Table of Contents

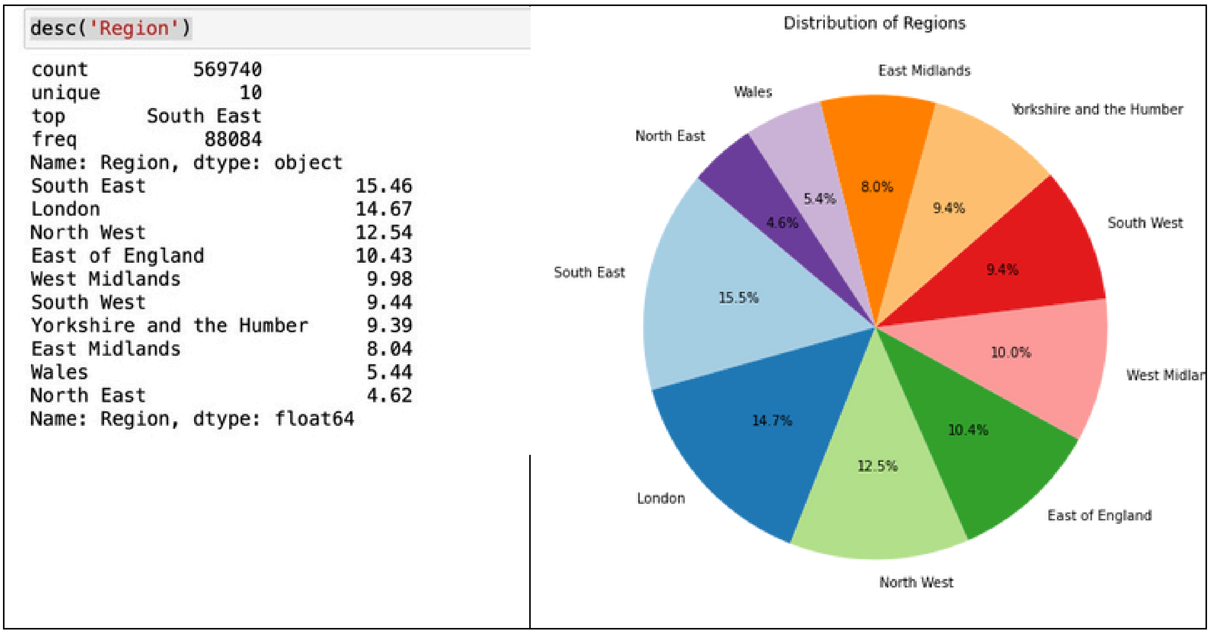
Summary

This report analyzes the ٢٠١١ UK census data to highlight key population trends، aiming to offer a concise overview of observed patterns.

١. Descriptiveanalytics

a. Basic Sta*s*cs

Figure١.١: Distribution of population in the Regions

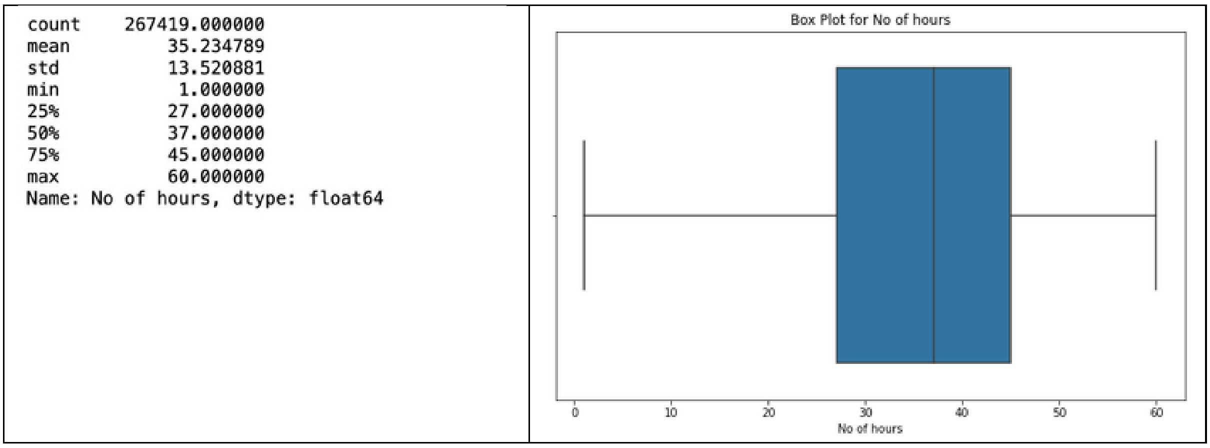


```
desc('Region')
```

```
count           569740
unique              10
top         South East
freq             88084
Name: Region, dtype: object
South East                  15.46
London                      14.67
North West                  12.54
East of England             10.43
West Midlands                9.98
South West                   9.44
Yorkshire and the Humber     9.39
East Midlands                8.04
Wales                        5.44
North East                   4.62
Name: Region, dtype: float64
```

Among the ١٠ distinct regions in the UK، the Southeast boasts the largest population، with approximately ٨٨،٠٨٤ residents، constituting around ١٥.٥٪ of the total population. In contrast، the North East exhibits the lowest population، accounting for approximately ٤.٦٢٪.

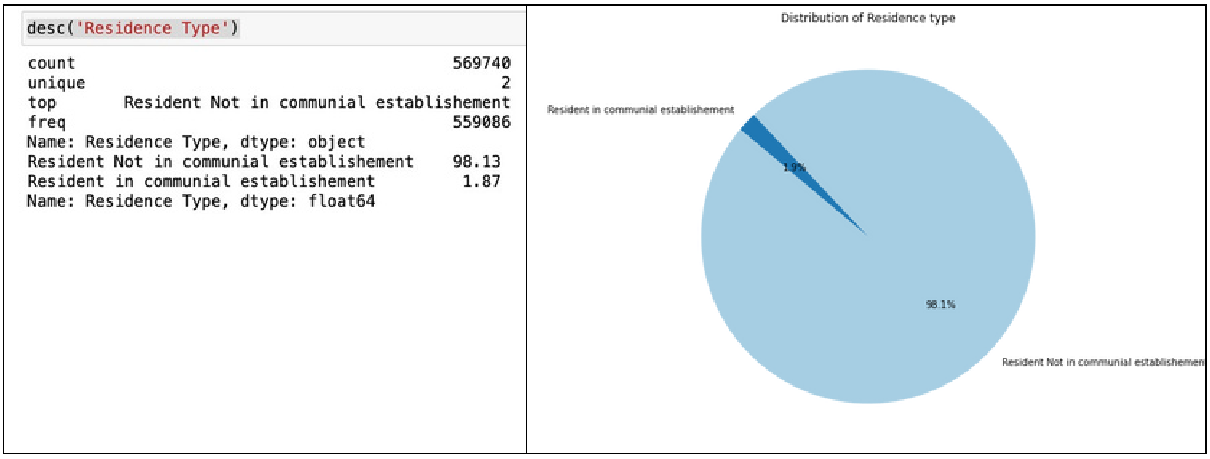
Figure١.٢: Trends on No. of hours worked



```
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
```

On average، individuals worked approximately ٣٥.٢٣ hours، with an average deviation of about ١٣.٥٢ hours. The number of hours worked ranges from ١ to ٦٠. The median value is ٣٧، indicating that half of the individuals worked less than ٣٧ hours، and half worked more than ٣٧ hours. Examining quartiles، ٢٥٪ of the individuals worked ٢٧ hours or fewer، signifying that ٧٥٪ of the population worked more than ٢٧ hours. Additionally، ٧٥٪ of the individuals worked ٤٥ hours or fewer، implying that only ٢٥٪ of the population worked more than ٤٥ hours.

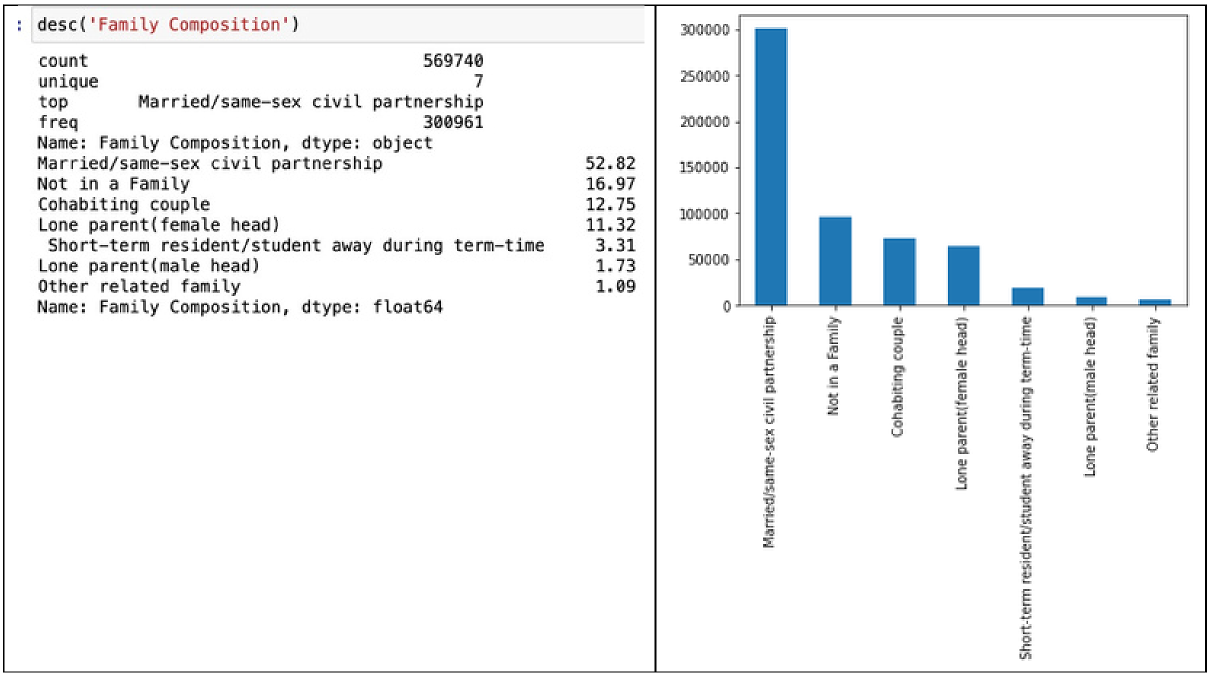
Figure ١.٣: Distribution of Residence Type



```
desc('Residence Type')

count                                           569740
unique                                               2
top          Resident Not in communial establishement
freq                                            559086
Name: Residence Type, dtype: object
Resident Not in communial establishement    98.13
Resident in communial establishement         1.87
Name: Residence Type, dtype: float64
```

The population is categorized into two types of residences، with ٩٨٪ of individuals living outside communal establishments and only ١.٨٧٪ residing in such establishments.

