DATA, CENTER OF EXCELLENCE TEAM REPORT

Prepared by Devon Walshe

Table of Contents

INTRODUCTION

GCC DATA TEAM PROJECT OVERVIEW

The reports in this document represent work completed over 4 years by the data, centre of excellence team (data team).

The data team supports the council-wide transformation programme, with a specific focus on data analytics, evidence generation and building business cases for service development.

The 9 included reports cut across the council service family, in most cases working in multidisciplanary teams, including domain expertise from the service area, technical expertise from the data team and senior management oversight.

The collaborative, exploratory methodology was a new approach to solving existing and well understood business challenges, and we hope its effectiveness is demonstrated in this collection of reports.

PROJECT SYNOPSES

BENEFITS AUTO ENTITLEMENT

Removing the need for low income families to manually apply for Free School Meal and Clothing Grant benefits. Families identified in existing council records and automatically paid out their benefits.

ASSISTED HOUSE GARDEN MAINTENANCE

Review of a free, council supplied service to tend gardens for the elderly or infirm. Identifying who in the database is still elligible.

LANDLORD REGISTRATION

Identifying expired landlords and landlords who should be on the register, but aren't. Using the council tax and housing benefits datasets.

HOUSING STOCK ESTIMATE

A spin off from the landlord registration making use of a more sophisticated tenure analysis to estimate the available housing stock.

LES BIN REPLACEMENT

Supporting the bin replacement programme with a data model to generate a sustainable record of refuse locations and materials in the city.

CORDIA WORKFORCE OPTIMISATION

Partnership programme between Strathclyde University, Glasgow City Council and Cordia implementing an algorithmic approach to staff scheduling for the large number of care workers delivering home care.

WORKFORCE PLANNING

Analysis of HR information to understand a detailed picture of staff attrition across the council and aid in the long term planning of staffing needs.

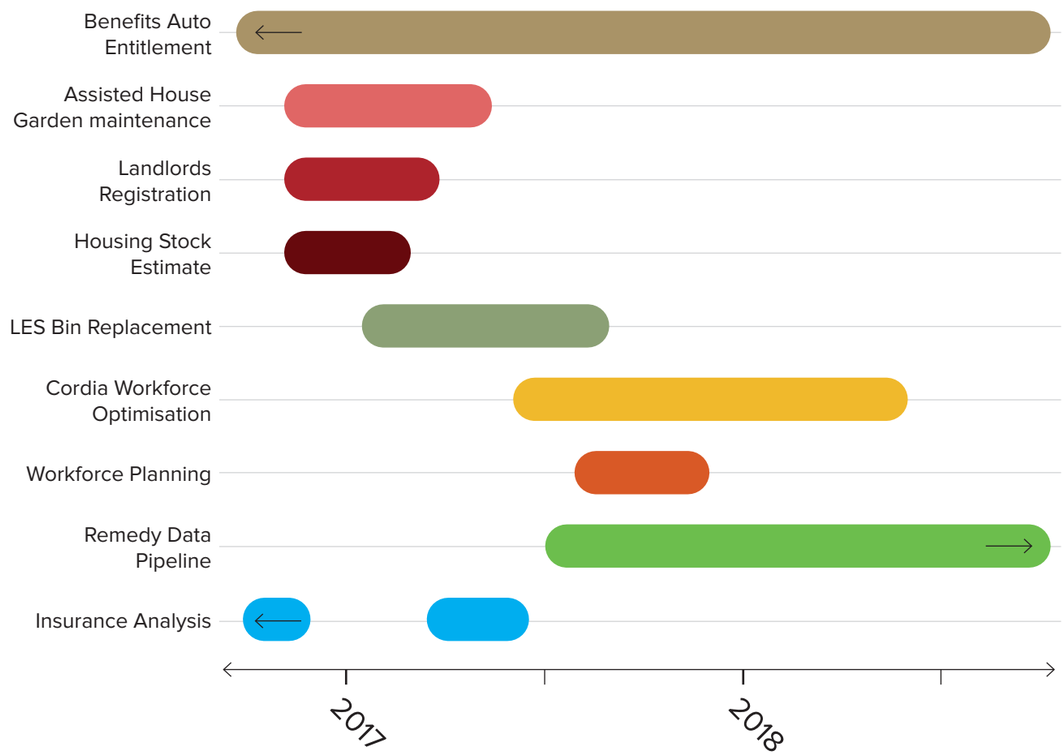
REMEDY DATA PIPELINE

Building infrastructure to create a highly accessible historical record of remedy service calls for business analysis

INSURANCE ANALYTICS

Analysis of insurance data to highlight trend and identify areas of interest in an effort to reduce the overall insurance premiums paid by the council.

PROJECT TIMELINES



IDENTIFIED SAVINGS

Savings	Low	High
Benefits Auto Entitlement	£6,345,270	£6,345,270
Assisted House Garden Maintenance	£458,490	£1,782,450
Landlords Registration	£735,691.37	£1,226,003.67
Cordia Workforce Optimisation	£3,233,480	£9,700,440
Workforce Planning	£848,492	£6,345,270
Insurance Analysis	£316,488.60	£515,243.50
Totals	**£ 11,937,911.97**	**£ 25,914,677.17**

OBSERVATIONS

» The effort required to generate impact from analysis is systematically under-accounted for. Changing a process requires a lot more time and people's effort than most people anticipate.

» The necessary materials for change are already present, taking a new approach often served as the catalyst for motivating teams to view an old problem with a fresh perspective.

» A significant observed barrier to making change was management buy in, with a variety of interests often competing with the evidence generated by the analysis.

» None of the projects would have been possible without being able to work closely and collaboratively with operational staff who held detailed knowledge about the subject matter. This insight is integral to identifying underlying issues that may not be obvious through reporting or statistical accounts.

CAVEATS

» All of the projects in this document were part of an effort to deliver evidence for making savings in a service area, not deliver those savings.

» The identified savings scenario offer a best estimate based on the available data for each project, with an effort to make the calculations transparent. They are not hard figures.

» The narrative content for each of the reports was written faithfully from the experience of working on them, but may contain some oversights or inconsistencies.

BENEFITS AUTO ENTITLEMENT

CONTEXT

» **Poverty leadership panel was looking to understand coverage of school clothing grant and free school meals for dependent children on benefits in Glasgow, particularly ralting to in-work poverty.**

» **Existing system relied on an application process, managed by a team in CBS and co-ordinating with Education.**

» **The question was asked, how many elligible children aren't represented?**

At a glance

Automation of an existing process to disburse clothing grant and free school meals to eligible dependent children in Glasgow.

Involved a lengthy research and test process, ultimately gaining support from the chief executive and put into practice.

£6,345,270 in realised benefit to residents as a result of additional funding from the 'pupil equity fund' from Scottish Government

PEOPLE

Lesley Haddow
Poverty Leadership Partnership

Chris Thompson
Finance

Melanie Dewar
Customer Business Services

Michelle McClung
Customer Business Services

Devon Walshe
Data Scientist (Data Team)

THE BRIEF

1. **Identify all dependent children in households on housing benefit, and prepare their information for a payment process to automate delivery of Free School Meal and Clothing Grant**

DATA SOURCES

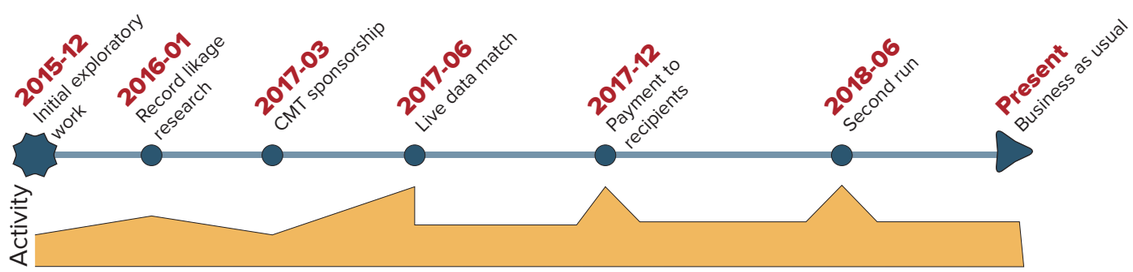
» SEEMIS - full school register for children attending in Glasgow. Including names, addresses, school information.

» Academy - records for individuals with claims for Council Tax, Housing Benefit and Council tax reduction. Includes detailed inforamtion on incomes, household members, tenure and types of benefits in receipt

TOOLS

» R - Statistical computing platform. Used for data processing, matching, analysis and visualization.

TIMELINE OF WORK



PROCESS

STAGE 1 - INITIAL DISCOVERY

Project setup

An initial conversation between National Service Scotland (NSS) and the Poverty Leadership Panel (PLP) led to the data services team getting involved with the data matching process.

Issues with data sharing with NSS meant internal investigation of eligibility was an attractive option and we were given the go ahead to

The initial methodology was to take all dependent children from the housing benefit dataset, and match with those already in receipt of free school meals (FSM) or clothing grant (CG), using exact matches between their name and date of birth.

All non-matched students were our cohort of elligible but not in receipt dependents.

Initial results

Initially the results looked good, but upon closer investigation, the numbers of matched students against the school roll were way off expectations.

We discovered that there were major data integrity issues between the two datasets. Exact matching between name and date of birth was not going to work on account of the names and in many cases dates of birth being enterred differently or incorrectly between the two systems.

As a result, we started investigating probabalistic data matching, or record linkage, a methodology where records are matched based on similarity.

STAGE 2 - RECORD LINKAGE & APPROVAL

Record linkage

Setting up the datasets for record linkage was a more involved process, and required significant computational resources to calculate similarity between the records.

