Data Analytics on the 2011 UK Census

Intelligent Data and Text Analytics

Coursework 1

Student Number: 2089114

Course: MSc Data Analytics

Year: 2024

Word count: 2,994

Table of Contents

This report will analyse the 2011 UK Census, a nationwide survey carried out across the United Kingdom to collect data about members of the population and their characteristics. This census provides rich comprehensive data about the population. Data analysis will be performed on this census on various levels, through means of classification, regression, association rule mining and clustering. To become familiarised with the data, see Table 1.

Demographic Information

Table 1.

Basic statistics for categorical data

Variable	Mode	Frequency		Percentage	
Region	South-East	South-East	88084	South-East	15.460385
		London	83582	London	14.670200
		North-West	71436	North-West	12.538351
		East of England	59411	East of England	10.427739
		West Midlands	56875	West Midlands	9.982624
		South-West	53774	South-West	9.438340
		Yorkshire and the Humber	53471	Yorkshire and the Humber	9.385158
		East Midlands	45782	East Midlands	8.035595
		Wales	30976	Wales	5.436866
		North-East	26349	North-East	4.624741
Residence Type	Resident in a communal establishment	Resident in a communal establishment	559086	Resident in a communal establishment	98.130024
		Not a resident in a communal establishment	10654	Not a resident in a communal establishment	1.869976
Sex	Female	Female	289172	Female	50.755081
		Male	280568	Male	49.244919
Student	No	No	443203	No	77.790396
		Yes	126537	Yes	22.209604
Country of Birth	UK	UK	485645	UK	85.239758
		Non-UK	77291	Non-UK	13.566013
		No code required	6804	No code required	1.194229
Religion	Christian	Christian	333481	Christian	58.532137
		No religion	141658	No religion	24.863622
		Not stated	40613	Not stated	7.128339
		Muslim	27240	Muslim	4.781128
		Hindu	8213	Hindu	1.441535
		No code required	6804	No code required	1.194229
		Sikh	4215	Sikh	0.739811
		Jewish	2572	Jewish	0.451434
		Buddhist	2538	Buddhist	0.445466
		Other religion	2406	Other religion	0.422298
Family Composition	Married/Equivalent	Married/equivalent	300961	Married/ equivalent	52.824271
		Not in a family	96690	Not in a family	16.970899
		Cohabiting	72641	Cohabiting	12.749851
		Lone Parent Family (female head)	64519	Lone Parent Family (female head)	11.324288
		No code required	18851	No code required	3.308702
		Lone Parent Family (male head)	9848	Lone Parent Family (male head)	1.728508
		Other related family	6230	Other related family	1.093481
Population Base	Usual Resident	Usual Resident	431868	Usual Resident	98.091403

		Student living away from home during term-time Short-term resident	6730 1673	Student living away from home during term-time Short-term resident	1.528604 0.379993
Age	0-15	0-15 35-44 45-54 25-34 16-24 55-64 65-74 75+	106832 78641 77388 75948 72785 65665 48777 43704	0-15 35-44 45-54 25-34 16-24 55-64 65-74 75+	18.751009 13.802963 13.583038 13.330291 12.775125 11.525433 8.561274 7.670867
Marital Status	Single	Single Married Widowed Divorced Separated but legally married	221084 151210 30430 29285 8262	Single Married Widowed Divorced Separated but legally married	50.215435 34.344756 6.911652 6.651585 1.876571
Health	Very good health	Very good health Good health Fair health Bad health Very bad health No code required	198777 140516 64163 23137 6874 6804	Very good health Good health Fair health Bad health Very bad health No code required	45.148783 31.915797 14.573524 5.255172 1.561311 1.545412
Economic Activity	No code required	No code required Employee Retired Student Self-employed Unemployed Long-term sick/disabled Looking after home/family Full-time student Other	112618 109049 97480 24756 19538 18109 17991 17945 12717 10068	No code required Employee Retired Student Self-employed Unemployed Long-term sick/disabled Looking after home/family Full-time student Other	25.579246 24.768608 22.140909 5.622900 4.437721 4.113148 4.086347 4.075899 2.888448 2.286773
Occupation	No code required	No code required Elementary Administrative/Secretarial Professional Sales and Customer service Skill Trades Caring, Leisure and Other Service Associate Professional and Technical Process, Plant, Machine Operatives Managers/Directors/ Senior Officials	149984 48140 40886 37790 32291 31190 28919 26039 23599 21433	No code required Elementary Administrative/Secretarial Professional Sales and Customer service Skill Trades Caring, Leisure and Other Service Associate Professional and Technical Process, Plant, Machine Operatives Managers/Directors/Senior Officials	34.066291 10.934175 9.286553 8.583350 7.334346 7.084273 6.568454 5.914312 5.360108 4.868138
Industry	No code required	No code required Wholesale and retail trade Mining Human health + social work Real estate Education Transport + storage Accommodation and food service Construction	149984 51473 37759 35658 32146 29555 21381 20184 18800	No code required Wholesale and retail trade Mining Human health + social work Real estate Education Transport + storage Accommodation and food service Construction	34.066291 11.691208 8.576309 8.099103 7.301412 6.712911 4.856327 4.584449 4.270097

