research questions - How does UK government grantmaking vary by theme or topic? How have these topics changed over time? The title and description fields provide a text description of each grant which might be useful for this. - Do different government departments fund the same organisations? You can use organisation identifiers included to track which organisations are funded by two or more grantmakers. - How well do grants reflect the priorities and aims of the government. Insights from this dataset will help us to understand grantmaking in the UK better and how to improve it.
Project Data	The grant data files are linked to from the 360Giving data registry, under "Cabinet Office" and "Department for Digital, Culture, Media and Sport": https://data.threesixtygiving.org/ You may want to use the 360Giving Datastore to access the data: https://www.threesixtygiving.org/data/360giving-datastore/ - if you are interested in the wider corpus of data published under the 360Giving standard we can provide access to the datastore, or you can explore it through the GrantNav tool: https://grantnav.threesixtygiving.org/
Other comments	You might want to look at variation by funders, or by award year, or by geography. The 360Giving standard website describes the fields found in the data: http://standard.threesixtygiving.org/en/latest/#. You'll need to take into account that the data may be incomplete or of varying quality. Please note that the organisation identifiers (used to uniquely identify organisations that receive grants) are not always complete or correct, please contact us if you run through difficulties with them.

Organisation Details

Organisation name	360Giving
Main contact name	David Kane

Main contact position	Product Lead
Main contact email	david.kane@threesixtygiving.org

Project 6: Predicting the Likelihood of a Patient becoming Stranded

Project name	**Predicting the Likelihood of a Patient becoming Stranded**
Project owner / sponsor	Dr Jamie-Leigh Chapman, NHS Wrightington, Wigan and Leigh Foundation Trust
Project description	Stranded patients can be identified as those with a length of stay (LOS) of seven days or more. Some proportion of stranded patients will be in hospital for a clinical reason however some patients become stranded due to delays in the system. A delayed discharge occurs when a patient, clinically ready for discharge, cannot leave hospital because the other necessary care, support, or accommodation for them is not readily accessible and/or funding is not available, for example to purchase a care home place. If a patient could be flagged as being at high risk of becoming stranded upon admission, or during the start of their stay, then there is an opportunity to prevent a delay before it occurs. <u>Task</u> • Explore the dataset to try and understand the reasons for patients becoming stranded. • Build and evaluate a model which predicts the probability of a patient becoming stranded. • Suggest additional data which could be used to improve your model.

Project Data	<u>Data</u>
	
	The NHS R-Community Datasets package has been created to help NHS, Public Health and related analysts/data scientists learn to use R.
	
	Within the R package is a dataset called stranded_data:
	
	`library(NHSRdatasets)`
	`data(stranded_data)`
	
	You can find the definitions of the variables in the data set from the help file:
	
	`?stranded_data`

Organisation Details

Organisation name	NHS Wrightington, Wigan and Leigh Foundation Trust
Main contact name	Dr Jamie-Leigh Chapman

Main contact position	Senior Data Scientist

Project 7: Substance Regulations – Volume and Frequency

Project name	Substance Regulations – Volume and Frequency
Project owner / sponsor	Yordas Group
Project description	Yordas Group is a company which assists clients in maintaining chemical substance regulatory compliance. A core part of this is recording regulatory information against substances, then monitoring for any changes. In order to steer our monitoring efforts, it is useful to know how frequently we can expect substances to update, particularly when it comes to "hot topic" substances that are getting a lot of regulatory attention. The objective is to combine substance and update datasets in order to test assumptions we have about internal relationships, and to identify the most significant factors when identifying "hot topic" regulated substances. Relevant features that we assume have relationships are: - the total number of regulations affecting a substance; - how long the substance has been regulated for; - the frequency of updates the substance gets (time averaged); - and the frequency of updates at this present moment. We would also be interested in identifying any other features within the data that have a relationship with the current frequency of updates.
Project Data	2 datasets: one containing substance regulatory information; the other containing a schedule of updates to those substances. These exports will be provided in CSV and either JSON or XLS formats (TBC)
Other comments	The substance data is proprietary, and NDAs will be required, as discussed.

Organisation Details

Organisation name	Yordas Group
Main contact name	Jonathan Holding
Main contact position	Data Scientist
Main contact email	j.holding@yordasgroup.com

Project 8: Player - On Ball Value Model

Project name	Player - On Ball Value Model
Project owner / sponsor	Burnley FC Analysis Department
Project description	The aim of the project is to attribute a total 'on ball value' for players in specific events i.e. passing, crossing, shooting, tackling, interceptions etc. By doing so we can ascertain player performance in what we regard to be KPIs in certain positions on the pitch…Once total values are collected they can be averaged out and ranked by percentile to give us a player score, which is comparable across the league. Visualising the data is important to see quick comparisons & highlight strong/weak areas
Project Data	The project data comes from StatsBomb and is currently held in Excel but can be pulled through Python/R
Other comments	Example visual can be provided

Organisation Details

Organisation name	Burnley Football Club
Main contact name	Daniel Morrison
Main contact position	Assistant First Team Analyst
Main contact email	d.morrison@burnleyfc.com

Project 9: Does previous load effect match day physical performance?