The details and results of that are discussed in more detail in the appendix CG_FSM_record_linkage_research_findings_2017_07_05 included with this document.

PROCESS

CMT approval

The research findings were presented to the CMT, and a suggestion was made to automate the delivery of the education benefits to all students, rather than a supplimentary payment to the ones identified.

With this brief the scope of the project grew and we started organising a new business process which fit under the transformation objectives of the council.

STAGE 3 - TEST & LIVE RUN

To ensure that everything went smoothly with the disbursement, we undertook a trail run of the data match using the older data we already had.

The process involved preparing the datasets for the record linkage, building the pairs that are to be compared, scoring each record based on its similarity, then manually checking the high scoring records to classify them as either true positvies or false positives.

This process was completed in conjunction with Melanie Dewar's help in CBS to complete the manual check and organise the automation of payments.

The trial run was successful and in July 2016, the live exercise was run with a fresh cut of data from academy and SEEMIS.

The resulting cohort from the data match was passed on to CBS, and payments deliverd in December 2016.

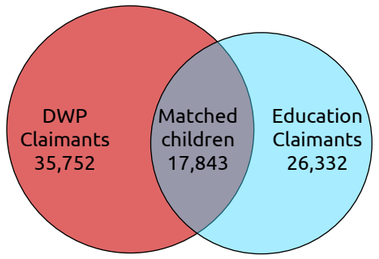
BUSINESS AS USUAL

Since the intial proof of concept, the process has now been integrated into CBS business as usual on an annual basis. In June new data is pulled from academy and SEEMIS and matched by the Data Team, processed by CBS and payment made to the recipients.

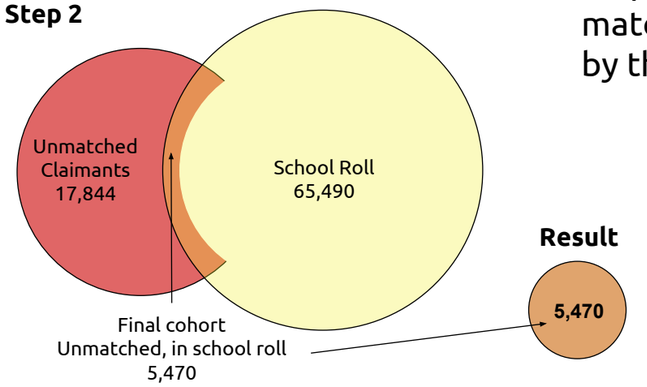
FINDINGS

MATCH PROCESS

Match Step 1



Match Step 2



- 2 Step process
- Process can be repeated with new data at any time
- Manual validation required for each step to confirm the matches discovered by the algorithm

RESULTS TABLE

Cohort	Figure
DWP total dependents	35,752
EDU total dependents	26,332
School Roll	65,490
DWP + EDU exact matches	12,373
DWP + EDU fuzzy matches	5,470
Combined matches	17,843
Remainder	17,844
Remainder + School Roll fuzzy matches	5,128

IMPACT

EVIDENCE GENERATED

» Automating an existing manual process using readily avialable internal datsets is possible and financially beneficial

» Significant gap between eligible parties and those who choose to apply

» Anecdotal evidence drawn from support calls received through the disbursment process indicated that the main reason for not applying was lack of information about the benefits available to them.

IDENTIFIED SAVINGS

Cost of the school day

Scottish Government programme that provides additional funding to schools based on the number of pupils in receipt of Free School Meals.

The 'Pupil Equity Fund' disburses £1,200 per student on FSM to the school, multiplied by 5,470 additional elligible students.

£6,564,000

Additional funding to schools

Process improvement

3 FTE roles were recovered as a result of automation, the staff moved to a team where they were more needed and continued to support any activity related to FSM/CG.

+ £81,900

Benefit to council

Additional cost for Clothing Grant disbursement

This includes the cost of paying out to the new recipients, as well as raising the amount paid from £47 to £52

- £300,630

Council additional cost

£6,345,270

Combined overall benefit less cost to council

CHALLENGES

» Large, interdepartmental effort

» Many technical and organisational hurdles to overcome

» Gaining approval and momentum behind the project

» Organising the appropriate oversight for a change in process

RECOMMENDATIONS

» Investigate methods to reduce the need for manual input to the process, further automation

» Establish and maintain a database containing existing eligible recipients

» Automate data access from academy / SEEMIS

» Expand the automated entitlement and management to other services / processes.

ASSISTED HOUSE GARDEN MAINTENANCE

CONTEXT

» **Review of AHGM, exploring cost saving & income generating opportunities**

» **AHGM is a free service offered by GCC offering free garden maintenance to households without anyone of able body to maintain the garden**

» **Project initiated with the aim to investigate introducing a charge for certain classes of recipients, which then expanded to an in depth examination of eligibility of the recipients**

At a glance

Using statistical analysis to investigate eligibility for AHGM

Between 2,696 and 10,485 out of 14,883 potentially uneligible for service

Between £485,490 and £1,782,450 per annum in identified savings to the council

At an estimated cost of £20,000 to contact service users to reapply, only 117 households would need to come off the list to make it worth while

PEOPLE

Sharon McKechnie
Transformation Manager (LES)

Thomas McMenamin
(LES)

Iain Langlands
GIS Manager (LES)

Richard Oswald
Assistant Information Officer (LES)

Devon Walshe
Data Scientist (Data Team)

THE BRIEF

1. **Initial: examine the AHGM dataset to determine if its feasible to introduce a charge for some recipients. Specifically the number of households with members 70+ in receipt of partial Council Tax Reduction.**

2. **Revised: examine in detail the eligibility criteria for all recipients, looking at whether names match, able bodied individuals under 70 in the household and whether they are the registered council tax payers.**

DATA SOURCES

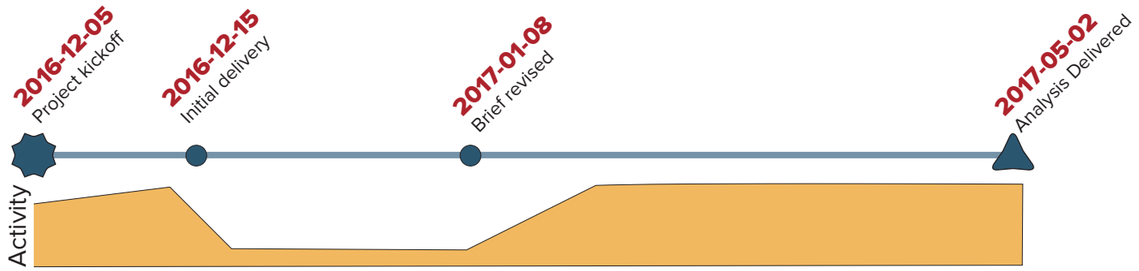
» LAMIS - Main data source for AHGM recipients. Includes information on name and address, criteria under which they were accepted, UPRN details

» Academy - records for individuals with claims for Council Tax, Housing Benefit and Council tax reduction. Includes detailed inforamtion on incomes, household members, tenure and types of benefits in receipt

TOOLS

» R - Statistical computing platform. Used for data processing, matching, analysis and visualization.

TIMELINE OF WORK



PROCESS

INITIAL BRIEF - SERVICE CHARGE

In the first instance, the aim was to understand how many recipients we might extend a service charge to, based on certain criteria such as partial council tax reduction, disability living alowance and attendance allowance.

To facilitate this a data match was required between Council Tax, Housing Benefit and the AHGM register.

Data Match

A data match is accomplished by comparing names and addresses between two datasets which aren't linked by identifiers, and producing a 'fuzzy' score for records which look like they match up.

Delivery

The outcomes of the match were delivered to stakeholders, and unexpectedly low. As a result, the project progressed to reviewing the eligibility criteria for the service.

SECONDARY BRIEF - ELIGIBLITY

As a result of the initial exercise, it was decided to investigate overall criteria to determine how many recipients were eligible for the service. Meanwhile we discovered that the AHGM database included UPRN's for the properties, which would allow us to provide exact matches between council tax, housing benefit and AHGM.

UPRN matching

A list of UPRN's linking CT/HB claims was supplied by the finance academy team. While simplifying the process of matching only 63% of properties were linked, leaving a sizeable hole in the analysis.

PROCESS

Classification

For the records which did match across, we were able to incorporate all of the household and individual level information from CT / HB to assess the elligiblity criteria for each household on AHGM. All four of the criteria were examined to determine eiligiblity:

- Registered Council Tax Payer
- Householder over 70 years of age
- Medical condition which prevents them from maintaining the garden
- No able bodied person in the household between 16 and 69

From these criteria each recipient was classfied into four categories based on the evidence pulled in from HB/CT:

- Probably elligible - Names match up, no-one on DLA or AA or below 70
- Questionable eligibliity - Names don't match up
- Not eligible - Someone under 70 in the house, with no disability.
- Unknown eligibility - Not enough data resulting from no match with CT/HB or doesn't meet or break any the criteria.

Handover

A dataset containing all of the matched information, along with matching criteria and classifications was supplied to the AHGM team in a handover session describing the process and the columns in the dataset.

