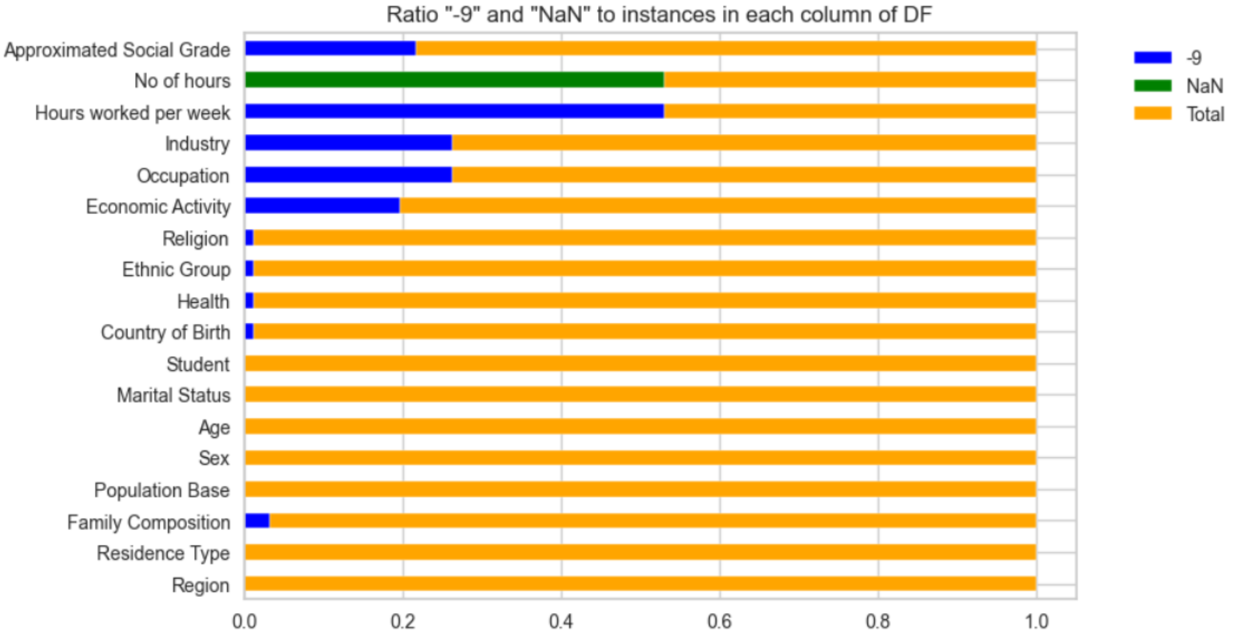
1. Descriptive analytics

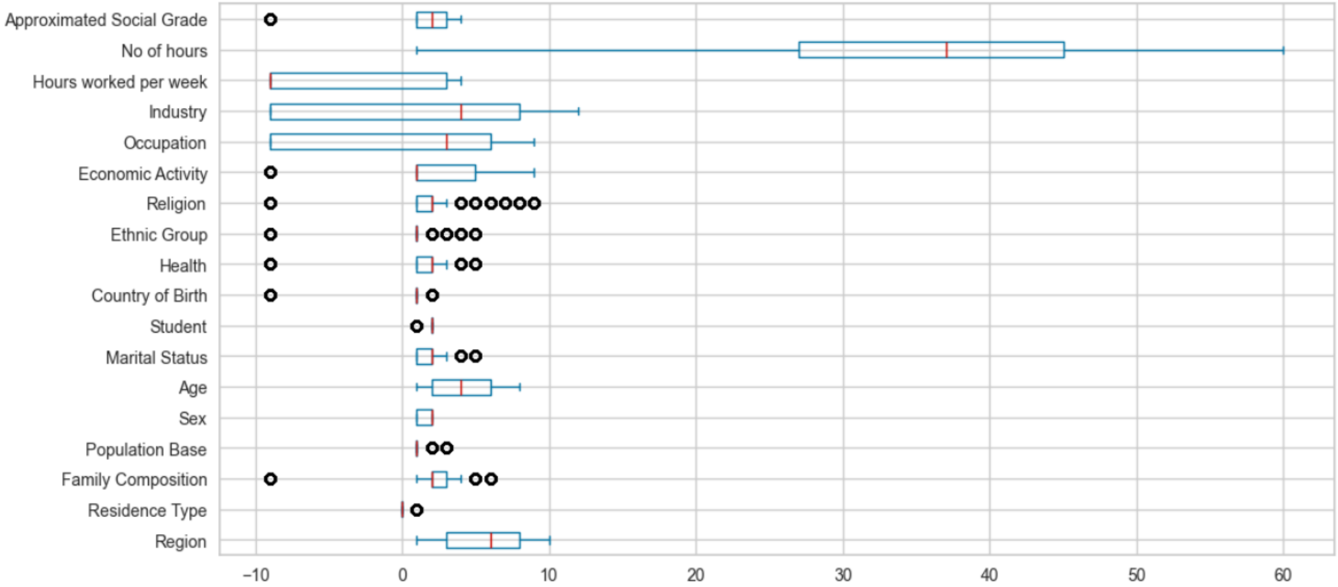
For analysis was taken base data characteristics of residents in communal establishments in ten regions countries.

The database is presented as a DataFrame containing 569,740 observations and consisting of 19 columns, so it is multidimensional. Most of the data is int64 the rest are objects and float64. The database contains a column 'Person ID' which is not of interest for analysis and which is not useful. This column will be removed from the database. The only column that contains continuous data is 'No of' hours'. Two columns 'Region' and 'Residence Type' contain nominal categorical data in the form strings. The remaining columns contain nominal data presented in the form of numbers. Columns 'Resident Type', 'Sex', 'Student', 'Country of birth' are binary categorical data types.