		Public administration	16347	Public administration	3.712940
		Other service activities	14728	Other service activities	3.345212
		Financial and insurance	9829	Financial and insurance	2.232489
		Agriculture/forestry/fishing	2427	Agriculture/forestry/fishing	0.551251
Hours Worked Per Week	No code required	No code required	302321	No code required	68.667026
		31-48 hours (full time)	60041	31-48 hours (full time)	13.637282
		16-30 hours (part time)	52133	16-30 hours (part time)	11.841116
		<=15 hours (part time)	25776	<=15 hours (part time)	5.854576
Approximated Social Grade	No code required	No code required	122855	No code required	27.904404
		C1	116234	C1	26.400558
		DE	104003	DE	23.622496
		C2	51281	C2	11.647599
		AB	45898	AB	10.424943

The key insights into this reveal the following, majority of the sample:

- Reside in South-East of the UK, while the North-East is the least represented.
- Live in communal establishments (hospitals, care homes. Prisons, defence bases, boarding schools and student’s halls of residence).
- Are females (only slightly more than males)
- Are 0-15 years of age (other age groups are evenly distributed)
- Not students
- Of Christian faith
- Most of the sample are born in the UK (‘No code required’ here may refer to individuals in communal establishments where information such as country of birth was not recorded, or simply missing data)
- In a married/equivalent family composition.
- Primarily usual residents
- Single or married
- Very good/good health
- Economic activity is distributed similarly across ‘No code required,’ ‘Employee,’ and ‘Retired,’ (‘No code required’ in this context may refer to those with no jobs or source of income, such as full-time students, retirees or volunteering/unpaid work)
- Reported ‘No code required’ for their occupation – possibly unemployed/students/children/dependent individuals or missing data.
- Reported ‘No code required,’ for industry information – possibly unemployed individuals/retirees/students or simply missing data
- Reported ‘No code required’ for the ‘Hours-worked-per-week’ variable - possibly due to people not in the working force or missing data.
- Approximated Social Grade responses were distributed somewhat evenly across ‘No code required,’ C1 and DE, referring to lower-middle to working class population.

Table 2.

Basic statistics for numerical variable

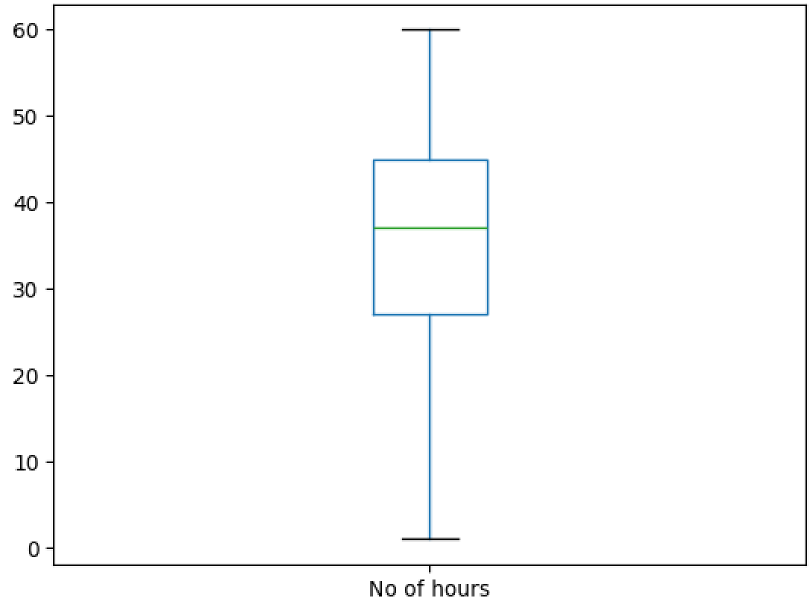
	Mean	Standard Deviation	Minimum Value	Lower Quartile (25%)	Middle Quartile (50%)	Upper Quartile (75%)	Maximum value
Number of Hours	35.23	13.52	1.00	27.00	37.00	45.00	60.00

To visualise this, see Figure 1.

Figure 1:

Boxplot demonstrating the number of hours worked by respondents

<Axes: >



As expressed in Table 2, on average, respondents worked approximately 35 hours a week. A minimum value of 1 implies little working hours, likely due to part-time or volunteering roles. 25% of the respondents work 27 hours or less, 50% work at least 37 hours a week, 75% work 45 hours or less. Maximum value of 60 hours indicates some are working substantial longer hours than the average, likely a full-time position. A large range of 59, suggests the census includes individuals working various hours.