Project name	Does previous load effect match day physical performance?
Project owner / sponsor	Matt Challoner – Burnley FC first team sport science department.
Project description	What are the aims of the project? **To ascertain whether previous load from across different time frames effect physical performance in subsequent matches.** What research question(s) should be addressed? **Do chronic loads over 7, 14, 21 and 28 days effect acute match physical performance metrics?** **Do chronic loads over 7, 14, 21 and 28 days effect acute match physical performance metrics where there are no mid-week games?** **Do chronic loads over 7, 14, 21 and 28 days effect acute match physical performance metrics where there are two or more games in a week?** What insights would be valuable? **The club seeks to understand how the prescription of the different physical metrics effects physical performance within matches.** **Are the club getting the workload right to maintain physical performance in the Premier league?** **Are the club getting the training workload right during fixture congestion (2 or more games in a week)?**
Project Data	What data will be analysed? **Physical external load on total distance run, high speed running, sprint distance, accelerations, decelerations, high metabolic distance load (HMLD) in training and matches.** **3 seasons of training and match data available** In what format is it currently held**?** **CSV file**

Organisation Details

Organisation name	Burnley Football Club
Main contact name	Matt Challoner
Main contact position	Performance coach and data analyst Burnley FC first team.
Main contact email	M.Challoner@burnleyfc.com

Project 10: Weather trends against railway on-time performance

Project name	Weather trends against railway on-time performance
Project owner / sponsor	Avanti West Coast Scheduling
Project description	Avanti West Coast operate the rail franchise that handles journey's on England's West Coast Mainline (including between Lancaster and London). AWC's scheduling department are interested in analysing the impact of weather trends against their trains' performance i.e. are certain weather patterns likely to increase delays to train journeys. The project will consider the impact of extreme weather events on train services on the day and following days; and the impact of seasonal weather on train performance. Further details to follow.
Project Data	

Organisation Details

Organisation name	Avanti West Coast
Main contact name	Scheduling Department
Main contact position	
Main contact email	

Project 11: Estimating current value of domestic property in England and Wales

Project name	Estimating current value of domestic property in England and Wales
Project owner / sponsor	Dr Ian Hopkinson
Project description	Estimating the current value of domestic properties What insights would be valuable? • Key aim is to get an estimate of value in the absence of a sale of the property; • Understanding where the Land Registry House Price Index works/does not work would be good; • An additional parameter of interest would be the rebuild cost of the building; • Some visualisation would be fun;
Project Data	Land Registry Price Paid data (CSV format) Other data which might be useful: Land Registry House Price Index (CSV) DCLG Energy Certificate data (CSV) ONSPD (CSV) OS Codepoint Open (CSV)
Other comments	

Organisation Details

Organisation name	GB Group
Main contact name	Dr Ian Hopkinson
Main contact position	Senior Data Scientist
Main contact email	Ian.hopkinson@gbgplc.com

Project 12: NHS Covid Predictions

Project name	NHS Covid Predictions
Project owner / sponsor	Dr Vishnu Chandrabalan, Lancashire Teaching Hospitals NHS Foundation Trust
Project description	The project will involve combining information from several datasets and prediction models as below: - Forecast number of new COVID cases in the community at the LTLA level using historical data (incidence, demographics, vaccination, lockdown level, etc.) -> completed. - Use GIS data to identify neighbours of every LTLA and use neighbour data as covariates -> completed - Use 2019 Hospital Episode Statistics data on the proportion of emergency admissions at each NHS trust from catchment LTLAs – data in public domain – I can provide this. - Use all of the above to predict number of COVID admissions to every NHS trust (based on COVID related factors in the corresponding catchment LTLAs). Lancashire Teaching Hospitals have setup a sandbox Kubernetes computing cluster on Azure (https://jupyter-lander.uksouth.cloudapp.azure.com/) for students to use this for the duration of the project. Alternatively, students can use their own devices. The Azure platform can support Python, R (either in a Jupyter environment or RStudio), Julia and Octave. Python will be preferred and supported from an environment POV but happy for students to use whatever tools they wish to use. No immediate value for the models themselves at LTHTR unless COVID22 becomes a thing. But, it will be a very useful test of the Azure platform for us and how it helps with a collaborative team project. No governance issues as all data is in the public domain.
Project Data	See above
Other comments	"I can set up an environment for this project with the data and give them a quick intro to what this is about. It will be challenging and will involve linking multiple sizable datasets - not for the light-hearted and for the large part they will be slaying dragons on their own after the initial meeting."