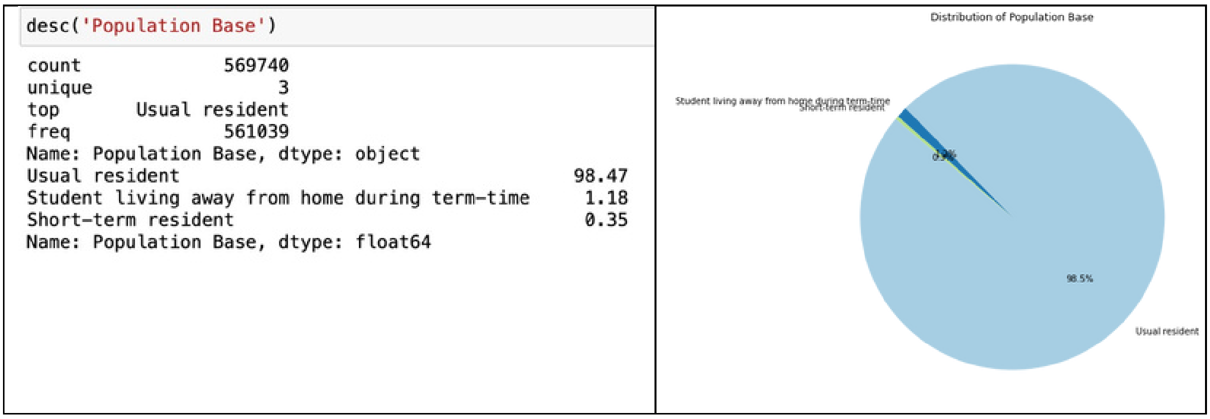
Figure ١.٤: Distribution of Family Composition



```
: desc('Family Composition')

 count                              569740
 unique                                  7
 top       Married/same-sex civil partnership
 freq                               300961
 Name: Family Composition, dtype: object
 Married/same-sex civil partnership                              52.82
 Not in a Family                                                 16.97
 Cohabiting couple                                               12.75
 Lone parent(female head)                                        11.32
  Short-term resident/student away during term-time               3.31
 Lone parent(male head)                                           1.73
 Other related family                                             1.09
 Name: Family Composition, dtype: float64
```

There are seven different groups that describe the makeup of families. The predominant group is Married/Same-Sex Civil Partnership، with ٥٢.٨٪ of the population in either a marriage or in a registered same-sex civil partnership.

Figure ١.٥: Population Base

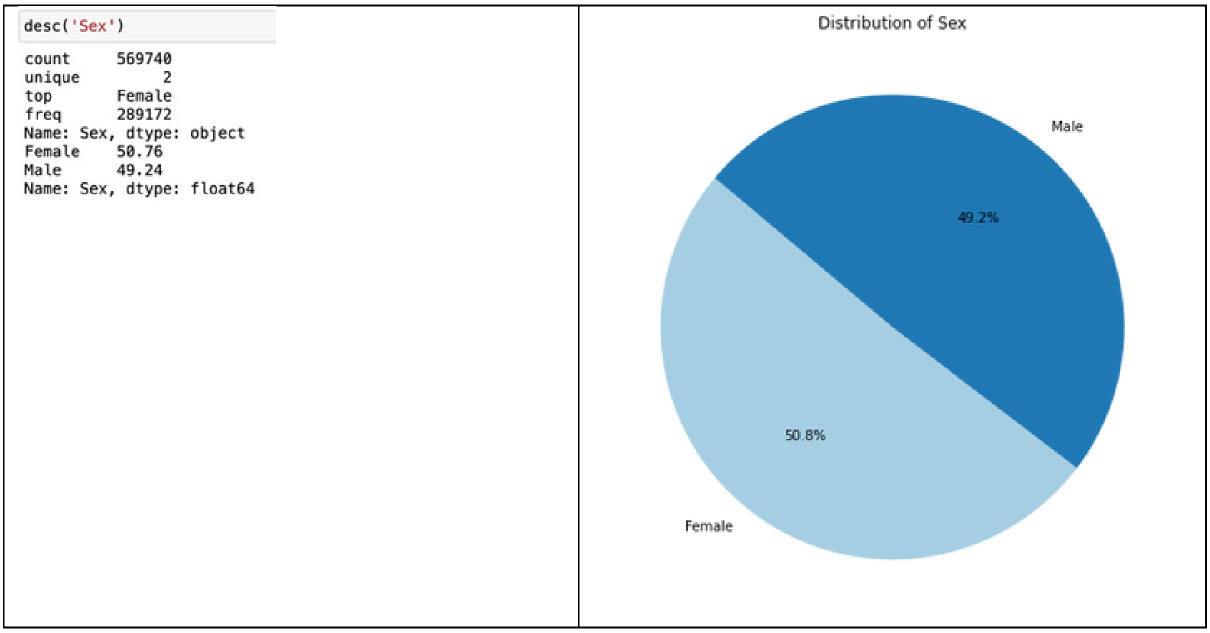


```
desc('Population Base')

count              569740
unique                  3
top        Usual resident
freq               561039
Name: Population Base, dtype: object
Usual resident                                     98.47
Student living away from home during term-time      1.18
Short-term resident                                 0.35
Name: Population Base, dtype: float64
```

Distribution of Population Base
Student living away from home during term-time
Short-term resident
98.5%
Usual resident

The population base has three different categories. The most common one being Usual Resident، with ٩٨٪ of the population reporting to be usual residents and ١٪ indicating that they are students living away during term time.

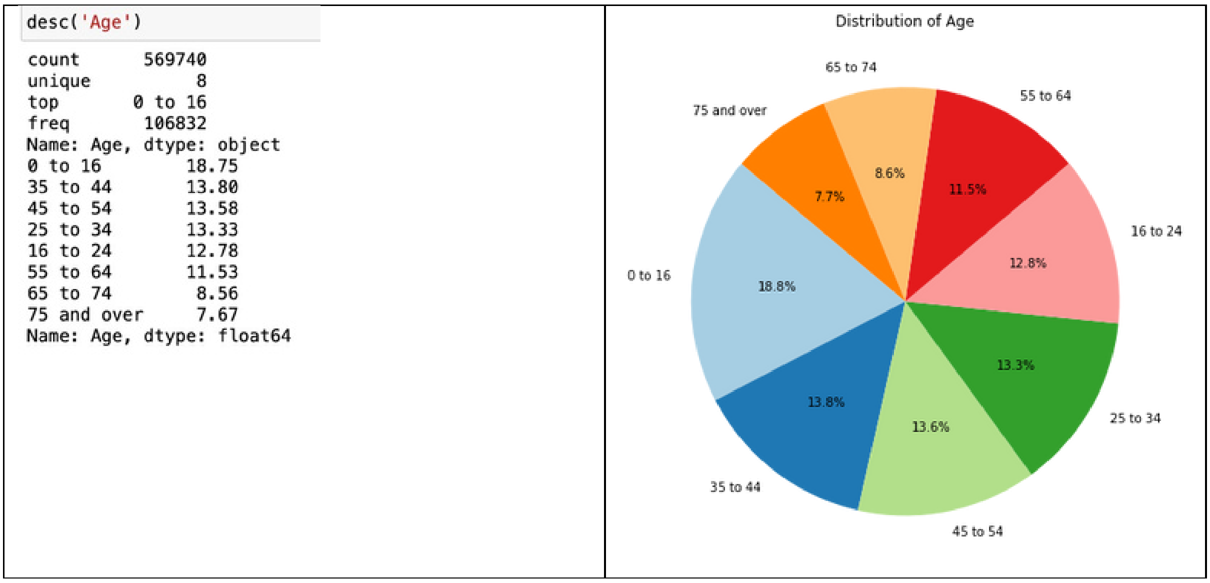
Figure ١.٦: Sex



```
desc('Sex')

count     569740
unique         2
top       Female
freq      289172
Name: Sex, dtype: object
Female    50.76
Male      49.24
Name: Sex, dtype: float64
```

The gender for the population is divided into two distinct categories. Females account for ٥٠.٨٪ of the population while male account for ٤٩.٢٤٪.

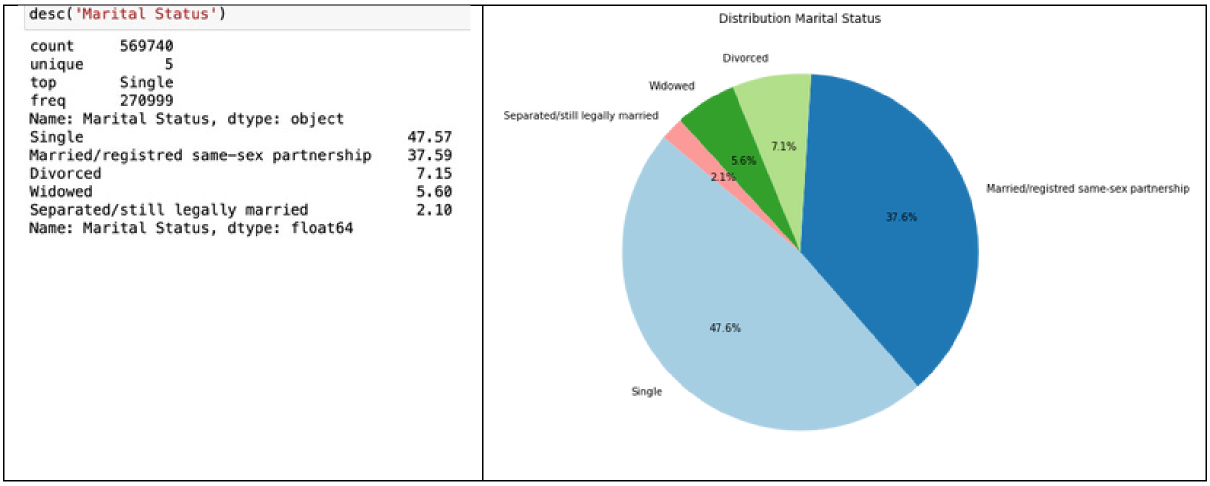
Figure ١.٧: Age



```
desc('Age')

count        569740
unique            8
top          0 to 16
freq         106832
Name: Age, dtype: object
0 to 16          18.75
35 to 44         13.80
45 to 54         13.58
25 to 34         13.33
16 to 24         12.78
55 to 64         11.53
65 to 74          8.56
75 and over       7.67
Name: Age, dtype: float64
```

The population's age distribution is segmented into eight distinct categories، with the most prevalent category being ٠ to ١٦، constituting ١٨.٨٪ of the population. In contrast، the age group of ٧٥ and over represents the smallest proportion، accounting for ٧.٦٧٪.

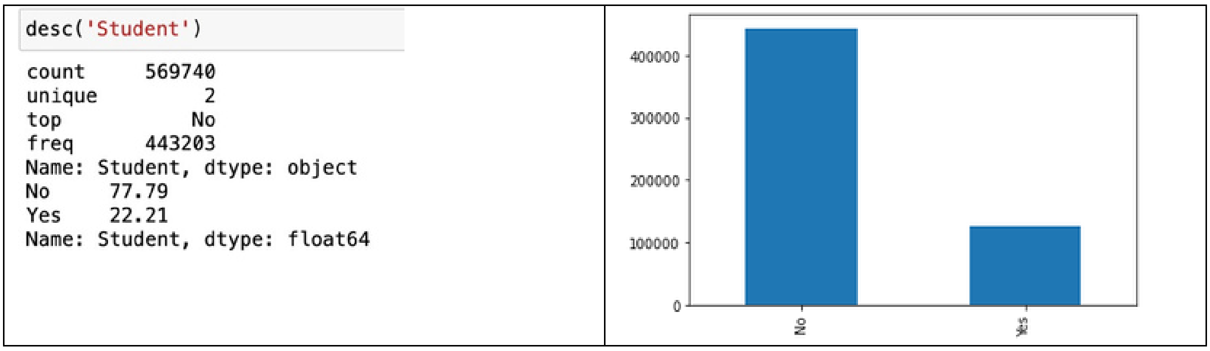
Figure ١.٨: Marital Status



```
desc('Marital Status')

count       569740
unique           5
top         Single
freq        270999
Name: Marital Status, dtype: object
Single                                    47.57
Married/registred same-sex partnership    37.59
Divorced                                   7.15
Widowed                                    5.60
Separated/still legally married            2.10
Name: Marital Status, dtype: float64
```

The relationship statuses of UK residents are distributed among five distinct categories. Notably، the most prominent category is "single،" comprising ٤٧.٦٪ of the population، while only ٢.١٠٪ are either separated or still legally married.