FINDINGS

ELIGIBILITY CLASSIFICATION TOTALS

0

4,904
Unknown eligibility

» All recipients start the classification process as unknown

» 80% of all unknowns don't match across to CT/HB so its not possible to make a judgement on eligiblility

» Around 20% of cases match, but eligiblity criteria are unclear based on missing data

4,398
Probably eligible

» Recipient household's surnames match between CT and HB, and everybody in the household is over 70, or any under 70's are on Disability Living Allowance or Attendance Allowance

7,500

2,884
Questionable eligibliity

» Cases when the surnames are mismatched between LAMIS and CT or HB

2,697
Not eligible

» If there is somone in the household under 70 with no disability, they are not eligible for service. Because this can only be infered from the HB dataset, we can be more certain of non-elligibility

15,000

IMPACT

EVIDENCE GENERATED

» **Combined the AHGM dataset with Council Tax, Housing Benefit and Council Tax Reduction**

» **Used available information to make an informed analysis of eligibility amongst recipients based on everything we know as a council**

» **Identified a troubling gap of > 7,000 UPRNs in the Council Tax dataset**

IDENTIFIED SAVINGS

Assumptions

Assumption	Variable
Visits per year	17
Workers per visit	1
Hours per visit	1
Rate per hour	£10
Number of gardens	14883

Details supplied in appendices

Scenario 1
Conservative
£458,490
per annum

» All QE are eligible
» All UE are eligible
» NE not eligible

Scenario 2
Realistic
£1,045,670
per annum

» Half of QE are eligible
» UE eligibility split proportionally
» NE not eligible

Scenario 3
Generous
£1,782,450
per annum

» No QE are eligible
» No UE are eligible
» NE not eligible

CHALLENGES

» Reliability of the information. Combining multiple disparate sources means that there will inevitably be a margin of error.

» With such a large cohort, its challenging to validate individual results

» Currently the council is missing > 7,000 UPRN's for council tax and HB, meaning any analysis using matched UPRN's will necessarily miss a number of households

» The process for developing the classification conditions weren't done in conjunction with the management team

RECOMMENDATIONS

» Undertake a compliance exercise with CBS

» Contact all registered recipients at their address and request information about the household

» At an estimated costing of £20,000 for the exercise, only 117 households would need to be struck from the list to make it worthwhile based on the current assumptions

LANDLORDS REGISTRATION

CONTEXT

» **Landlord team working to identify expired landlords which should be chased up**

» **Landlords pay a registration fee once every 3 years and yearly fees for each property**

» **An open question was - how many landlords whose property registrations 'expired' each year should still be on the list?**

» **A financial opportunity exists in finding expired or unregistered landlords and bringing them onto the register.**

» **Properties can be assessed based on information found in council tax, housing benefit and the landlord register.**

At a glance

Review of existing methodology to identify expired landlords

Started from from the beginning using fresh data from the landlords register, academy and the corporate address gazateer.

identified a large gap in our holdings of accurate property records.

Generated a qualified list of expired landlords to target for compliance exercises.

PEOPLE

Duncan Thompson
(DRS)

Jenny O'Hagan
Head of Service (DRS)

Brian Carroll
Housing Intervention & Support (DRS)

Devon Walshe
Data Scientist (Data Team)

THE BRIEF

1. **Determine the number of expired landlords who are still privately letting their flats and therefore eligible for a charge**

2. **Investigate the tenure of all properties to identify properties which may be privately let but not on the landlord register.**

DATA SOURCES

» Landlord Register - Main data source for landlords. Managed by the Scottish government with access to local officials to undertake their work.

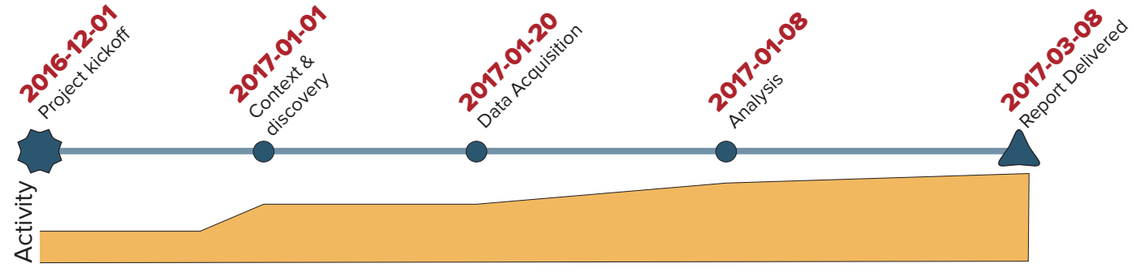
» Academy - records for individuals with claims for Council Tax, Housing Benefit and Council tax reduction.

» Corporate Address Gazateer, the canonical UPRN records for properties.

TOOLS

» R - Statistical computing platform. Used for data processing, matching, analysis and visualization.

TIMELINE OF WORK



PROCESS

REVIEW OF EXISTING PROCESS

The project started with an extensive review of the current process for estimating expired landlords.

The national landlord register clearly identifies expiry dates and there is an existing notification system in place to let landlords know they are up for renewal or if they are expired and encouraging them to keep their registration up to date.

Additionally, any landlords with an expired status are cross referenced with Academy (benefits and council tax management system) which for properties on housing benefit provides a tenure category that can be used as a reference to show properties that are privately let.

One issue that immediately presented itself was that the extracts for both the landlord register and academy that were being used were stale and had been passed around multiple parties.

STARTING FROM SCRATCH

Fresh extracts of the Landlord Register, Council Tax and Housing Benefit were sourced to provide the clearest and up to date picture of the current situation.

The detail of the technical process is included in the appendix report: LLR_update_DW_2017_03_08

UPRN MATCHING

The properties on the LLR, HB and CT datasets were matched using UPRN. This was an imperfect match on accord of both Academy and LLR datasets having out of date UPRN's which have been periodically matched by the property team at DRS. However the most up to date picture of UPRN's across these datasets was compiled and used to match the properties across.

TENURE ESTIMATION

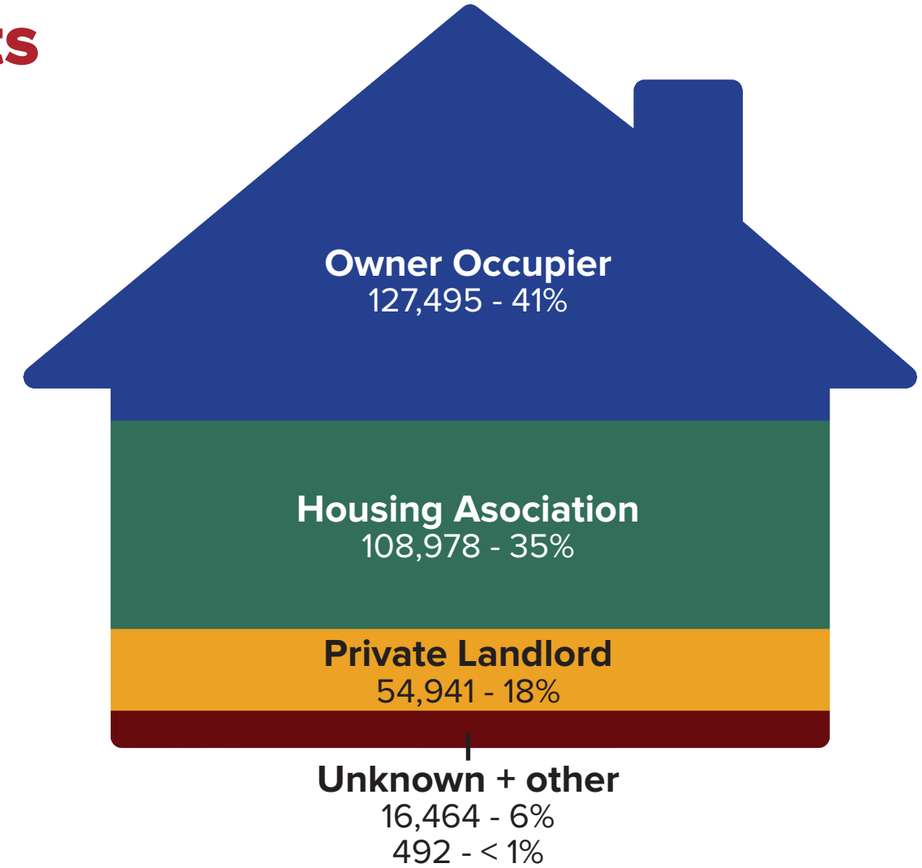
For all the properties that did not have a corresponding HB record, and therefore a difinitive tenure attribute, we went through a process of estimating tenure from additional information contained within the CT dataset.

The process followed a standard procedure used by the team that oversees academy in Finance.

FINDINGS

Tenure analysis results

308,370
Combined properties



UPRN gap analysis

Dataset	Total records	UPRN's held	UPRN's Matched	Unmatched
Council Tax	307,422	300,764	299,309	**8,113**
Housing Benefit	98,359	96,399	96,399	**1,960**
Landlords Register	55,753	60,056	48,970	**6,783**

IMPACT

EVIDENCE GENERATED

» Qualified list of expired landlords

» Unqualified list of unregistered landlords

» Significant and unsustainable gap in the UPRN coverage between various council data assets.

IDENTIFIED SAVINGS

Assumptions	
Assumption	Variable
Registration fee	55
Property fee	11
Landlords	31,366
Avg. props / landlord	1.59
Payment period (years)	3

Details supplied in appendices

Scenario 1
Conservative
£735,691.37
per annum
-£2,411.66
» Including PLL certain, strong

Scenario 2
Conservative realistic
£950,800.52
per annum
+£194,696.99
» Including PLL certain, strong, moderate

Scenario 3
Realistic
£1,067,352.90
per annum
+£311,249.57
» Including PLL certain, strong, moderate
» 10% weak OO, 15% expired, 5% unknown

Scenario 4
Realistic generous
£1,166,670.65
per annum
+£410,567.32
» Including PLL certain, strong, moderate
» 15% weak OO, 30% expired, 10% unknown

Scenario 5
Generous
£1,226,003.67
per annum
+£509,900.34
» Including PLL certain, strong, moderate
» 20% weak OO, 45% expired, 15% unknown

CHALLENGES

» Reliability of the information. Combining multiple disparate sources means that there will inevitably be a margin of error.