Figure 2:

Contingency table

Occupation	Female	Male
Administrative and Secretarial	42636	10618
Associate Professional and Technical	18999	25938
Caring, Leisure and Other Service	30872	6425
Elementary	30731	27752
Managers/Directors/Senior Officials	14473	25315
No code required	76619	73365
Process, Plant and Machine Operatives	7103	27714
Professional	33431	30680
Sales and Customer Service	26853	11670

Skill Trades	7455	41091

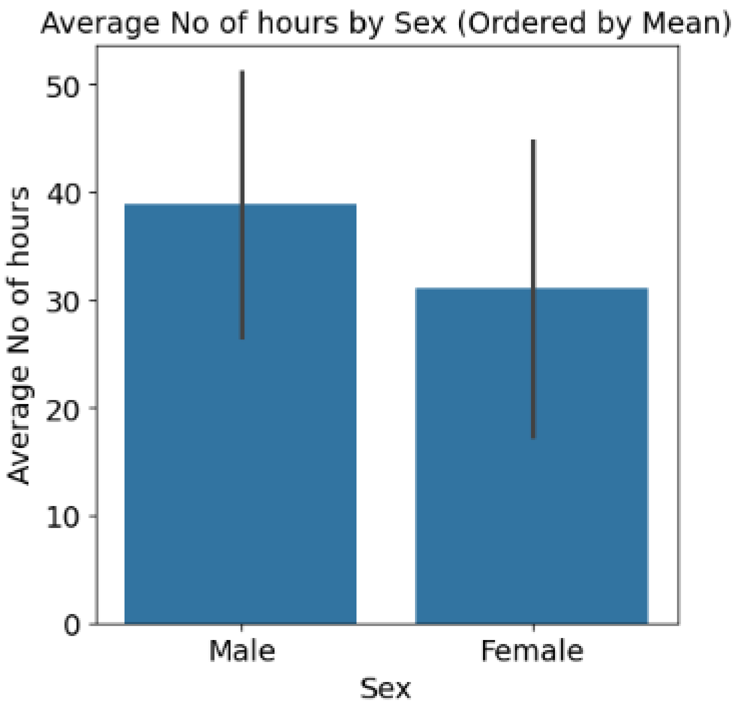
Figure 2 reveals the occupation of caring, leisure and other services is dominated by female workers. Similarly, in administrative/secretarial roles, as well as sales and customer service with greater female workers. Some occupations are largely male dominated, specifically roles in Managers/Directors/Senior Officials, Process, Plant and Machine Operative roles and jobs in Skill Trades.

A chi-square test of independence yielded a p-value of 0.0, suggesting a strong significant relationship between Occupation and Sex.

Given these findings, it raised the question if males tend to work longer hours than females, see Figure 3:

Figure 3.

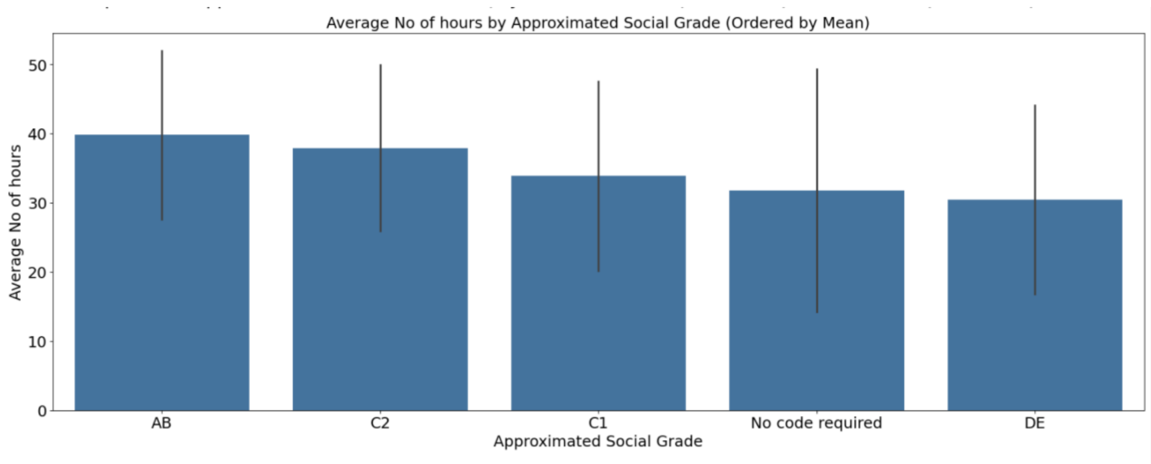
Average no of hours by sex



On average, males tend to work greater hours than females, error bars for females are larger than errors bars for males, showing there is more variability in hours worked for females than for males. This suggests females may typically be more involved in jobs requiring less strict shift hours, such as part-time and flexible jobs.

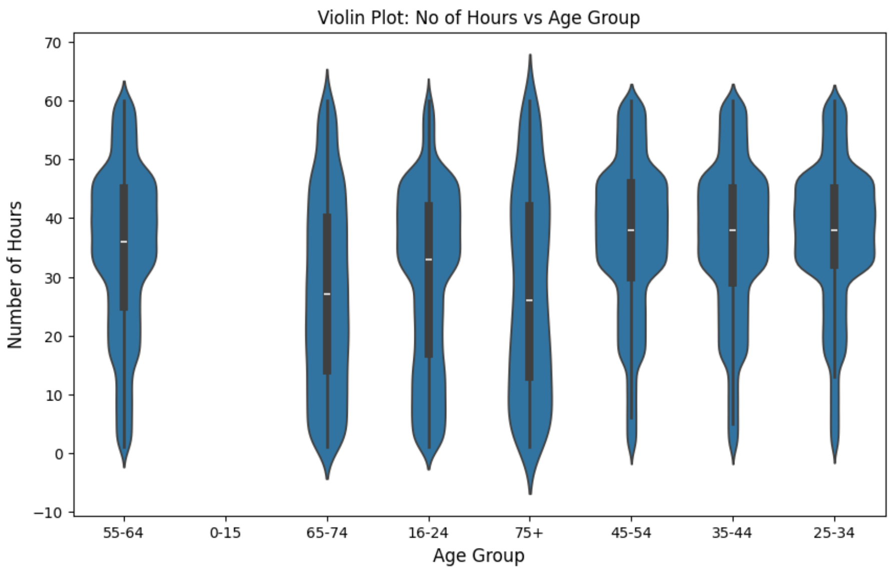
Figure 4:

Bar chart representing average weekly hours worked against social grades (AB = Higher & intermediate managerial, administrative, professional occupations, C1 = Supervisory, clerical & junior managerial, administrative, professional occupations, C2 = Skilled manual occupations, DE = Semi-skilled & unskilled manual occupations, Unemployed and lowest grade occupations)



Majority of the sample belong to AB and C2 social grades, dedicating around 40 hours of work per week. C1 has a slightly lower average than C2, while 'No code required' and DE have the lowest average hours worked per week. The error bars in C1 and DE are slightly larger than that of AB and C2, suggesting a wider range of work hours within C1 and DE, whilst more demanding jobs in AB and C2 appear to have less variability, suggesting more consistency in hours worked in this social grade.

Figure 5:



Ages 25-64 have the most consistent work patterns distributing consistently around 30-40 hours a week. The median, which typically lies between 30-40 hours emphasises this. There is more variability among groups such as 16-24, 65-74 and 75+, suggesting more diverse lifestyles, with some working very few or many hours, including retired individuals, students and people working part-time.

Classification

Classification was performed on the "Approximate Social Grade" variable, consisting of 5 categories, see Table 1. Naïve Bayes, K-Nearest Neighbour (KNN) and Logistic Regression classified the "Approximate Social Grade" variable and can be interpreted by its precision, recall and F1-score.

Table 3.

Classification with Naïve Bayes on Approximate Social Grade

	Precision	Recall	F1_score
C1	0.30	0.05	0.09
C2	0.31	0.21	0.25
DE	0.24	0.05	0.08
No code required	0.29	0.96	0.45
AB	1.00	0.19	0.32

C1 and DE both have low precision, recall and F1 scores, suggesting the Naïve Bayes struggles classifying these groups. C2 shows stronger recall and F1, indicating better classification. Classification in the 'No code required' group revealed a high recall and F1-score, suggesting more accurate classifications to this group, but a low precision suggests there are possible false positives. Although there is a perfect precision score for AB, it's lower recall suggests misclassifications for individuals in this group.

Table 4

Classification with KNN on Approximate Social Grade

	Precision	Recall	F1_score
C1	0.43	0.60	0.50
C2	0.61	0.65	0.63
DE	0.46	0.37	0.41
No code required	0.55	0.46	0.50
AB	0.97	0.92	0.94

KNN performs exceptionally well in accurately classifying individuals to the AB group with high F1-Score, precision and recall. Accurate classification to C2 group is moderate with reasonable precision and recall, whereas classifying people to C1 and 'No code required' show difficulty in correct classifications with lower F1 scores, precision and recall. Results for DE highlight the weakest performance for the KNN classifier.

Table 5

Classification with Logistic Regression on Approximate Social Grade

	Precision	Recall	F1_score
C1	0.49	0.36	0.41
C2	0.65	0.76	0.70
DE	0.52	0.36	0.43
No code required	0.52	0.60	0.55
AB	1.00	1.00	1.00

Logistic Regression classifier demonstrates perfect classification in the AB group, identification of C2 group also performs very well. Identifying individuals in 'No code required' displays moderate performance, although allocations to C1 and DE show weaker performance with lower F1-Scores, recall and precision.

Naïve Bayes' precision, recall and F1-score are reportedly low for categories C1, DE and AB; it may have a high recall for 'No code required' category, but the KNN still dominated for the AB group with high precision, recall and F1. Nevertheless, KNN struggles classifying individual's into C1 and DE social grades. Logistic Regression demonstrates perfect classification for the AB group, moderate performance for C2 and 'No code required,' but it struggled with C1 and DE. Deciding if Logistic Regression or KNN is the better model here is difficult, hence additional metric analysis was conducted.

<u>Additional Metric Analysis</u>

Table 6.

Mean Absolute Error (MAE) values for each classifier against CM and CV:

	KNN	**Logistic Regression**
Cross Validation	0.575	0.585

KNN is slightly more reliable with fewer misclassifications, only slightly outperforming Logistic Regression

Table 7.

Mean Squared Error (MSE) values for each classifier against CM and CV:

	KNN	**Logistic Regression**
Cross Validation	1.179	1.198

Both KNN and Logistic Regression are better at estimating social grades of individuals to their actual social grade, KNN only slightly performs more reliably than Logistic Regression.

Table 8.

Root Mean Squared Error (RMSE) values for each classifier against CM and CV:

	KNN	**Logistic Regression**
Cross Validation	1.086	1.095

Approximated social grade predictions vary more from actual grades with Logistic Regression while the KNN varies the least.

Table 9.

Accuracy values for each classifier against CM and CV

	KNN	**Logistic Regression**
Cross Validation	0.660	0.663

Logistic Regression only slightly outperforms KNN in accurately classifying social grades.

