Intelligent Data And Text Analytics.

COURSEWORK 1

UP2225522

Descriptive Analytics

Table 1a

```
Column: Region
Number of unique values: 10
Unique values:
South East                  88084
London                      83582
North West                  71436
East of England             59411
West Midlands               56875
South West                  53774
Yorkshire and the Humber    53471
East Midlands               45782
Wales                       30976
North East                  26349
Name: Region, dtype: int64
```

- The highest population is in the "South East"
- "London" closely follows
- The lowest population is in the "North East"

Table 1b

```
Column: Residence Type
Number of unique values: 2
Unique values:
Non-Communal    559086
Communal         10654
Name: Residence Type, dtype: int64
```

Most individuals are in Non-Communal settings while Communal settings have a significantly lower population.

Table 1c

```
Column: Family Composition
Number of unique values: 7
Unique values:
Married             300961
None                 96690
Cohabiting           72641
Lone Female          64519
No code required     18851
Lone Male             9848
Other                 6230
Name: Family Composition, dtype: int64
```

- The majority are classified as "Married" followed by "None") and "Cohabiting"
- "Other" and "Lone Male" categories have the lowest representation.

Table 1d

```
Column: Population Base
Number of unique values: 3
Unique values:
Usual resident    561039
Student             6730
Short-term          1971
Name: Population Base, dtype: int64
```

The data primarily represents "Usual residents" with a smaller representation of "Students" and "Short-term" residents.

Table 1e

```
Column: Sex
Number of unique values: 2
Unique values:
Female    289172
Male      280568
Name: Sex, dtype: int64
```

The gender distribution is almost balanced, with slightly more females than males.

Table 1f

```
Column: Age
Number of unique values: 8
Unique values:
0 to 15        106832
35 to 44        78641
45 to 54        77388
25 to 34        75948
16 to 24        72785
55 to 64        65665
65 to 74        48777
75 and over     43704
Name: Age, dtype: int64
```

The age group "0 to 15" has the highest representation, while "75 and over" has the lowest.

Table 1g

```
Column: Marital Status
Number of unique values: 5
Unique values:
Single      270999
Married     214179
Divorced     40713
Widowed      31898
Seperate     11951
Name: Marital Status, dtype: int64
```

"Single" individuals have the highest representation followed by "Married".

Table 1h

```
Column: Student
Number of unique values: 2
Unique values:
No     443203
Yes    126537
Name: Student, dtype: int64
```

A majority of individuals are not students compared to those who are.

Table 1i

```
Column: Country of Birth
Number of unique values: 3
Unique values:
UK                  485645
Non-UK               77291
No code required      6804
Name: Country of Birth, dtype: int64
```

The majority were born in the "UK", with a smaller proportion born outside the UK.

Table 1j

```
Column: Health
Number of unique values: 6
Unique values:
Very good health    264971
Good health         191743
Fair health          74480
Bad health           24558
Very bad health       7184
No code required      6804
Name: Health, dtype: int64
```

A significant number report "Very good health", while a smaller number report "Bad" or "Very bad health."

Table 1k

```
Column: Ethnic Group
Number of unique values: 6
Unique values:
White                                483477
Asian and Asian British               42711
Black or Black British                18786
Mixed                                 12209
No code required                       6804
Chinese or Other ethnic group          5753
Name: Ethnic Group, dtype: int64
```

"White" is the dominant ethnic group followed by "Asian and Asian British" and "Black or Black British."

Table 1l

```
Column: Religion
Number of unique values: 10
Unique values:
Christian            333481
No religion          141658
Not stated            40613
Muslim                27240
Hindu                  8213
No code required       6804
Sikh                   4215
Jewish                 2572
Buddhist               2538
Other religion         2406
Name: Religion, dtype: int64
```

"Christian" is the most reported religion followed by "No religion."

Table 1m

```
Column: Economic Activity
Number of unique values: 10
Unique values:
Active Employee          216024
No code required         112618
Inactive Retired          97480
Active Self-employed      40632
Inactive Student          24756
Active Unemployed         18109
Inactive Sick             17991
Inactive Homecare         17945
Active Student            14117
Inactive Other            10068
Name: Economic Activity, dtype: int64
```

The majority are "Active Employees" , and a significant number have "No code required."

Table 1n

```
Column: Occupation
Number of unique values: 10
Unique values:
No code required                              149984
Professional                                   64111
Elementary                                     58483
Administrative                                 53254
Skilled Trades                                 48546
Associate                                      44937
Managers, Directors and Senior Officials       39788
SSales and Customer Service                    38523
Caring, Leisure and Other Service              37297
Process, Plant and Machine                     34817
Name: Occupation, dtype: int64
```

"No code required" is the most common occupation, followed by "Professional" and "Elementary."

Table 1p

```
Column: Industry
Number of unique values: 13
Unique values:
No code required                                                                                                  149984
Wholesale and retail trade; Repair of motor vehicles and motorcycles                                               68878
Mining and quarrying; Manufacturing; Electricity, gas, steam and air conditioning system; Water supply             53433
Real estate activities; Professional, scientific and technical activities; Administrative and support service activities  49960
Human health and social work activities                                                                            49345
Education                                                                                                          40560
Transport and storage; Information and communication                                                               35240
Construction                                                                                                       30707
Accommodation and food service activities                                                                          25736
Public administration and defence; compulsory social security                                                      24908
Other community, social and personal service activities; Private households employing domestic staff; Extra-territorial organisations and bodies  20256
Financial and insurance activities; Intermediation                                                                 16776
Agriculture, forestry and fishing                                                                                   3957
Name: Industry, dtype: int64
```

A large number have "No code required" for industry, with significant representation in wholesale/retail, health, and education sectors.

Table 1q

```
Column: Hours worked per week
Number of unique values: 5
Unique values:
No code required                     302321
Full-time: 31 to 48 hours worked     153937
Part-time: 16 to 30 hours worked      52133
Full-time: 49 or more hours worked    35573
Part-time: 15 or less hours worked    25776
Name: Hours worked per week, dtype: int64
```

A substantial number have "No code required" for hours worked, and the majority work full-time.

Table 1r

```
Column: Approximated Social Grade
Number of unique values: 5
Unique values:
C1: Supervisory, clerical & junior managerial, administrative    159642
No code required                                                 124103
DE: Semi-skilled & unskilled manual occupations, Unemployed      123739
AB: Higher & intermediate managerial, administrative              82320
C2: Skilled manual occupations                                    79936
Name: Approximated Social Grade, dtype: int64
```

The social grade "C1" has the highest representation, followed by "No code required." "AB" and "C2" also have notable representation.

In summary, the dataset shows a wide range of factors influencing individuals' lives. The regional distribution indicates higher populations in the South East and London, while the North East has a comparatively smaller population. Family-wise, married individuals form the majority, and the dataset notably captures the social dynamics of cohabiting and lone individuals. The age distribution shows a concentration in younger age groups, with a significant portion in the 0 to 15 range. In terms of marital status, a substantial number of individuals report being single. Employment-wise, a majority are active employees, and the dataset provides insights into various occupational and industrial sectors. A large proportion of individuals report good health. Lastly, there are a couple of columns with is fluctuate with “No code required” which are related to econometrics. This is because a large amount of under-working-age or economically inactive.

Table 2a: Before mapping

	Person ID	Region	Residence Type	Family Composition	Population Base	Sex	Age	Marital Status	Student	Country of Birth	Health	Ethnic Group	Religion	Economic Activity	Occupation	Industry
0	7394816	E12000001	H	2	1	2	6	2	2	1	2	1	2	5	8	2
1	7394832	E12000001	H	3	1	2	1	1	2	1	2	1	1	-9	-9	-9
2	7394719	E12000001	H	2	1	1	7	2	2	1	1	1	2	5	8	2
3	7394840	E12000001	H	1	1	2	6	4	2	1	3	1	2	5	9	5
4	7394711	E12000001	H	2	1	1	1	1	1	1	1	1	1	-9	-9	-9

Table 2b: After mapping

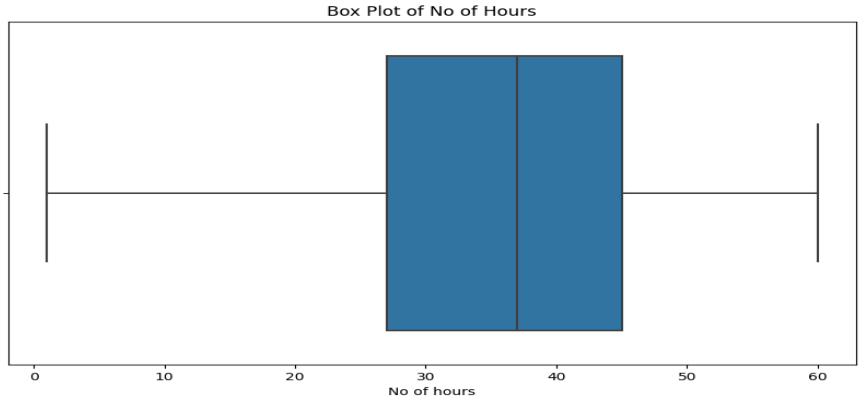
	Person ID	Region	Residence Type	Family Composition	Population Base	Sex	Age	Marital Status	Student	Country of Birth	Health	Ethnic Group	Religion	Economic Activity	Occupation	In
0	7394816	North East	Non-Communal	Married	Usual resident	Female	55 to 64	Married	No	UK	Good health	White	Christian	Inactive Retired	Process, Plant and Machine	Mini qua Manufac Ele
1	7394832	North East	Non-Communal	Cohabiting	Usual resident	Female	0 to 15	Single	No	UK	Good health	White	No religion	No code required	No code required	N re
2	7394719	North East	Non-Communal	Married	Usual resident	Male	65 to 74	Married	No	UK	Very good health	White	Christian	Inactive Retired	Process, Plant and Machine	Mini qua Manufac

Data mapping of numerical categorical variables to text based on the 2011 Census variable list.

Table 3: Summary statistics of numerical variable

```
count    267419.000000
mean         35.234789
std          13.520881
min           1.000000
25%          27.000000
50%          37.000000
75%          45.000000
max          60.000000
Name: No of hours, dtype: float64
```

Figure 1



The dataset contains information on the number of working hours per week for 267,419 individuals with a mean number of hours worked of 35.23 hours per week and a standard deviation of 13.52 hours, indicating a variation in working hours across the sample with a central tendency of around 37 hours and the majority of individuals work between 27 and 45 hours per week. The maximum recorded working hours in the dataset is 60 hours per week.