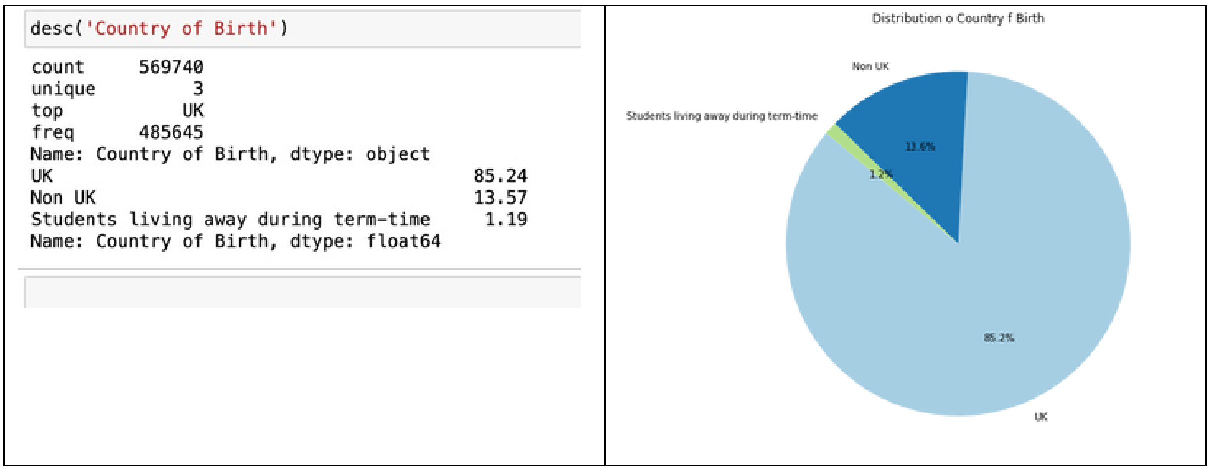
Figure ١.٩: Student



```
desc('Student')

count      569740
unique          2
top            No
freq       443203
Name: Student, dtype: object
No     77.79
Yes    22.21
Name: Student, dtype: float64
```

In the UK، ٧٧.٨٪ of individuals are non-students، while ٢٢.٢١٪ are students.

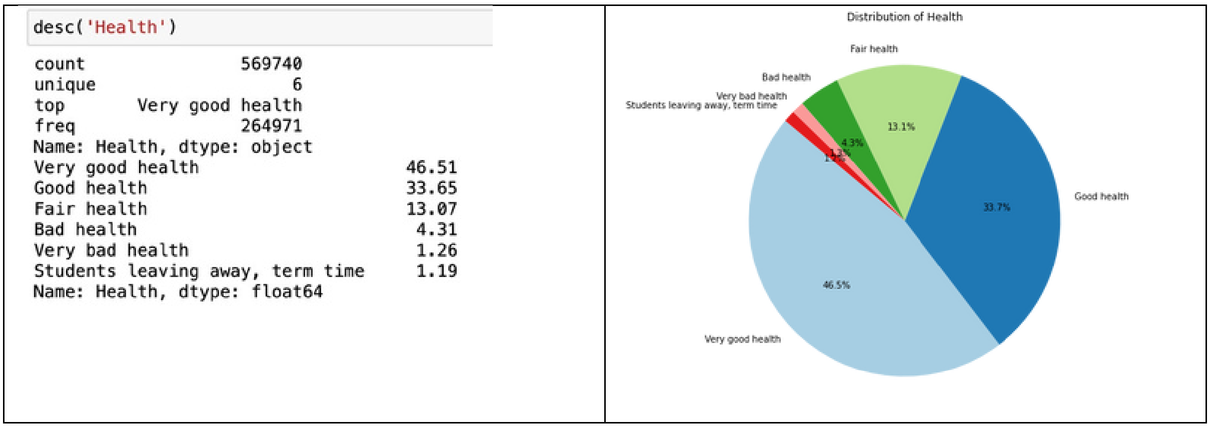
Figure ١:١٠: Country of Birth



```
desc('Country of Birth')

count      569740
unique          3
top            UK
freq       485645
Name: Country of Birth, dtype: object
UK                                     85.24
Non UK                                 13.57
Students living away during term-time   1.19
Name: Country of Birth, dtype: float64
```

Residence in the UK are categorized into three groups based on their place of birth. Approximately ٨٥.٢٪ were born in the UK، while ١٣٪ were born outside the UK.

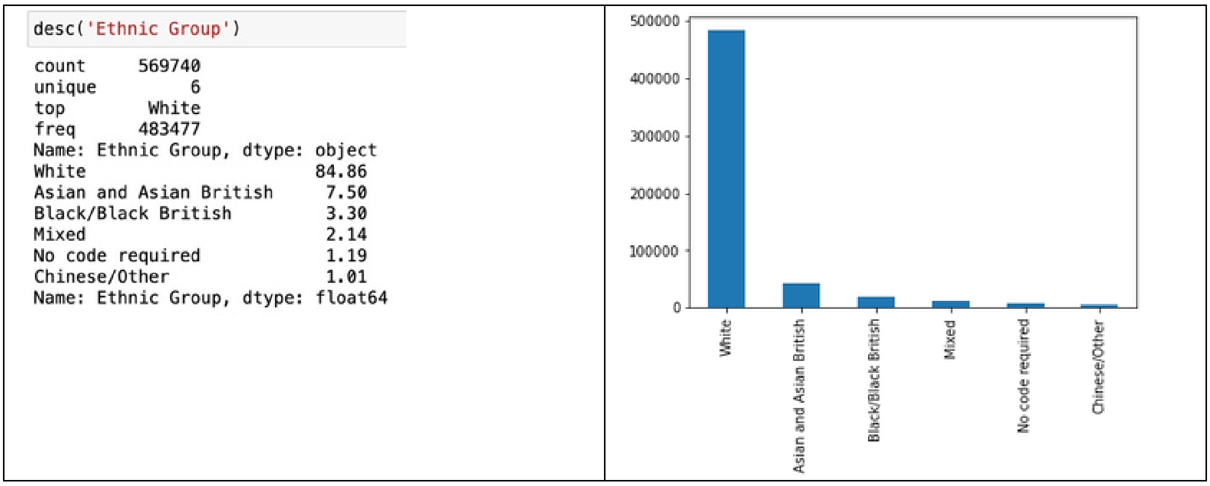
Figure ١.١١: Health



```
desc('Health')

count                   569740
unique                       6
top           Very good health
freq                    264971
Name: Health, dtype: object
Very good health                   46.51
Good health                        33.65
Fair health                        13.07
Bad health                          4.31
Very bad health                     1.26
Students leaving away, term time    1.19
Name: Health, dtype: float64
```

There are six distinct health statuses، with ٤٦.٥٪ of individuals reporting very good health، andonly ١.٢٦٪ indicating very bad health.

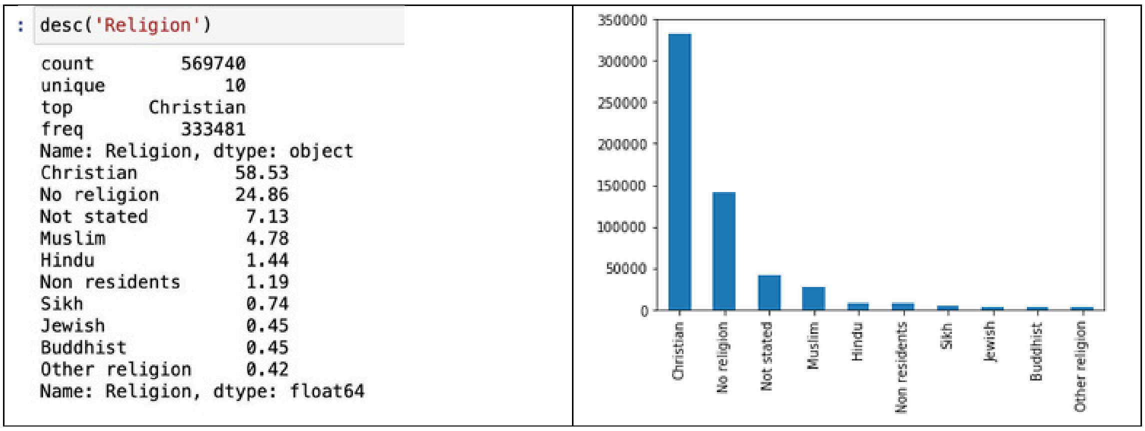
Figure ١.١٢: Ethnic Group



```
desc('Ethnic Group')

count      569740
unique          6
top         White
freq       483477
Name: Ethnic Group, dtype: object
White                      84.86
Asian and Asian British     7.50
Black/Black British         3.30
Mixed                       2.14
No code required            1.19
Chinese/Other               1.01
Name: Ethnic Group, dtype: float64
```

There are six distinct ethnic groups in the UK. The most common one is white as ٨٤.٩٪ of individuals in the UK identify as being white while ١٪ are of Chinese ethnic group.

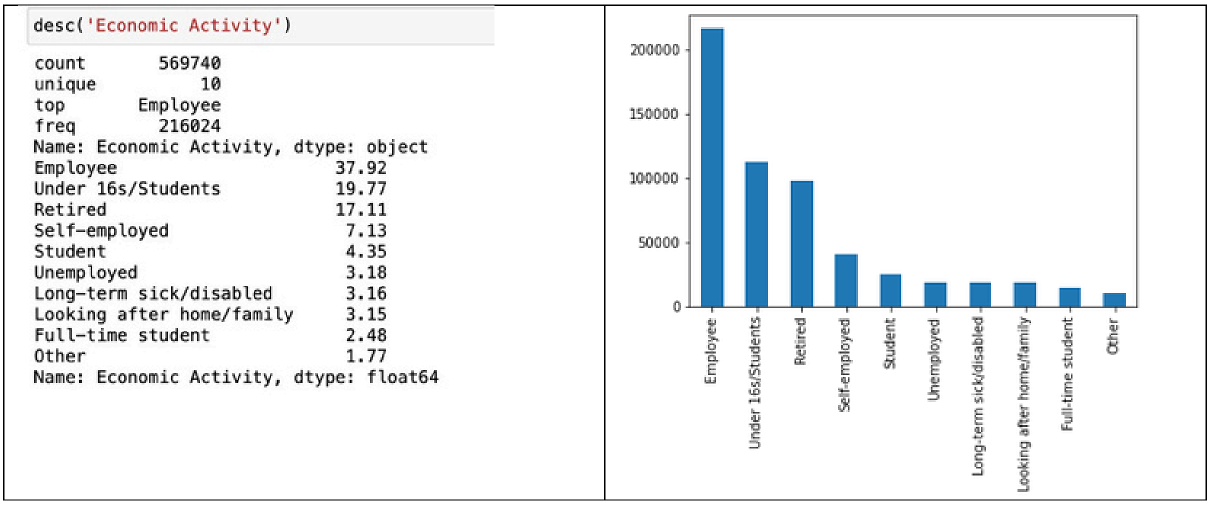
Figure ١.١٣: Religion



```
: desc('Religion')

count          569740
unique             10
top         Christian
freq           333481
Name: Religion, dtype: object
Christian          58.53
No religion        24.86
Not stated          7.13
Muslim              4.78
Hindu               1.44
Non residents       1.19
Sikh                0.74
Jewish              0.45
Buddhist            0.45
Other religion      0.42
Name: Religion, dtype: float64
```

There are ten distinct religious affiliations in the UK. Christianity stands out as the most prevalent religion، with ٥٨.٥٪ of individuals identifying as Christians، while only ٠.٤٥٪ and ٠.٤٢٪ follow Buddhism and other religions، respectively.