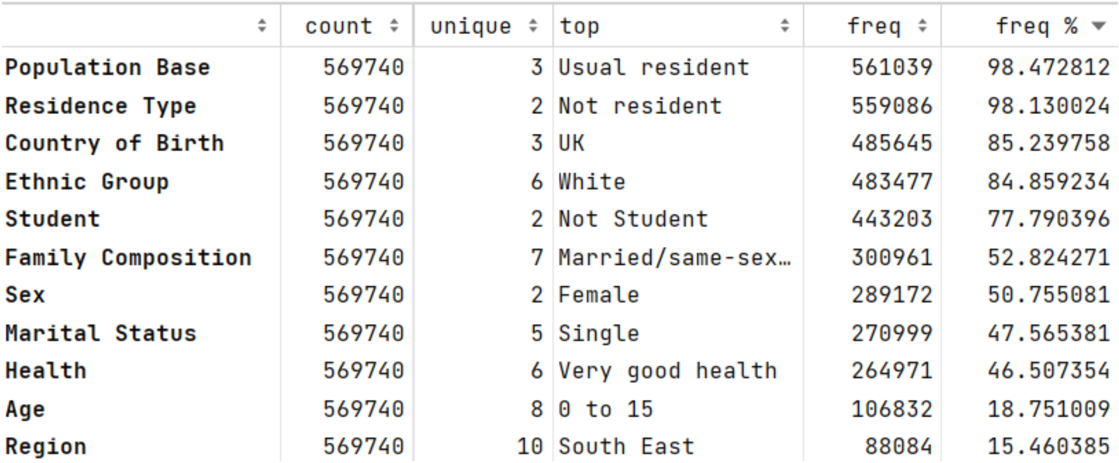
Checking the DF for missing data showed that the column 'No of hours' contains 302,321 instances with missing data. In addition, from the description of the database in the documents attached to it and a brief overview of the contents of the database, a large amount of data is visible that does not fall into any of the described categories and is marked as '-9'. From this we can assume that category '-9' does not contain useful information and may be missing data. Below in the graph is **'Ratio '-9' and 'NaN' to instances in each column of DF'** you can see a visual comparison of the missing data to the total amount of data.



Outliers can be observed on a Boxplot type graph. It can be seen that emissions in the negative direction are most likely due to the presence of the '-9' category. Also, this same category probably influenced the minimum values of features 'Hours worked per week', 'Industry', 'Occupation'.

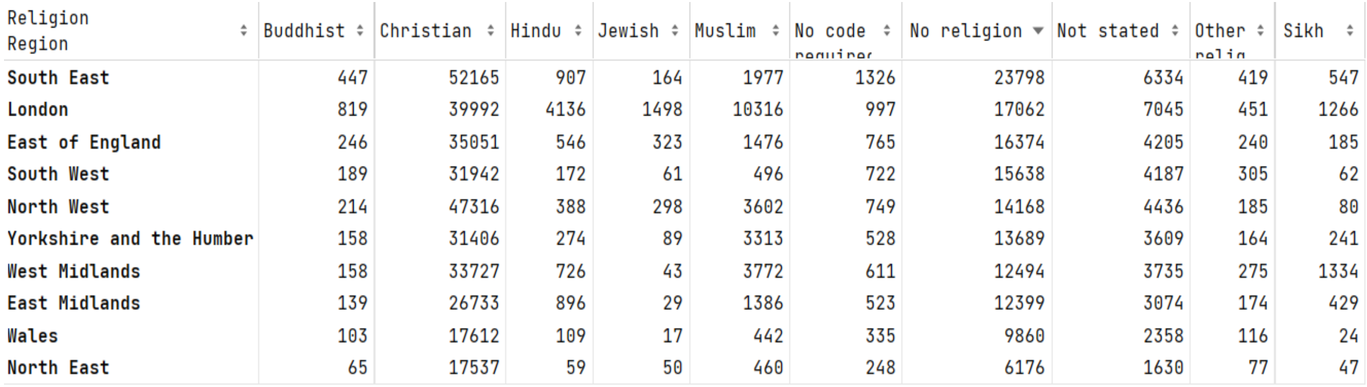


Using the function .describe() the most common categories in the database were found out. The most frequent are such categories as 'Usual resident', 'Not resident', 'Country of Birth – UK', 'White', 'Not Student'.