Organisation Details

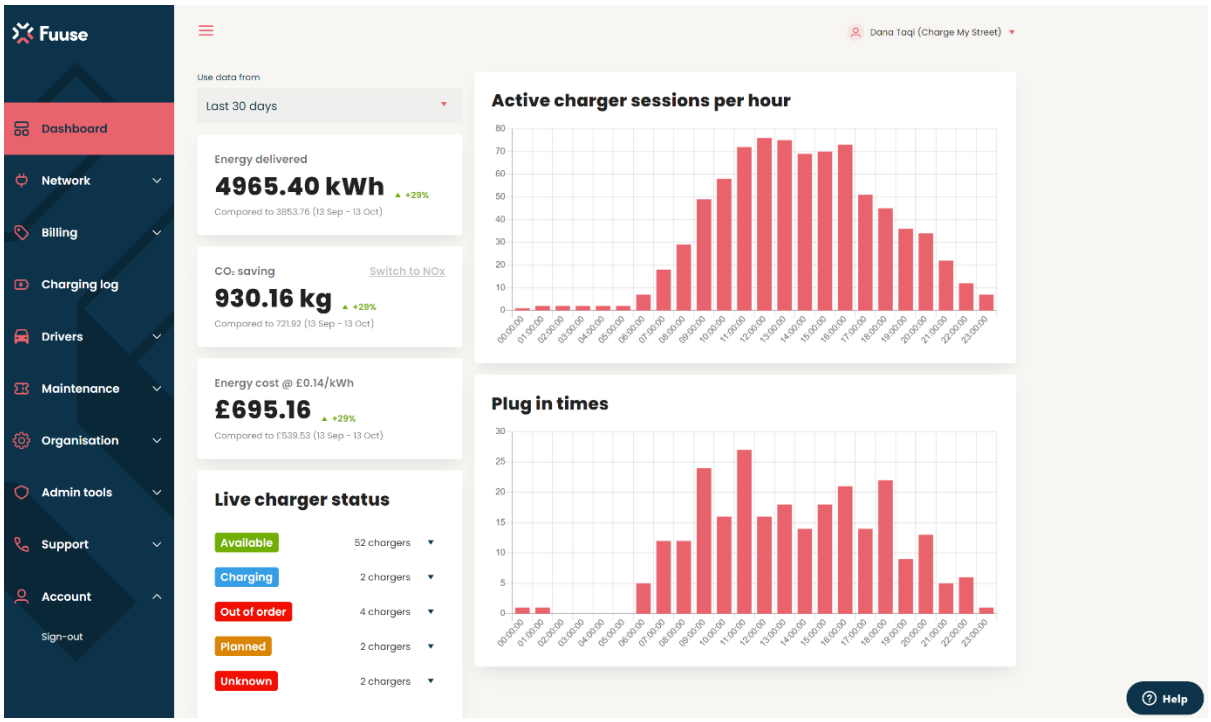
Organisation name	Lancashire Teaching Hospitals NHS Foundation Trust
Main contact name	Dr Vishnu Chandrabalan
Main contact position	
Main contact email	

Project 13: Electric Vehicle Charging Dashboard Analytics

Project name	Electric Vehicle Charging Dashboard Analytics
Project owner / sponsor	Miralis Data
Project description	This project will entail examining electric vehicle charging data to determine if further insights could be gained that would be valuable to service users. The aim of the project is to determine if further information could be added to the service user dashboard (see diagram) that would improve users' perception of the service. Additional insights may be derived from available service data alone or in combination with publicly available datasets (e.g. population density in the area surrounding a charging point – to predict likelihood of charging point availability at certain times)
Project Data	Service use data, public sources.
Other comments	

Organisation Details

Organisation name	Miralis Data
Main contact name	Dr Will Maden
Main contact position	Data Scientist
Main contact email	



Project 14: Factors affecting length of stay in Mental Health wards

Project name	Factors affecting length of stay in Mental Health wards
Project owner / sponsor	Phil Horner, Director of Information, NHS Lancashire and South Cumbria NHS Foundation Trust
Project description	This project will examine the length of stay on NHS mental health wards within our region to determine if factors can be identified that could be used to predict the patient's likely length of stay. Factors to be considered will include diagnosis, mode of admission, patient's address, patient's personal status (age, sex, ethnic group, occupation etc.). Through exploration of the data it is hoped that the team will be able to develop a model to help predict patients' likely length of stay within the ward.
Project Data	The project will be based on admissions data to mental health wards in our region.
Other comments	There may be a short delay to the start of this project as the data is anonymised and sent.

Organisation Details

Organisation name	Lancashire and South Cumbria NHS Foundation Trust
Main contact name	Phil Horner
Main contact position	
Main contact email	