Figure 2

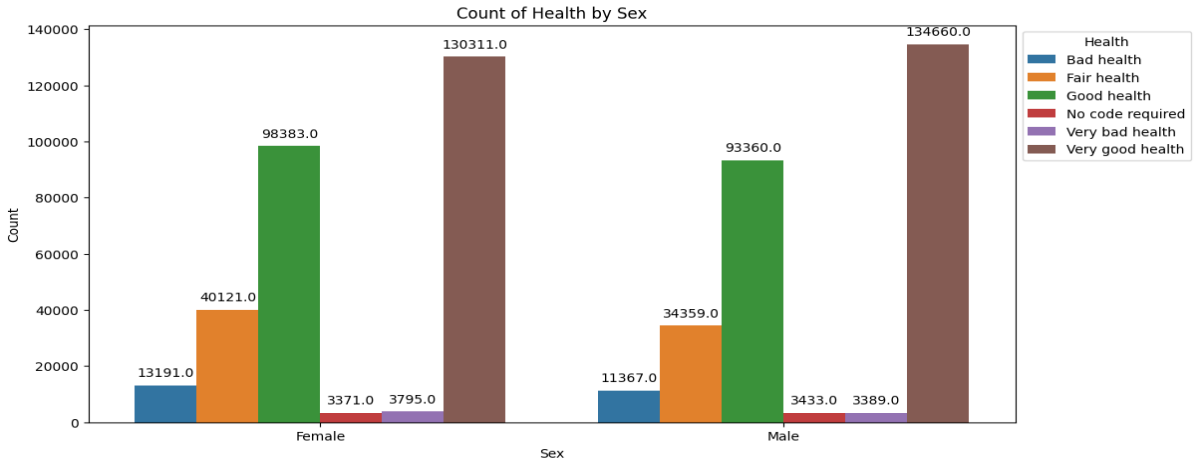


Figure 3

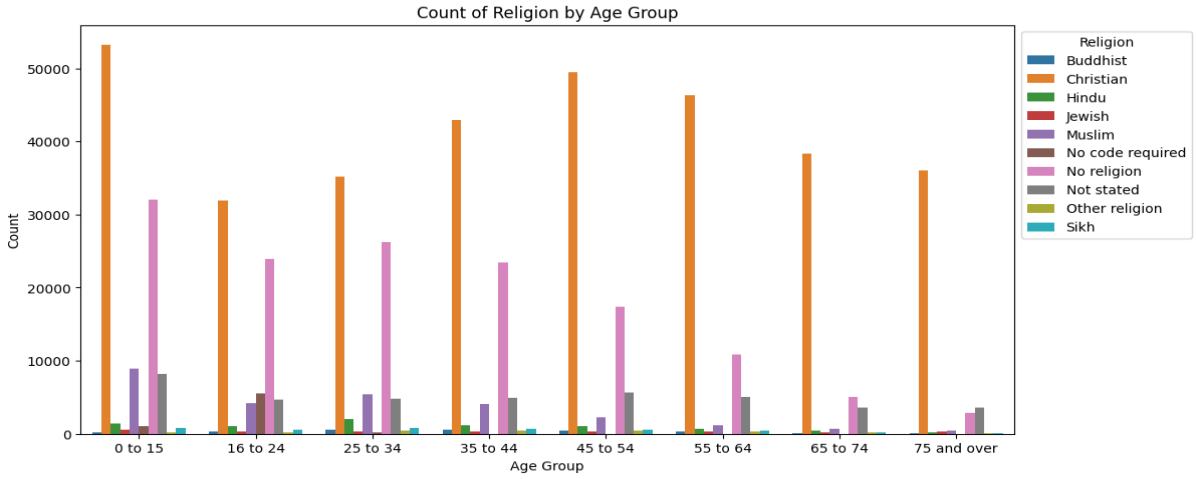


Figure 4

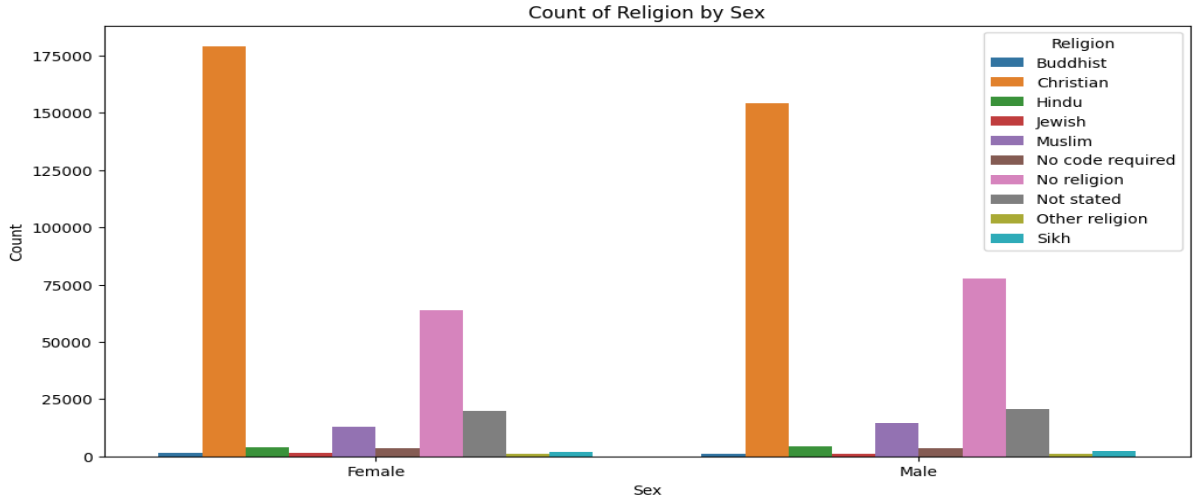


Figure 5a

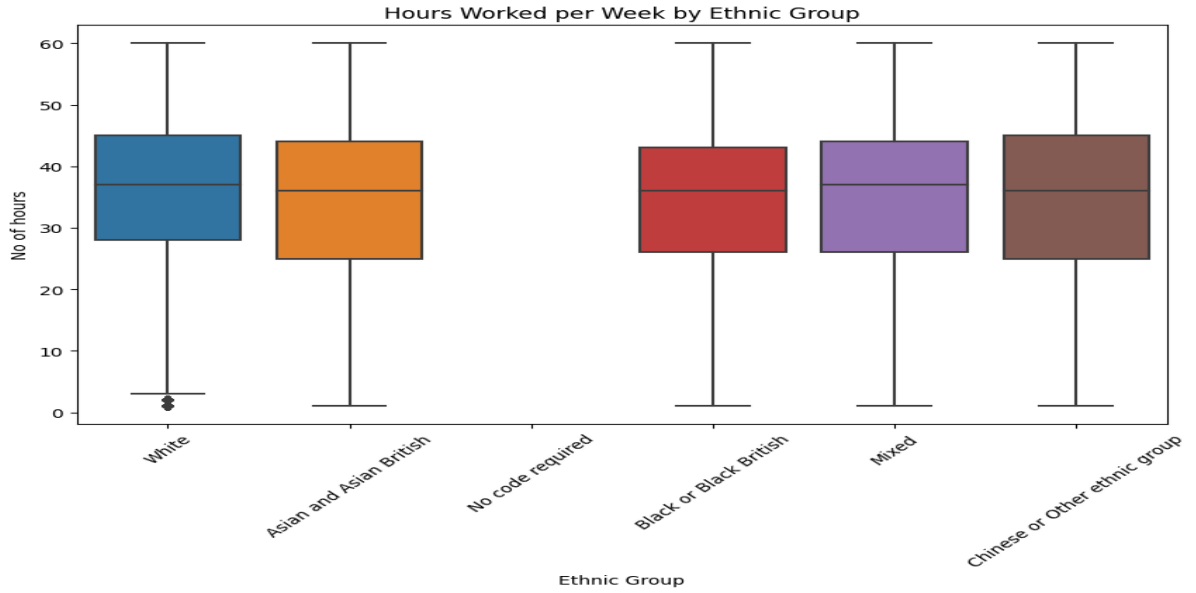


Figure 5b

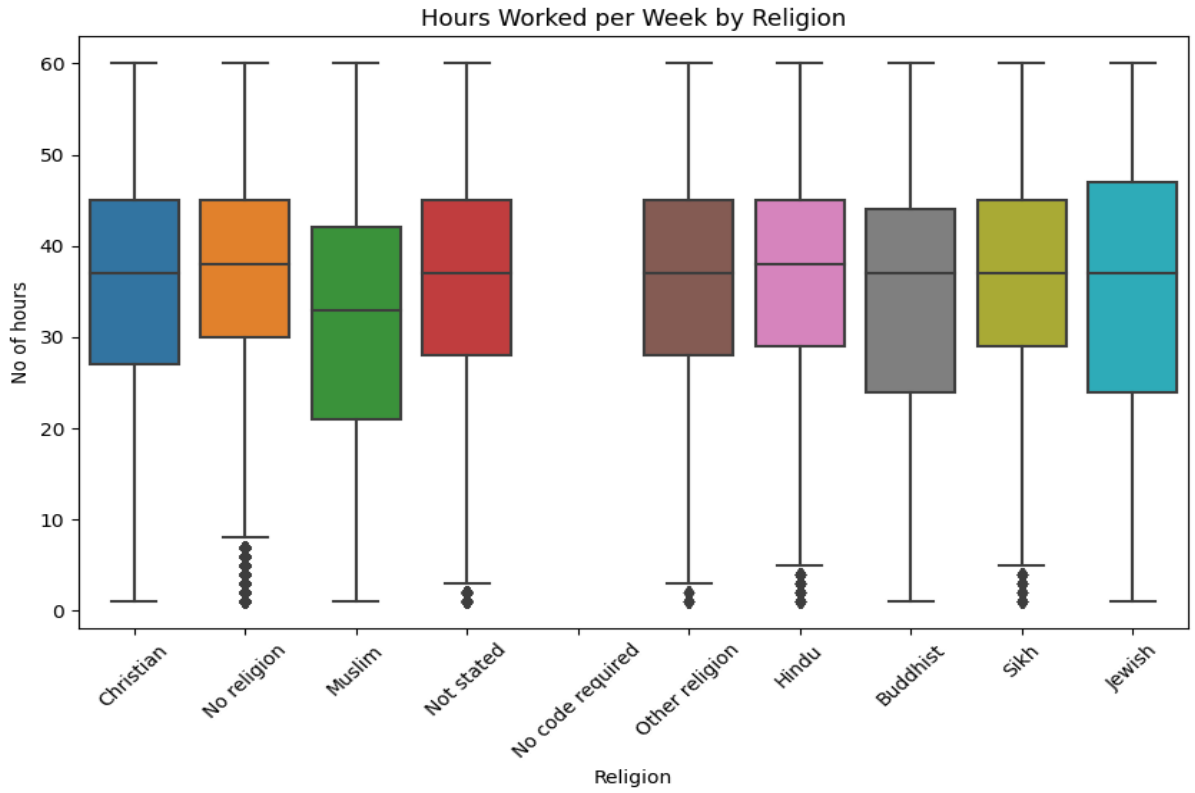
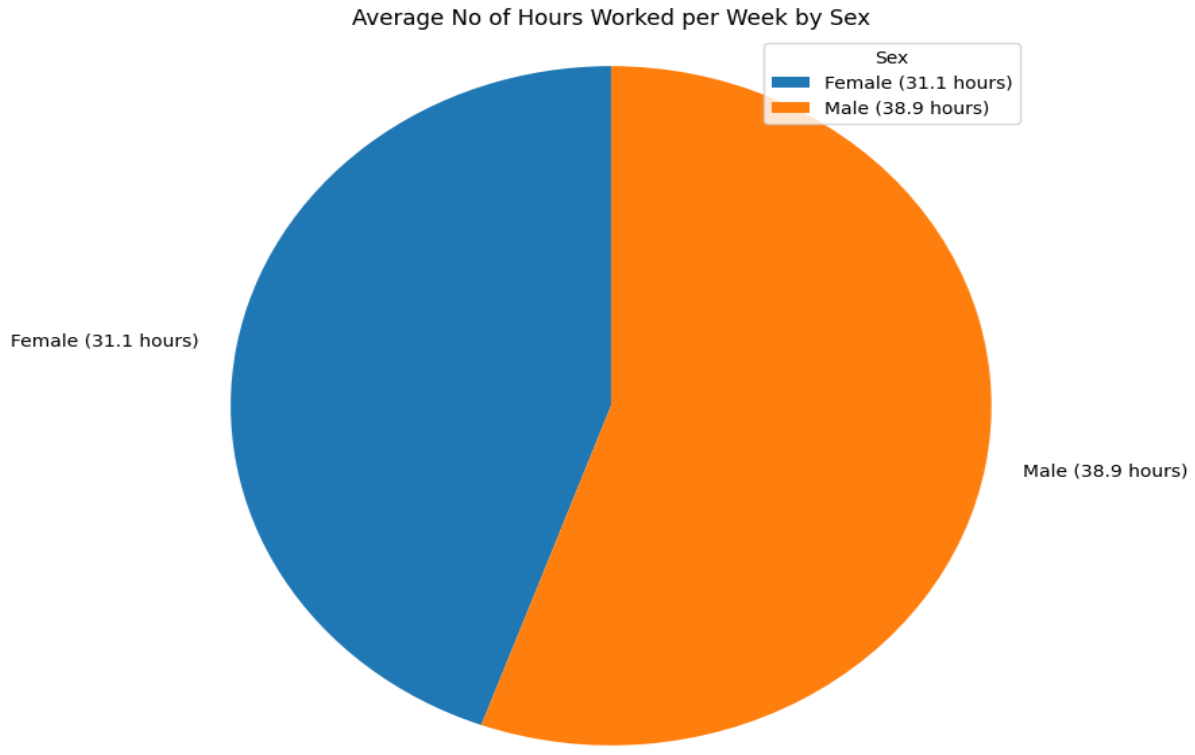