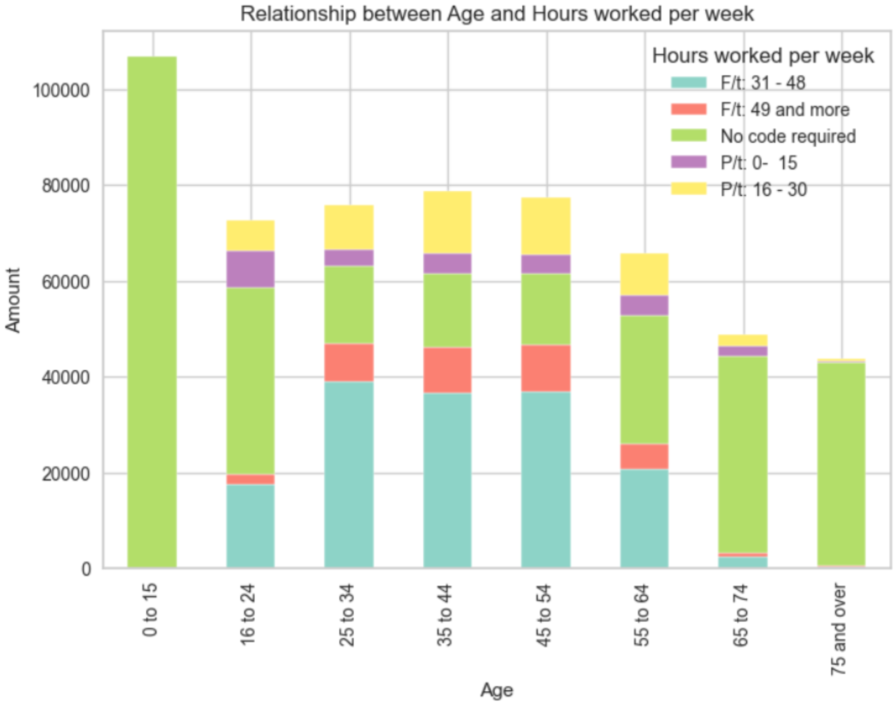
⇕	count ⇕	unique ⇕	top ⇕	freq ⇕	freq % ▾
Population Base	569740	3	Usual resident	561039	98.472812
Residence Type	569740	2	Not resident	559086	98.130024
Country of Birth	569740	3	UK	485645	85.239758
Ethnic Group	569740	6	White	483477	84.859234
Student	569740	2	Not Student	443203	77.790396
Family Composition	569740	7	Married/same-sex…	300961	52.824271
Sex	569740	2	Female	289172	50.755081
Marital Status	569740	5	Single	270999	47.565381
Health	569740	6	Very good health	264971	46.507354
Age	569740	8	0 to 15	106832	18.751009
Region	569740	10	South East	88084	15.460385

Using crosstab, you can observe statistics on the distribution of the most popular religions by region. For example, London and South East are leaders in the number of people living in government institutions and professing the most popular religions.

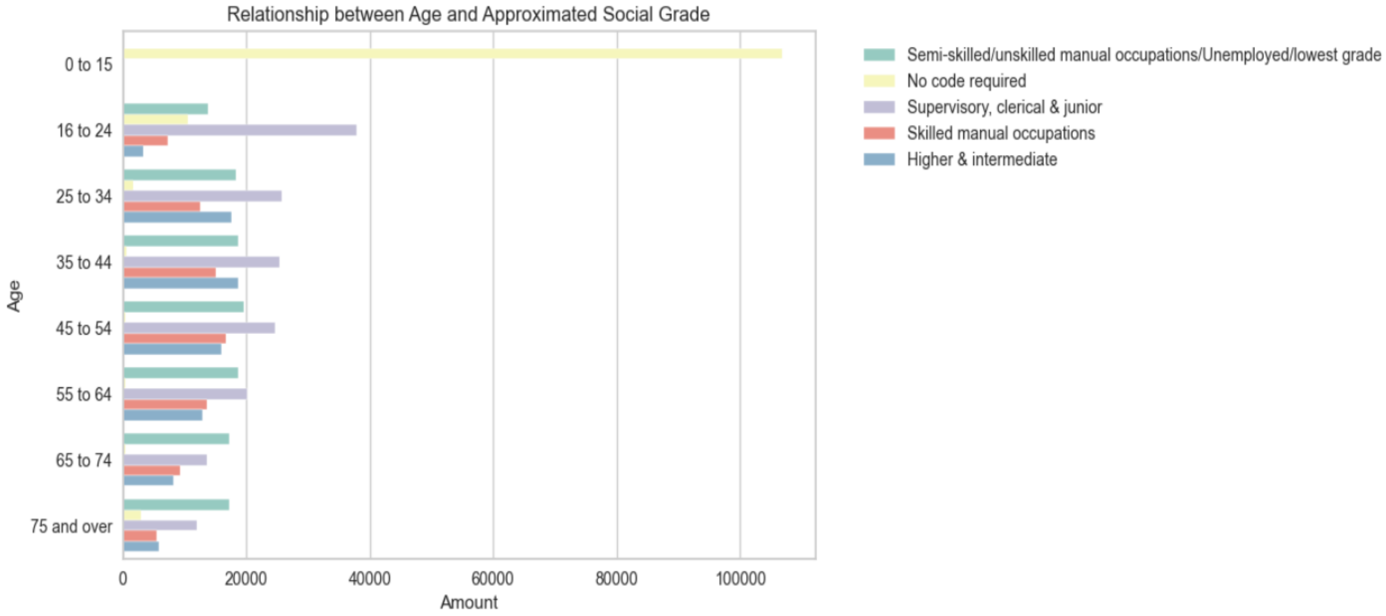


Religion Region ⇕	Buddhist ⇕	Christian ⇕	Hindu ⇕	Jewish ⇕	Muslim ⇕	No code requirec ⇕	No religion ▾	Not stated ⇕	Other relig ⇕	Sikh ⇕
South East	447	52165	907	164	1977	1326	23798	6334	419	547
London	819	39992	4136	1498	10316	997	17062	7045	451	1266
East of England	246	35051	546	323	1476	765	16374	4205	240	185
South West	189	31942	172	61	496	722	15638	4187	305	62
North West	214	47316	388	298	3602	749	14168	4436	185	80
Yorkshire and the Humber	158	31406	274	89	3313	528	13689	3609	164	241
West Midlands	158	33727	726	43	3772	611	12494	3735	275	1334
East Midlands	139	26733	896	29	1386	523	12399	3074	174	429
Wales	103	17612	109	17	442	335	9860	2358	116	24
North East	65	17537	59	50	460	248	6176	1630	77	47

