FRS-Data Analysis Report

Joseph Arber

14/08/2020

Project Aim

The aim of this analysis was to explore the main themes and trends behind the welfare payments system in the UK. However, given the size of the dataset, this specific projects timeframe and intended audience, the scope of the analysis was narrowed down to focus on a few key questions. Notably, the analysis carried out tried to gain a better understanding through two dimensions:

- **1. Qualitative Understanding:** What are the main drivers of welfare claims, and what are the key reasons for failure to meet debt repayments?
- **2. Quantitative Understanding:** Is housing benefit enough to sustain rent?

Specifically, this paper wanted to invesigate what welfare and benefit features included in the dataset are most correlated with a users inability to manage debt repayments. Alongside this, the analysis also sought to determine whether the cost living, housing and rental costs matched the amount a user recieves in benefit payments such as Universal Credit, Housing Allowance and several others.

Data

This report captures the key analysis captured from the 'FRS' dataset. The dataset is drawn from the 2017 version of the Family Resources Survey (FRS), a nationally representative cross-sectional survey of UK households. The attached datset includes only those households from the FRS that are eligible for Universal Credit, according to the Policy in Practice policy micro-simulation engine (the Benefit and Budgeting Calculator). The dataset includes household information on demographics, earnings and other income drawn from the raw FRS data, as well as benefit eligibility (under both Universal Credit and the legacy system) as modelled by the Policy in Practice engine.

The dataset has **4545** unique observations and **36** feature columns. A glimpse of the features in the dataset is given below.

Data Cleaning

Some basic data manipulation was done at the start. This included removing the unwanted columns from the dataset as well as dealing any missing values encountered. As such, the following variables were removed:

- Child ages
- Age group (band) and Age group Universal Credit

Summary and Descriptive Statistics | Early Stage EDA

Standard Deviations

It is also useful to get an understanding of how the features in this dataset are dispersed. The standard deviations and mean values for the data are displayed below.

```
## Descriptive Statistics
## df
## N: 4545
##
##                                        Mean    Std.Dev        Min          Max
## ------------------------------------ ---------- ---------- ---------- -----------
##                           age_partner      12.50      19.70       0.00       65.00
##                             age_user      39.73      13.99      17.00       65.00
##         child_tax_credit_eligibility     188.51     263.21       0.00     1596.66
##                           child5_age       0.04       0.50       0.00       13.00
##                             earnings     491.82     727.21       0.00     5380.24
##               ESA_JSA_IS_eligibility     232.72     247.16       0.00     1339.87
##           estimated_household_costs    1131.76     540.40     280.12     4895.07
##          housing_benefit_eligibility     167.65     225.61       0.00     2059.37
##                                   ID   96184.76   55566.25      51.00   191053.00
##                     income_aftercosts     391.48     753.13   -2849.75    10653.30
##                          monthly_rent     332.39     339.45       0.00     3813.00
##                  nondependent_income     199.00     715.00       0.00    11667.00
##                nonmeanstested_income      34.09     108.20       0.00     1396.17
##                   number_of_children       0.91       1.21       0.00        7.00
##                numberof_nondependants       0.58       0.98       0.00        5.00
##                               savings     266.91    1143.13       0.00    15194.20
##                takehomeincome_legacy    1324.24     817.46      98.23     5500.10
##                   takeshomeincome_UC    1266.69     838.82      41.30     5760.14
##                       UC_eligibility     634.06     454.79       0.90     3314.89
##                            weighting    1266.34     900.65     211.00    28412.00
##       working_tax_credit_eligibility      51.47     126.98       0.00     1231.20
```

The mean values for the variables in the dataset are also displayed. An interesting finding was that "child tax credit" has is higher on average than "housing benefit". Notably, the mean for "income after costs" is approximatley £391. However, the variation (standard deviation) is £753, suggesting that there is disparity between those with the highest incomes and those with the lowest.