» Currently the council is missing > 7,000 UPRN's for council tax and HB, meaning any analysis using matched UPRN's will necessarily miss a number of households

RECOMMENDATIONS

» Work with CBS to undertake a compliance exercise with all landlords showing as privately letting but not on LLR or expired.

» Review the methodology used to estimate tenure across the council

» Close the UPRN gap betwen CAG and Academy

HOUSING STOCK ESTIMATE

CONTEXT

» DRS produces an annual report of residential housing stock across the city

» Some discrepencies between the landlord team and analysis were identified

» The data team was asked to investigate the issues and see what could be done to clarify the discrepancies

At a glance

In tandem to work with the landlord registration team, it was identified that our housing stock estimate might need some attention

Developed a new approach to estimating housing stock tenure thorughout the city

Identified a large gap in UPRN's held between council systems

PEOPLE

Jan Freeke
Service Development

Brian Carrol
Housing Regeneration Services

Duncan Thomson
Housing Regeneraton Services

Devon Walshe
Data Scientist (Data Team)

THE BRIEF

1. **Review the discrepancies between data holdings in the landlord team and the results of the housing stock estimate**

2. **Investigate a new methodology for calculating housing stock and improve the transparency of the process**

DATA SOURCES

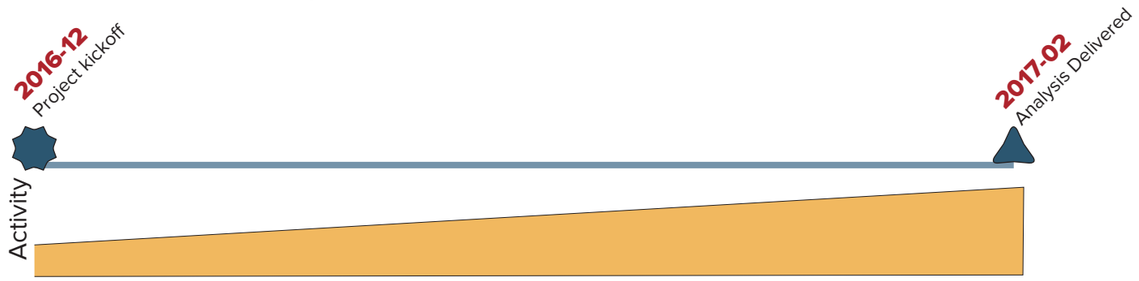
» Academy – HB and CT
» Landlord Register

TOOLS

» R for data analysis and processing

TIMELINE OF WORK

December 2016 - February 2017



PROCESS

IDENTIFIYING THE CAUSE OF THE DISCREPANCY

The process of investigating the discrepancy between the landlord team and the numbers coming out of the housing stock estimate closely followed the landlord registration project.

The main issue was that the housing stock estimate took a sample of tenure across the whole city and extrapolated tenure rates from there.

it was quickly discovered that the size of the sample, 150 properties, was insufficiently large to produce a reliable estimate in this manner, giving a margin of error of roughly 3,000 properties, or potentially 6,000 wrongly identified properties either side of the numbers given.

Additionally, the tenure figures kept by the landlord team were based on many years of accumulated assumptions. After a large review of tenure that had happened a few years previous, each year these were adjusted marginaly based on other reports.

As a result, the overall picture of tenure within the council had become out of sync with the numbers coming from the raw data.

With this in mind the data team decided to return to the raw data and build up tenure estimates from there.

STARTING FROM SCRATCH

UPRN MATCHING

The properties on the landlord register, housing benefit and council tax datasets were matched using UPRN. This was an imperfect match on accord of both Academy and LLR datasets having out of date UPRN's which have been periodically matched by the property team at DRS. However the most up to date picture of UPRN's across these datasets was compiled and used to match the properties across.

TENURE ESTIMATION

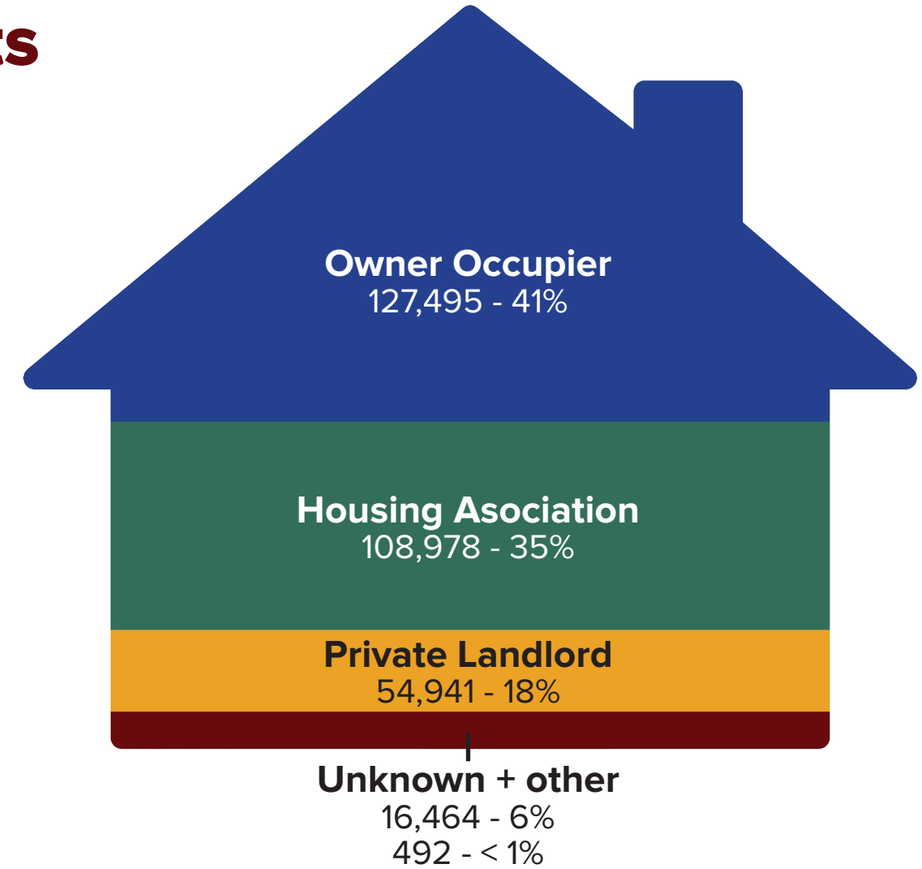
For all the properties that did not have a corresponding HB record, and therefore a difinitive tenure attribute, we went through a process of estimating tenure from additional information contained within the CT dataset.

The process followed a standard procedure used by the team that oversees academy in Finance.

FINDINGS

Tenure analysis results

308,370
Combined properties



UPRN gap analysis

Dataset	Total records	UPRN's held	UPRN's Matched	Unmatched
Council Tax	307,422	300,764	299,309	**8,113**
Housing Benefit	98,359	96,399	96,399	**1,960**
Landlords Register	55,753	60,056	48,970	**6,783**

IMPACT

EVIDENCE GENERATED

» Qualified analysis of tenure for residential properties in Glasgow

» Improved list of UPRNs matched between the academy system and the landlord register

» Increased the transparency of how housing tenure is calculated

CHALLENGES

» Working between two teams with differing views

» Entrenched process for calculating tenure that was biased towards agreeing with existing views as opposed to facts coming from the data

RECOMMENDATIONS

» Implement the new tenure estimation as a business as usual process

» Close the gap in UPRN’s for properties in systems thoughout the council

LES BIN REPLACEMENT

CONTEXT

» **LES project to upgrade the street and back-court collected bins and surrounding areas**

» **Data team asked to support with some data challenges they were facing**

» **Involved in planning for a data collection exercise to determine the positioning and acessibility of the various back courts**

» **Previous data collection exercise was completed in 2011, not fit for purpose in current context.**

» **Project well funded with a high profile**

At a glance

Asked to support an LES transformation project to survey current bin capacity and replace bin stock

Reviewed existing data holdings and surveys

Significant data integrity issues between the holdings

Generated an updated data model for maintaining a sustainable database of backcourt information

PEOPLE

Kevin Howell
Land and Environmental Services

Rolf Matthews
Land and Environmental Services

Tony Boyle
Land and Environmental Services

Oliver Penman
Land and Environmental Services

Devon Walshe
Data Scientist (Data Team)

THE BRIEF

1. **Take a data driven approach to understanding the bin replacement project needs and understanding the back courts**

2. **Support the team managing the process in collecting and processing new backcourt survey**

DATA SOURCES

» Back Court Survey 2011

» Updated sample survey of bin locations and capacity from 2016

» Corporate address gazetteer

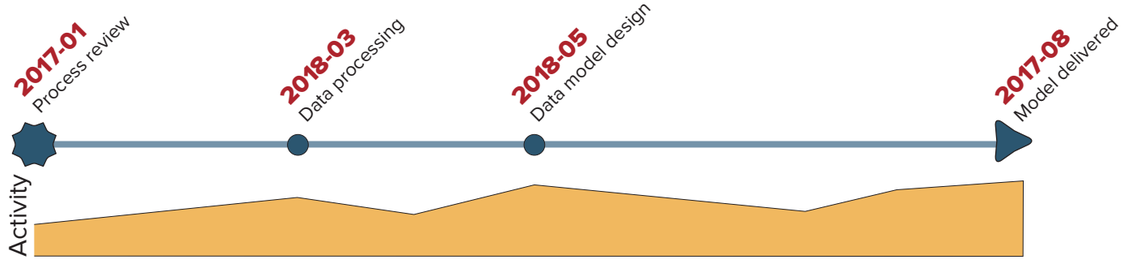
TOOLS

» R for data analysis

» SQL data modeling tool for survey data model

TIMELINE OF WORK

July 2017 - December 2017



PROCESS

REVIEW OF EXISTING RESOURCES

Datasets

The data team was supplied with two datasets used in assessing the current state of the backcourts throughout the city.