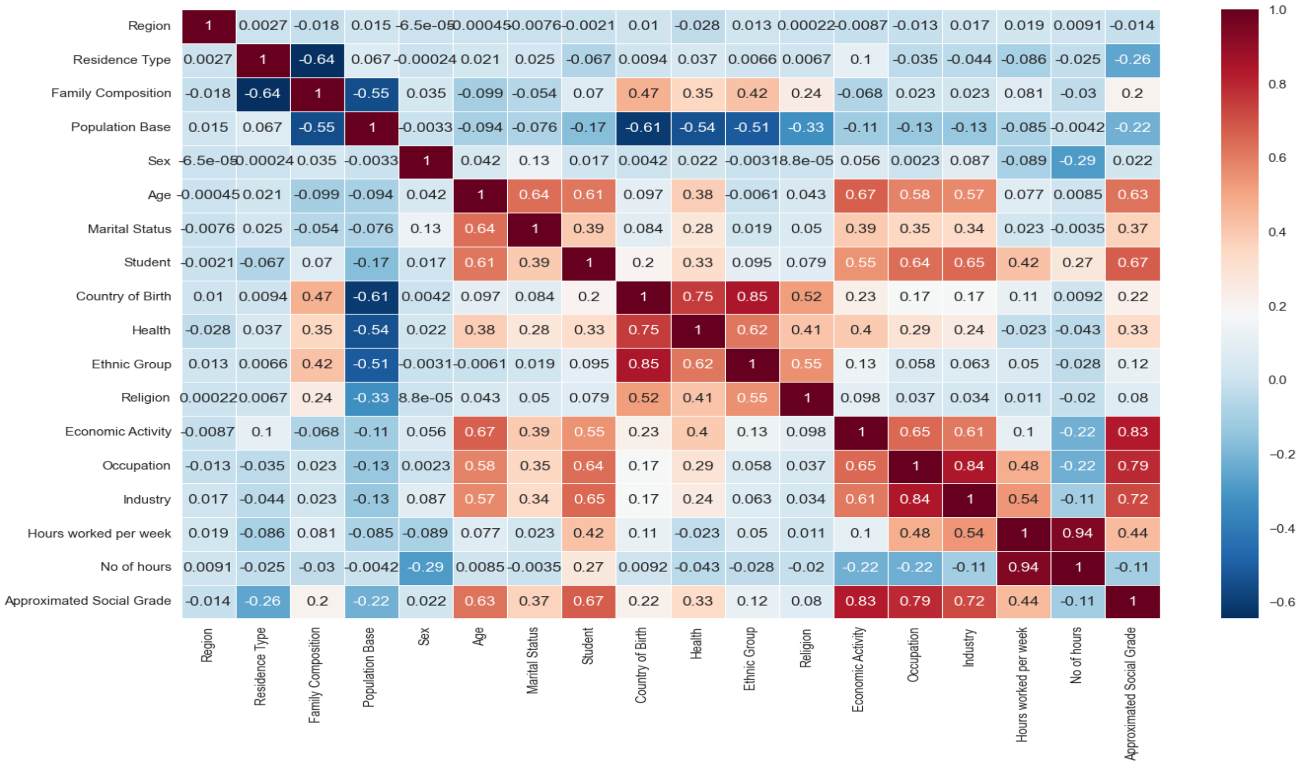
On the scatterplot **'Realization between 'Age' and 'Hours worked per week'** you can see the interesting fact . Previously there was an assumption that the category 'No code required ' or '-9' may be defined as missing data. Although now the scatterplot shows that approximately 50% of the 'No code' required refer to ages '0-15' and '75 and above'. This means that children and older people do not work and it is impossible to indicate the number of hours worked per week. As for the presence of '-9' in other age categories, we can conclude that people are not working. Therefore, category '-9' cannot be considered as missing data.



Such same conclusions also confirm the plot 'Relationship between 'Age' and 'Approximated Social Grade'. And in this case the category 'No code required' (-9) applies entirely to ages '0-15' years.



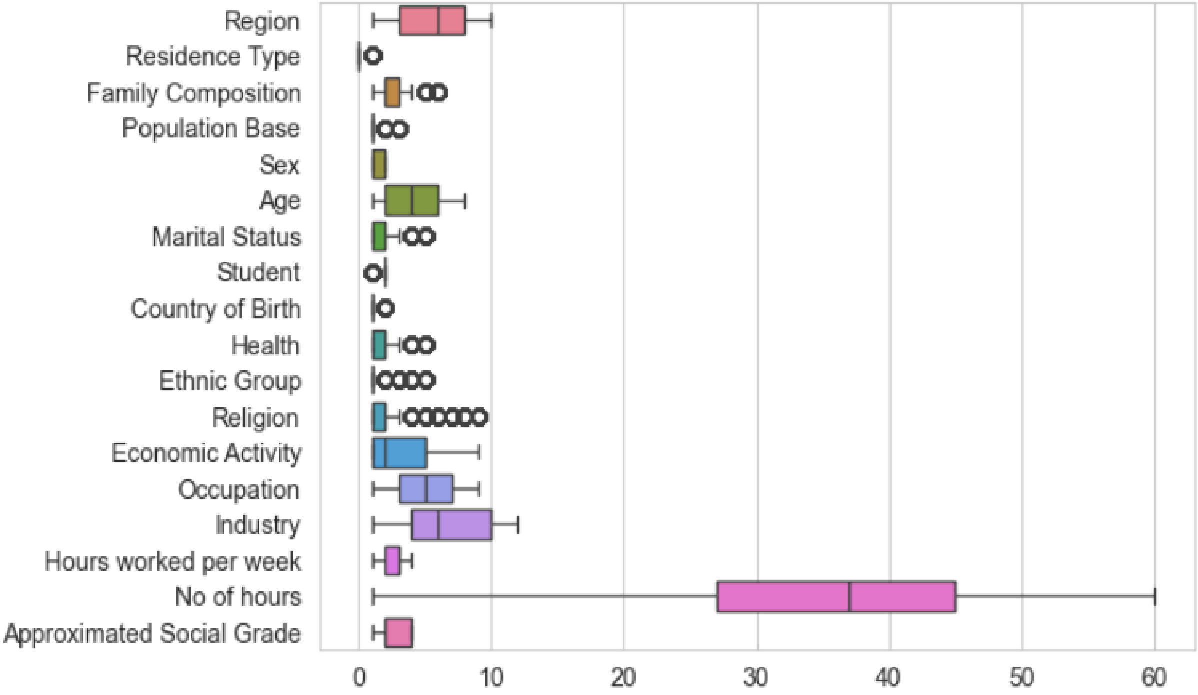
As can be seen from the HeatMap, the largest positive correlation is observed between the two attributes 'Hours worked per week' and 'No of hours'. Based on all the observations on these two attributes, it can be assumed that they are categories with repeated data. Their difference is that the number of hours in 'Hours worked per week' is divided into groups of categories, and in another - the number of hours is presented as continuous data. Missing data in 'No of hours' and category '-9' are most likely also duplicates of each other with missing data.