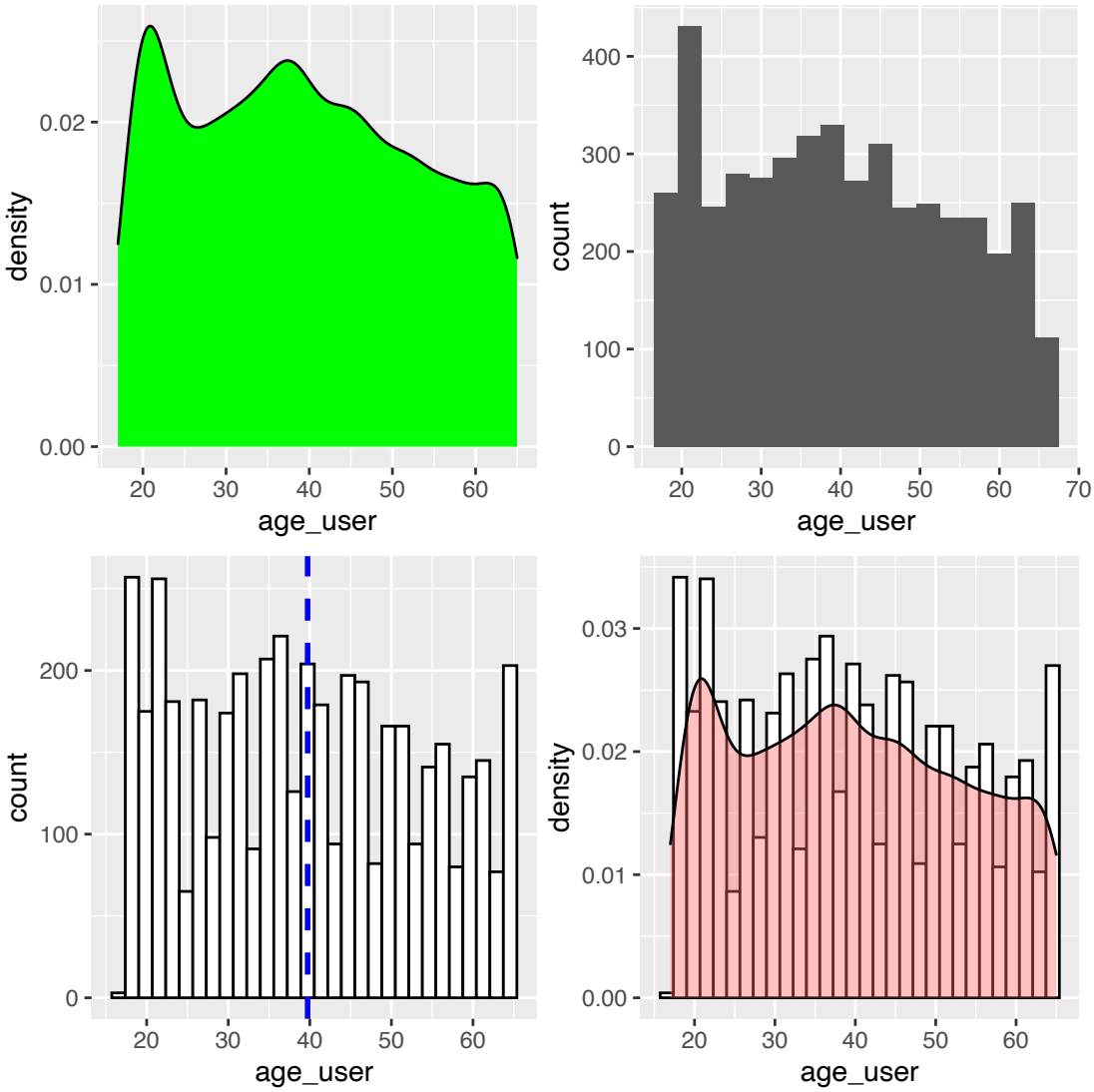
Data Distribution | EDA Continued

Variable 1: One-way Feature Analysis of "User Age"

After exploring the descriptive statistics of the dataset, it made sense to focus on the data distribution and structure of some of the key variables in the larger dataset. It was also helpful to explore the structure of some of the demographic variables in the dataset. Below we have grouped five demographic variables as "key columns of interest", doing this give a good overview of the data.

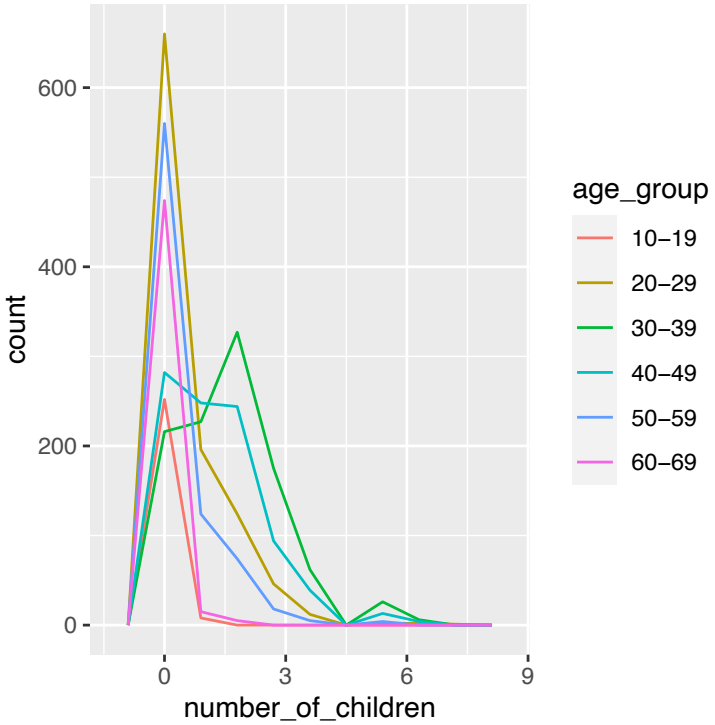
Table 1: Columns of Interest

age_user	gender_user	number_of_children	tenure	disability
22	female	0	Owner-occupier	Not disabled
43	female	2	Social tenant	Not disabled
49	male	2	Owner-occupier	Not disabled
61	male	0	Social tenant	ESA only
35	female	1	Private tenant	Not disabled
25	male	0	Social tenant	Not disabled