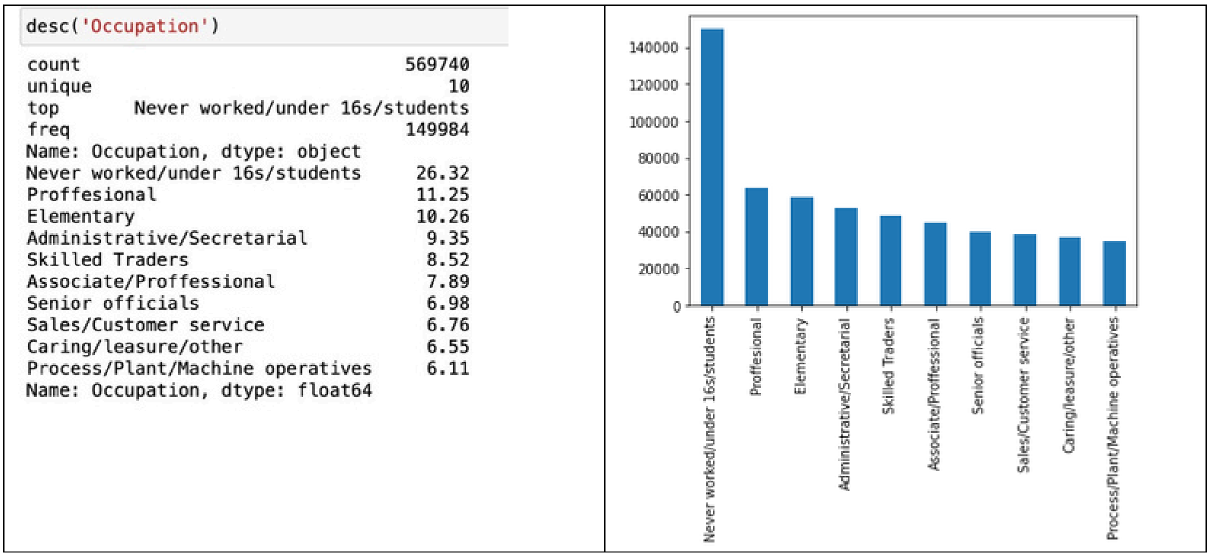
Figure ١.١٤: Economic Activity



```
desc('Economic Activity')

count        569740
unique           10
top        Employee
freq         216024
Name: Economic Activity, dtype: object
Employee                     37.92
Under 16s/Students           19.77
Retired                      17.11
Self-employed                 7.13
Student                       4.35
Unemployed                    3.18
Long-term sick/disabled       3.16
Looking after home/family     3.15
Full-time student             2.48
Other                         1.77
Name: Economic Activity, dtype: float64
```

Individuals participate in ١٠ diverse economic activities in the UK. The predominant economic activity reported is employment، with ٣٧.٩٪ of the population being employed.

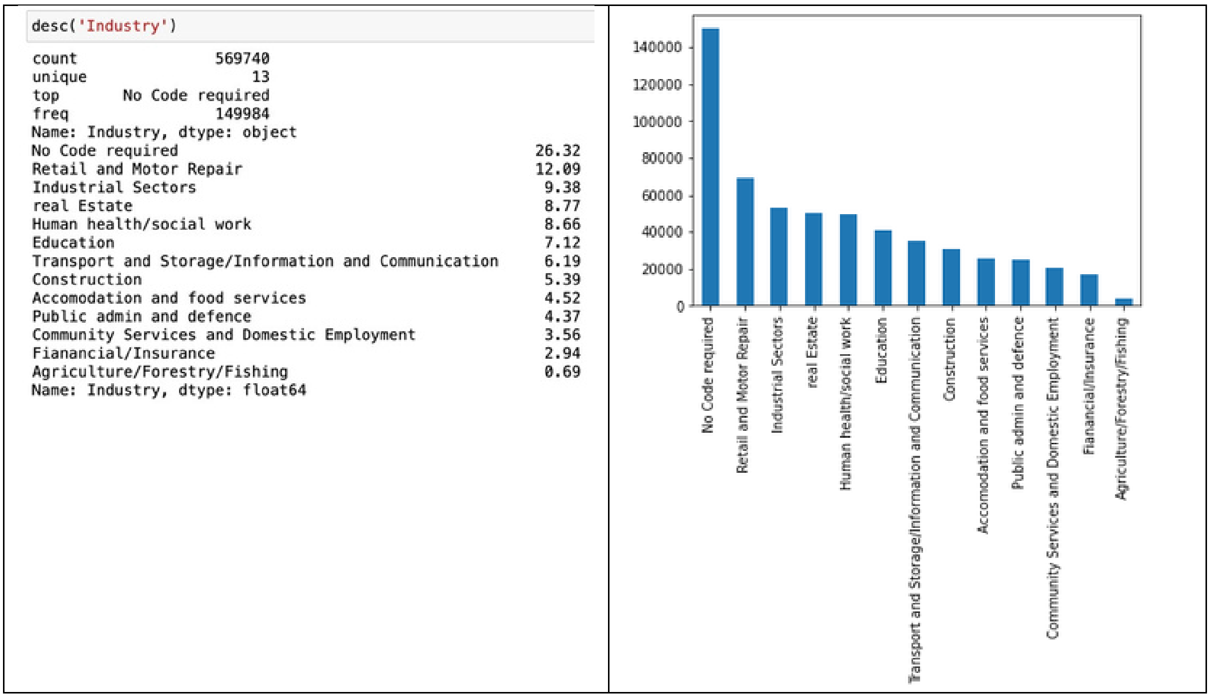
Figure ١.١٥: Occupation



```
desc('Occupation')

count                                 569740
unique                                    10
top       Never worked/under 16s/students
freq                                  149984
Name: Occupation, dtype: object
Never worked/under 16s/students        26.32
Proffesional                           11.25
Elementary                             10.26
Administrative/Secretarial              9.35
Skilled Traders                         8.52
Associate/Proffessional                 7.89
Senior officials                        6.98
Sales/Customer service                  6.76
Caring/leasure/other                    6.55
Process/Plant/Machine operatives        6.11
Name: Occupation, dtype: float64
```

Among the ١٠ occupation categories، the one most identified with by people is "Never worked/under ١٦s/Students،" representing ٢٦.٣٪ of the population. Conversely، ٦٪ reported working in the process/plant/machine operatives sector.

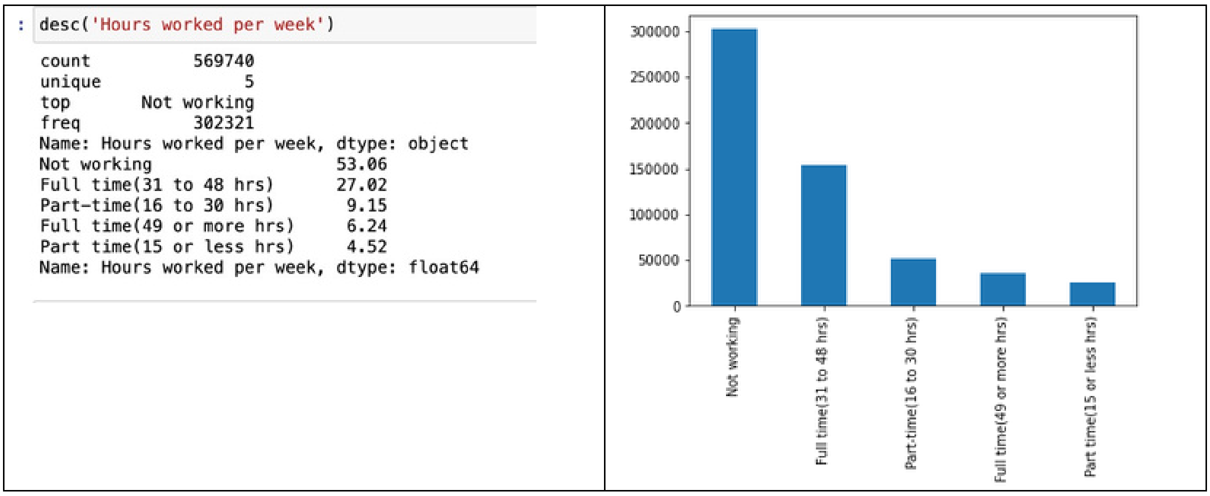
Figure ١.١٦: Industry



```
desc('Industry')

count                   569740
unique                      13
top          No Code required
freq                    149984
Name: Industry, dtype: object
No Code required                                          26.32
Retail and Motor Repair                                   12.09
Industrial Sectors                                         9.38
real Estate                                                8.77
Human health/social work                                   8.66
Education                                                  7.12
Transport and Storage/Information and Communication        6.19
Construction                                               5.39
Accomodation and food services                             4.52
Public admin and defence                                   4.37
Community Services and Domestic Employment                 3.56
Fianancial/Insurance                                       2.94
Agriculture/Forestry/Fishing                               0.69
Name: Industry, dtype: float64
```

The industries are categorized into ١٣ unique sectors. The sector with the highest number of people employed is the retail and motor repair industry، encompassing ١٢٪ of the population. In contrast، only ٠.٦٩٪ of individuals work in the agriculture/forestry/fishing industry.

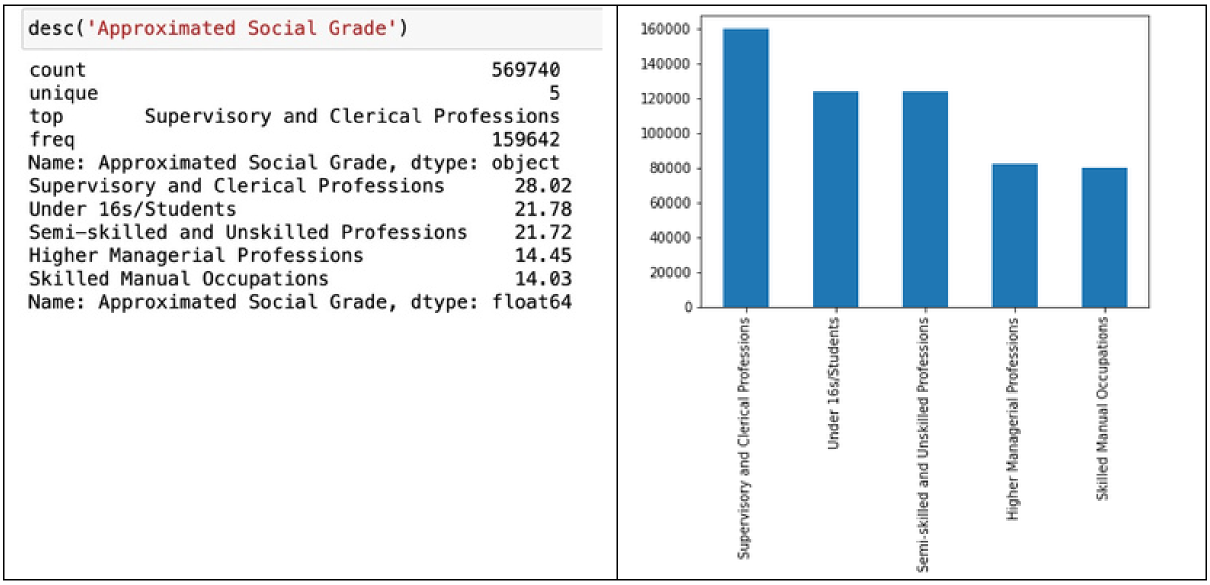
Figure ١.١٧: Hours worked per week



```
: desc('Hours worked per week')

  count            569740
  unique                5
  top         Not working
  freq             302321
  Name: Hours worked per week, dtype: object
  Not working                   53.06
  Full time(31 to 48 hrs)       27.02
  Part-time(16 to 30 hrs)        9.15
  Full time(49 or more hrs)      6.24
  Part time(15 or less hrs)      4.52
  Name: Hours worked per week, dtype: float64
```

The weekly working hours are segmented into four distinct categories. Just over half (٥٣٪) of the overall population indicated that they are not working. The smallest proportion، at ٤.٥٪، corresponds to individuals working part-time، specifically ١٥ hours or less.