2. Classification for the “Approximate d Social Grade” attribute:

Before carrying out the classification, it was decided to carry out preliminary preparation of the data, pre-processing.

For this purpose, the category '-9' was replaced by 'NaN ', after which the boxplot no longer showed emissions in the negative direction.



Next, duplicate rows were removed. Values 'NaN' into attributes 'No of hours', 'Hours worked per week' are replaced to 0. For the 'No of hours' replacing missing values with 0 seems logical, since non-working persons work exactly 0 hours. Regarding the attribute 'Hours worked per week' then replacing it with 0 is not the best way, but it also looks logical.

For the remaining instances, the decision was made to remove all lines containing NaN .

Attributes 'Region', 'Residence Type', 'Family Composition', ' Population Base ' were removed as those that either had little correlation with other data or that duplicated information from other attributes.

The database prepared in this way will be used for all tasks, in some cases undergoing additional pre-processing.

Immediately before classification, a prepared copy of the DF was created, from which rows with a value of 0 were removed. In my opinion, this should improve the model.

To classify the attribute 'Approximated Social Grade' three classifiers were selected:

- Decision tree classifier
- Random Forest classifier
- Linear Discriminant Analysis

Decision tree classifier

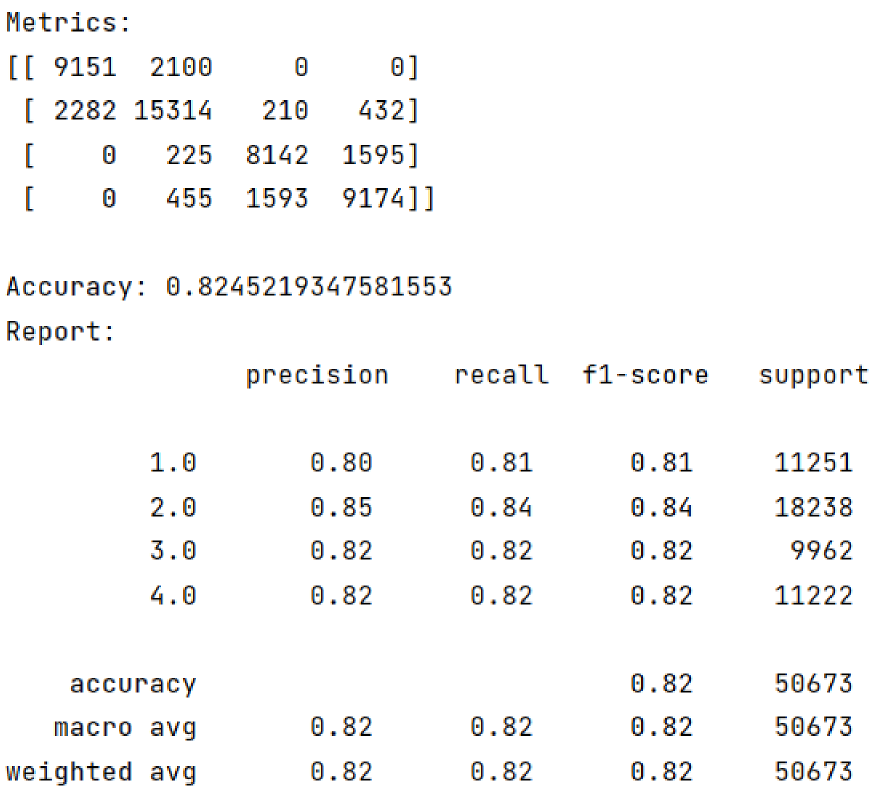
After classifying the 'Approximate Social Grade' attribute, result metrics were obtained. The overall accuracy of the classifier is 82.45%, which means that the model correctly predicts the class for approximately 82.45% of the instances.

Class 1.0 (precision: 80%, recall: 81%, F1 score: 81%) has balanced performance.

Class 2.0 (precision: 85%, recall: 84%, F1 score: 84%) is well predicted with high precision and recall.

Classes 3.0 and 4.0 (precision: 82%, recall: 82%, F1 score: 82%) also show balanced performance.

The model exhibits good overall performance across all classes, with a balanced trade-off between precision and recall.



```
Metrics:
[[ 9151  2100     0     0]
 [ 2282 15314   210   432]
 [    0   225  8142  1595]
 [    0   455  1593  9174]]

Accuracy: 0.8245219347581553
Report:
               precision    recall  f1-score   support

         1.0       0.80      0.81      0.81     11251
         2.0       0.85      0.84      0.84     18238
         3.0       0.82      0.82      0.82      9962
         4.0       0.82      0.82      0.82     11222

    accuracy                           0.82     50673
   macro avg       0.82      0.82      0.82     50673
weighted avg       0.82      0.82      0.82     50673
```

Random Forest classifier

The overall accuracy of the model is 83.65%, which means that the model correctly predicts the class for approximately 83.65% of the instances.