Table 10.

Area Under the Curve (AUC) values for each classifier against CM and CV

	KNN	**Logistic Regression**
Cross Validation	0.754	0.754

Both classifiers have the same AUC of 0.754 (across CM and CV), implying each classifier has a similar ability in distinguishing between social grades, regardless of their differences in other performance.

Overall, Naïve Bayes performs the worst, so it is not suitable as a model in predicting social grade levels. KNN produces the lowest measures in MAE, MSE and RMSE, thus if the primary objective is to minimise error rates and ensure more reliable predictions of approximated social grade levels, KNN is the ideal candidate due to its low error metrics. Although, Logistic Regression achieves the highest accuracy, so if the priority is accuracy is predicting social grades, Logistic Regression will be most preferable. Although Logistic Regression outperforms KNN in accuracy by a very little amount, KNN performs just as well.

As a result, KNN is the most suitable classifier for predicting Approximate Social Grade

Regression

Regression was performed on the numeric variable "Number of Hours", referring to the number of hours worked per week. Linear Regression (LR) and Regression Tree (RT) were applied. The strength of. the model's predictions were evaluated by the MAE, MSE, RMSE, but also R^2 score, and adjusted R^2.

Table 11.

Comparison of metrics across LR and RT as a predictor of 'Number of Hours' (rounded to 3 d.p)

	Linear Regression	Regression Tree
Mean Absolute Error	0.530	0.397
Mean Squared Error	0.539	0.462
Root Mean Squared Error	0.734	0.680
R^2	0.134	0.257
Adjusted R^2	0.134	0.257

RT has a lower MAE (0.397) than LR (0.530), suggesting LR makes fewer errors in predictions on average. It further outperforms LR, as seen in the MSE, where RT has a value of 0.462 as opposed to LR with a value of 0.539. The RMSE for RT (0.680) is smaller than that of the LR model (0.734), indicating fewer errors in predicting the number of hours individuals work. An R^2 value of 0.257 for RT means it can explain 25.7% of the variance in number of hours worked among individuals, while LR can only account for 13.4% of this variance. Matching adjusted R^2 values for both models suggest there was no overfitting.

In comparison, the RT consistently outperformed LR across all the metrics, with lower error rates in MAE, MSE and RMSE, highlighting that its predictions of hours worked per week are closer to the actual values. It explains more variance in number of hours worked across individuals than LR. A possible explanation may be that RT is able to better understand non-linear relationships, while LR's assumption of better suitability to linear data demonstrated its limits to non-linear data.

Nevertheless, both models do have relatively low R^2 values, the RT may outperform LR but there is still 74.3% of variance that is unexplained by RT, and 86.6% for LR. Neither provide a comprehensive overview in predicting the Number of Hours variable, so improvements are required.

Association Rule Mining

Association rule mining was conducted on the entire sample and all attributes.

Table 12:
Rule 1:

Items	Antecedent	Consequent	Support	Confidence	Lift
Not a Student, Married, UK, Married/equivalent, Not a Resident in communal establishment, White, Usual Resident	Married, Not a Resident in communal establishment, UK, Usual Resident	Not a student, White, Married/equivalent	0.294	0.946	2.581

Rule 1 shows 29.4% of the respondents are not a student, married, born in the UK, family dynamic is married/equivalent, not a resident in a communal establishment, ethnically White and a usual resident. If they are married, living outside a communal establishment, from the UK and a usual resident, it is 94.6% likely they are not a student, of White ethnicity and family composed as married/equivalent. A very high lift value strengthens the relationship between antecedent and consequent.

Table 13:
Rule 2:

Items	Antecedent	Consequent	Support	Confidence	Lift
White, Not a Student, Very good health, Usual Resident	Very good health, White	Not a Student, Usual Resident	0.268	0.685	0.882

Rule 2 implies 26.8% of the population are ethnically White, not a student, very good health and a usual resident. If they are of very good health and white, it is 68.5% likely they are not a student and are a usual resident. A low lift value below 1 implies a lack of relationship.

Table 14:
Rule 3:

Items	Antecedent	Consequent	Support	Confidence	Lift
Not a Student, Very good health, 'Usual Resident	Very good health, Usual Resident	Not a Student	0.309	0.669	0.860

Rule 3 suggests 30.9% of the sample are not a student, with very good health and a usual resident. If they have very good health and a usual resident it is 66.9% likely they are not a student. A slightly low lift value does not support this proposed association.

Table 15:
Rule 4:

Items	Antecedent	Consequent	Support	Confidence	Lift

Not a Student, Usual Resident, Single	Single	Not a Student, Usual Resident	0.255	0.540	0.695

Rule 4 states 25.5% of the population are not a student, a usual resident and single. If they are single, it is 54% likely they are not a student and are a usual resident. A moderately weak lift value of 0.695 indicates a weak association between antecedent and consequent.

Table 16:
Rule 5:

Items	Antecedent	Consequent	Support	Confidence	Lift
(27.0, 37.0, Not a Student, UK	Not a Student	(27.0, 37.0, UK	0.385	0.495	0.892

Rule 5 highlights 38.5% of the sample are aged 27-37, not a student and from the UK. If they are not a student, it is 49.5% likely they are aged 27-37 and from the UK. The lift value does not support any such association as it is below 1.