Figure: ١.١٨: Approximate Social Grade



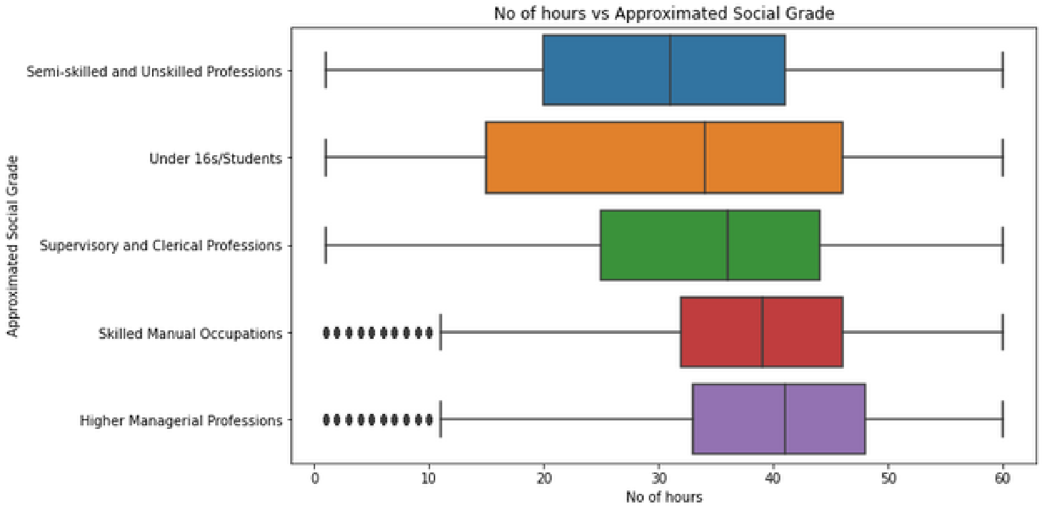
```
desc('Approximated Social Grade')

count                                       569740
unique                                           5
top          Supervisory and Clerical Professions
freq                                        159642
Name: Approximated Social Grade, dtype: object
Supervisory and Clerical Professions         28.02
Under 16s/Students                           21.78
Semi-skilled and Unskilled Professions       21.72
Higher Managerial Professions                14.45
Skilled Manual Occupations                   14.03
Name: Approximated Social Grade, dtype: float64
```

160000
140000
120000
100000
80000
60000
40000
20000
0
Supervisory and Clerical Professions
Under 16s/Students
Semi-skilled and Unskilled Professions
Higher Managerial Professions
Skilled Manual Occupations

There are five diverse groups representing approximations of social grades. The prevalent category is Supervisory/Clerical Professions، with ٢٨٪ of the population engaged in supervisory or clerical professions، while skilled manual work comprises ١٤٪ of the total.

b. Visualiza*on graphs

Figure ٢.١: Trends between approximated Social Grade and No. of hours.



Students and individuals under the age of ١٦ exhibit a diverse range of working hours، spanning from a minimum of ١٥ hours to a maximum of approximately ٤٧ hours. The distribution appears to be skewed towards fewer working hours.

In contrast، skilled manual occupations and higher managerial positions display a more

homogeneous

distribution with less variation in the hours worked. Both groups demonstrate a relatively

narrow range،

hovering between approximately ٣٣ and ٤٨ hours. The average hours worked for these

groups align

with their means، indicating an equal distribution of individuals working fewer or more than

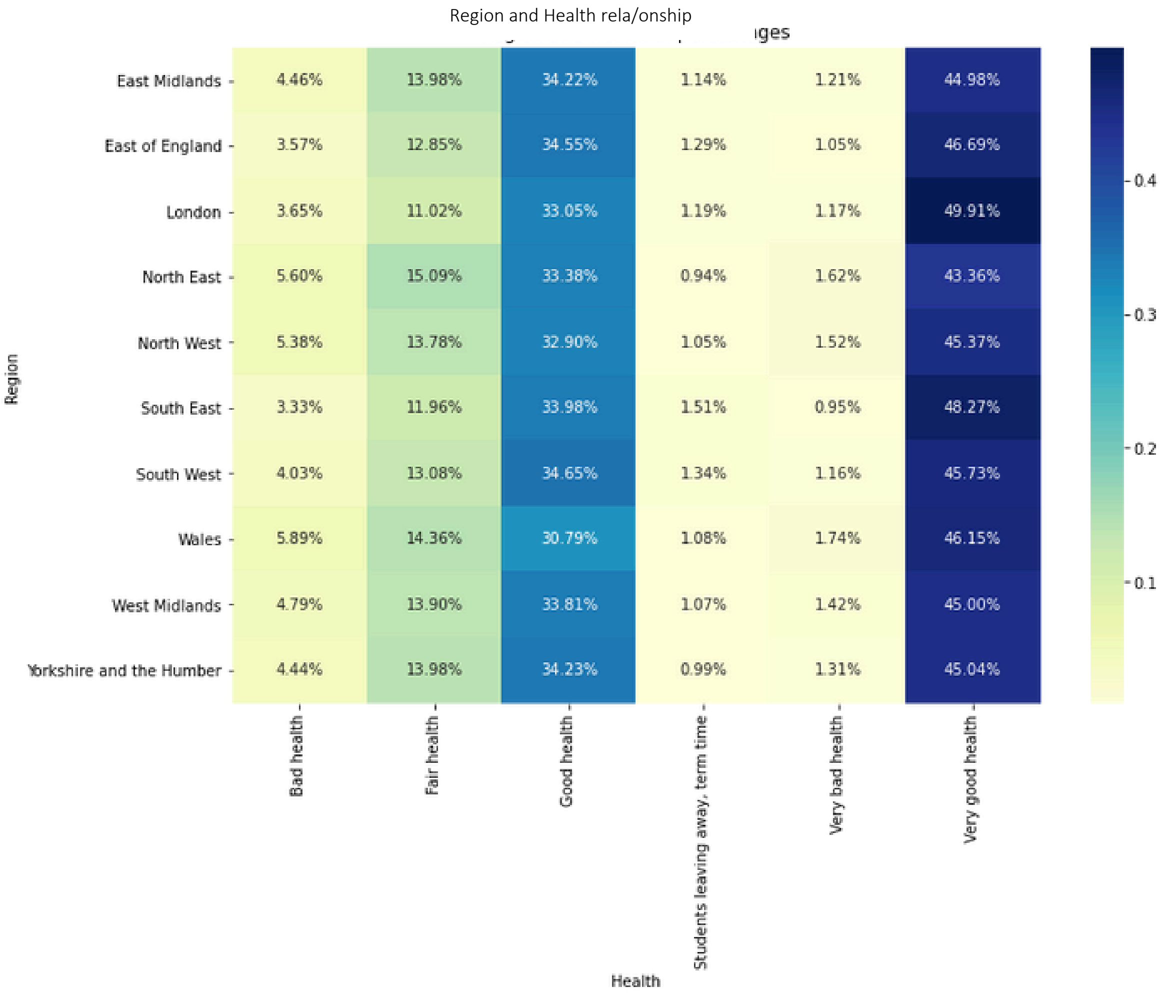
the average

hours. Overall، this suggests distinct patterns in working hours distributions based on

demographic and

occupational factors.

Figure ٢.٢: Trends between health status and geographic regions

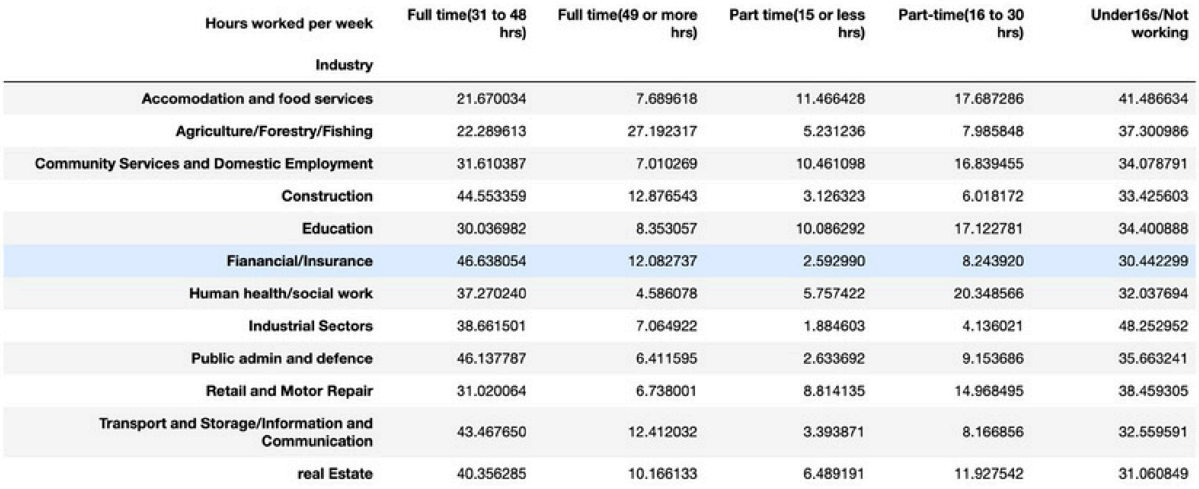


Overall، a substantial majority of residents in all ten regions reported being in very go
London and the Southeast exhibit the highest number of inhabitants in this category، while

the Northeast
region has comparatively fewer individuals reporting very good health.
To delve into specifics، London and the Southeast regions stand out with approximately ٥٠٪

and ٤٨٪
respectively declaring very good health. In these regions، approximately ٣٣٪ and ٣٥٪ of
individuals،
respectively، report being in good health.