Figure 6



The key insights from Figure 2 - 6 are as follows:

- Sex is not a determining factor for health as both sexes reported similar health conditions
- Christianity dominates all age groups and Sex followed by no religion
- Both sexes are mostly distributed in all religions similarity
- More males are no religion and more females are Christians
- The median number of working hours is consistent across different ethnic and religious groups meaning people from various ethnic and religious groups tend to work a similar number of hours.
- Most ethnic groups have similar variability except white people who have a higher spread of data.
- The box plot of the Muslim group has more data points below others indicating a greater number of individuals who work fewer hours compared to the other groups.
- The Jewish group has more data points above others indicating more people worked more hours compared to other groups.
- Both Muslim and Jewish group also shows a wider spread
- No religious group has the most outlier. This implies a greater variability in the data.
- On average males work more hours than females

Classification

The classification is implemented with the following clarification, assumptions and limitations:

- All -9 is replaced by 0
- Due to computing limitations only 1% of the data is sampled
- Approximate social grades will be referred to as:

```
1 : 'AB: Higher & intermediate managerial, administrative',
2 : 'C1: Supervisory, clerical & junior managerial, administrative',
3 : 'C2: Skilled manual occupations',
4 : 'DE: Semi-skilled & unskilled manual occupations, Unemployed',
0 : 'No code required'
```

Decision Tree

Figure 7

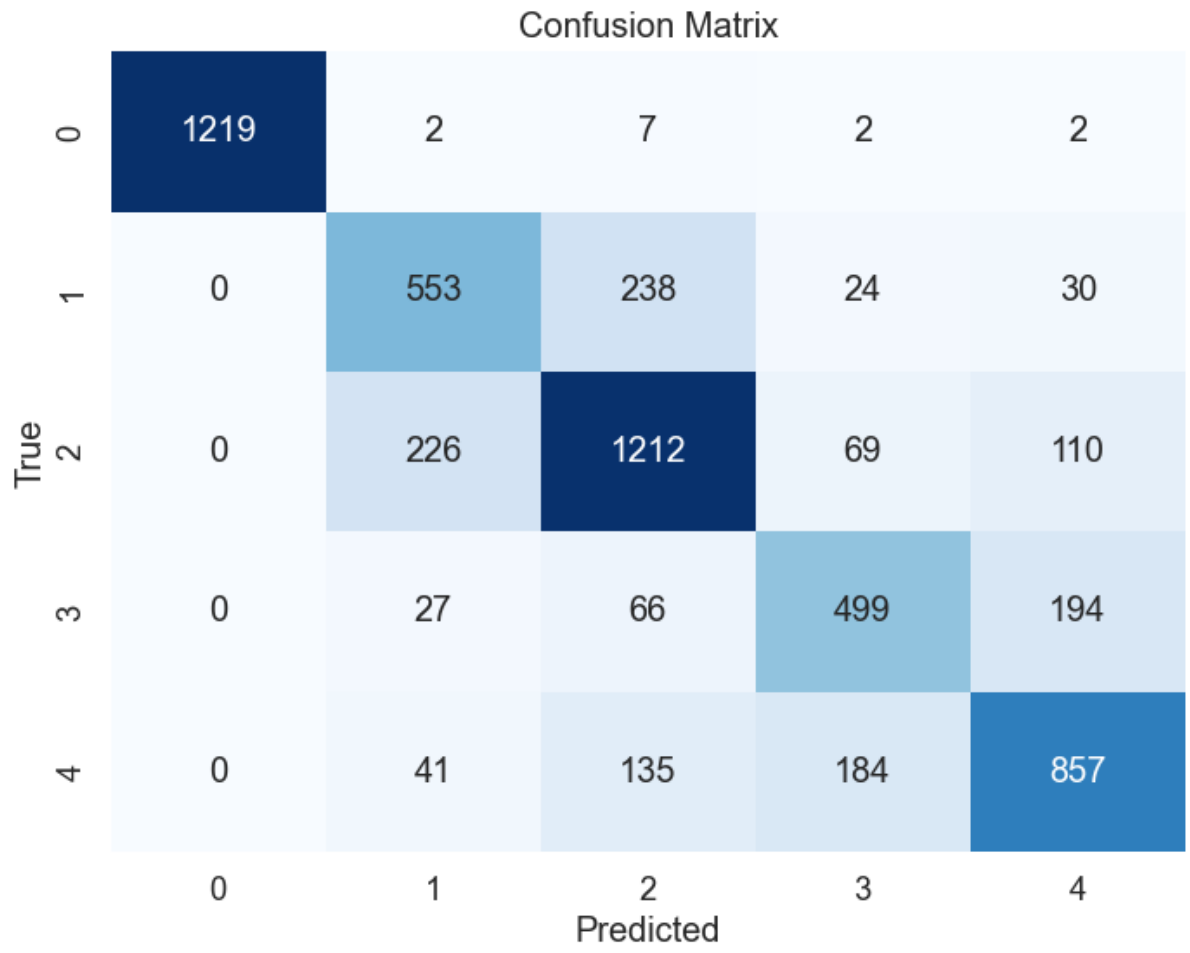


Table 4

```
Decision Tree - Cross-Validation Accuracy: 0.7618044584869229
Decision Tree - Cross-Validation Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.99      0.99      1232
           1       0.65      0.65      0.65       845
           2       0.73      0.75      0.74      1617
           3       0.64      0.63      0.64       786
           4       0.72      0.70      0.71      1217

    accuracy                           0.76      5697
   macro avg       0.75      0.75      0.75      5697
weighted avg       0.76      0.76      0.76      5697

Decision Tree - Cross-Validation F1 Scores: [0.75437521 0.74702247 0.73711354 0.75579185 0.73820622]
Decision Tree - Cross-Validation Average F1 Score with variance: (0.7465018578980145, 6.110452840159467e-05)
```

The Decision Tree model achieved a good score in every metric and performed especially well for classifying class 0.

The results showed an average F1-score of 0.7490, indicating consistent performance. Based on these metrics, this model performs very well.