Despite the user age data peaking slightly at the age 20 mark, the density graph showed us that the "user age" of benefits is fairly spread out. Overall there is really not any age over or underrepresented. In other words the data is "uniformally distributed". The histogram plots confirm this assumption. We also plotted a histogram which displays the mean (age = 40) for user age values which provides a good benchmark for understanding the variation in the data. Further analysis was conducted on the "user age" variable. In addition to this we also created a new variable, **"age_group"** in which users were be placed into respective bins based on their age. Users were placed into bins based on their age, with the intervals between **17** as the minimum age and **65** as the maximum in the sample set. A function was created to sequence the age data and fill each value to a correct bin.

We can see now see the data variation for users who have **multiple children**. The results are fairly intuitive as the age group, 30-39, is associated with the most number of children, however these insights are still useful nonethless.



Two-Way Feature Analysis: Categorical and Continous Variables

A brief examination was also done of the some of the other important demographic variables in the dataset. This included, **gender of user, number of children, household type, tenure and disability**. The results of which are highlighted below.

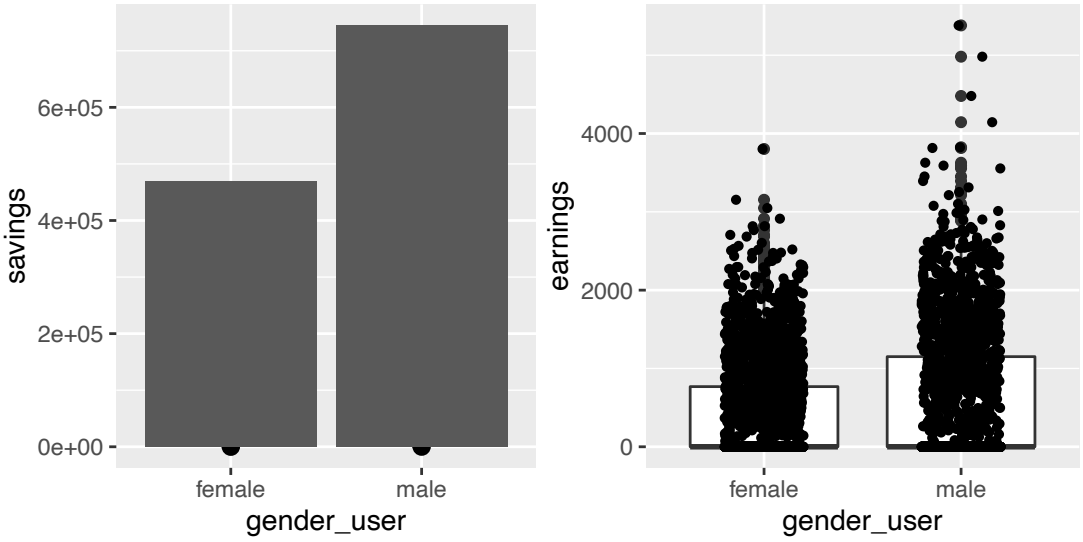
```
##      number_of_children           householdtype         tenure   disability  n
##   1:                  0 Couple without children Owner-occupier  DLA and ESA 17
##   2:                  0 Couple without children Owner-occupier     DLA only 10
##   3:                  0 Couple without children Owner-occupier     ESA only 19
##   4:                  0 Couple without children Owner-occupier Not disabled 80
##   5:                  0 Couple without children Private tenant  DLA and ESA  6
##  ---
## 119:                  6    Couple with children Owner-occupier Not disabled  1
## 120:                  6    Couple with children Private tenant Not disabled  4
## 121:                  6    Couple with children  Social tenant     DLA only  1
## 122:                  6    Couple with children  Social tenant Not disabled  6
## 123:                  7    Couple with children  Social tenant Not disabled  1
```

What is the relationship between gender/age and earnings/savings? The relationship between gender the welfare state is a complex topic, it is beyound the scope of this report to go into detailed analysis. Although it was examined whether there is a link between "gender" and "earnings". The boxplots below highlight the data distribution for male and female earnings. The plots indicate that on average males earn