Table 17:
Rule 6:

Items	Antecedent	Consequent	Support	Confidence	Lift
Married, Not a Resident in communal establishment, Usual Resident, Christian, Not a Student	Married, Not a Student	Christian, Usual Resident, Not a Resident in communal establishment	0.252	0.667	1.162

Rule 6 proposes 25.2% of the population are married, not a resident in communal establishment, a usual resident, Christian and not a student. If they are married and not a student, they are 66.7% likely to be Christian, a usual resident and living outside communal establishment. Lift value strengthens there is an association, given it is above 1.

Table 18:

Rule 7:

Items	Antecedent	Consequent	Support	Confidence	Lift
UK, Christian, Not a Student 'White	UK, Christian	Not a Student, White	0.416	0.797	1.155

Rule 7 suggests 41.6% of the sample are from the UK, Christian, not a student and White. If they are from the UK and Christian, they are 79.7% likely to not be a student and of White ethnicity. Lift value is above 1, therefore, supporting there is a strong association.

Table 19:

Rule 8:

Items	Antecedent	Consequent	Support	Confidence	Lift
Not a Resident in communal establishment, White Not a Student, Good health	White, Good health	Not a Student, Not a Resident in communal establishment	0.260	0.886	1.149

Rule 8 dictates 26% of the population are not a resident in a communal establishment, White, not a student and in good health. If they are White and with good health, they are 88.6% likely to not be a student and living outside communal establishment. Lift value above 1 indicates such a relationship exists.

Clustering

K-Means Clustering algorithm was applied to the entire dataset to define clusters based on similarity, mean values were standardised and mapped back to its corresponding categorical variable.

Table 20a.

K-Means Clustering on all attributes – standardised values

	Region	**Residence Type**	**Family Composition**	**Population Base**	**Sex**
Cluster 1	0.00302147	-0.04711457	0.13004111	0.11670853	-0.01065922
Cluster 2	-0.00671955	0.10477968	-0.2892028	-0.25955202	0.02370541
	Age	**Marital Status**	**Student**	**Country of Birth**	**Health**
Cluster 1	0.49230477	-0.3417143	-0.52570225	0.04234671	-0.2428996
Cluster 2	-1.09485314	0.75994993	1.16912692	-0.09417627	0.5401926
	Ethnic Group	**Religion**	**Economic Activity**	**Occupation**	**Industry**
Cluster 1	0.10712471	-0.10746843	-0.16202918	-0.06688598	-0.00919845
Cluster 2	-0.23823825	0.23900265	0.36034215	0.14874998	0.02045674
	Hours worked per week	**Approximated Social Grade**			
Cluster 1	-0.19584187	-0.37673261			
Cluster 2	0.43553933	0.83782833			

Figure 8.

Visual representation of the mean values per cluster

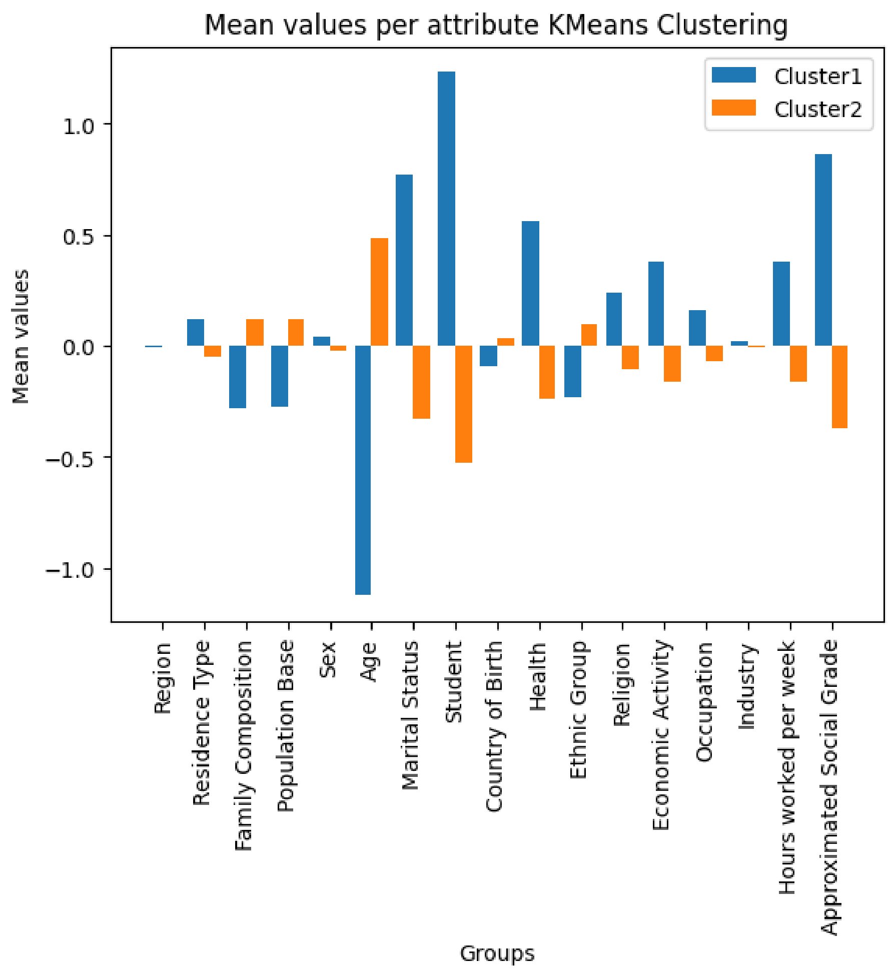


Table 20b.