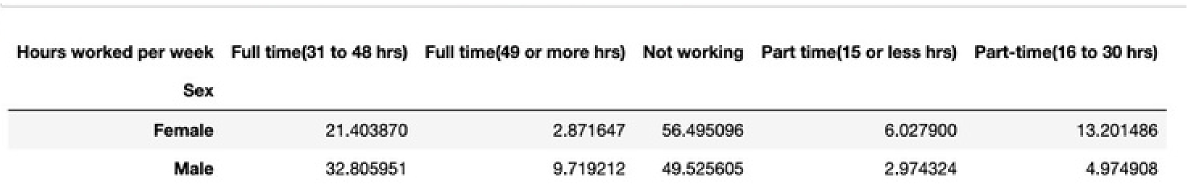
Figure ٢.٣: Table representing correlation between Industry and Hours worked per week



Hours worked per week Industry	Full time(31 to 48 hrs)	Full time(49 or more hrs)	Part time(15 or less hrs)	Part-time(16 to 30 hrs)	Under16s/Not working
Accomodation and food services	21.670034	7.689618	11.466428	17.687286	41.486634
Agriculture/Forestry/Fishing	22.289613	27.192317	5.231236	7.985848	37.300986
Community Services and Domestic Employment	31.610387	7.010269	10.461098	16.839455	34.078791
Construction	44.553359	12.876543	3.126323	6.018172	33.425603
Education	30.036982	8.353057	10.086292	17.122781	34.400888
Fianancial/Insurance	46.638054	12.082737	2.592990	8.243920	30.442299
Human health/social work	37.270240	4.586078	5.757422	20.348566	32.037694
Industrial Sectors	38.661501	7.064922	1.884603	4.136021	48.252952
Public admin and defence	46.137787	6.411595	2.633692	9.153686	35.663241
Retail and Motor Repair	31.020064	6.738001	8.814135	14.968495	38.459305
Transport and Storage/Information and Communication	43.467650	12.412032	3.393871	8.166856	32.559591
real Estate	40.356285	10.166133	6.489191	11.927542	31.060849

Overall، the financial and insurance sectors، as well as public administration and defense، showed the highest proportion of individuals working full time (٣١ to ٤٨ hours) both accounting for ٤٦٪ of all hours worked across those industries. Notably، the industrial sectors had the highest percentage of individuals indicating they were not working، as ٤٨٪ of individuals reported to be not working. This percentage was also the highest relative to individuals in other industries that indicated they are not working.

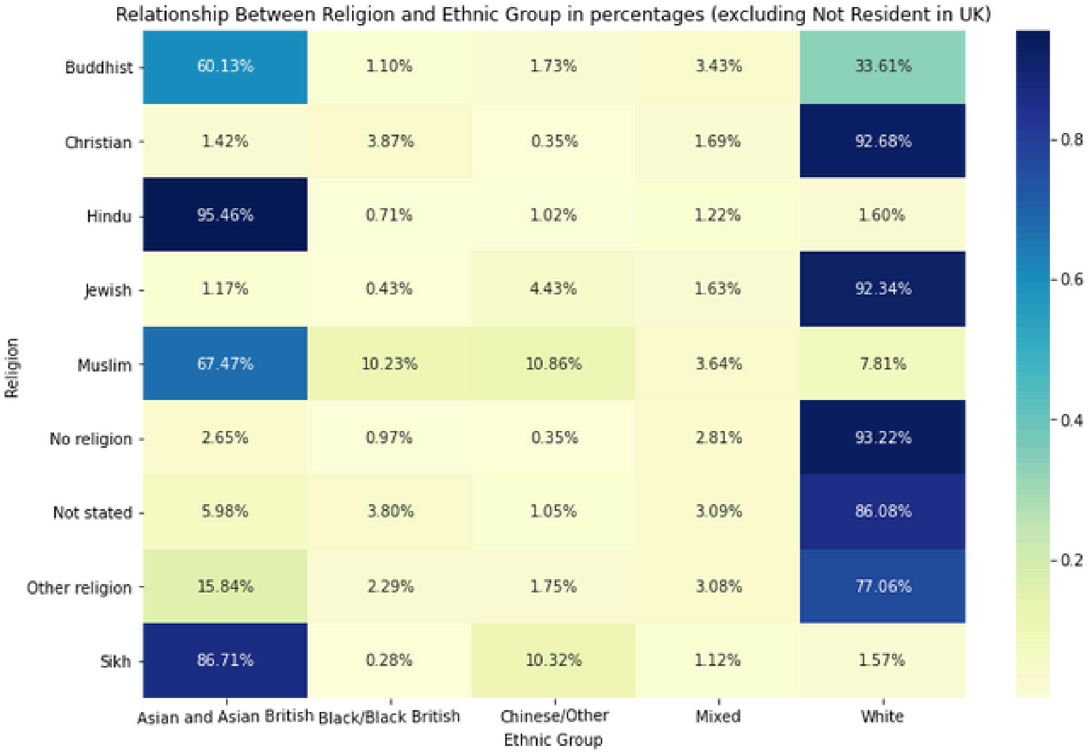
Figure ٢.٤: Table showing correlation between Sex and hours worked per week



Hours worked per week Sex	Full time(31 to 48 hrs)	Full time(49 or more hrs)	Not working	Part time(15 or less hrs)	Part-time(16 to 30 hrs)
Female	21.403870	2.871647	56.495096	6.027900	13.201486
Male	32.805951	9.719212	49.525605	2.974324	4.974908

The data reveals distinctive patterns in the distribution of working hours among women and men. Notably، ٥٦.٥٪ of women are not working relative to ٤٩.٥٪ of men، while ٣٢.٨٪ of men work full time between ٣١ and ٤٨ hours per week compared with ٢١.٤٪ of women in the same category. Additionally، approximately ١٣٪ of women are engaged in part-time employment، specifically working between ١٦ to ٣٠ hours per week، compared with ٤.٩٪ of men in the same category.

Figure ٢.٥: Trends between Religion and Ethnic group.



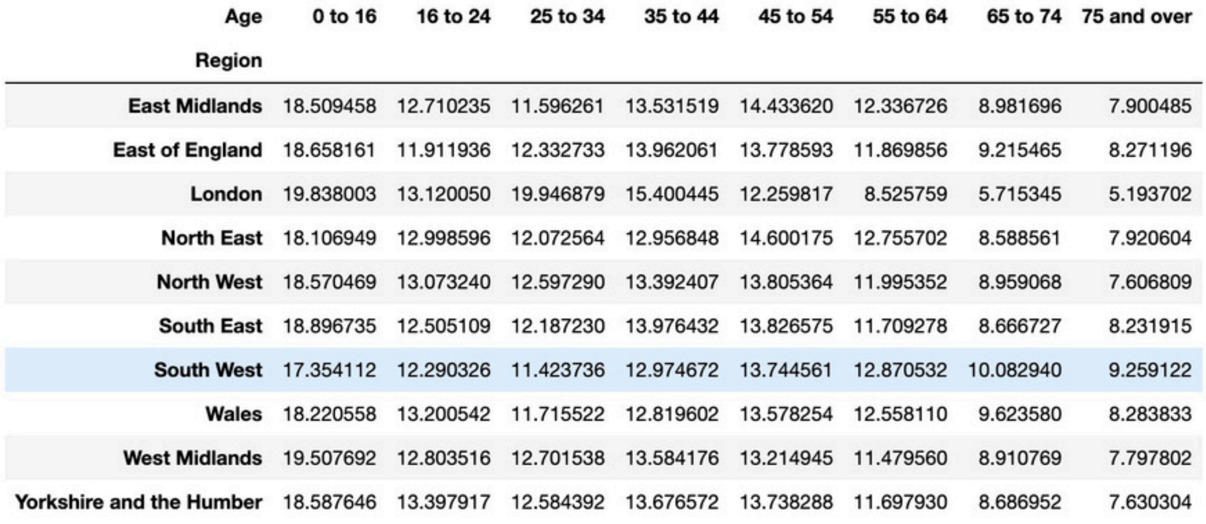
Hinduism and Sikhism are primarily associated with the Asian and Asian British ethnic groups، while Christianity، Judaism، and other religions prevail within the white ethnic group. A comparable trend is observed among those professing no religion، with a predominant presence in the white ethnic group. Sikhs، Hindus، Christians، and those with no religious affiliation collectively constitute a significant majority within specific ethnic groups.

In contrast، the Muslim community displays a diverse composition، incorporating

individuals from

various ethnic backgrounds. This diversity underscores the multicultural aspect of the Muslim religion، differing from Judaism، where adherence is predominantly tied to a single ethnic group.

Figure [illegible]: Correlation between Region and Age

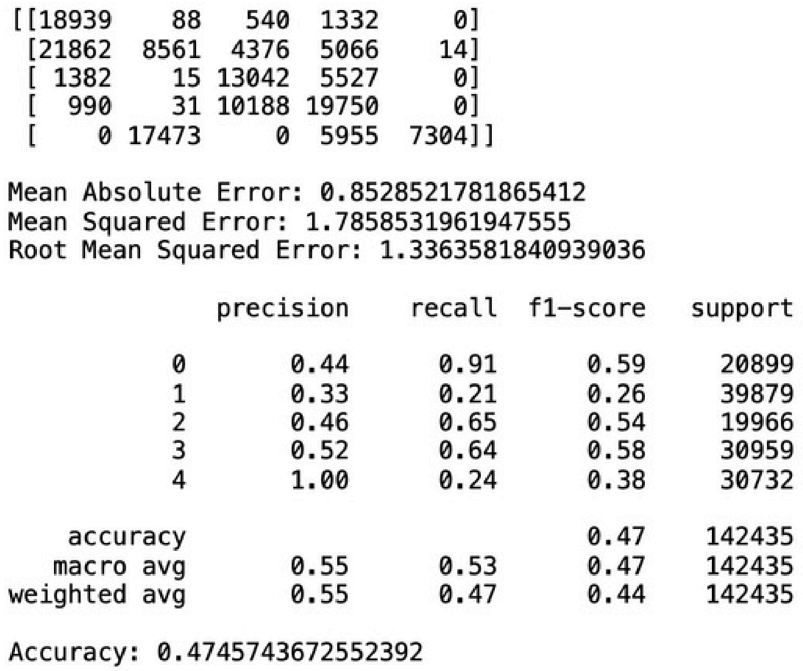


Age Region	0 to 16	16 to 24	25 to 34	35 to 44	45 to 54	55 to 64	65 to 74	75 and over
East Midlands	18.509458	12.710235	11.596261	13.531519	14.433620	12.336726	8.981696	7.900485
East of England	18.658161	11.911936	12.332733	13.962061	13.778593	11.869856	9.215465	8.271196
London	19.838003	13.120050	19.946879	15.400445	12.259817	8.525759	5.715345	5.193702
North East	18.106949	12.998596	12.072564	12.956848	14.600175	12.755702	8.588561	7.920604
North West	18.570469	13.073240	12.597290	13.392407	13.805364	11.995352	8.959068	7.606809
South East	18.896735	12.505109	12.187230	13.976432	13.826575	11.709278	8.666727	8.231915
South West	17.354112	12.290326	11.423736	12.974672	13.744561	12.870532	10.082940	9.259122
Wales	18.220558	13.200542	11.715522	12.819602	13.578254	12.558110	9.623580	8.283833
West Midlands	19.507692	12.803516	12.701538	13.584176	13.214945	11.479560	8.910769	7.797802
Yorkshire and the Humber	18.587646	13.397917	12.584392	13.676572	13.738288	11.697930	8.686952	7.630304

Clearly، a significant majority of the population in various regions falls within the ٠-١٦ age bracket. However، when analysing age distribution by region، London notably stands out with the highest concentration of individuals aged ٢٥ to ٣٤، making up ١٩.٩٪ compared to the average of ١٢٪ for this age group in other regions. Conversely، London exhibits the lowest percentage of people aged ٧٥ and over، accounting for ٥.٢٪ in contrast to the average of ٨٪ in other regions. This suggests a distinctive demographic trend، indicating a higher proportion of young professionals، around ٢٥-٣٤ years old، residing in London compared to other regions.