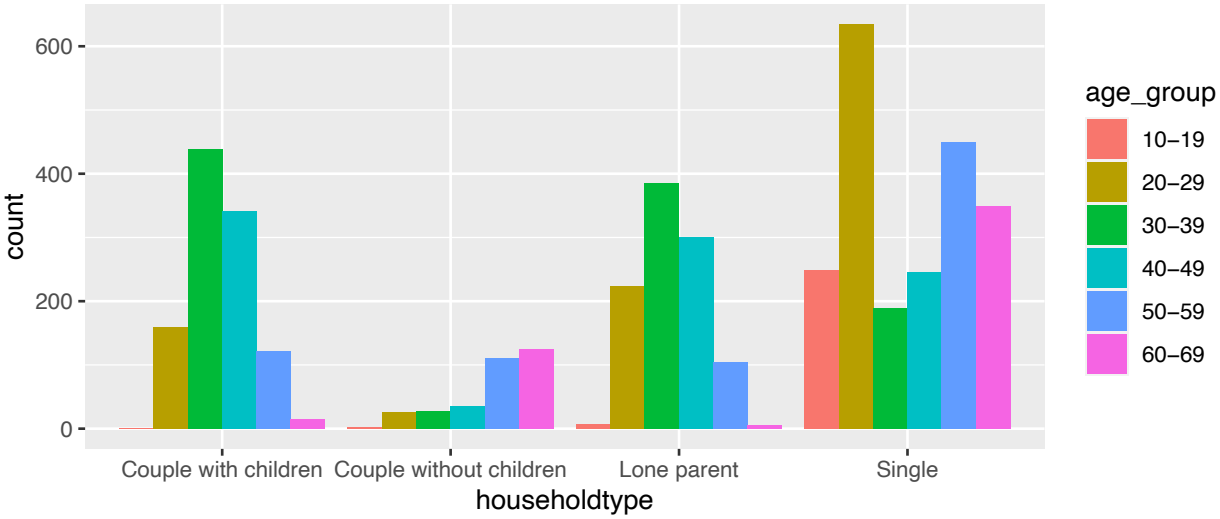
Table 2: Gender Split

Var1	Freq
female	2402
male	2143

slightly more and have greater savings than females. Although this may be due to the effect of some outlier values.



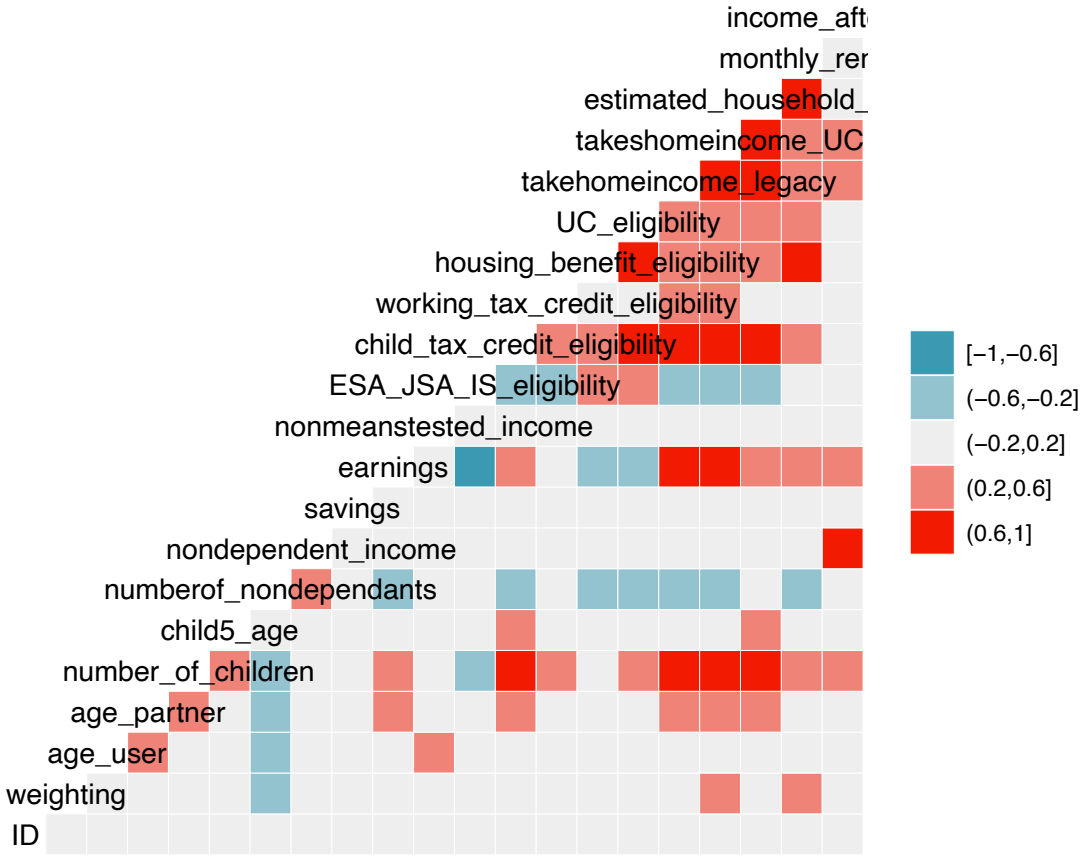
What is the data distribution at the household level?