The macro-average precision, recall, and F1-score are around 83%, suggesting balanced performance across all classes.

Class 1.0 (precision: 82%, recall: 81%, F1 score: 82%) - the model performs well in predicting class 1.0 with balanced precision and recall.

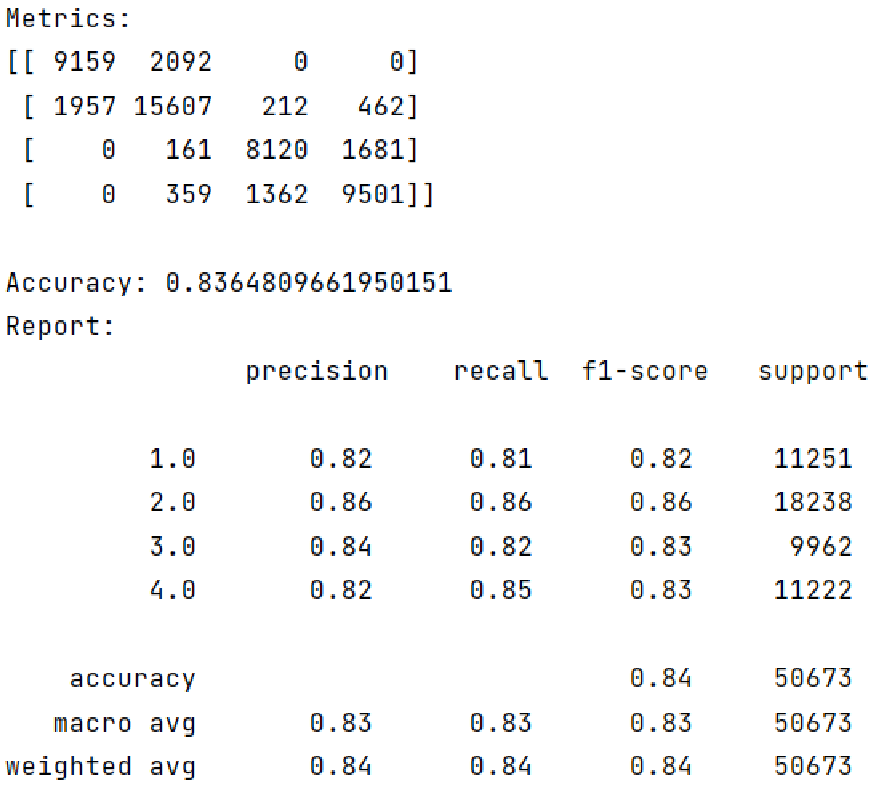
Class 2.0 (precision: 86%, recall: 86%, F1 score: 86%) - class 2.0 predicts well and demonstrates high precision and recall.

Class 3.0 (precision: 84%, recall: 82%, F1 score: 83%) - the model shows good performance in predicting class 3.0 with balanced precision and recall.

Class 4.0 (precision: 82%, Recall: 85%, F1 Score: 83%) - class 4.0 predicts well and has a balanced trade-off between precision and recall.

The weighted average precision, recall, and F1-score (84%) represent the overall performance measure considering the distribution of classes in the dataset.

The model exhibits good overall performance with balanced precision and recall across classes. High macro-average scores suggest that the model is effective in dealing with imbalances between classes. Analysis by class shows that the model performs well for each individual class, demonstrating its ability to make accurate predictions over a wide range of cases.



```
Metrics:
[[ 9159  2092     0     0]
 [ 1957 15607   212   462]
 [    0   161  8120  1681]
 [    0   359  1362  9501]]

Accuracy: 0.8364809661950151
Report:
               precision    recall  f1-score   support

         1.0       0.82      0.81      0.82     11251
         2.0       0.86      0.86      0.86     18238
         3.0       0.84      0.82      0.83      9962
         4.0       0.82      0.85      0.83     11222

    accuracy                           0.84     50673
   macro avg       0.83      0.83      0.83     50673
weighted avg       0.84      0.84      0.84     50673
```

Linear Discriminant Analysis

The overall accuracy of the model is 74.90%, which means it predicts the class correctly about 74.90% of the time.

The macro-average precision, recall, and F1-score are around 75%, suggesting balanced performance across all classes.

Class 1.0 (precision: 70%, Recall: 74%, F1 Score: 72%) - Class 1.0 predicts reasonably well, with a balanced trade-off between precision and recall.