Standardised values mapped back to its categorical meaning

	Region	**Residence Type**	**Family Composition**	**Population Base**	**Sex**
Cluster 1	Scotland	Resident	Cohabiting	Student living away during term time	Female
Cluster 2	Scotland	Not resident	Married/equivalent	Student living away	Female
	Age	**Marital Status**	**Student**	**Country of Birth**	**Health**
Cluster 1	35-44	Married	Not a student	Not born in the UK	Fair health
Cluster 2	0-15	Separated but legally married	A Student	Not born in the UK	Bad health
	Ethnic Group	**Religion**	**Economic Activity**	**Occupation**	**Industry**
Cluster 1	Chinese/Other ethnic group	Christian	Unemployed	Administrative & Secretarial	Financial & Insurance activities

Cluster 2	Black/Black British	Christian	Full-time student	Skill trades	Financial & Insurance activities
	Hours worked per week	**Approximated Social Grade**			
Cluster 1	Part time (16-30 hours)	C1			
Cluster 2	Part time (16-30 hours)	C2			

Hierarchical clustering was applied to the entire dataset to collate a tree of nested clusters based on similarity, like K-Means clustering, mean values were standardised and mapped back to its corresponding categorical variable.

Table 21a.

Hierarchical Clustering on all attributes

	Region	**Residence Type**	**Family Composition**	**Population Base**	**Sex**
Cluster 1	0.21636134	-0.13804384	1.43985716	0.11776432	0.9850107
Cluster 2	0.93892917	-0.13804384	0.14324745	0.11776432	1.0152174
	Age	**Marital Status**	**Student**	**Country of Birth**	**Health**
Cluster 1	1.81182162	1.6580582	-0.53432737	0.40141031	-0.69325558
Cluster 2	-0.89149957	0.75994993	-0.53432737	0.40141031	-0.69325558
	Ethnic Group	**Religion**	**Economic Activity**	**Occupation**	**Industry**
Cluster 1	0.39542183	-0.77609871	0.94360325	-0.97544702	-0.6033569
Cluster 2	0.39542183	1.91071878	-1.09361447	1.6757684	-1.521950
	Hours worked per week	**Approximated Social Grade**			
Cluster 1	0.24653177	0.6579581			
Cluster 2	-0.91048077	-0.06006826			

Figure 9.

Visual representation of the mean values per cluster

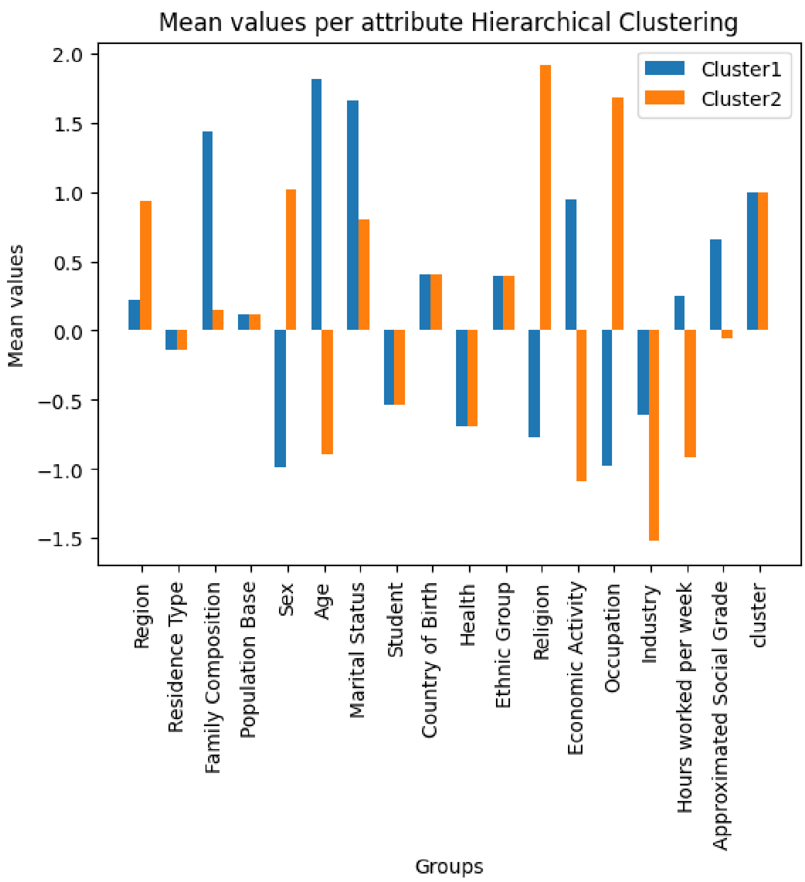


Table 21b.