The data distribution at the household level indicates that most welfare users in the sample set are "single", with the largest proportion of welfare users (this could include but not limited to universal credit, housing benefit, child care benefit) being single 20-29 year olds. At the other end of the scale, perhaps more intuitivley, couples without children are far less represented in the dataset.

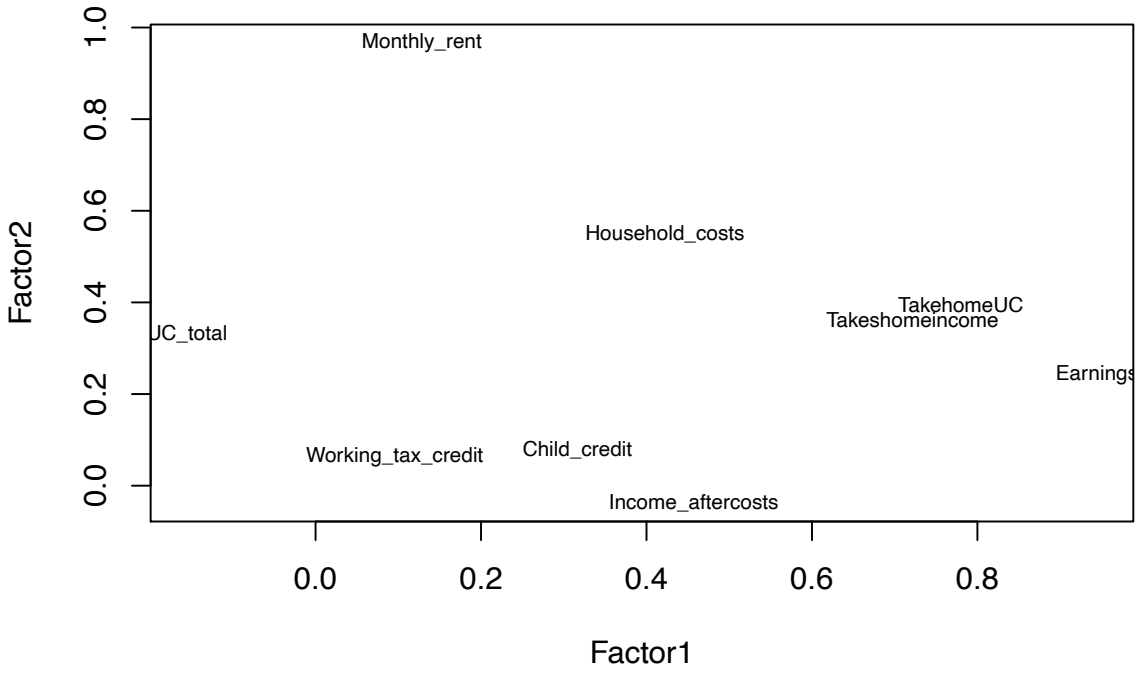
Feature Selection: Covariation and Correlation

If variation describes the behavior within a variable, covariation describes the behavior between variables. Covariation is the tendency for the values of two or more variables to vary together in a related way. The best way to spot covariation is to visualise the relationship between two or more variables. Before visualising the relationships between certain features in this dataset, we will first explore the correlation levels in the dataset. The correlation plot below highlights the highly correlated variables in the dataset.



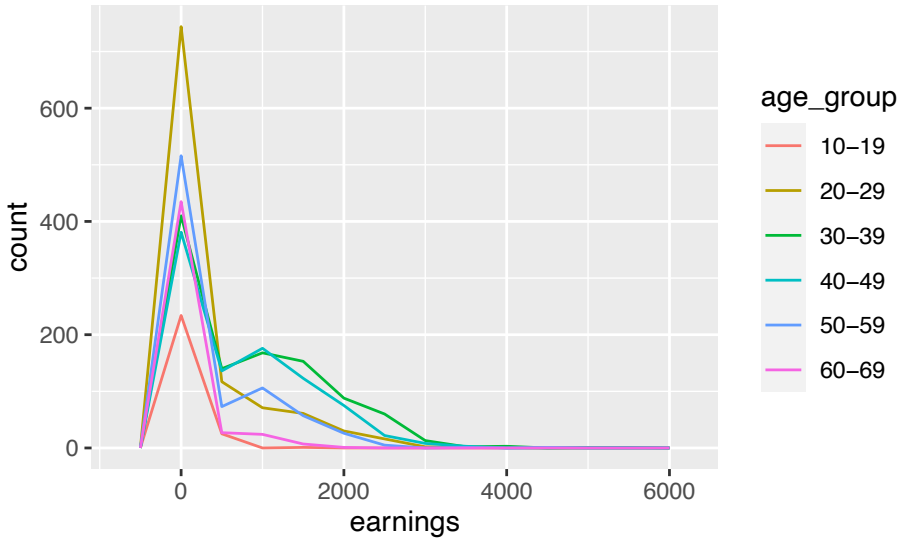
There are a number of intituitive findings we can take from the correlation plot. As expected more children is correlated with higher household costs, but having more children is also correlated with higher housing benefit.