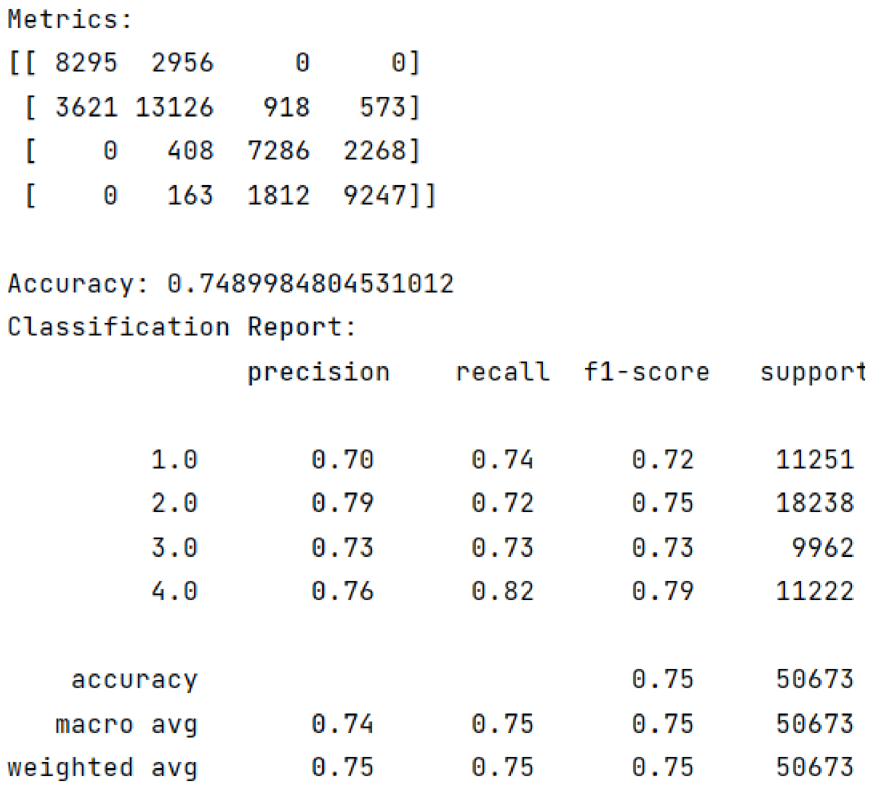
Class 2.0 (precision: 79%, Recall: 72%, F1 Score: 75%) - Class 2.0 shows good precision, but there is a slight imbalance in memorability.

Class 3.0 (precision: 73%, Recall: 73%, F1 Score: 73%) - The model provides consistent Class 3.0 performance with balanced precision and recall.

Class 4.0 (precision: 76%, Recall: 82%, F1 Score: 79%) - Class 4.0 is highly predictive, with a greater emphasis on recall.

The weighted average precision, recall, and F1-score (75%) represent the overall performance measure considering the distribution of classes in the dataset.

The model exhibits acceptable overall performance with balanced precision and recall across classes. Class-specific analysis shows that the model performs well for each individual class, but there are some differences in accuracy and recall. Higher macro-average scores suggest that the model is effective in handling imbalances between classes. Weighted averages provide a comprehensive assessment that takes into account the class distribution of the data set.



```
Metrics:
[[ 8295  2956     0     0]
 [ 3621 13126   918   573]
 [    0   408  7286  2268]
 [    0   163  1812  9247]]

Accuracy: 0.7489984804531012
Classification Report:
              precision    recall  f1-score   support

         1.0       0.70      0.74      0.72     11251
         2.0       0.79      0.72      0.75     18238
         3.0       0.73      0.73      0.73      9962
         4.0       0.76      0.82      0.79     11222

    accuracy                           0.75     50673
   macro avg       0.74      0.75      0.75     50673
weighted avg       0.75      0.75      0.75     50673
```

By comparing the results of all three classifiers, they can be ranked according to the following rating:

1. Random Forest classifier (83.65%)
2. Decision tree classifier (82.45.0 %)
3. Linear Discriminant Analysis (74.90%)

Random Forest classifier and Decision tree classifier have the best performance than Linear Discriminant Analysis.

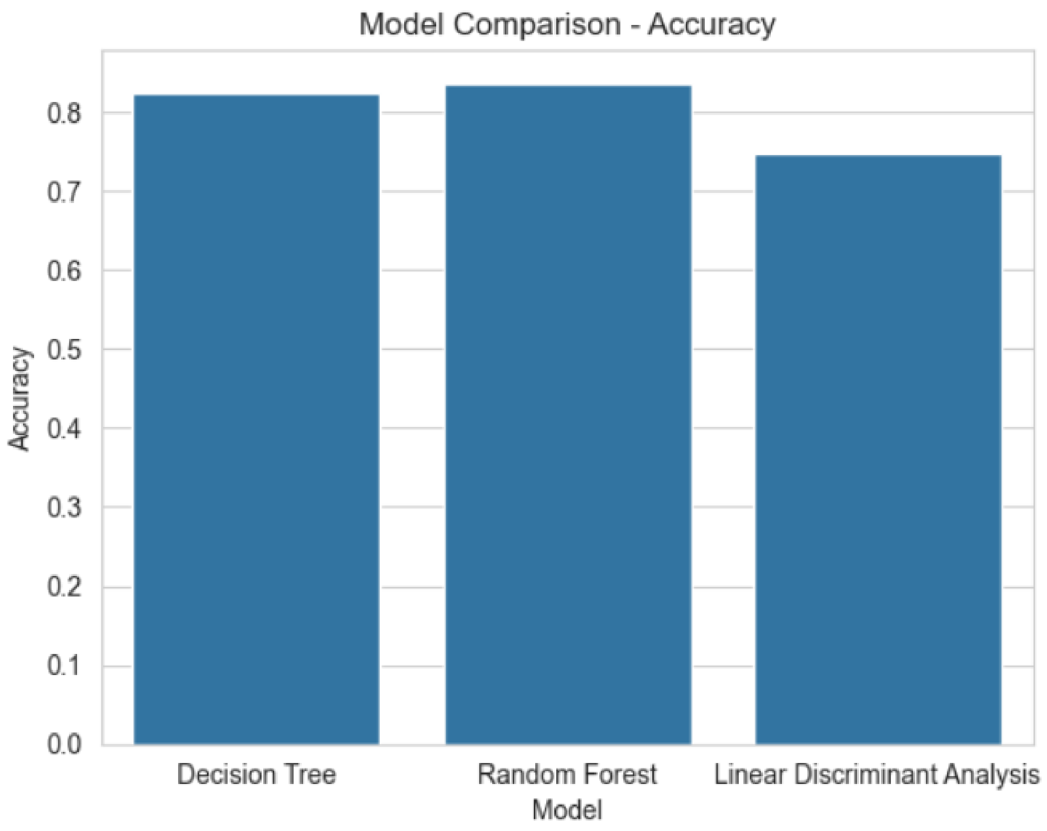
Random Forest classifier shows higher values, indicating better overall performance. Further follows Decision tree classifier, and Linear Discriminant Analysis has the lowest estimates.