KNN

Figure 8

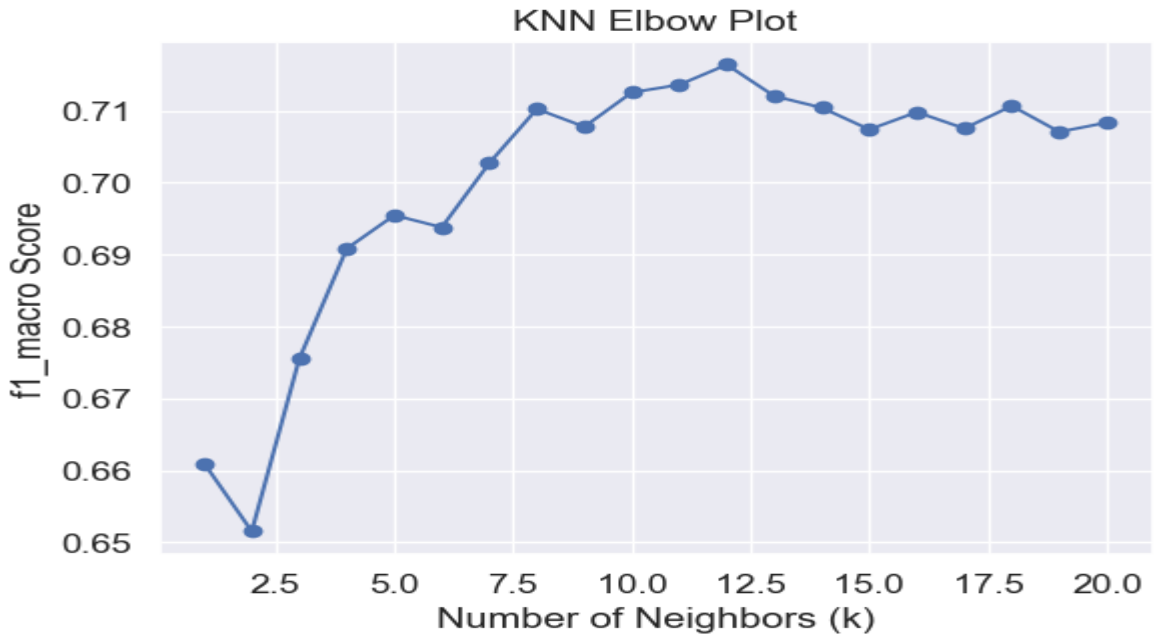


Figure 9

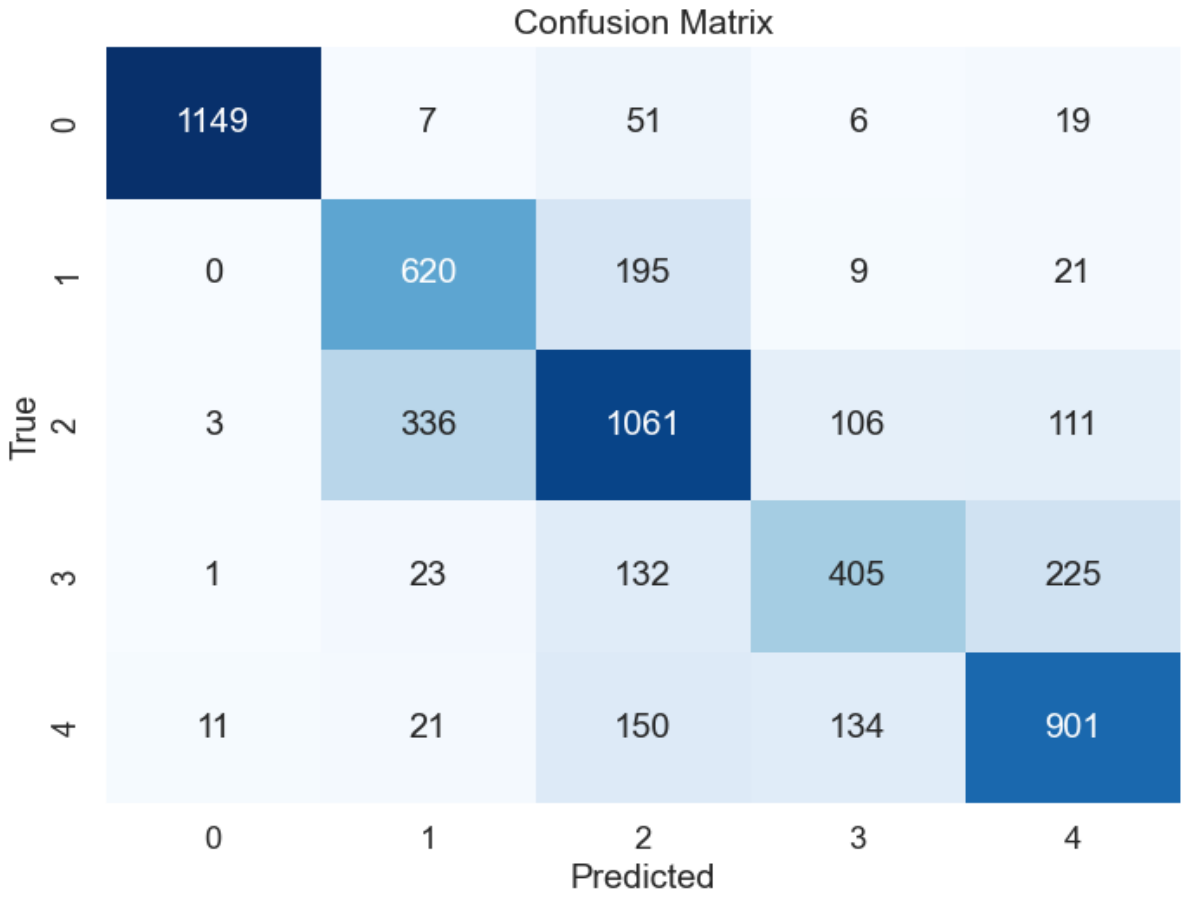


Table 5

```
Optimal k: 12
KNN - Cross-Validation Accuracy: 0.7259961383184133
KNN - Cross-Validation Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.93      0.96      1232
           1       0.62      0.73      0.67       845
           2       0.67      0.66      0.66      1617
           3       0.61      0.52      0.56       786
           4       0.71      0.74      0.72      1217

    accuracy                           0.73      5697
   macro avg       0.72      0.72      0.71      5697
weighted avg       0.73      0.73      0.73      5697

KNN - Cross-Validation F1 Scores: [0.70854423 0.72346883 0.70549575 0.73309522 0.72214451 0.68753403
 0.70308315 0.7227186  0.7289938  0.70968457]
KNN - Cross-Validation Average F1 Score with variance: (0.7144762688290444, 0.00017595146327251959)
```

The elbow plot indicates that the best value for K in this case is 12. The cross-validation results indicate that the model has good results in classifying 0 and not as good results for 3.

Logistic Regression

Figure 10

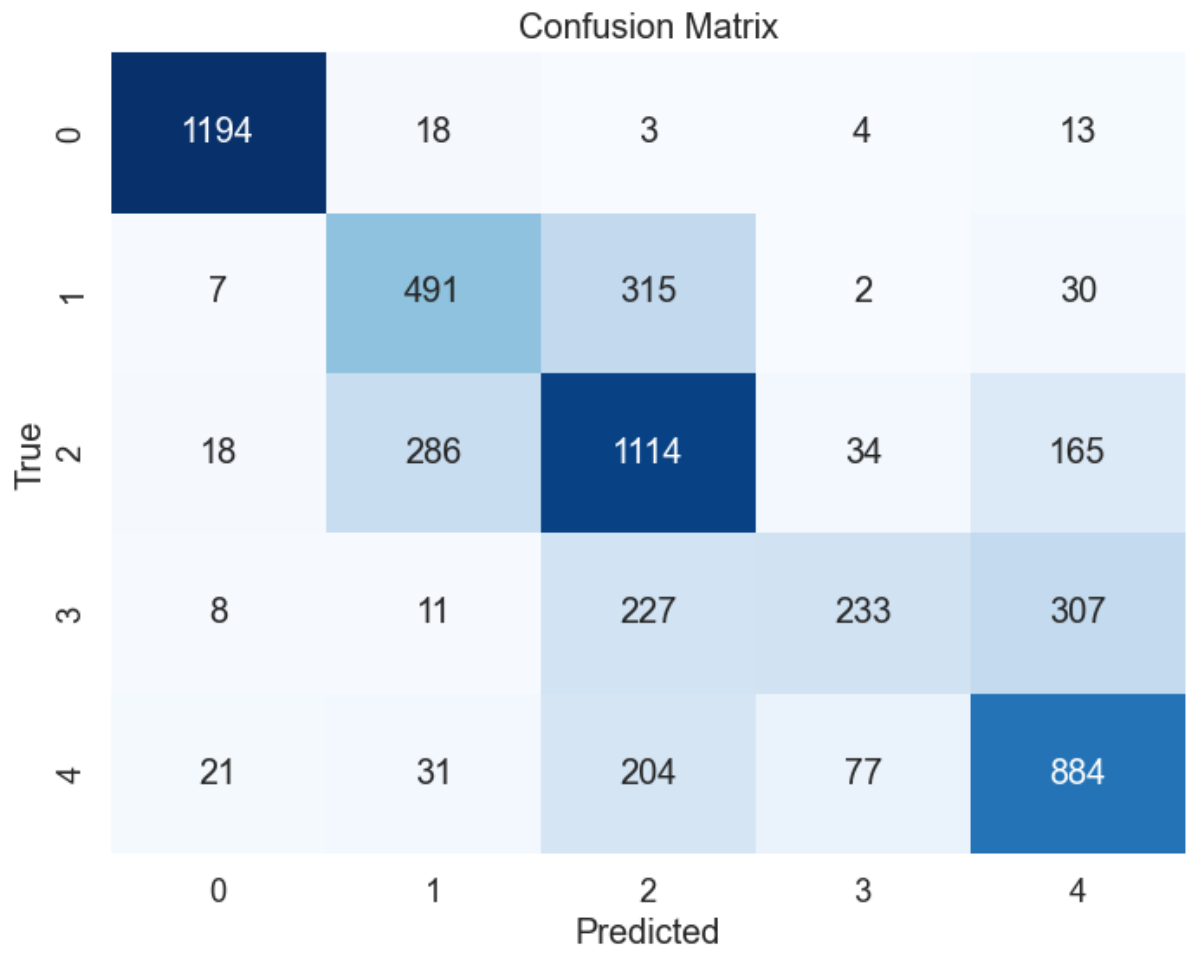


Table 6

```
Logistic - Cross-Validation Accuracy: 0.6873793224504124
Logistic - Cross-Validation Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.97      0.96      1232
           1       0.59      0.58      0.58       845
           2       0.60      0.69      0.64      1617
           3       0.67      0.30      0.41       786
           4       0.63      0.73      0.68      1217

    accuracy                           0.69      5697
   macro avg       0.69      0.65      0.65      5697
weighted avg       0.69      0.69      0.68      5697

Logistic - Cross-Validation F1 Scores: [0.59344481 0.64770924 0.70406302 0.64834535 0.64039585 0.6622356
 0.66982282 0.64774012 0.65622921 0.67115123]
Logistic - Cross-Validation Average F1 Score with variance: (0.6541137250747728, 0.0007086236056121197)
```

The Logistic Regression model achieved a respectable cross-validation score. Overall, this model does not perform as well compared to the others. This is because this model is more suitable for binary classification.

SVM

Figure 11

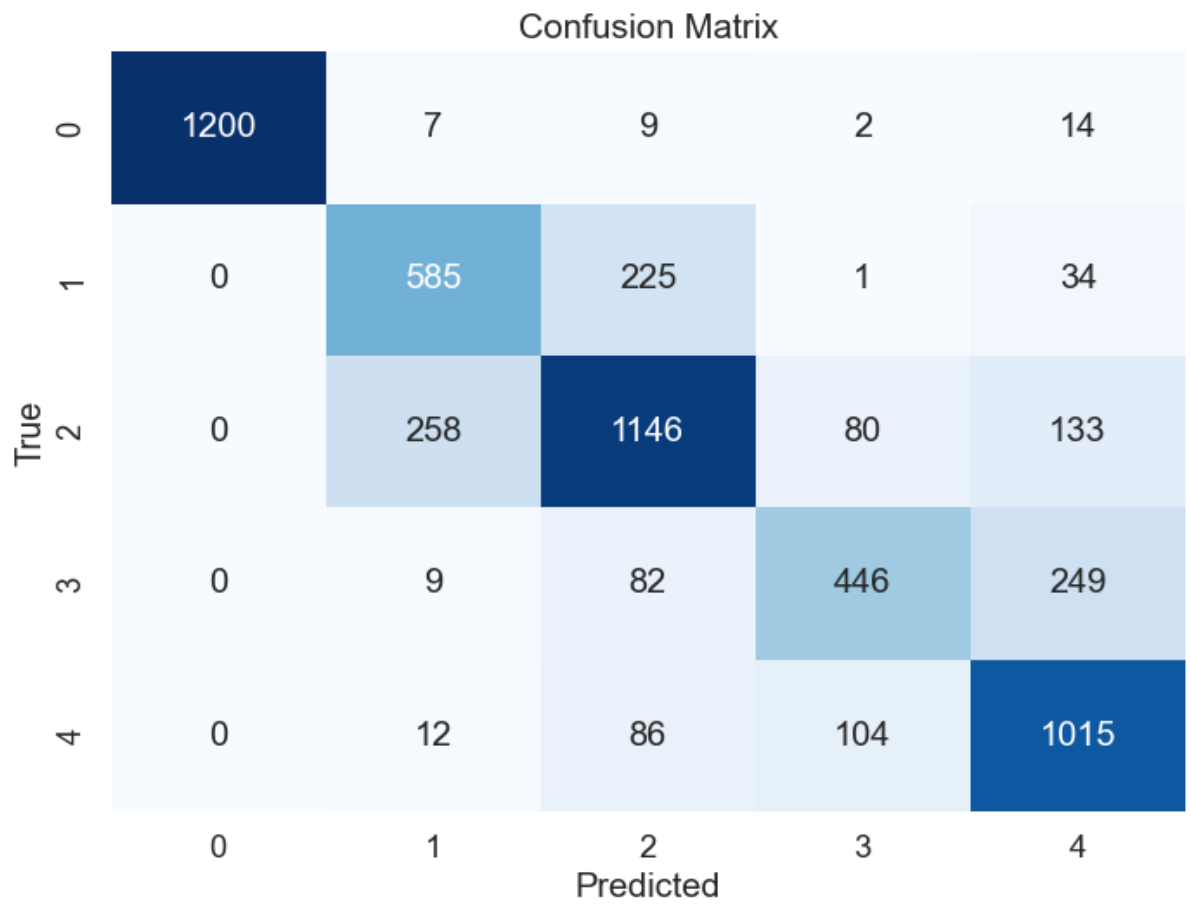


Table 7

```
SVM - Cross-Validation Accuracy: 0.7709320695102686
SVM - Cross-Validation Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.97      0.99      1232
           1       0.67      0.69      0.68       845
           2       0.74      0.71      0.72      1617
           3       0.70      0.57      0.63       786
           4       0.70      0.83      0.76      1217

    accuracy                           0.77      5697
   macro avg       0.76      0.76      0.76      5697
weighted avg       0.77      0.77      0.77      5697

SVM - Cross-Validation F1 Scores: [0.7844497  0.74273344 0.74002366 0.76182671 0.76718925 0.74425841
 0.75506736 0.74524262 0.76851246 0.75663403]
SVM - Cross-Validation Average F1 Score with variance: (0.7565937649863699, 0.00018076746938859868)
```

The SVM model also has a good cross-validation result and performs the best in classifying 0 not as well on 3.