Factor Analysis A simple Exploratory Factor Anlaysis model was deployed. The rationale for doing this being that it allows us to reduce the dimensionality of the data, which is vital for selecting features for the later models.



Analysis of the Continous variables | Continued

Using the correlation plot we can now begin to drill down on the continous variables in the dataset. This includes understanding whether benefit payments are evenly distributed across the different demographic segments as shown below.



Two Continous Variables

The first thing to try and understand is whether selected welfare payments cover or match a users expenses. It is beyound the scope of this anlaysis to quantify these relationships, but visualising the data is likley to provide us with a high-level understanding of how effective the welfare state system.

A number of the continous variables were analysed, but the main thing to establish was whether there was a relationship between costs and **rental vs benefit and credit eligibility**. As seen in the scatterplot above, an increase in household costs is (as expected) associated with an increase in univeral credit (designed to help with costs and to replace household benefit). Moreover, there is a slight variation between those struggling with debt and those not, particularly at the higher cost levels.

Hypothesis

Why are some users struggling to manage debt payments?

For most welfare users, household and monthly rental costs are likley to form the largest share of their expenditures. However, the structure of the household (single, couple, single with children), the type of tenure they occupy (private, social, landlord), and of course the number of children they have also likley to significantly influence a users ability to navigate the welfare state and manage their personal finances.

Some key questions to consider are:

- Is their monthly rent & household costs too high?
- Are they are a single parent?
- Do they recieve enough in benefit compared to their earnings? (E.g. Is the benefits system working as it should be)

Table 3: Monthly rent for users unable to manage debt

mean	min	max
415.488	0	3014.4

First it is useful to see which segments are more likley to struggle with debt. We do this by visualising the variation at the household structure level as well as the type of tenure a user occupies.

Evidently, users who are **social tenants** (occupy state housing) are far more likley to struggle with debt management than those in private accomodation. On a similar level, welfare users who are "single parents" also tend to struggle with debt payments more frequently.

Is there variation in the average rental and household costs between those managing debt payments and those not? To get a better idea, we summarised the mean values for the different costs and benefits by whether or not a user was managing debt payments.

The tables above capture the relationship between monthly rental costs and the amount a user is eligible for in housing benefit and their ability to manage debt payments. The results are important for two reasons:

- 1. It indicates that on average monthly rent for those who struggle with debt payments is substantially higher than the rent for those who can manage debt.
- 2. That users who struggle with managing debt, on average tend to receive more (annually) in state housing benefit than their counterparts.

Table 4: Monthly rent for users able to manage debt

mean	min	max
320.0142	0	3813

Table 5: Housing benefit eligibility for users unable to manage debt

mean	min	max
242.6732	0	1688.78

As shown below we are also able to see a breakdown of the demographic segments for rental and household costs.

tenure	gender_user	average_monthly_rent	housing_costs
Owner-occupier	female	137.29876	919.4248
Owner-occupier	male	88.30836	908.0669
Private tenant	female	621.48022	1414.0479
Private tenant	male	602.30926	1379.3855
Social tenant	female	433.01609	1246.9743
Social tenant	male	346.22252	1028.8865