The first was the original backcourt survey from 2011. The contents were the result of what appeared to be written answers to a questionnaire, with ~130 datapoints collected for the entire ~22K backcourt locations.

The second was an updated smaller sample of backcourt locations, with less descriptive information regarding the backcourts, but detailed numerical information regarding the capacity of bins for recycling and general waste.

Processing

The initial step was to process the original backcourt survey in a way that allowed us to interrogate the data it contained. For example many columns contained a single yes/no answer regarding a topic, which could be combined to create a single categorical column. Many columns were complete empty and removed and so on. Overall 130 + columns were reduced to 60+ and most of the fields were filled out for the 22,000+ observations.

The more recent updated survey of backcourts was also processed, with a smaller amount of manipulation necessary to generate machine readable data.

After the processing was completed, both datasets were combined using their street location, which functioned as an ID across both datasets.

ANALYSIS

Statistics were then pulled out of the backcourt survey using basic segmentation and visualisation to get a sense for the spread of characteristics for backcourts throughout the city.

In addition, we looked at the differential of capacity for the sites in the original backcourt survey versus the newly collected data.

PROCESS

CAPACITY GAP

The most notable result of the findings was a considerable discrepancy between the amount of bin capacity detailed in the original backcourt survey and the updated review in 2016. For the 600 sites that matched up between the two, there was a combined increase of 80,000 litres for the general waste and recycling.

Extrapolating that amount across the whole city, that would suggest two scenarios - 1) that the data in both cases was correct, and we had added 2.9 million litres of bin capacity over ~5 years, or 2) that the data collection in the original, more recent, or both survey's was incorrect.

Assuming that case 2 is the more likely, this empasised how important it was to establish a) the type of data you were collecting at the outset and b) monitor and control the the data in an ongoing basis.

ESTABLISHING A SUSTAINABLE DATA MODEL

Taking into account a pending new data collection exercise, and the issues with data integrity, we developed and proposed a new model for holding information about the backcourts. The model included detailed datapoints regarding the physical properties of all the elements in the backcourt, the relationship between them (for example bin housing, number of bins inside) as well as the overall condition of the backcourt.

The new model was was designed to facilitate a long lasting database of backcourts throughout the city, as opposed to a snapshot of information at the current time.

An overview of the model and technical description was supplied to the refuse and GIS team at LES.

IMPACT

EVIDENCE GENERATED

» Notable issues with data integrity from previous exercises and inconsistencies between surveys

» Large discrepancy in reported capacity

» No long term plan for holding a permanent, sustainable record on backcourt contents or condition

» New data model for generating a permanent, sustainable record of backcourt contents and condition

CHALLENGES

» Huge spending challenge, original bin replacement project scoped at 10x less than funds received

» Several teams working seperately on the same problem with little co-ordination

» Changing the direction of an already approved survey approach

RECOMMENDATIONS

» Implement a permanent, sustainable data model for the contents and condition of the backcourt as a live database that can be managed and maintained as an ongoing concern

» Integrate the current data holdings into the database and roll out improvements over time

CORDIA WORKFORCE OPTIMISATION

CONTEXT

» **Partnership programme of work between Cordia, transformation, Data lab and Strathclyde university with data team facilitation**

» **Joint funded by GCC and Datalab**

» **The aim is to develop a system to automatically schedule Cordias care work, reducing the amount of work needed to accomplish this manually and increasing the efficiency of schedules for carers and service users**

At a glance

Supporting a partnership programme between Cordia, GCC and Strathclyde University

Developing an optimisation system for staff scheduling of care workers.

Helped bridge the gap between operational and research objectives, and manage the governance process.

10% efficiency savings equates to £335,000 savings per annum

PEOPLE

Karim Arkitunali
Strathclyde University

Annalisa Riccardi
Strathclyde University

Kimberly Hose
Corida

Mateusz Polnik
Strathclyde University

Neil Brown
Datalab

Janette Hughes
Strathclyde University

Devon Walshe
Data Scientist (Data Team)

THE BRIEF

1. **Develop an optimization system to generate schedules for Cordia workers.**

2. **Support development through the governance and operational requirements**

DATA SOURCES

» Caresafe database from Cordia, using previous schedules, visits, locations and people information

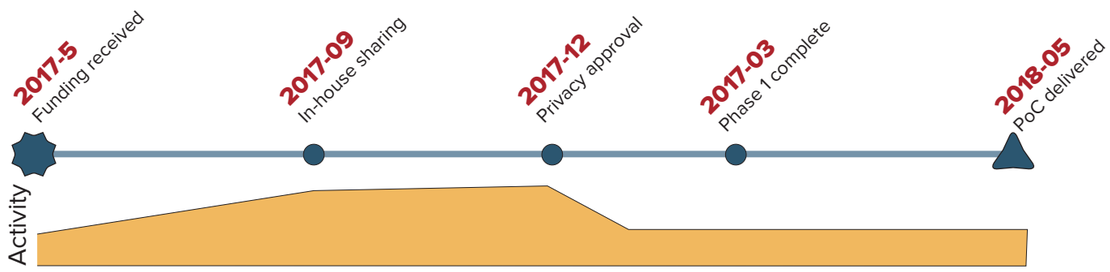
TOOLS

» Google optimization tools:

https://developers.google.com/optimization/

TIMELINE OF WORK

May 2017 – May 2018



PROCESS

PROJECT KICKOFF

Initially the data team was brough in to understand the scope of the project and help align the perspectives between the researchers proposing the optimisation implementation and the operational environment at Cordia.

After a few planning sessions were attended a full proposal was supplied to data lab, authored in tandem with Cordia, Strathclyde and the Data Team, which resulted in an opportunity to pitch to the Data Lab panel for project funding.

The pitch was successful and the project funded.

SCOPE OF THE PROJECT

The projects aim was to optimise the scheduling of Cordia's 2000+ care workers in their daily shift routines. This required access to historical data from Cordia's servers about schedules and care workers locations and how these schedules are currently applied manually by cordia staff.

Multiple travelling salesperson approach

The data from Cordia provided the input to an optimisation approach proposd by the researches at Strathclyde called the 'multiple travelling salesperson problem'. The approach centers around multiple agents, each of which have to make multiple visits to various locaitons within a set amount of time. The optimisation algorithm takes the locations of the agents and locations they need to visit, and attempts to minimise the time spent over the whole period and making all the necessary stops.

Integrating with cordia operations

Aside from developing and testing the algorighm's efficacy within the context of Cordia's care schedule, an additional challenge was planning how this could be used to deliver operational savings

Considering the funding and time limitations of the project, it was decided to position the exercise as a proof of concept to demonstrate savings, leading to additional funding for full implementation in the future.

To facilitate this, the project delivery was an application that received availability from care workers, and visits that the service users required, and outputs a schedule to demonstrate the results of the algorithm.

PROCESS

SUPPORTING DEVELOPMENT

Privacy approval

The first organisational challenge in kicking off the project was getting all of the data governance and privacy paperwork in order, allowing cordia to share the necessary data to power the optimisation algorithm.

The nature of the data in question was particularly sensitive, including personally identifiable information for both carers and service users. As a result there were a number of specific considerations with regards to data security that were above and beyond the norm.

This process took a little longer than expected. It was agreed between governance, cordia and GCC that a 'data processing' agreement was necessary and reporting established on risks and mitigations.

This was accomplished in concert between the parties and sign off granted before the winter holidays 2017.

In house data exploration

Without being able to share the data with Cordia, the project was at risk of losing much of its proposed schedule. To alleviate this it was agreed that one of the researchers could start investigating the dataset and building his tools on site at GCC. This work was accomodated by the data team and was a significant benefit to the project.

Data delivery

Following governance approval, we had to organise the transfer of data to the Strathclyde premises and conduct checks on the setup.

The data was to be housed on static servers in a lab in Strathclyde with no internet connection. In addition we had named individuals with access to the data and a retention policy that expired at the end of the project.

All of these measures were checked during a site visit, tieing up the governance process and allowing the development work to carry on undisturbed.

FINDINGS

EVIDENCE GENERATED

» Working in partnership on cutting edge research to solve operational problems is possible and achievable

» Governance and privacy concerns are manageageable, even in the GDPR environment

» Multidisciplinary approach is a key factor in success

IDENTIFIED SAVINGS

Assumptions	
Assumption	Variable
Scheduling staff reduction	5 x FTE
Optimisation efficiency gains	5-15%
Care Staff	4,000
Salary	£16,167.40

Scenario 1 **5%** **£3,233,480** per annum
- » 5% reduction
- » 200 FTE

Scenario 2 **10%** **£6,466,960** per annum
- » 10% reduction
- » 400 FTE

Scenario 3 **15%** **£9,700,440** per annum
- » 15% reduction
- » 600 FTE

CHALLENGES

» Aligning perspectives

» Dealing with governance challenges

» Applying the findings to an operational setting in Cordia

RECOMMENDATIONS

» Ensure that future funding is available for implementing the research findings

» Investigate other operational areas where external research could impact day to day business

WORKFORCE PLANNING

CONTEXT

» **Proposed by transformation as part of the 2017 summer internship programme hosted in the data team.**

» **Corporate HR looking to understand staff attrition and where resources may be needed in over the near, mid and long terms.**

» **Aim for the internship was to offer more detailed analysis and information to inform decision making at the executive level.**

» **A further exercise was completed by the data team to produce a full report on attrition and possible avenues for further work.**

At a glance

Undertook a deep exploratory analysis of HR data to understand attrition rates across the council

Generated new reporting possibilities to inform policy decisions

Identified relationships between reward and progression and leaving posts.