Neural Network

Figure 12

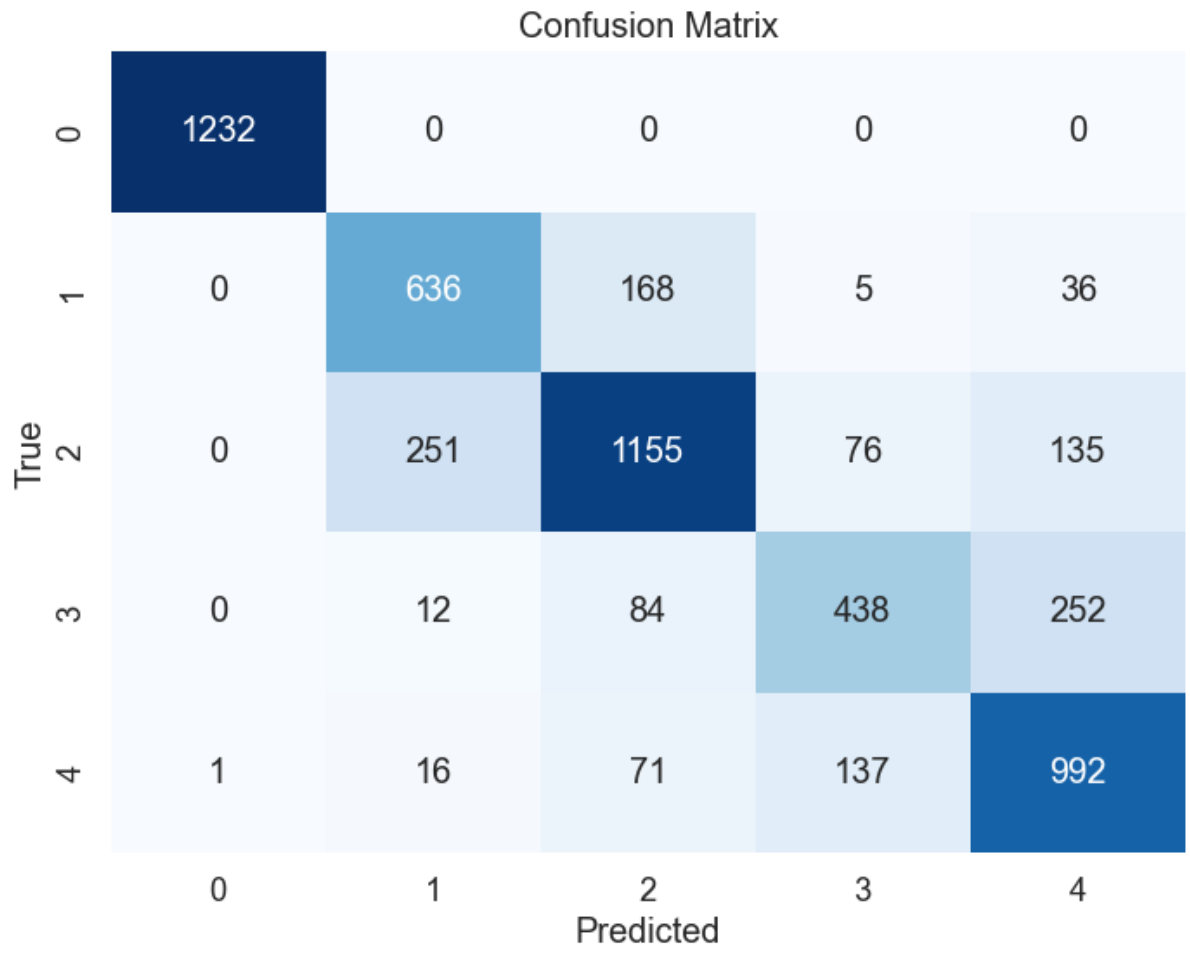


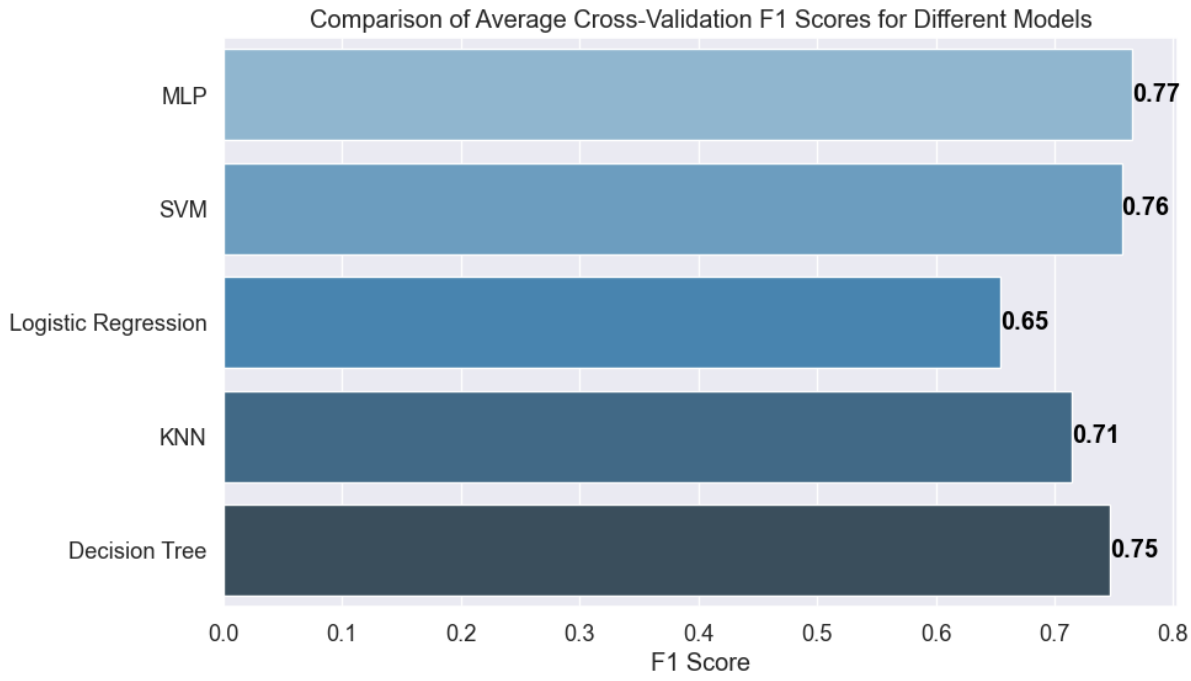
Table 8

```
MLP - Accuracy: 0.7816394593645778
MLP - Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1232
           1       0.70      0.75      0.72       845
           2       0.78      0.71      0.75      1617
           3       0.67      0.56      0.61       786
           4       0.70      0.82      0.75      1217

    accuracy                           0.78      5697
   macro avg       0.77      0.77      0.77      5697
weighted avg       0.78      0.78      0.78      5697

MLP - Cross-Validation F1 Scores: [0.78542919 0.74309365 0.76336977 0.76816017 0.79204677 0.72704177
 0.76487813 0.77133312 0.77738755 0.76223281]
MLP - Cross-Validation Average F1 Score with variance: (0.7654972935928697, 0.00032809979494674275)
```

The MLP model yields the best cross-validation result and is the only model that classified 0 all correctly. It also has the same pattern as other models in not predicting class 3 well.

Comparison

Figure 13



All models perform well in classifying 0 (no code required) and have problems classifying 3 (C2: Skilled manual occupations). This means people with class 0 have similar characteristics which all models can pick up and class 3 is the opposite.

In summary, the MLP performs the best with the highest F1 score. This shows that the model performs well with the balance between precision and recall. Following closely behind, the SVM also demonstrated solid performance. The Decision Tree and KNN model, while trailing slightly, still displayed respectable results. Lastly, the Logistic Regression model

recorded the lowest F1 score among the models evaluated, which suggests that it may not be the optimal choice for this specific task. Logistic Regression is typically suited for binary classification problems and may not perform as strongly in situations where more complex relationships within the data need to be captured.

Regression

The Regression is implemented with the following assumptions and limitations:

- All -9 is replaced by 0
- Null values in the “No of hour” are replaced with 0

Multiple Linear Regression

Figure 14

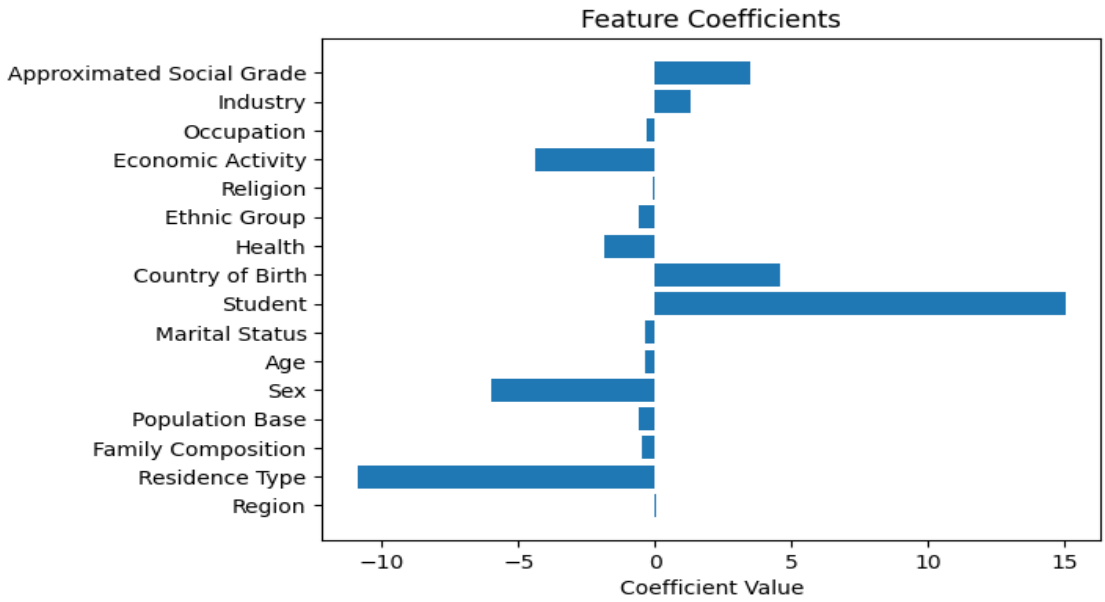


Table 9

```
{'LR - Mean Absolute Error': 10.999564229624966,
 'LR - R-squared': 0.5256607431545115,
 'LR - Intercept': 11.072428487769084}
```

The LR model predicts working hours per week with an average error of approximately 11 hours meaning it is likely to predict the working hours off by 11 hours. The R-squared indicates that the model explains 53% of the variability in working hours. While not perfect, it suggests a moderate fit, capturing more than half of the variation in the data.