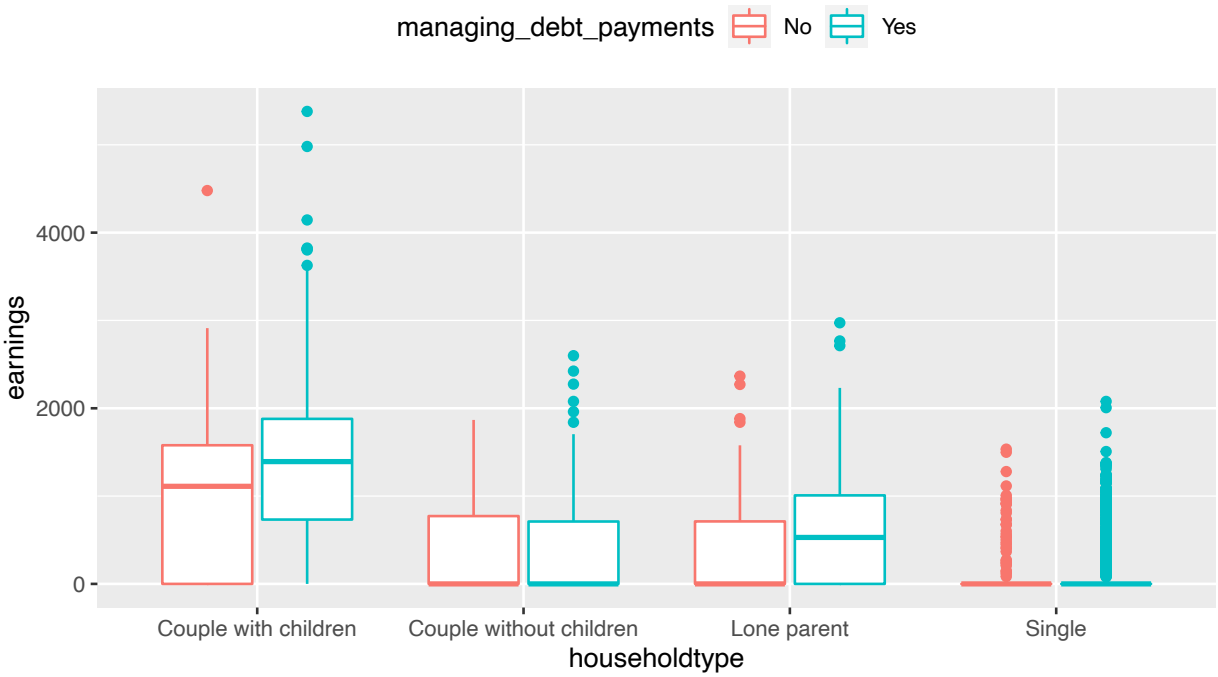
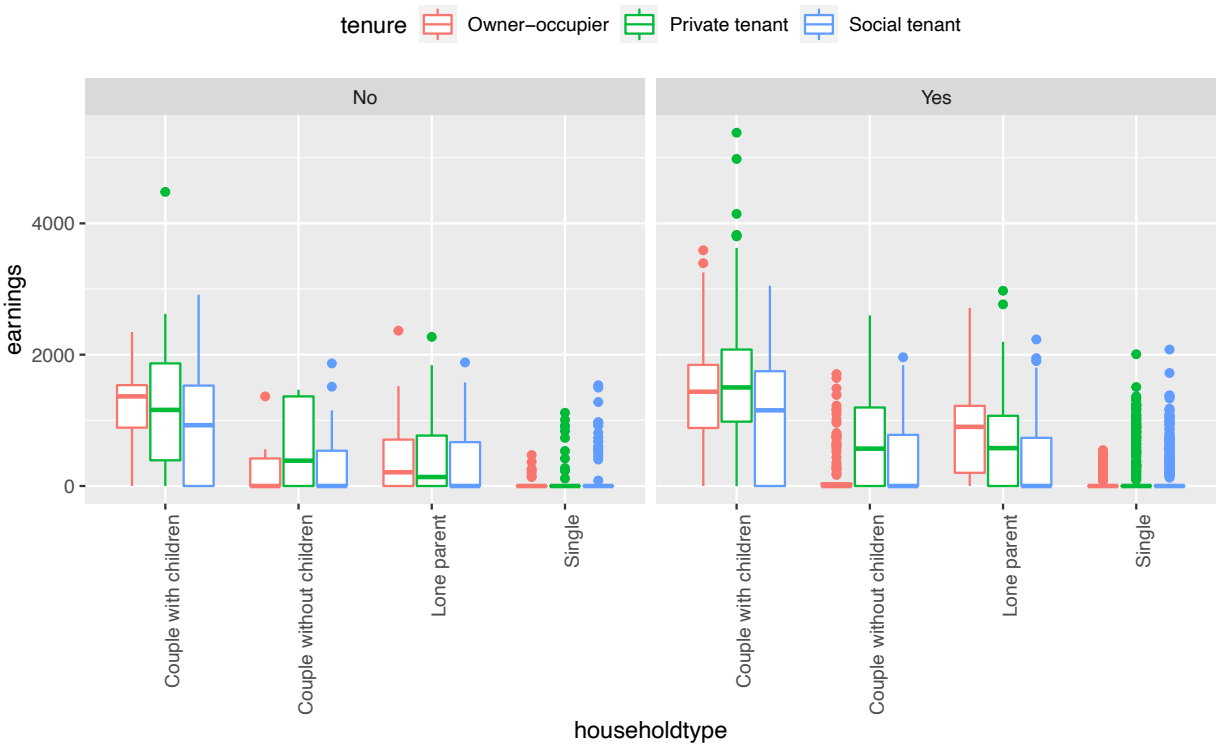