Despite budgetary conditions, staff numbers are increasing council wide.

PEOPLE

Catherine Kirwan
Corporate Human Resources

Marika Salta
Data Chief Executives Intern

Devon Walshe
Data Scientist (Data Team)

THE BRIEF

1. **Analyse HR data to report on attrition rates based on job role and department**

2. **Research a predictive approach to forecast attrition further in the future.**

DATA SOURCES

» SAP and business warehouse

» All current staff

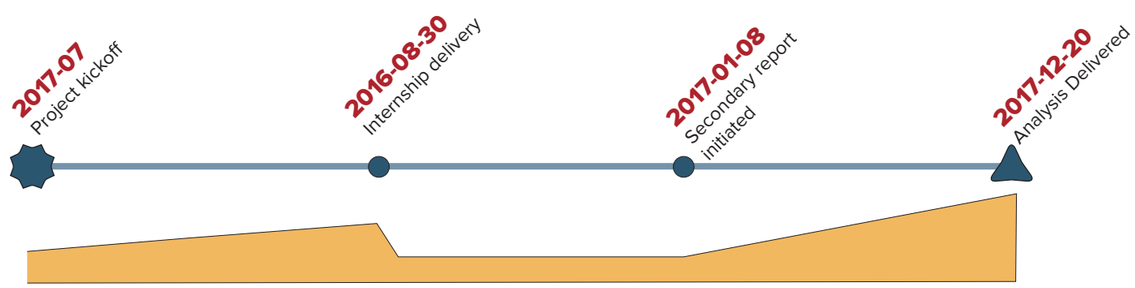
» Leavers from 2008 onwards

TOOLS

» R for data analysis and visualisation

TIMELINE OF WORK

July 2017 – December 2017



PROCESS

RESEARCH & DATA COLLECTION

To better understand the context during the internship programme, we understook stakeholder interviews with Catherine Kirwan and Derek Maitland at Corporate HR.

These sessions aimed at building up a qualitative understanding of the work that had already been accomplished, a defined view of the problem, and what resources were available to begin the process of investigation.

Two years worth of data compiled for the ongoing attrition project was forwarded to the data team, comprising of a list of sta with elds indicating ‘leaver’ or ‘active’ status and characteristic information relating to job context.

Following the initial analysis and modelling, a second tranche of information was supplied by the Business Intelligence team in CBS with a longer time frame of leavers, and additional detail for each staff record.

PROCESSING THE DATA

Once the data was supplied, we developed a pipeline to perform a number of transformations to it so that it could be used in an analytical setting.

This includes:

- **String normalisation - converting all text in the data to a standardized format**
- **Joins - merging datasets on a single record so all the information is in one place**
- **Remove duplicates**
- **Numerical conversion - converting any text based numbers into numeric representations**

PROCESS

EXPLORATORY DATA ANALYSIS & FEATURE ENGINEERING

Exploratory data analysis and feature engineering are repeatable, ongoing processes to enrich, understand and extract value from a dataset.

The quality and mix of information in the supplied sets meant that we were able to add many interesting new variables that describing details in the dataset that weren’t previously apparent.

Factor Compression

The dataset contained a number of collumns with a large number of categories. One of the ways to make better use of these categories this is called factor compression where you map a new smaller set of categories to the larger ones and work with those. For example the payscale column contained 87 categories on account of them listing the weekly hours which is listed elsewhere. Using factor compression these categories were reduced to 16.

Progression

Looking at salaries and grade over time, we developed the concept of progression for each staff member. By taking each staff’s start and ending salaries and grades, you could calculate a statistic for their average progression for year which could be compared against entries.

These averate yearly progression figures were normalized against contracted circumstances to put everyone on a level playing field.

Reward Metric

With the aim of understanding why some sta might stay and some might leave, we went a step further with the concept of a reward metric, or a measure of how well rewarded a sta member was relative to their circumstances.

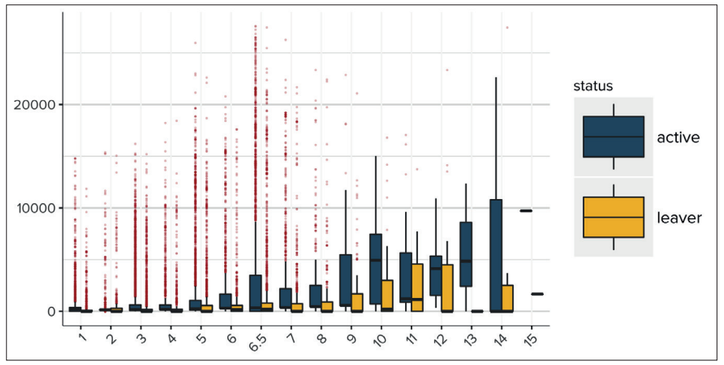
The hunch was that under-rewarded sta might have a higher probability of leaving.

To generate this metric, a lightweight linear regression model to predict salary was implemented, using tenure, age, grade, service, and their fulltime work ratio as the input varialbes to the model.

This model was then used to predict the ‘expected’ salary for each sta member, and the di erence between the ‘expected’ salary from the model and the actual salary reults in our new reward metric.

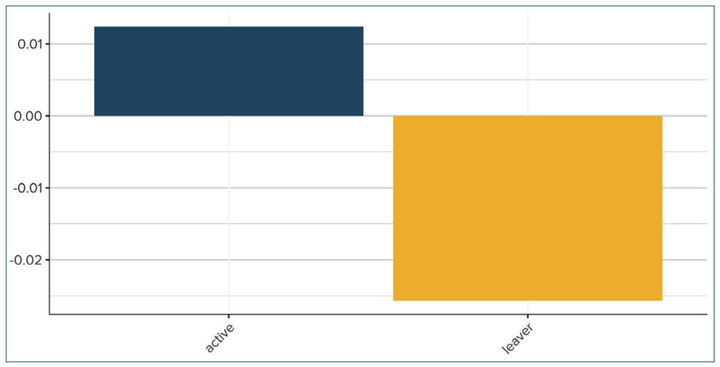
FINDINGS

Salary progression by grade for active staff vs leavers



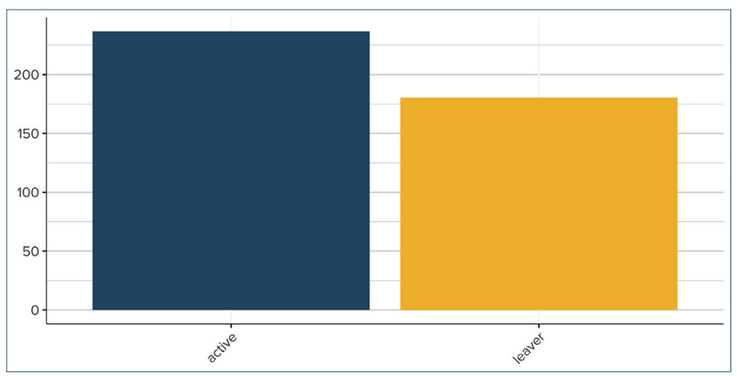
This chart demonstrates a clear relationship between salary progression and the decision to leave. As grade level increases, this relationship becomes more prominent. Simply put, those who progress less than the average are more likely to leave.

Mean reward metric, active vs leavers



Looking at the predicted salary for active vs leavers, this relationship becomes even clearer, overwhelmingly leavers are rewarded less than their peers.

Mean salary progression by year, active vs leavers

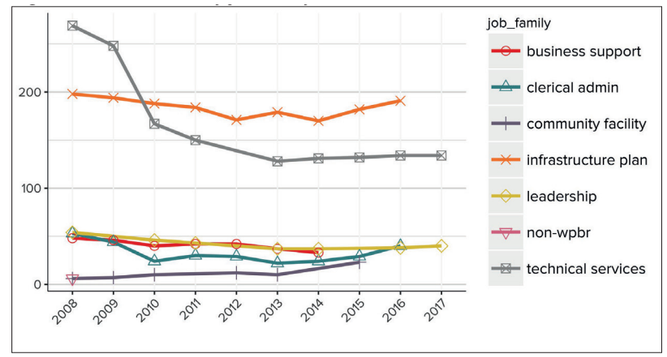


This aggregates salary progression over the entire council. Reinforcing the causal relationship between progression and the decision to stay on.

FINDINGS

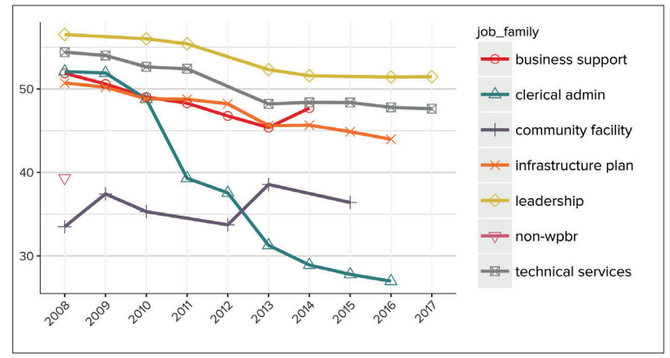
Total DRS workforce by job family from 2008

This illustrative chart shows total staff numbers across job families in a single service. Its easy to see the relative numbers of staff in the service, and how they have changed over time.