Standardise Multiple Linear Regression

Figure 15

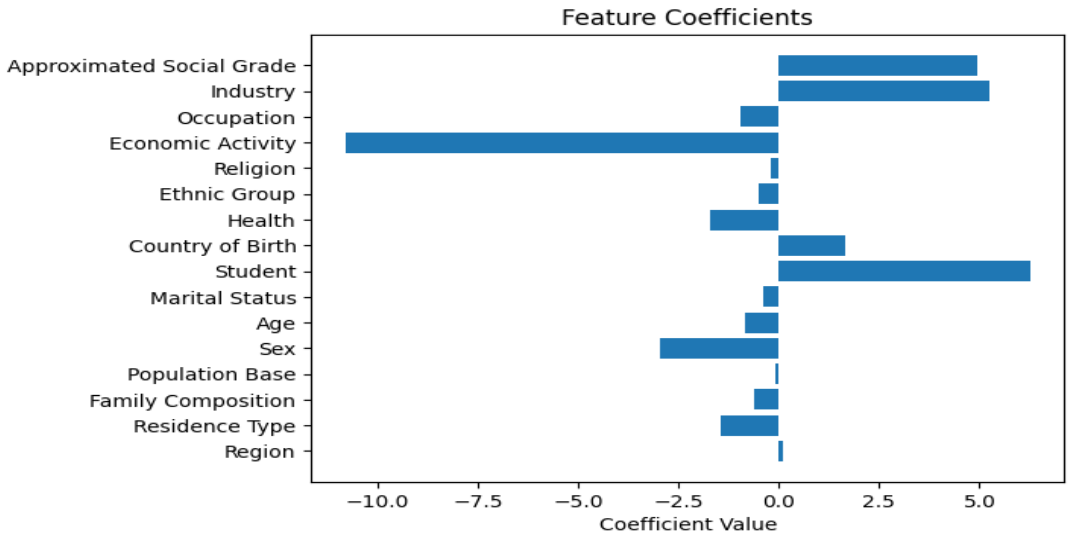


Table 10

```
{'Scaled_lr - Mean Absolute Error': 11.007560259520744,
 'Scaled_lr - R-squared': 0.5268442417886853,
 'Scaled_lr - Intercept': 16.529778934250707}
```

The standardized LR model had a similar performance to the non-standardized model, with moderate fit capturing over half of the variation.

Decision Tree Regression

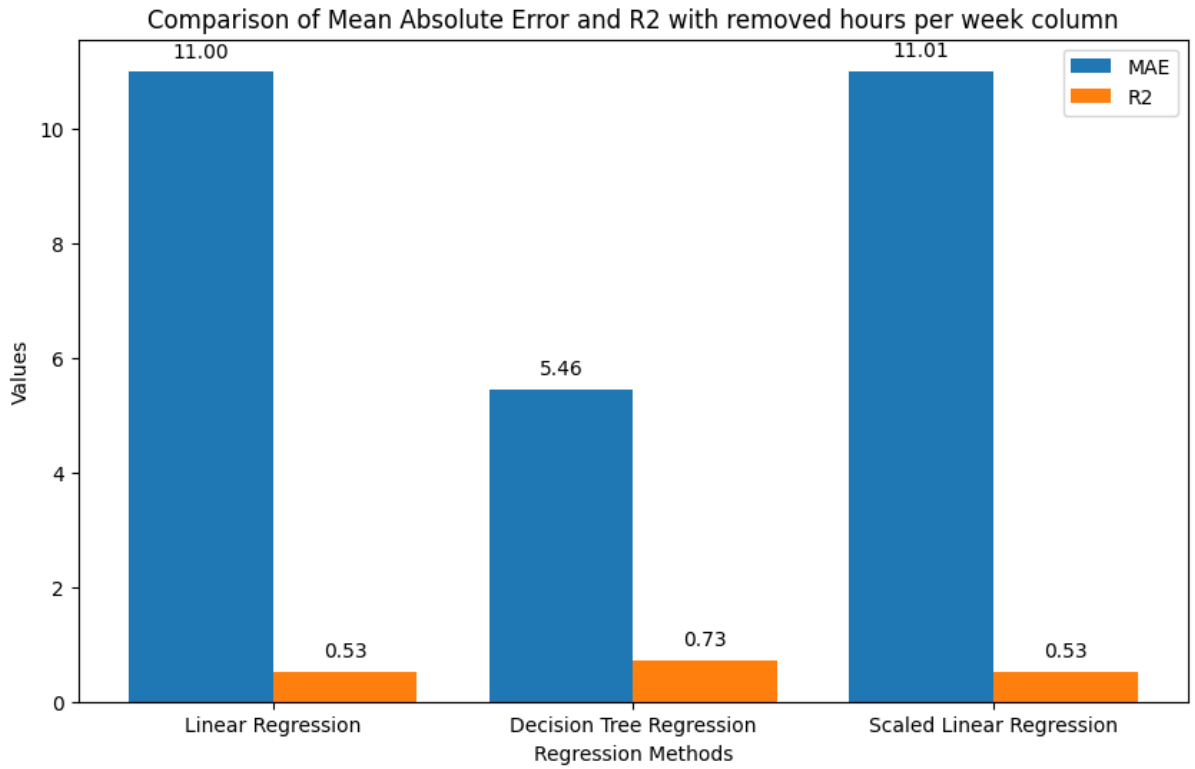
Table 11

```
{'Mean Absolute Error': 5.458144003677429,
 'R-squared': 0.7274799233738509,
```

The Decision Tree Regression model gives a mean absolute error of about 5.46 hours meaning it is likely to predict the working hours off by 5 hours. The R-squared value suggests a strong fit, explaining approximately 73% of the variability in working hours. This indicates a relatively accurate prediction with a substantial portion of the variance captured by the model.

Analysis and Comparison

Figure 16



The LR model showed that its predictions had a higher MAE, indicating that, on average, its predictions deviated by this considerable amount from the actual values. The R-squared value indicates that it explained approximately half of the data's variance, which is a moderate fit. The scaled version of LR presented nearly identical results, suggesting that scaling had little impact on performance.

The Decision Tree Regressor yielded significantly lower MAE, indicating that its predictions were closer to the actual values. The R-squared also was higher signifying a better fit. This suggests that the Decision Tree Regressor outperformed the linear regression models in this context.

The feature coefficient plot indicates that the best predictors for LR are “Approximate Social Grade”, “Industry”, “Student” and “Country of Birth”. On the other hand, “Economic activity”, “Sex” and “ Resident type” are the opposite.

Association Rule Mining

The table and figure below show the results after the data of the select column has been replaced with the text value based on the 2011 census variable list.

Table 12

	Person ID	Region	Residence Type	Family Composition	Population Base	Sex	Age	Marital Status	Student	Country of Birth	Health	Ethnic Group	Religion
0	7394816	North East	H	2	1	Female	55 to 64	2	2	1	Good health	White	Christian
1	7394832	North East	H	3	1	Female	0 to 15	1	2	1	Good health	White	No religion
2	7394719	North East	H	2	1	Male	65 to 74	2	2	1	Very good health	White	Christian
3	7394840	North East	H	1	1	Female	55 to 64	4	2	1	Fair health	White	Christian
4	7394711	North East	H	2	1	Male	0 to 15	1	1	1	Very good health	White	No religion

Figure 17

```
# Select the columns for analysis
selected_columns = ['Sex', 'Ethnic Group', 'Age', 'Region', 'Religion', 'Health']
```

Table 13

	Items	Antecedent	Consequent	Support	Confidence	Lift
0	{0 to 15, Very good health, White}	{Very good health}	{0 to 15, White}	0.118150	0.254047	1.731926
1	{0 to 15, Very good health, White}	{0 to 15, White}	{Very good health}	0.118150	0.805473	1.731926
2	{0 to 15, Very good health}	{Very good health}	{0 to 15}	0.146698	0.315431	1.682207
3	{0 to 15, Very good health}	{0 to 15}	{Very good health}	0.146698	0.782350	1.682207
4	{0 to 15, Very good health, White}	{Very good health, White}	{0 to 15}	0.118150	0.302167	1.611472

Table 14: Lift more than 1

	Items	Antecedent	Consequent	Support	Confidence	Lift
0	{0 to 15, Very good health, White}	{Very good health}	{0 to 15, White}	0.118150	0.254047	1.731926
1	{0 to 15, Very good health, White}	{0 to 15, White}	{Very good health}	0.118150	0.805473	1.731926
2	{0 to 15, Very good health}	{Very good health}	{0 to 15}	0.146698	0.315431	1.682207
3	{0 to 15, Very good health}	{0 to 15}	{Very good health}	0.146698	0.782350	1.682207
4	{0 to 15, Very good health, White}	{Very good health, White}	{0 to 15}	0.118150	0.302167	1.611472
5	{0 to 15, Very good health, White}	{0 to 15}	{Very good health, White}	0.118150	0.630101	1.611472
6	{No religion, Very good health, White}	{No religion}	{Very good health, White}	0.123976	0.498623	1.275219
7	{No religion, Very good health, White}	{Very good health, White}	{No religion}	0.123976	0.317066	1.275219
8	{No religion, White, Male}	{No religion}	{White, Male}	0.127679	0.513518	1.233549
9	{No religion, White, Male}	{White, Male}	{No religion}	0.127679	0.306705	1.233549
10	{Good health, Christian, White, Female}	{Good health, Christian}	{Female, White}	0.102341	0.507662	1.174330
11	{Good health, Christian, White, Female}	{White, Female}	{Good health, Christian}	0.102341	0.236738	1.174330
12	{Good health, Christian, White, Female}	{Good health, White, Female}	{Christian}	0.102341	0.685501	1.171153
13	{Good health, Christian, White, Female}	{Christian}	{Good health, Female, White}	0.102341	0.174847	1.171153
14	{No religion, Very good health}	{Very good health}	{No religion}	0.133466	0.286979	1.154211
15	{No religion, Very good health}	{No religion}	{Very good health}	0.133466	0.536793	1.154211
16	{Christian, White, Female}	{Christian}	{Female, White}	0.291552	0.498106	1.152227
17	{Christian, White, Female}	{White, Female}	{Christian}	0.291552	0.674423	1.152227
18	{No religion, Very good health, White}	{No religion, White}	{Very good health}	0.123976	0.534883	1.150105
19	{No religion, Very good health, White}	{Very good health}	{No religion, White}	0.123976	0.266573	1.150105

Table 15: Lift more than 1 and confidence more than 80

	Items	Antecedent	Consequent	Support	Confidence	Lift
1	{0 to 15, Very good health, White}	{0 to 15, White}	{Very good health}	0.118150	0.805473	1.731926
29	{Good health, Christian, White}	{Good health, Christian}	{White}	0.188447	0.934788	1.101575
31	{No religion, White, Male}	{No religion, Male}	{White}	0.127679	0.934436	1.101160
32	{Good health, Christian, White, Female}	{Good health, Christian, Female}	{White}	0.102341	0.933257	1.099770
34	{No religion, White}	{No religion}	{White}	0.231781	0.932210	1.098537
37	{55 to 64, White}	{55 to 64}	{White}	0.107270	0.930724	1.096786
38	{No religion, White, Female}	{No religion, Female}	{White}	0.104102	0.929494	1.095336
41	{No religion, Very good health, White}	{No religion, Very good health}	{White}	0.123976	0.928894	1.094629
45	{Christian, White, Female}	{Christian, Female}	{White}	0.291552	0.927113	1.092531
46	{Christian, White}	{Christian}	{White}	0.542453	0.926760	1.092115
49	{Christian, White, Male}	{Christian, Male}	{White}	0.250900	0.926351	1.091633
60	{Very good health, Christian, White, Female}	{Very good health, Christian, Female}	{White}	0.121443	0.906246	1.067941
62	{Very good health, Christian, White}	{Very good health, Christian}	{White}	0.234324	0.905449	1.067002
64	{Very good health, Christian, White, Male}	{Very good health, Christian, Male}	{White}	0.112881	0.904594	1.065993
73	{White, Fair health}	{Fair health}	{White}	0.117087	0.895663	1.055469
74	{South East, White}	{South East}	{White}	0.137905	0.891989	1.051140
77	{North West, White}	{North West}	{White}	0.111800	0.891665	1.050758
78	{45 to 54, White}	{45 to 54}	{White}	0.120978	0.890655	1.049567
107	{Good health, White, Female}	{Good health, Female}	{White}	0.149294	0.864570	1.018829
108	{Good health, White}	{Good health}	{White}	0.290501	0.863187	1.017198

Six rules are chosen from Table 18 and the analysis of the selected results:

1. The first rule suggests that individuals who are "White" and within the age group "0 to 15" are highly likely to be in "Very good health" (Support: 0.118, Confidence: 0.805). This implies a strong association between these factors.