Standardised values mapped back to its categorical meaning

	Region	Residence Type	Family Composition	Population Base	Sex
Cluster 1	Wales	Resident in communal establishment	Cohabiting	Student living away during term time	Female
Cluster 2	West Midlands	Not resident	Cohabiting	Student living away during term time	Female
	Age	**Marital Status**	**Student**	**Country of Birth**	**Health**
Cluster 1	65-74	Divorced	Not a student	Not born in the UK	Good health
Cluster 2	0-15	Separated but legally married	A student	Not born in the UK	Good health
	Ethnic Group	**Religion**	**Economic Activity**	**Occupation**	**Industry**

Cluster 1	Chinese/Other ethnic group	No religion	Student	Professional	Accommodation and food service activities
Cluster 2	Chinese/Other ethnic group	Sikh	No code required	Elementary	Mining

	Hours worked per week	**Approximated Social Grade**
Cluster 1	Part time (16-30 hours)	C2
Cluster 2	Part time (<=15 hours)	C1

Key interpretations and Comparison

Demographics

In K-Means, Cluster 1 represents the population that are mainly aged 35-44, married and in cohabiting arrangements, while Cluster 2 describes the sample is predominantly aged 0-15 years, typically separated but legally married and are students.

In Hierarchical clustering, age is more differentiated, Cluster 1 highlights 65-74 year olds, likely divorced, while Cluster 2 consists of those aged 0-15 years, particularly full-time students.

Hierarchical clusering offers a more complex and nuanced explanation for the variable relationship across age groups, suggesting age is not evenly distributed throughout clusters but entwined with other attributes such as marital status and job type.

Health

Both clustering models show contrasting trends of health status. In K-Means, Cluster 1 represents those with fair health, while Cluster 2 identifies those in bad health.

In Hierarchical Clustering, health is commonnly good for both clusters, contrasting K-Means.

A feature of hierarchical clustering is that it does not assume fixed cluster boundaries, while K-Means does, implying health status may be dependent on other attributes, hence the difference in Cluster 1 and 2 for K-Means but not Hierarchical.

Ethnic Group and Religion

K-Means proposes majority of the population are of Chinese/Other ethnic group and of Christian faith (Cluster 1), while Cluster 2 suggests they are Black/Black British and also of Christian faith.

For Hierarchical clustering, groups are also majorly Chinese/Other ethnic group, although religion varies compared to K-Means; Cluster 1 dictates the population mainly has no religious affiliation, while Cluster 2 highlights those with Sikh faith.

It is possible hierarchical clustering can reveal more nuanced interactions between individual's religion and ethnic background, while K-Means may not be able to perform as well in this manor.

Economic Activity

Both Clusters represent population from the financial and insurance industry, with different economic activities. Cluster 1 highlights those unemployed but somewhat involved in administrative and secretarial occupations, while Cluster 2 represents those in skilled trades roles.

Hierarchical is more complex, where Cluster 1 captures those with professional occupations in accommodation and food service sectors, conversely, Cluster 2 presents those in elementary roles within mining industry.

Given it is unlikely individuals can be unemployed but involved in administrative and secretarial roles, it suggests K-Means cannot convey the nuances of occupations, industry and economic activity as well as hierarchical clustering can, where this model is possibly reflecting influences from other variables on occupational grouping.

Hours Worked per Week:

With K-Means, the population is suggested to be primarily working part-time (16-30 hours per week) in Cluster 1, and less than or equal to 15 hours in Cluster 2.

In Hierarchical Clustering, Clusters are reportedly the same as K-Means. K-Means performed just as well as Hierarchical Clustering to identify similarities in work patterns.

Approximated Social Grade

With K-Means, social grade is varied, Cluster 1 represents most individuals belong to C1, characterised by lower-middle class workers, while Cluster 2 suggest most individuals belong to C2, representing skilled manual workers.

Hierarchical Clustering shows C2 is the most common social grade in Cluster 1, and C1 is the most common social grade in Cluster 2, the complete opposite to K-Means.

Hierarchical clustering possibly identified attributes social grade is dependent on, such as economic activity, occupation and industry worked in, therefore proposed a different cluster formation to K-Means. Since it does not assume cluster boundaries that are predefined, it is possible Hierarchical Clustering recognises these interdependencies and relationships.

Overall comparison

K Means typically predefines its cluster boundaries, simplifying data into distinct clusters. It excels in providing clear-cut groups but underperforms in capturing complex, nuanced relationships between variables unlike hierarchical clustering.

Hierarchical Clustering is more flexible and provides more of an in-depth detailed analysis. It's sensitivity to clustering boundaries are not predefined and less sensitive than that of K-Means, allowing this model to outperform K-Means in capturing complex interrelationships and interdependencies between variables.

In summary, K-Means would be more sutiable in creating well-defined, larger clusters from major demographic features such as age, marital status and occupation, while Hierarchical Clustering is more effective at revealing intricate and subtle relationships, which is very suitable for datasets with many multi-category dimensions, such as the dataset used for this report. Therefore, Hierarchical Clustering was the most suitable and effective model for clustering in this context.

Reference:

Navlani, A. (2024, August 11). *Python logistic regression tutorial with Sklearn & Scikit*. DataCamp. https://www.datacamp.com/tutorial/understanding-logistic-regression-python

seaborn.violinplot — seaborn 0.13.2 documentation. (n.d.). https://seaborn.pydata.org/generated/seaborn.violinplot.html