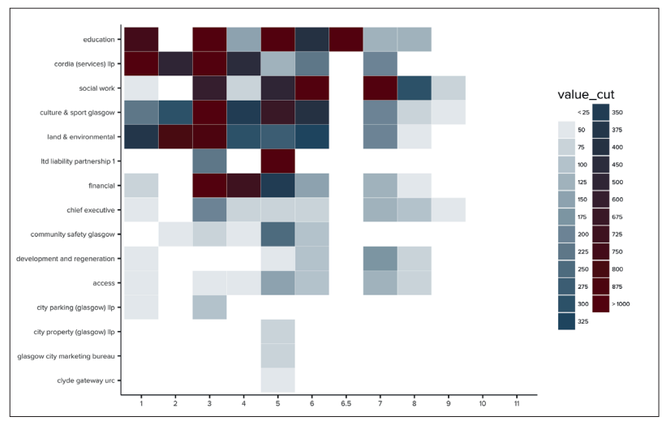
Mean age by job family from 2008, DRS

One of the aspects of staff demographics that was originally hypothesised was that we had an aging workforce. In DRS we can see that the trend is the opposite, with mean age trending downards in all job families.



Heatmap of grade distribution by service

Another illustrative chart to quickly get a sense for relative levels of seniority across the council family.



IMPACT

EVIDENCE GENERATED

» The purported hiring freeze isn't going so well, staff totals on average increasing across the entire council

» Total salary bill has increased by £38m from 2014-2016, with Education, Glasgow Life and Access as main contributers.

» More staff starting than leaving, and decrease in leavers contributes to increased staff bill more than new starts.

» Using the attrition reporting, HR can use more detailed information in order to forward policy and the hiring freeze agend

SAVINGS IDENTIFIED

Bulk staff reduction

£848,492
per annum @ .1%

£4,242,458
per annum @ .5%

£8,484,916
per annum @ 1%

£21,212,290
per annum @ 2.5%

Starts / leavers reduction

Reduce new starts by 10% and increase leavers by 10% year on year

year	2018	2019	2020	2021	2022
starts	2,353 -10%	2,117 -10%	1,906 -10%	1,715 -10%	1,544 -10%
leavers	1,707 +10%	1,877 +10%	1,878 +10%	2,272 +10%	2500 +10%
starter salary (millions)	£46.99	£42.30	£38.07	£34.26	£30.83
leaver salary (millions)	£34.10	£37.51	£41.26	£45.39	£49.39
position (millions)	-£12.90	-£4.78	+£3.20	+11.13	-£19.09

CHALLENGES

» Large volume of data to work with and prepare

» Open briefs can be challenging to work with

» Understanding the difference between the high level view and the operational reality of staffing policy and factors behind the trends

RECOMMENDATIONS

» Use the reporting analysis to inform policy decisions and build business cases.

» Take a longer term view for policy decisions

» Make some policy adjustments to remedy the currently worsening situation for payroll budget

REMEDY DATA PIPELINE

CONTEXT

» **All inbound calls or service requests from residents handled through a management system - Remedy**

» **Historical data going back to 2003 is held in archives and accessible via the Business Objects system**

» **Some of the live and historical data is used for reporting, but access is limited throughout the council and the contents of the underlying data is often misunderstood.**

» **The data team decided to work with the dataset to establish it as a more sustainable and accessible resource**

At a glance

Internal green field project to automate the backup of a significant council data asset

Developed a near realtime data pipeline which backed up 15 years worth of analytics material in the council

Positive proof of concept for other valuable and inaccessible datasets.

PEOPLE

Devon Walshe
Data Scientist (Data Team)

Steven Livingston Perez
Data Team

Stephen Sprott
Data Team

THE BRIEF

1. **Establish the remedy archive as an accessible and searchable database for live and historical analysis of council activities.**

DATA SOURCES

» Remedy Business Objects Universe

TOOLS

» Python – Data pipeline

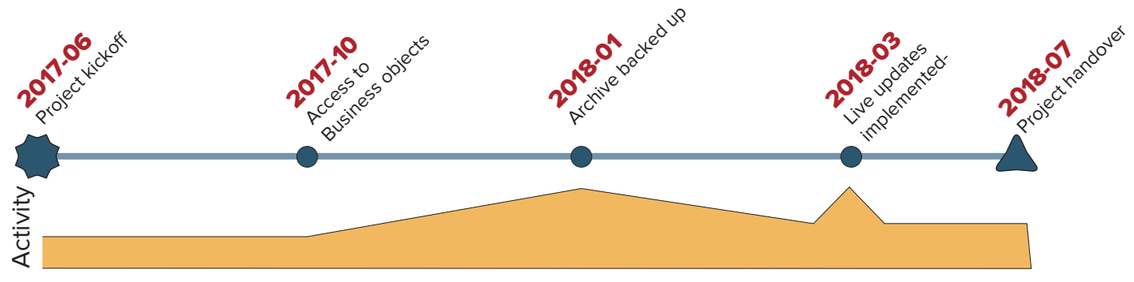
» Business Objects Query as a Web Service – API to business objects

» Postgres – database containing remedy calls pulled from business objects

» Azure cloud services

TIMELINE OF WORK

June 2017 - July 2018



PROCESS

INVESTIGATING THE ARCHIVE

The first step for the team was getting access to the archive. In the first instance we were granted access to the Business Objects web interface, which would allow some filtering and subsetting of the data, and the ability to export CSV files for further analysis.

These static datasets pulled from Business Objects formed the basis of an understanding of what data was contained within the source database, what we were working with and how we might design an analysis database around that.

This exploratory analysis resulted in a decision to leave the data as it was, not to do any specific procesing before we pulled the data out. The reasoning being once we had a copy of the raw data in a more accessible location, we could then process and manipulate it as required.

DESIGNING A DATA PIPELINE

As the source data is constantly updated and changing, we needed to think about building a pipeline that could be run at regular intervals, transporting new and updated calls from its base location into our database.

To help us build this pipelines, we received access to an existing system called ‘Query As a Web Service’, that would allow us to access the source data via computer programs that we would write speicfically for this purpose.. The web service allowed for the creation of an ‘application programming interface’ or API from which we could connect to the original data source and retrieve archived, new or updated call data, and then transport it into our database, stored on Microsoft Azure cloud services.

Step 1 - Archive Retrieval

The historical call data for Remedy was held in a seperate database within Business Objects, and the first task was to build a programme that could access the records going back to 2003 and populate the analysis database with them.

The main challenge to overcome here was that only a small amount of data could be transported per API call. The programme was designed in such a way that you could specify a start and end date and time for the section of data you wanted, and it would automate the batch requests in a size that the API could handle.

PROCESS

Step 2 - Merging the live and archive database

Once the Remedy archive was successfully backed up into the analysis database, the next challenge was to merge the live database, which had a slightly different set of columns from the archive.

This was overcome by creating an intermediary database that included all the columns from the archive and live database and importing both databases.

Once all of the data from the live database was joined with the archive, we had a single, merged source of analyitics material for Remedy calls dating back to 2003.

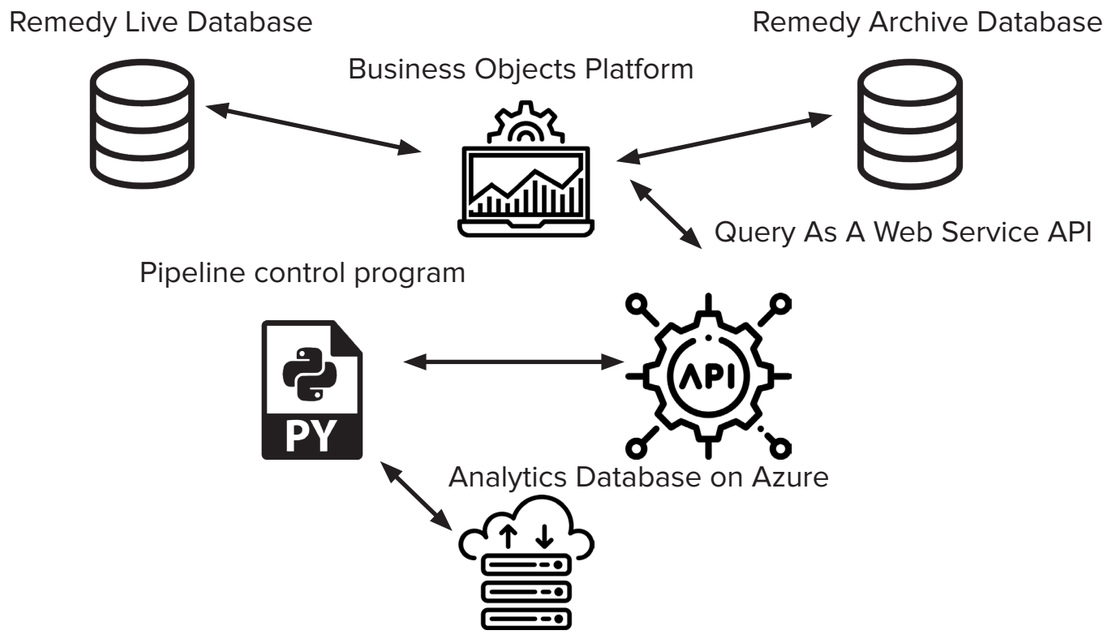
Step 3 - Realtime updates

The final step was to have the database automatically refresh itself with new data.

This was accomplished by modifying our pipeline application to periodically query the live API for new or updated call data. The information then would pass just as before from the source database to our analytics database.

At this point we now had a live database which was self updating, so any new calls to the council would be available for analysis, along with all historical calls.