٢. Classification

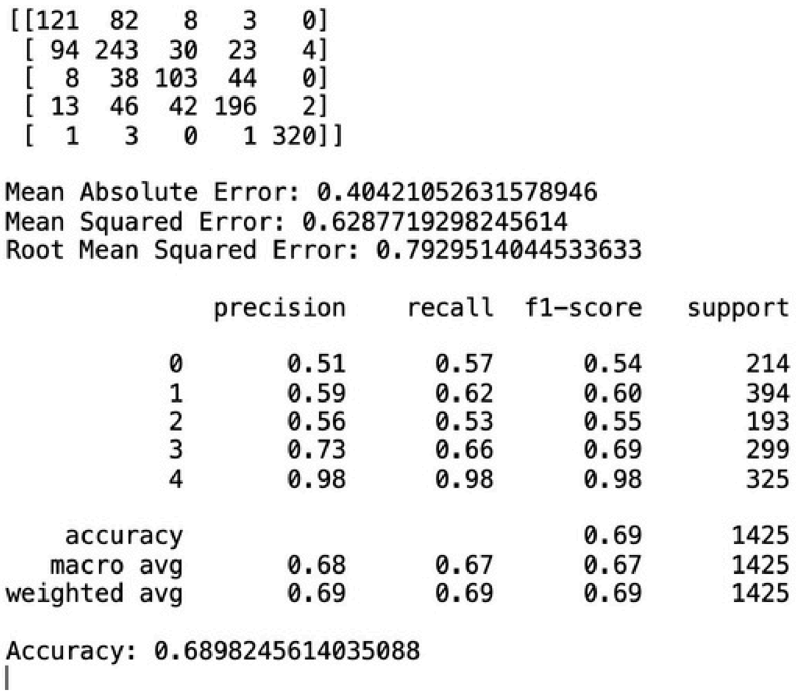
a. Naïve Bayes



```
[[18939    88   540  1332     0]
 [21862  8561  4376  5066    14]
 [ 1382    15 13042  5527     0]
 [  990    31 10188 19750     0]
 [    0 17473     0  5955  7304]]

Mean Absolute Error: 0.8528521781865412
Mean Squared Error: 1.7858531961947555
Root Mean Squared Error: 1.3363581840939036

              precision    recall  f1-score   support

           0       0.44      0.91      0.59     20899
           1       0.33      0.21      0.26     39879
           2       0.46      0.65      0.54     19966
           3       0.52      0.64      0.58     30959
           4       1.00      0.24      0.38     30732

    accuracy                           0.47    142435
   macro avg       0.55      0.53      0.47    142435
weighted avg       0.55      0.47      0.44    142435

Accuracy: 0.4745743672552392
```

The Naïve Bayes model encounters difficulties in precisely forecasting classes ١، ٢، and ٣، as indicated by lowprecision، recall، and F١-score values. Notably، while class ٤ exhibits a precision of ١.٠٠، its recallismerely ٠.٢٤، indicating adeptness in identifying true positives but overlooking a significant numberof actual instances. In summary، the Naïve Bayes algorithm achieves an accuracy of only ٤٧٪، fallingbelow the level of chance or random guessing.

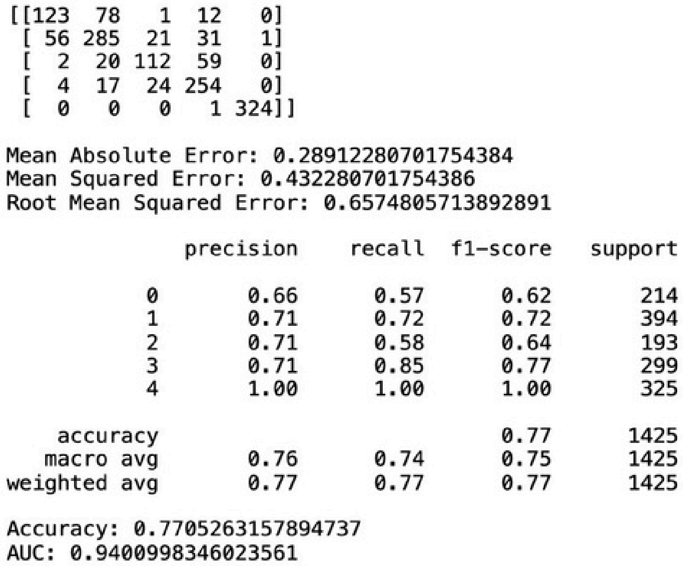
b. K-Nearest Neighbor



```
[[121  82   8   3   0]
 [ 94 243  30  23   4]
 [  8  38 103  44   0]
 [ 13  46  42 196   2]
 [  1   3   0   1 320]]

Mean Absolute Error: 0.40421052631578946
Mean Squared Error: 0.6287719298245614
Root Mean Squared Error: 0.7929514044533633

              precision    recall  f1-score   support

           0       0.51      0.57      0.54       214
           1       0.59      0.62      0.60       394
           2       0.56      0.53      0.55       193
           3       0.73      0.66      0.69       299
           4       0.98      0.98      0.98       325

    accuracy                           0.69      1425
   macro avg       0.68      0.67      0.67      1425
weighted avg       0.69      0.69      0.69      1425

Accuracy: 0.6898245614035088
```

The K-Nearest Neighbor model shows moderate performance with a ٦٩٪accuracy. It exhibits sensible precision، recall، and F١-score for class ١ and ٣ and an average score forclass ٠. Notably، it performs well for class ٤، achieving high performance score of ٩٨٪ in all performance metrics، indicating accurate predictions for that category. The model، though not perfect (approximately ٦٩٪ accuracy)، demonstrates a reasonable understanding of patterns in the data، as it accurately predicts over two-thirds of the predictions.

c. SVM



```
[[123  78   1  12   0]
 [ 56 285  21  31   1]
 [  2  20 112  59   0]
 [  4  17  24 254   0]
 [  0   0   0   1 324]]

Mean Absolute Error: 0.28912280701754384
Mean Squared Error: 0.432280701754386
Root Mean Squared Error: 0.6574805713892891

              precision    recall  f1-score   support

           0       0.66      0.57      0.62       214
           1       0.71      0.72      0.72       394
           2       0.71      0.58      0.64       193
           3       0.71      0.85      0.77       299
           4       1.00      1.00      1.00       325

    accuracy                           0.77      1425
   macro avg       0.76      0.74      0.75      1425
weighted avg       0.77      0.77      0.77      1425

Accuracy: 0.7705263157894737
AUC: 0.9400998346023561
```

The SVM model performs well with an accuracy of ٧٧٪، indicating that it makes correct predictions over ٧٠٪ of the time. It shows a perfect performance metrics (١)،for class ٤. It also performs well in classes ١، ٢ and ٣ with an average of ٧٥٪ across all performance metrices. In general، the model is effective at understanding patterns.

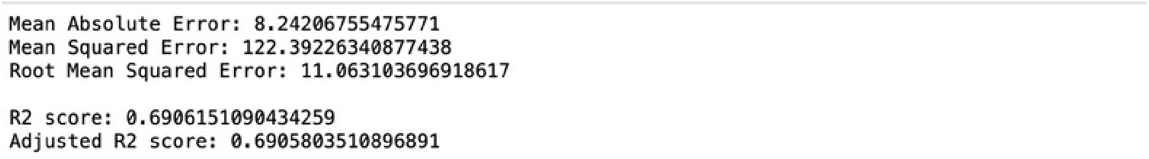
Figure ٢.١ Performance evaluation of the models

Clasification model	Accuracy Score	Macro-avg: Precision	Macro-avg: Recall	Macro-avg: F-١ score
Naïve Bayes	٤٧٪	٠.٥	٠.٥	٠.٤
K-Nearest Neighbor	٦٩٪	٥	٣	٧
Support Vector Machhine	٧٧٪	٠.٧٦	٠.٧٤	٠.٧٥
		٨	٧	٧

In summary، the Support Vector Machine (SVM) algorithm performs better than Naïve Bayes and K-NN as it achieves higher accuracy، which means it makes fewer mistakes in its predictions and scores best in the performance metrics of precision، recall and F١ score as illustrated in table ٢.١.

٣. Regression

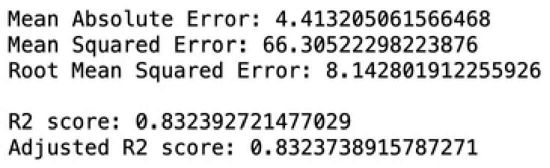
a. Linear regression



```
Mean Absolute Error: 8.24206755475771
Mean Squared Error: 122.39226340877438
Root Mean Squared Error: 11.063103696918617

R2 score: 0.6906151090434259
Adjusted R2 score: 0.6905803510896891
```

With a Mean Absolute Error (MAE) of ٨.٢٤، the model's predictions for hours worked per week، on average، deviate by approximately ±٨ hours. Given a mean of ٣٥ hours، this difference represents an entire working day، significantly deviating from typical work hours. The R٢ score of ٠.٦٨ indicates the model correctly predicts ٦٨٪ of attempts، suggesting accuracy slightly above two-thirds. Overall، the model performs poorly and is ineffective in predicting the number of hours.

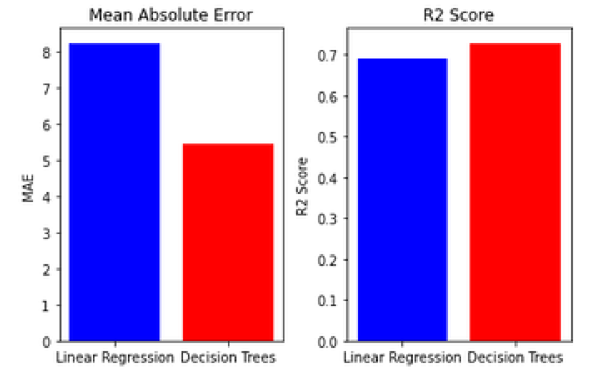
b. Regression Tree



```
Mean Absolute Error: 4.413205061566468
Mean Squared Error: 66.30522298223876
Root Mean Squared Error: 8.142801912255926

R2 score: 0.832392721477029
Adjusted R2 score: 0.8323738915787271
```

The regression tree model typically deviates by approximately ±٤ hours when predicting work hours. With an R٢ score of around ٠.٨٣ (out of ١)، the model accurately predicts about ٨٣٪ of the hours worked، indicating a small error. Overall، it demonstrates effectiveness in predicting the number of hours worked.

Figure ٣.١: Linear regression and regression tree comparison

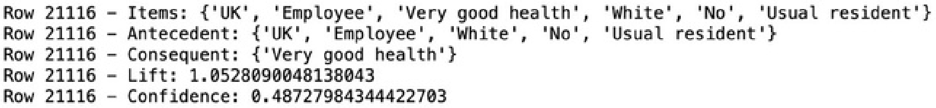


Regression model	R٢ Score	Mean Absolute Error(MAE)
Linear Regression	٦٩٪	٨
Decision Tree	٨٣٪	٤

Linear regression struggles to accurately predict the number of hours worked، with an ٨-hour deviation while the regression tree model performs relatively better، with only a ٤-hour deviation.