Random Forest classifier exhibits the best overall performance and would be the preferred model for practical use.

Decision tree classifier is moderately accurate.

Linear Discriminant Analysis with lower accuracy and an F1 rating may require further optimization before practical implementation.

The models exhibit varying levels of performance from moderate to good. Random Forest classifier seems to be the most useful for practical use, given its higher accuracy and better overall performance.



3. Regression on the 'No of hours' attribute

To apply regression of the 'No of hours' two Linear models were selected Linear Regression and MLPRegressor. A pre-processed database without missing values was loaded as a DataFrame, with attributes that showed little or no correlation removed. Additionally, StandardScaler() was applied to the database and the duplicate 'Hours worked per week' attribute was removed.

Linear Regression

- Average absolute error (MAE): 0.4894

- Mean square error (MSE): 0.3842

- Root mean square error (RMSE): 0.6199.

- R2 score: 0.6167

The R2 estimate of 0.6167 indicates that the linear regression model explains about 61.67% of the variance in the data.

The standard deviation of 0.6199 represents the average error between the predicted and actual values.

The model's performance is average, providing a reasonable but not optimal fit to the data.

MLPRegressor (neural network)

- Average absolute error (MAE): 0.3366

- Mean square error (MSE): 0.2395

- R2 score: 0.7611

The R2 estimate of 0.7611 suggests that MLPRegressor captures approximately 76.11% of the variance, indicating better performance than linear regression.

Lower MAE and MSE values (0.3366 and 0.2395, respectively) indicate reduced errors compared to linear regression.

The performance of the model is good and it shows a better fit to the data.

Comparing the results of the two models, we can confidently say that MLPRegressor outperforms linear regression in all respects, which indicates excellent forecasting capabilities. MLPRegressor also captures underlying patterns in the data more effectively.

Both models can be considered suitable for practical use, but MLPRegressor is preferred due to better performance.

In this context, MLPRegressor, being a neural network model, demonstrates better predictive performance compared to linear regression.

4. Association rule mining

For the Association rule mining, Apriori algorithm was applied. For the algorithm to work, a pre-processed Census DataFrame was loaded, after which a new DataFrame was generated with a description of the categories according to the attached documentation.

To search for interesting rules, filters " Lift " > 1 were applied and duplicates were removed. 175 lines left to select interesting rules.

5 interesting rules were selected and presented in the table below:

No.	index	Antecedent	Consequent	Lift
1.	6	Employee, Male, White	Not Student, F/t: 31-48	1.241
2.	150	Female, White	Christian	1.125
3.	286	Country of Birth: UK	Approximated Social Grade: Supervisory, clerical & junior, White	1.098
4.	1082	F/t: 31-48	Very good health, White	1.018
5.	1258	Female, Not Student	Country of Birth: UK	1.006

1. The lift value of 1.241 suggests that the probability of the consequent (Not a student, Full/time: 31–48 hours) is 1.241 times higher when the antecedent (Employee, Male, White) is present compared to when the antecedent is not present.

2. A lift value of 1.125 indicates a positive relationship between being Female, White, and Christian religion. The likelihood of becoming a Christian is 1.125 times higher if the previous conditions are met.

3. The lift value of 1.098 suggests a positive association between being born in the UK and reported social status and ethnicity. The odds of having this social class and ethnicity are 1.098 times higher for people born in the UK.

4. The lift value of 1.018 indicates a small positive association between working Full time at ages 31–48 and very good health among whites. People in this age and employment category are 1.018 times more likely to be in very good health and have white ethnicity.

5. The lift value of 1.006 suggests a very weak positive association between being a Female Non-student and being born in the UK. Women who are not students are 1.006 times more likely to be born in the UK.

These rules provide insight into the potential relationships between different attributes in a data set.

5. Clustering

For data clustering, two popular algorithms widely used in machine learning, DBSCAN and K- Means, were chosen. Both of these algorithms are used for clustering but they differ in many ways. While K-Means splits data into clusters based on their similarity, DBSCAN groups data based on density.

Unfortunately, my research led to the fact that it was found out that at the moment there are no exact metrics and indicators of the performance of the DBSCAN algorithm.

After clustering

DBSCAN

Cluster 0: 58689 instances

Cluster 1: 10578 instances

Cluster 2: 2408 instances

Cluster 3: 712 instances

Cluster -1: 91 instances

Silhouette Score: 0.2124113715985302

Cluster 0 is the largest, covering 64.53% of the instances.

Clusters 1, 2 and 3 cover 11.64%, 2.66% and 0.79% of cases, respectively.

Cluster -1 (Noise) contains 0.09% of instances .

KMeans

Cluster 0: 64.726773

Cluster 2: 30.931232

Cluster 1: 4.341995

Silhouette Score: 0.21

Cluster 0 covers 64.73% of the instances.

Clusters 1 and 2 cover 4.34% and 30.93% of cases, respectively.

Both algorithms identify a dominant cluster that covers a significant portion of the instances.

Silhouette scores are similar (0.2124 for DBSCAN and 0.21 for K-Means), indicating a moderate level of cluster cohesion.

DBSCAN creates more clusters (including noise), providing higher granularity.

K - Means creates fewer clusters, potentially leading to more generalized groupings.

Both algorithms provide meaningful clustering, with DBSCAN offering more detailed information about irregular cluster structures.