PIPELINE ARCHITECTURE



IMPACT

EVIDENCE GENERATED

» Established a new, scaleable method for analysis to be completed on the remedy dataset.

» Demonstrated that its possible to build on existing infrastructure using modern, flexible toolsets.

» First large scale use of cloud services for holding council data asesets.

CHALLENGES

» Working with an unweildy legacy system to extract data

» Large volumes of histoical data

» Operating with internal network resources to and from the cloud

RECOMMENDATIONS

» Replicate the process for other significant data holdings

» Open up analytics tooling for business intelligence staff in the council

» Build new services on top of the combined dataset.

INSURANCE ANALYTICS

CONTEXT

» Data Services team asked to support a data review exercise in 2016 relating to insurance for the transformational agenda

» Large insurance premiums were being paid to the supplier. If we could better predict or understand our liablity, we could budget against the claims internally to reduce premiums by being proactive.

» Additionally the insurance dataset offered insight into the relative levels of liability across departments in the council.

» Later in Summer 2017, additional work completed on the motor insurance dataset to make a case for business process changes.

At a glance

Asked to support transofmation business case relating to insurance claims

Undertook an exploratory data anlysis and reporting process looking at where savings could be identified or opportunities for better management across services and in managing claims

Uncovered an overall reduction in claims and several areas where savings could be identified

PEOPLE

Una Scouller
Insurance Manager

Janice Timoney
Finance Manager

Cyril Dyer
Data Services

Devon Walshe
Data Scientist (Data Team)

THE BRIEF

1. **Look at the last few years of insurance claims and see if there is any particular insights which would help reduce the number of claims or the amounts of payments**

2. **Specific issues included our lack of analysis on the figures or efforts to reduce claims**

3. **Review and comment on data management**

DATA SOURCES

» Data coming from the insurance system

» Public Liability - public claims

» Motor Liability - claims relating to our vehicle fleet

» Employers Liability - employee claims

TOOLS

» R for data analysis and visualisation

TIMELINE OF WORK

September 2015 – June 2017

PROCESS

EXPLORATORY ANALYSIS

Retreving / processing data

In the first instance the data services team was supplied with three datasets covering a five year period:

- Public liablility claims
- Employers liability claims
- Motor liablity claims

Each of these datasets was investigated column by column, with basic processing steps taking place, including converting text values to numeric, adding new calculated columns, reducing the factor levels for categorical columns, and other techniques to get them into a format that was ready for analysis.

DESCRIPTIVE STATISTICS

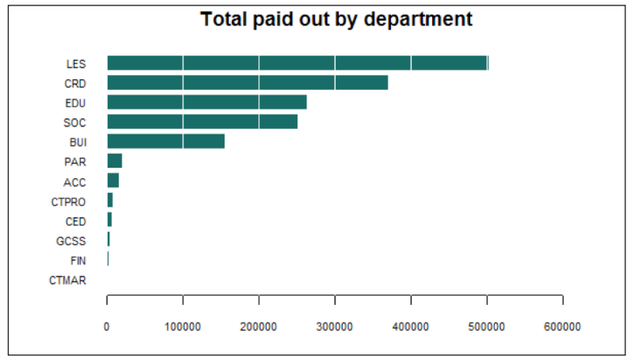
Each of the datasets was investigated to understand a few key factor - their trends over time, which categories of claim were the most costly, which were avoidable, whether certain departments were particularly prone to claim types et cetera.

The analysis was completed using a straightforward descriptive statistical approach, segmenting and aggregating the figures into groups. For the initial analysis in 2015, these results were compiled into a report and suppled to the transformation team for the purposes of understanding the exposure. In the second run for motor liability in 2017, the analysis was compiled into a series of visualisations for use in a later business case

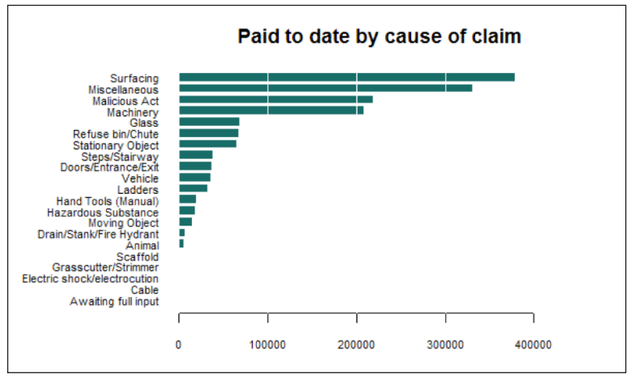
FINDINGS

TREND ANALYSIS

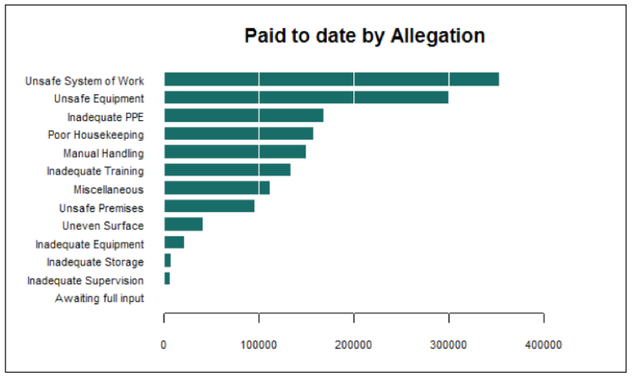
5 departments make up the overwhelming majority of claims paid out. £1,547,437 vs. £63,331.56



Surfacing, Machinery and Malicious Acts are significant cost contributor categories in addition to “Miscellaneous”. Interestingly, Cordia, LES and Soclail Work are all overwhelmingly responsible for these amounts, respectively.

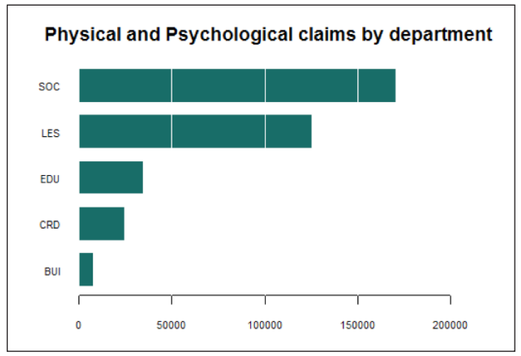


Unsafe equipment and work are the primary causes of claims contributing the better part of £1m in claims over the 5 year period.

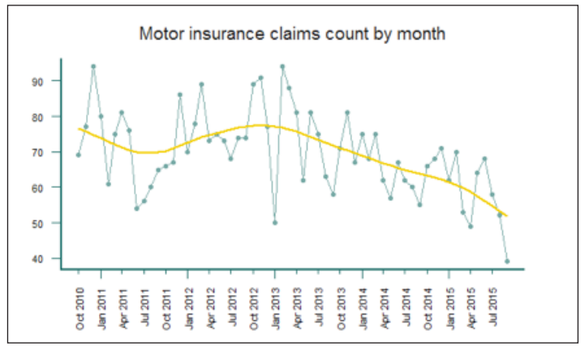


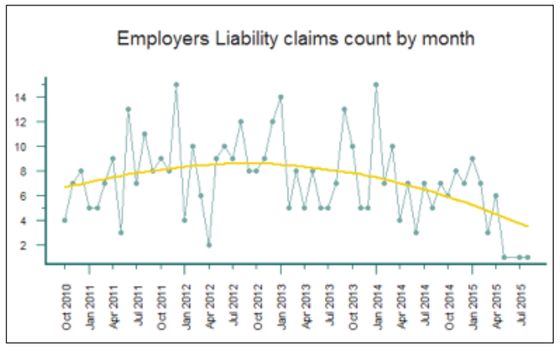
FINDINGS

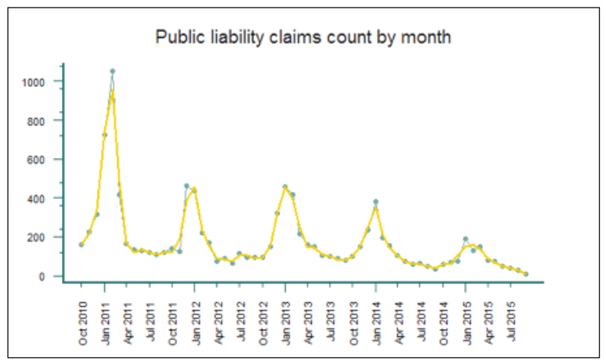
Non injury related claims were primarily issues relating to workplace stress, specifically in LES and Social work.



Looking at the overal trend for Motor, Public Liability and Employers Liability claims, by count, we discovered a consistant downwards trend in all three. This trend was also mirrored in the amount paid out totals in each three, but that could have been on account of claims not yet being settled.







IMPACT

EVIDENCE GENERATED

» LES responsible for a significant number and size of insurance claims, as expected due to nature of work

» Confirmed intuitive understandings of how claims were spread out through the council

» Edge cases responsible for vast majority of cash output to claims, standard reporting on averages misses the detail of the claims

» Not enough information collected on claims, for example how they progress, background of claimants.

» Road surfacing responsible for £515,243.60 per annum in public liability claims.

IDENTIFIED SAVINGS

£515,243.50
per annum

» Eliminating public liability claims caused by surfacing issues

£316,488.60
per annum

» Reduce motor claims by 65% (as seen in West Lothian) by introducing telematics

CHALLENGES

» Not enough substantive data relating to the context of the claims

» Missing information about how claims are settled

» Only operating on 5 years of data

RECOMMENDATIONS

» Implement more data driven decision making into the process of managing claims

» Collect more data to enable a forward looking statistical approach

» Use quantitative reporting on claims to encourage services to update processes or build business cases for change, especially in Cordia, Social Work and LES