2. The second rule indicates that individuals who are "Christian," and in "Good health" are strongly associated with just being "White" (Support: 0.188, Confidence: 0.935). This suggests that these attributes often co-occur together.

3. The third rule demonstrates a similar pattern, where individuals who are "Male," and have "No religion" are highly associated with just being "White" (Support: 0.128, Confidence: 0.934).

4. The fourth rule extends the previous finding by adding "Female" to the mix. It shows that individuals who are "Female," "Christian," in "Good health" are highly associated with just being "White" (Support: 0.102, Confidence: 0.933).

5. The fifth rule reveals that individuals who are "Christian" are associated with just being "White" (Support: 0.542, Confidence: 0.927). This indicates a strong association between these two attributes.

6. The sixth rule highlights that individuals who are in "Very good health" and have "No religion" are highly associated with just being "White" (Support: 0.124, Confidence: 0.929).

These association rules provide insights into the relationships between various attributes. The results also reflect how a large portion of the population affects this data mining technique as shown in the table that lots of combinations with high lift and confidence are always associated with being white. It's important to note that high confidence levels indicate a high likelihood of the consequent attribute given the antecedent attributes, making these rules useful for decision-making and targeted actions.

Clustering

Clustering is implemented with the following assumptions and limitations:

- All -9 is replaced by 0
- Data is then standardised
- Due to computing limitations only 1% of the data is sampled for hierarchical clustering

Table 16a

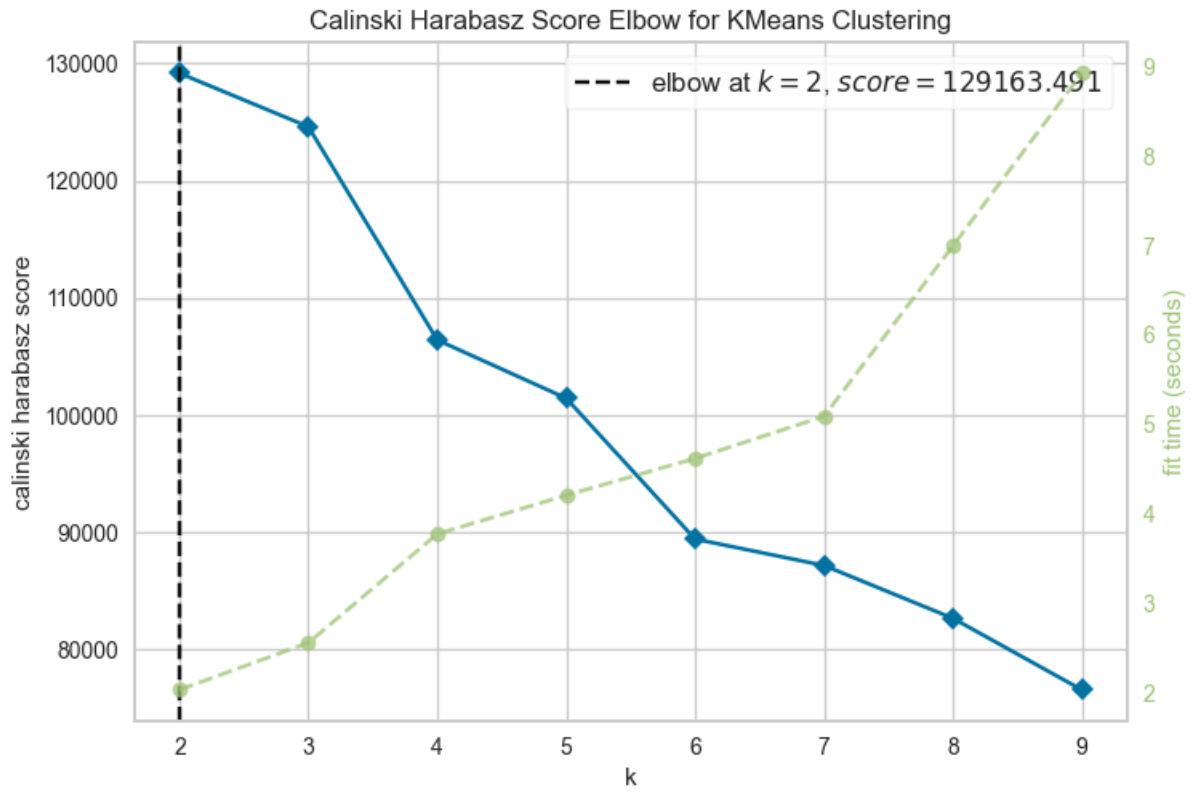
	Region	Residence Type	Family Composition	Population Base	Sex	Age	Marital Status	Student	Country of Birth
count	5.697400e+05	5.697400e+05	5.697400e+05	5.697400e+05	5.697400e+05	5.697400e+05	5.697400e+05	5.697400e+05	5.697400e+05
mean	-8.700014e-17	1.835783e-16	1.903627e-16	6.138898e-16	8.460564e-17	1.301510e-16	-5.492383e-17	1.995416e-16	-1.875691e-16
std	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00
min	-1.793840e+00	-7.244075e+00	-1.827682e+00	-1.177643e-01	-1.015217e+00	-1.342053e+00	-7.610541e-01	-1.871512e+00	-3.089466e+00
25%	-1.027515e+00	1.380438e-01	-2.450360e-01	-1.177643e-01	-1.015217e+00	-8.914996e-01	-7.610541e-01	5.343274e-01	-3.401406e-01
50%	1.219718e-01	1.380438e-01	-2.450360e-01	-1.177643e-01	9.850107e-01	9.607496e-03	1.278190e-01	5.343274e-01	-3.401406e-01
75%	8.882965e-01	1.380438e-01	5.462871e-01	-1.177643e-01	9.850107e-01	9.107146e-01	1.278190e-01	5.343274e-01	-3.401406e-01
max	1.654621e+00	1.380438e-01	2.920257e+00	1.245627e+01	9.850107e-01	1.811822e+00	2.794438e+00	5.343274e-01	2.409185e+00

Table 16b

Health	Ethnic Group	Religion	Economic Activity	Occupation	Industry	Hours worked per week	No of hours	Approximated Social Grade
5.697400e+05	5.697400e+05	5.697400e+05	5.697400e+05	5.697400e+05	5.697400e+05	5.697400e+05	5.697400e+05	5.697400e+05
2.115141e-17	-8.520427e-17	1.596333e-18	-3.671566e-17	5.886478e-17	-1.149360e-16	-5.587165e-16	-5.395605e-16	-3.970878e-17
1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00	1.000001e+00
-1.888134e+00	-1.549717e+00	-1.163526e+00	-9.970600e-01	-1.158152e+00	-1.201585e+00	-8.724132e-01	-8.321093e-01	-1.401689e+00
-8.188565e-01	-3.564589e-01	-7.030639e-01	-5.925055e-01	-1.158152e+00	-1.201585e+00	-8.724132e-01	-8.321093e-01	-6.989253e-01
2.504208e-01	-3.564589e-01	-2.426015e-01	-5.925055e-01	-1.957604e-01	-2.062965e-01	-8.724132e-01	-8.321093e-01	3.838595e-03
2.504208e-01	-3.564589e-01	-2.426015e-01	1.025713e+00	7.666311e-01	7.889920e-01	1.158720e+00	9.792127e-01	7.066024e-01
3.458252e+00	4.416572e+00	2.980635e+00	2.643931e+00	1.729023e+00	1.784281e+00	1.835764e+00	2.186761e+00	1.409366e+00

K-Means Clustering

Figure 18



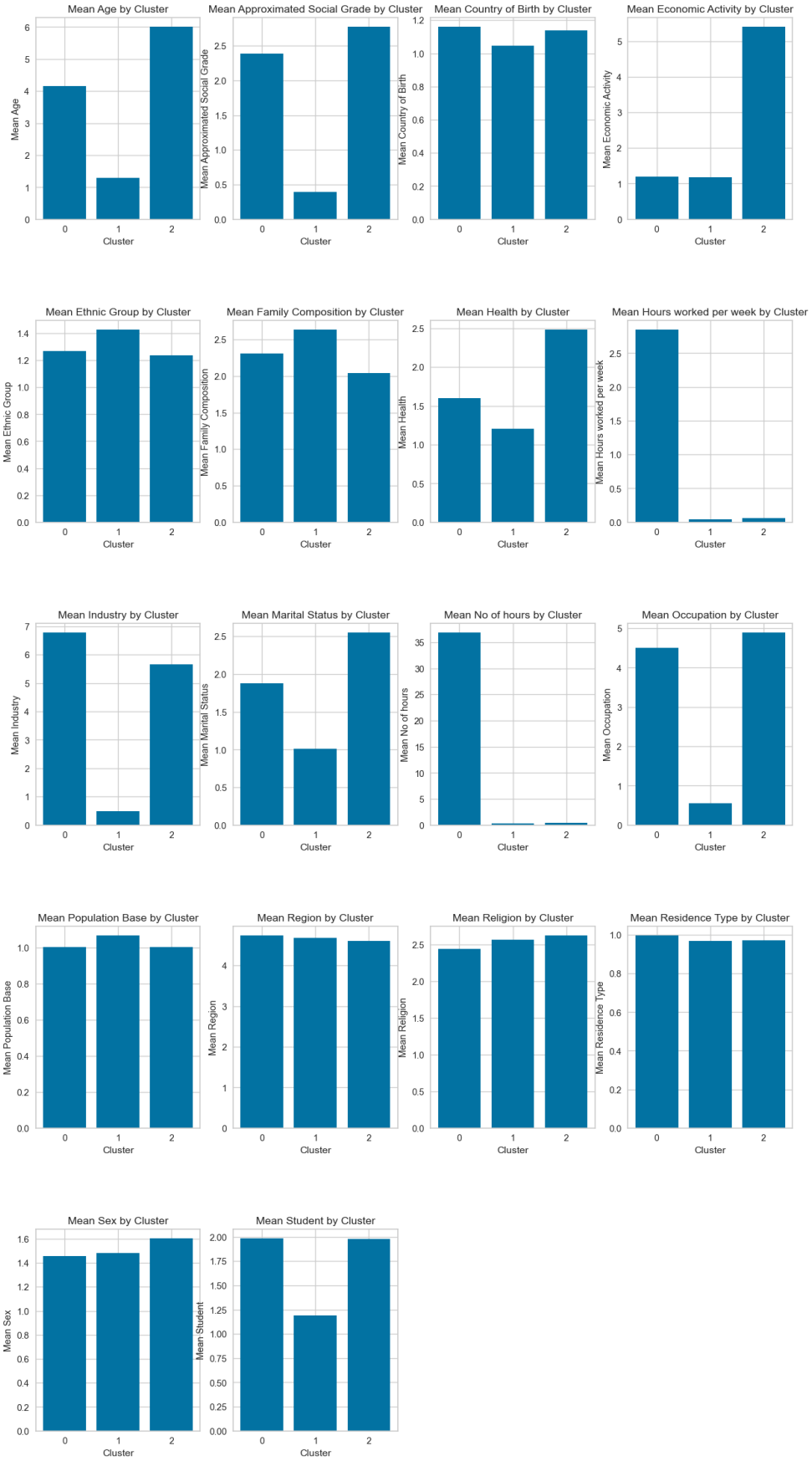
An elbow plot is implemented to find the optimum number of K. In this case, 3 is chosen as it is the point before the score drops drastically.