٤. Association rule mining

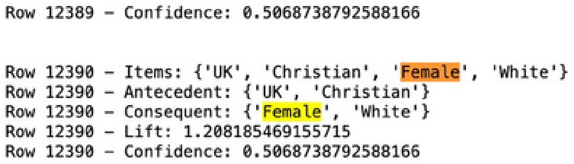
1.



```
Row 21116 - Items: {'UK', 'Employee', 'Very good health', 'White', 'No', 'Usual resident'}
Row 21116 - Antecedent: {'UK', 'Employee', 'White', 'No', 'Usual resident'}
Row 21116 - Consequent: {'Very good health'}
Row 21116 - Lift: 1.0528090048138043
Row 21116 - Confidence: 0.48727984344422703
```

People who were born in the UK، are employed، are of white ethnicity، not students(“No”) and are usual residents are associated with being in very good health.

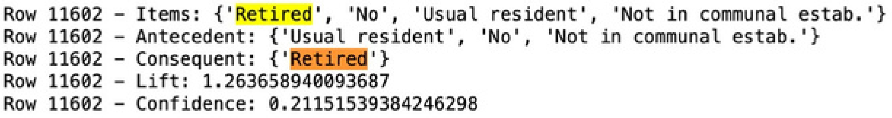
٢.



```
Row 12389 - Confidence: 0.5068738792588166

Row 12390 - Items: {'UK', 'Christian', 'Female', 'White'}
Row 12390 - Antecedent: {'UK', 'Christian'}
Row 12390 - Consequent: {'Female', 'White'}
Row 12390 - Lift: 1.208185469155715
Row 12390 - Confidence: 0.5068738792588166
```

If someone has the UK as their country of birth and they ascribe to the Christian religion، they are likely to be a female and of white ethnicity.

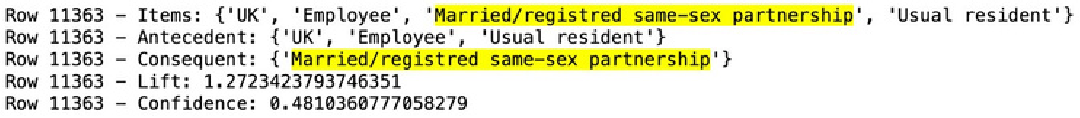
٣.



```
Row 11602 - Items: {'Retired', 'No', 'Usual resident', 'Not in communal estab.'}
Row 11602 - Antecedent: {'Usual resident', 'No', 'Not in communal estab.'}
Row 11602 - Consequent: {'Retired'}
Row 11602 - Lift: 1.263658940093687
Row 11602 - Confidence: 0.21151539384246298
```

If someone is a usual resident، not a student and not in a communal establishment، they are likely to be retired.

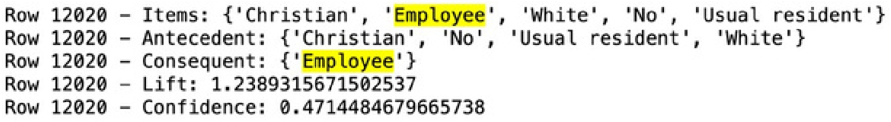
٤.



```
Row 11363 - Items: {'UK', 'Employee', 'Married/registred same-sex partnership', 'Usual resident'}
Row 11363 - Antecedent: {'UK', 'Employee', 'Usual resident'}
Row 11363 - Consequent: {'Married/registred same-sex partnership'}
Row 11363 - Lift: 1.2723423793746351
Row 11363 - Confidence: 0.4810360777058279
```

If somebody’s country of birth is the UK، are employed and a usual resident، they are likely to be married or in a registered same sex partnership.

٥.

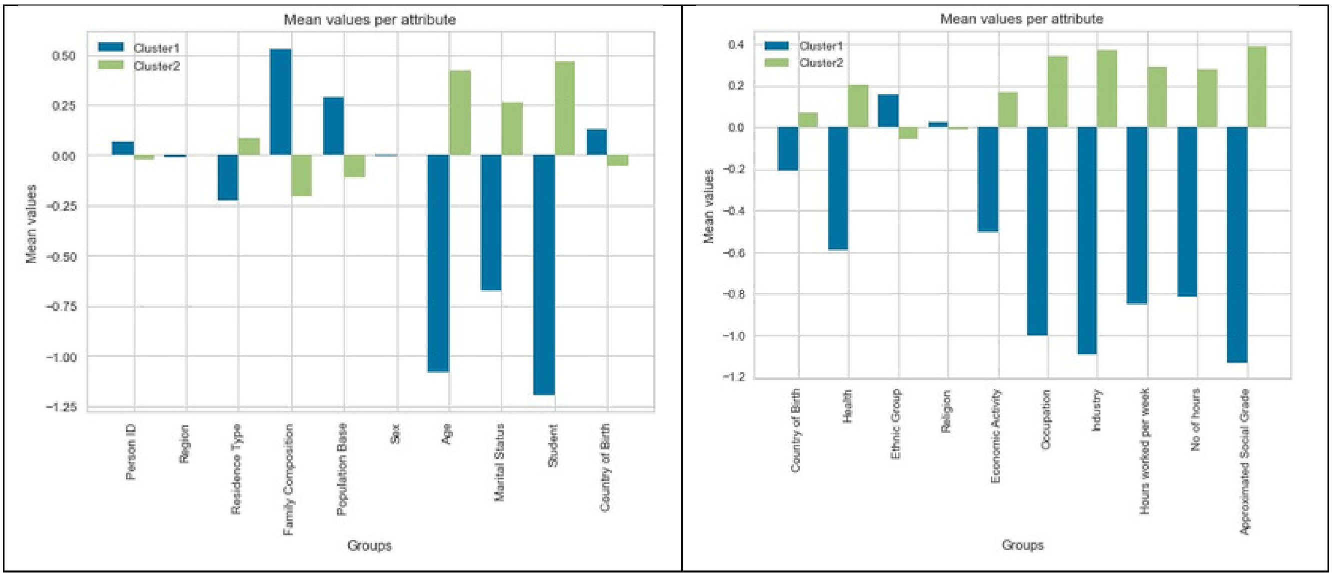


```
Row 12020 - Items: {'Christian', 'Employee', 'White', 'No', 'Usual resident'}
Row 12020 - Antecedent: {'Christian', 'No', 'Usual resident', 'White'}
Row 12020 - Consequent: {'Employee'}
Row 12020 - Lift: 1.2389315671502537
Row 12020 - Confidence: 0.4714484679665738
```

If an individual’s religion is Christianity، and they are not a student(“No”)، are a usual resident and are of white ethnicity، they are likely to be employed.

٥. Clustering

Figure ٣.١: K-means clustering

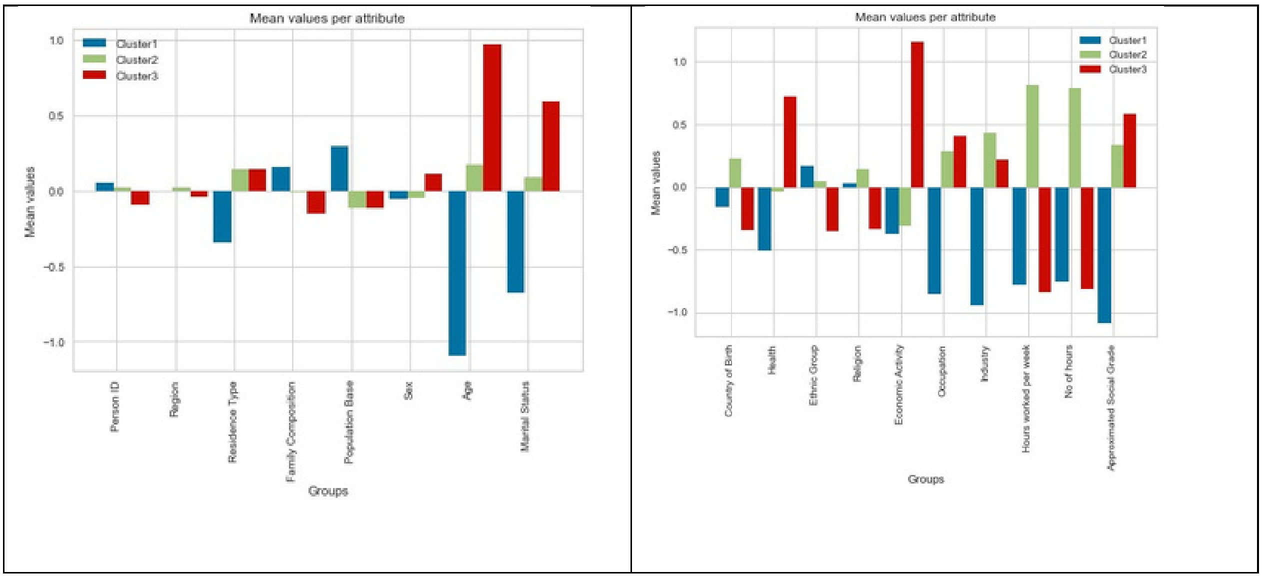


K-means		
Data point Age	Cluster ١	Cluster ٢
Economic	٠ to١٦،٣٤ to ٤٤	Older people (٦٥ and over، ٦٥ to ٧٤)
Activity	Employed، Retired	Unemployed، long term sick /disabled
Occuaption	Proffessional، elemenatary	Sales، caring and machine operatives
Industry	Retail and motor repair، Industrial sectors and real estate	Agricuature، finance and community s ervice
	Size: (٤٢٤٥٧٢) ٧٥٪	Size: (١٤٥١٦٨)٢٥٪

Cluster ١ comprises a larger portion (٧٥٪) compared to Cluster ٢ (٢٥٪)، revealing an uneven distribution. This implies that Cluster ١ has substantial coverage، and a significant proportion of the dataset aligns with the characteristics of Cluster ١.

The clustering demonstrates significance by effectively segregating groups according to anticipated patterns. For instance، employees are distinctly grouped apart from the unemployed، and a similar pattern is evident in age groups. Also، a silhouette score of ٠.٢٤ is okay because it means the clusters are somewhat separated from each other.

Figure ٣.٢: Hierarchical clustering



Hierarchical			
Data point	Cluster ١	Cluster ٢	Cluster ٣
Age	٠to١٦،٣٥to ٤٤	Olderpeople(٦٥ and over)	٢٥to٣٤، ١٦ to ٢٤
Economic Activity	Unemployed، Long term sick /disabled،	Looking after family، full time student	Employee، Retired
Occuaption	Proffessional، elemenatary	Machine operatives، caring sales	Skilled traders، Associate /proffessional
Industry	Retail and motor repair، Industrial sectors and real estate	human health، education	Agriculture،Finance، community service
	Size: (٧٧١٤) ٢٧٪	Size:(١٤٤٩٨) ٥١٪	Size: (٦٢٧٥) ٢٢٪

Cluster ٢ constitutes a larger proportion (٥١٪) compared to both cluster ١ (٢٧٪) and cluster ٣ (٢٢٪)، emphasizing an uneven distribution. This clustering is considerable as it efficiently segments and distinguishes groups based on anticipated patterns. For instance، self-employed، unemployed، and employed individuals are allocated to three distinct clusters each.

Both clustering methods show resemblances in how attributes are organized into identical

clusters e.g.،

in age data point، similar age groups are grouped together (٠ to ١٦، ٣٤ to ٤٤)، and older
people (٦٥ and
over، ٦٥ to ٧٤). A similar pattern is evident in the industry and occupation data points.