Table 6: Housing benefit eligibility for users able to manager debt

mean	min	max
156.4841	0	2059.374

Table 7: Comparing Means of Different Features

managing_debt_payments	avg_costs	sd_costs	avg_inc_pc	sd_inc_pc	avg_earnings	sd_earnings	avg_children	sd_children	n
No	1192.906	549.8140	327.7649	526.5107	407.7422	639.0823	1.0560272	1.233842	589
Yes	1122.657	538.4658	400.9630	780.8759	504.3335	738.6914	0.8895349	1.202897	3956



As we are interested in how values for a given variable differ between two groups. For example, does the size of household costs incured by a user affect the users ability to manage to debt payments. The summary statistics are given above, which suggest users that struggle with debt may, on average, have larger household and rental costs than users that do not struggle with debt. This is all good, but there are likley to be several factors that influence debt problems. Whether it is the size of a users earnings, or the number of a children a household has, there are clearly a number of significant variables.

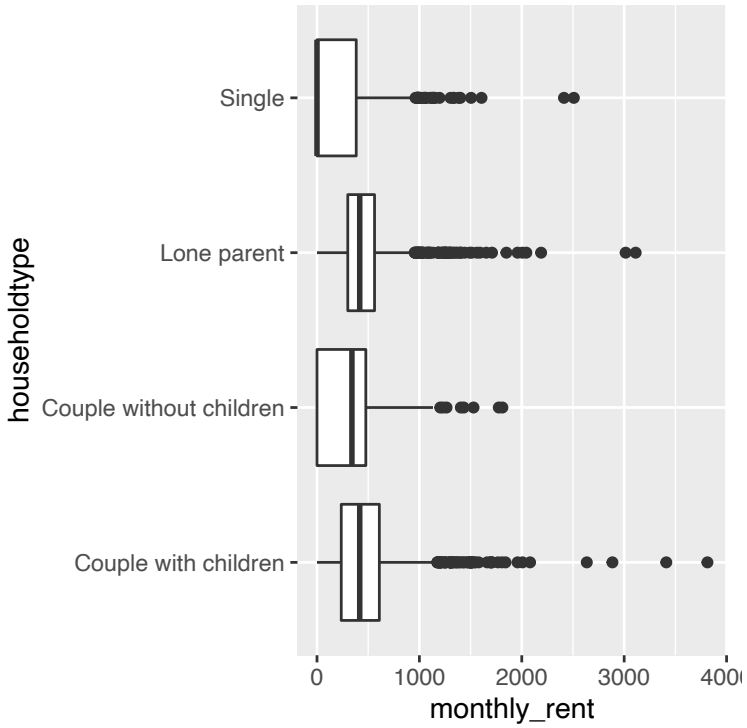
However, the question remains, how certain are we that there is a difference in distances between the two user groups? Assuming this data is a sample of a larger population data set,there is some uncertainty (margin of error) around the estimated mean produced by a sample. If the margin of error is large then we cannot state that any differences in mean values between these groups are statistically significant. Therefore we would need to test for the difference in group means by using the two-sample test statistic.

```
##
##  Welch Two Sample t-test
##
## data:  monthly_rent by managing_debt_payments
## t = 6.7354, df = 803.04, p-value = 3.116e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   67.64939 123.29834
## sample estimates:
```

Table 8: Comparing Means of Three or More Groups

householdtype	avg_rent	sd_rent	avg_cost	sd_cost	avg_house_credit	sd_house_credit	n
Couple with children	472.8876	391.5399	1676.2240	502.8267	144.4956	245.6102	1076
Couple without children	322.7532	323.8592	1034.1228	418.8108	165.9471	216.7280	328
Lone parent	449.1574	333.1886	1289.5378	439.4214	253.6323	252.3250	1025
Single	205.8706	258.8338	793.6043	315.5226	138.0457	189.3964	2116

```
##  mean in group No mean in group Yes
##          415.4880           320.0142
```



We can turn to analysis of variance (ANOVA) to assess these hypotheses.

```
##                  Df    Sum Sq  Mean Sq F value Pr(>F)
## householdtype     3  69116895 23038965   230.2 <2e-16 ***
## Residuals      4541 454471885   100082
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Evaluation and Conclusions

We have seen the difficulty in distinguishing what influences debt managment in the welfare state system. The statistical tests that were run attempted to demonstate the variation in the experiences that users who struggle with debt have compared to user which do not. However, with no real certainty can we claim statistical significance, as this stage more modelling and more accurate feature seleciton is required which is beyound the scope of this analysis.