The model is then implemented on the census data set to give labels to each row. It is either 0, 1 or 2.

Figure 22 suggests Cluster 0 tends to have a slightly higher average age and social grade, indicating a potentially older and higher socioeconomic status population. Cluster 1 gives a younger average age and a lower social grade, suggesting a younger and possibly economically diverse group.

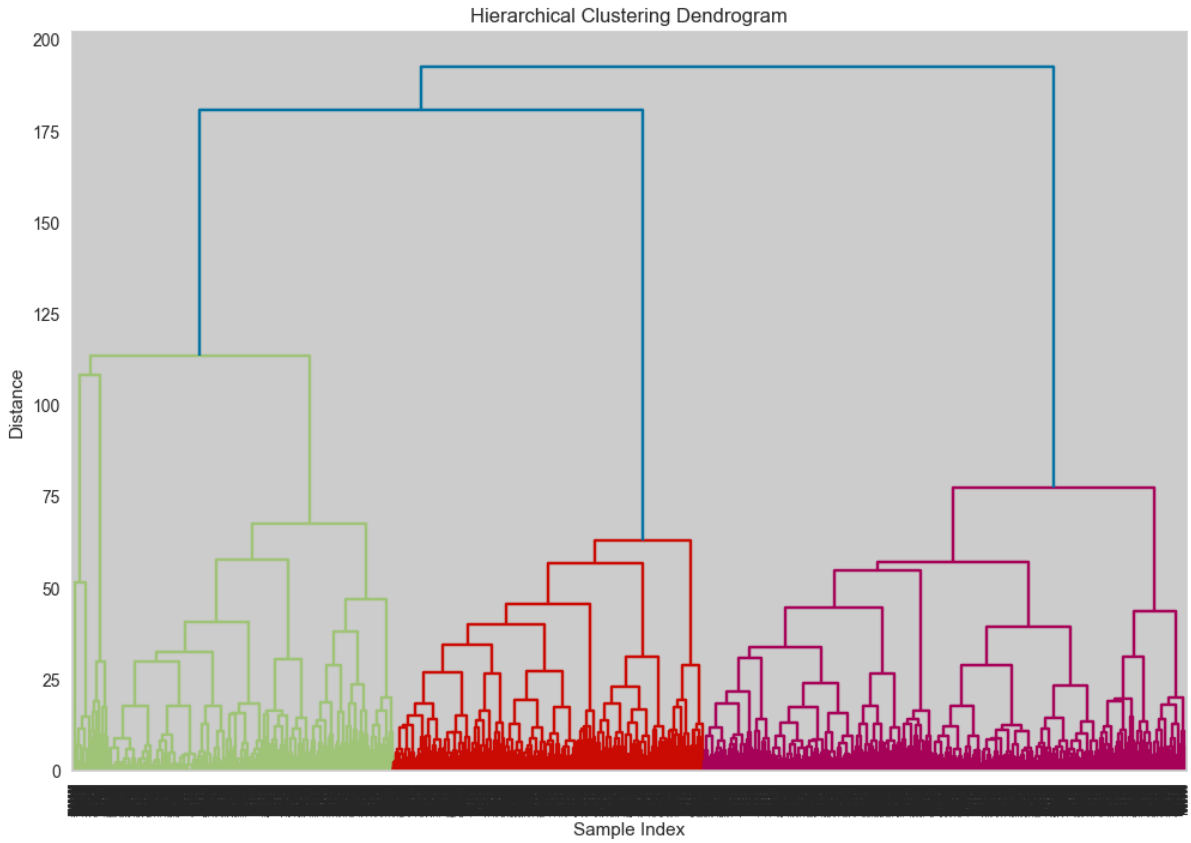
Cluster 2 has the highest average age and an even higher social grade, signifying an older and potentially more affluent population. Most of the people with working hours are segmented into Cluster 0 meaning there is a cluster of working and no working people.

Figure 19


Mean Age by Cluster
Mean Age
Cluster
Mean Approximated Social Grade by Cluster
Mean Approximated Social Grade
Cluster
Mean Country of Birth by Cluster
Mean Country of Birth
Cluster
Mean Economic Activity by Cluster
Mean Economic Activity
Cluster
Mean Ethnic Group by Cluster
Mean Ethnic Group
Cluster
Mean Family Composition by Cluster
Mean Family Composition
Cluster
Mean Health by Cluster
Mean Health
Cluster
Mean Hours worked per week by Cluster
Mean Hours worked per week
Cluster
Mean Industry by Cluster
Mean Industry
Cluster
Mean Marital Status by Cluster
Mean Marital Status
Cluster
Mean No of hours by Cluster
Mean No of hours
Cluster
Mean Occupation by Cluster
Mean Occupation
Cluster
Mean Population Base by Cluster
Mean Population Base
Cluster
Mean Region by Cluster
Mean Region
Cluster
Mean Religion by Cluster
Mean Religion
Cluster
Mean Residence Type by Cluster
Mean Residence Type
Cluster
Mean Sex by Cluster
Mean Sex
Cluster
Mean Student by Cluster
Mean Student
Cluster

Hierarchical Clustering

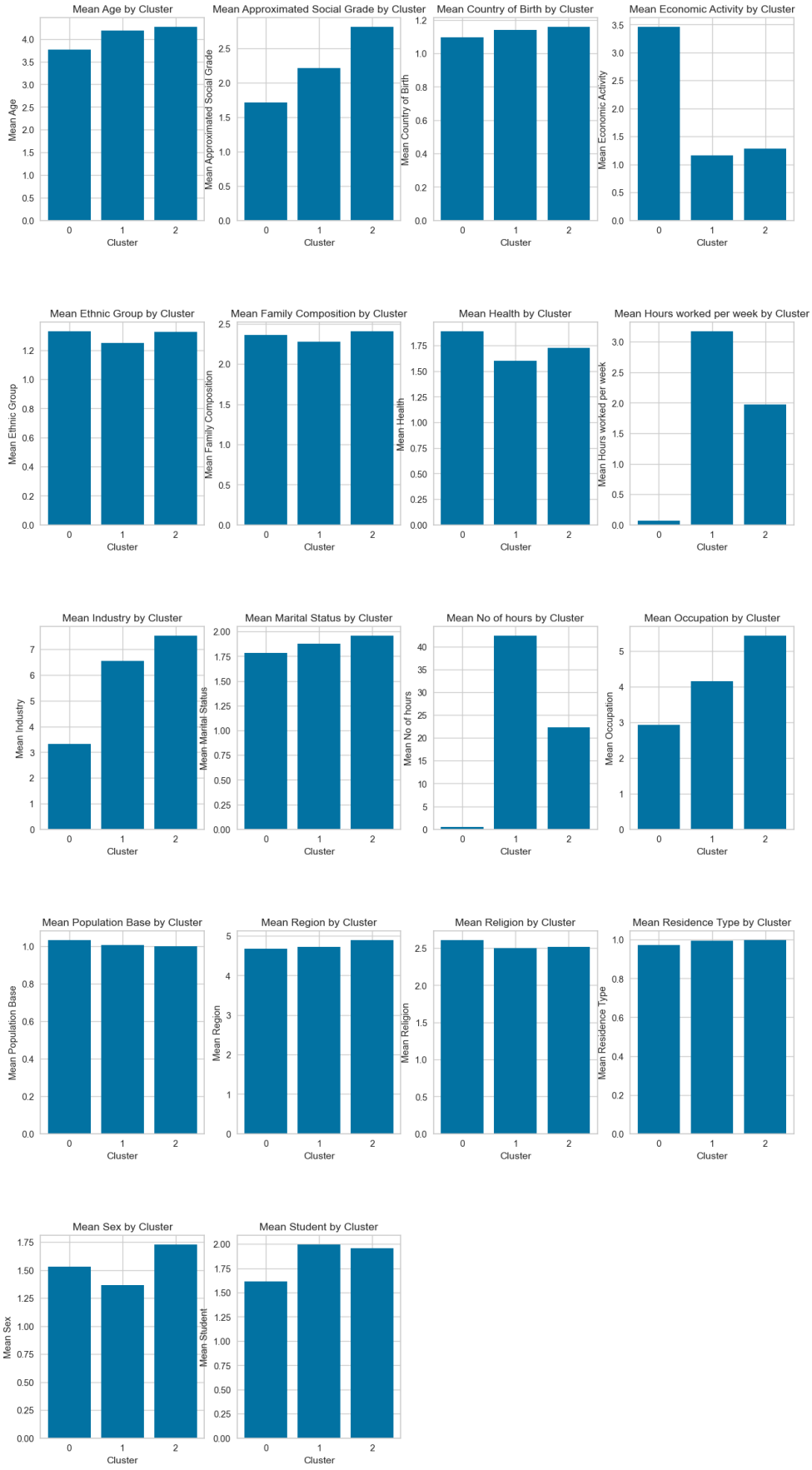
Figure 20



Based on the dendrogram the number of clusters is also 3 as the cut-off point in regards to the Y axis is before the 4th cluster is formed.

Figure 21 suggests Cluster 0 tends to have slightly lower average values for most attributes, indicating a more conservative or balanced profile. Cluster 1 shows variations, with higher values for attributes like "Hours worked per week" and "No of hours," suggesting a potentially more working-class or labour-intensive profile. Cluster 2 shows generally higher values for attributes like "Occupation" and "Industry," signifying a more professionally oriented or industrious group.

Figure 21


Mean Age by Cluster
Mean Age
Cluster
Mean Approximated Social Grade by Cluster
Mean Approximated Social Grade
Cluster
Mean Country of Birth by Cluster
Mean Country of Birth
Cluster
Mean Economic Activity by Cluster
Mean Economic Activity
Cluster
Mean Ethnic Group by Cluster
Mean Ethnic Group
Cluster
Mean Family Composition by Cluster
Mean Family Composition
Cluster
Mean Health by Cluster
Mean Health
Cluster
Mean Hours worked per week by Cluster
Mean Hours worked per week
Cluster
Mean Industry by Cluster
Mean Industry
Cluster
Mean Marital Status by Cluster
Mean Marital Status
Cluster
Mean No of hours by Cluster
Mean No of hours
Cluster
Mean Occupation by Cluster
Mean Occupation
Cluster
Mean Population Base by Cluster
Mean Population Base
Cluster
Mean Region by Cluster
Mean Region
Cluster
Mean Religion by Cluster
Mean Religion
Cluster
Mean Residence Type by Cluster
Mean Residence Type
Cluster
Mean Sex by Cluster
Mean Sex
Cluster
Mean Student by Cluster
Mean Student
Cluster

Comparison

Figure 22

Cluster 0 is characterized by higher values for features like "Age" and "Occupation," suggesting a potentially older and professionally oriented group. Cluster 1 shows lower values for most attributes, indicating a more balanced or conservative profile. Cluster 2 shows higher values for attributes like "Economic Activity" and "Industry," signifying a more industrious or economically active segment.

The Hierarchical algorithm also produced three clusters. Cluster 0 has slightly lower values for most attributes, indicating a potentially more balanced or conservative profile. Cluster 1 displays higher values for features like "Occupation" and "Industry," suggesting a more professionally oriented or industrious group. Cluster 2, in this case, exhibits variations, with higher values for attributes like "No of hours," implying a potentially more labour-intensive or working-class profile.

The choice between these algorithms depends on the specific context and goals. K-means clustering seems to focus more on variations in attributes like age and occupation, while Hierarchical clustering highlights variations in attributes related to hours worked and labour intensity. Depending on the application, one algorithm may be more suitable than the other for identifying distinct demographic segments.

Word count (Excluding tables and figures name and title): 2134