IDTA Coursework 1

UP954123

Table of contents

Task 1: Descriptive Analytics

Part A

The analysis started by loading and preprocessing the dataset. Approximately 302,321 entries were missing in the 'No of hours' attribute out of 569,740 records. To ensure data integrity and prevent loss, the 'backwardfill' method was applied after experimenting with various techniques. Experimental results are visualized in Figure 1's box plots.

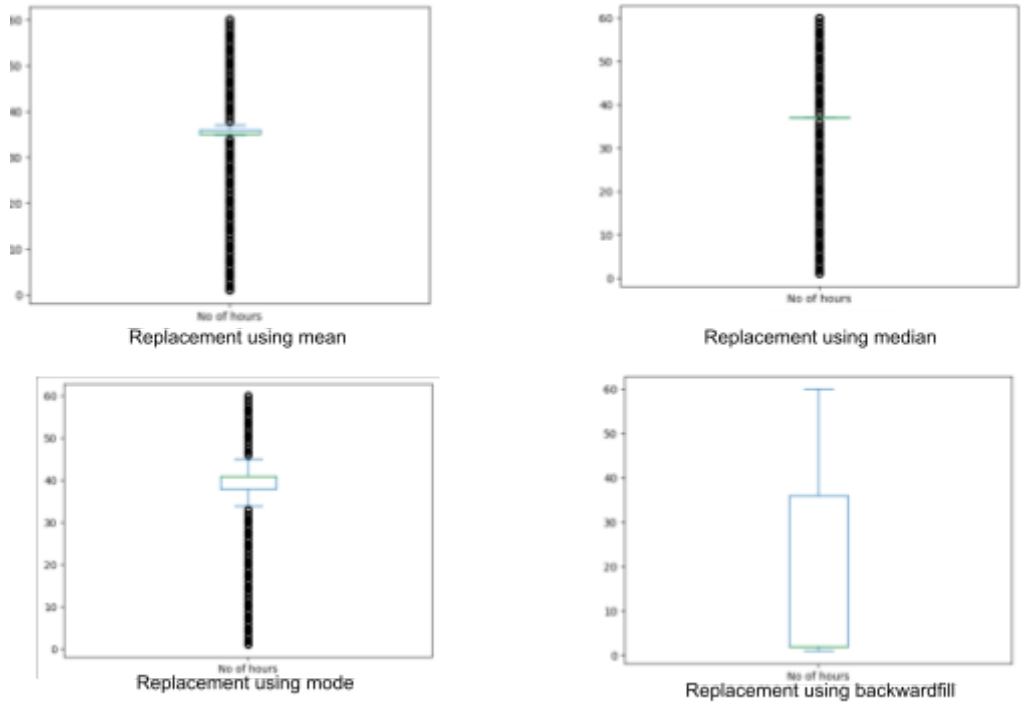


Figure1: Missing values experimentation

Next, categorical data, initially represented with codes, was transformed into meaningful names for a more insightful analysis. Figure 2 presents key findings from the descriptive analysis of all categorical attributes.

	count	unique	top	freq
Region	569740	10	South East	88084
Residence Type	569740	2	Not a resident in a communal establishment	559086
Family Composition	569740	7	Married/same-sex civil partnership couple family	300961
Population Base	569740	3	Usual resident	561039
Sex	569740	2	Female	289172
Age	569740	8	0-15	106832
Marital Status	569740	5	Single (never married or never registered a sa...	270999
Student	569740	2	Not a Student	443203
Country of Birth	569740	3	UK	485645
Health	569740	6	Very good health	264971
Ethnic Group	569740	6	White	483477
Religion	569740	10	Christian	333481
Economic Activity	569740	10	Economically active: Employee	216024
Occupation	569740	10	No code required (Aged under 16 or students or...	149984
Industry	569740	13	No code required (Aged under 16 or students or...	149984
Hours worked per week	569740	5	No code required (Aged under 16 or students or...	302321
Approximated Social Grade	569740	5	C1	159642

Figure 2: Descriptive analysis of categorical attributes

Figure 2 reveals a dataset with no missing values across attributes, ensuring completeness. Moreover, 82% of the attributes present more than two categories, indicating a high level of detail. Notably, certain attributes exhibit significant data imbalance. For instance, 'Residence Type' has two categories, with one dominating at 97%. Similarly, 'Population Base' and 'Country of Birth' show imbalances with one category representing 98% and 85% of the population, respectively. 'Ethnic Group' further demonstrates data imbalance, with one category encompassing 85% of the population.

	count	mean	std	min	25%	50%	75%	max
No of hours	569740.0	17.599421	18.997615	1.0	2.0	2.0	36.0	60.0

Figure 3: Descriptive analysis of numeric attribute

Figure 3 provides a comprehensive numeric analysis of the 'No of hours' attribute, which contains 569,740 records without any missing values, ensuring data completeness. The average weekly working hours in the dataset stand at 16.5, reflecting a diverse range due to a high standard deviation. The range spans from 1 to 60 hours, with some individuals working minimal hours and others a full 60. The percentiles reveal that a quarter of individuals work less than 2 hours ('25%'), half work less than 2 hours ('50%'), and 75% work less than 36 hours ('75%'). These findings suggest a prevalence of shorter work periods, with a smaller segment working longer hours.

Part B

Visualization 1:

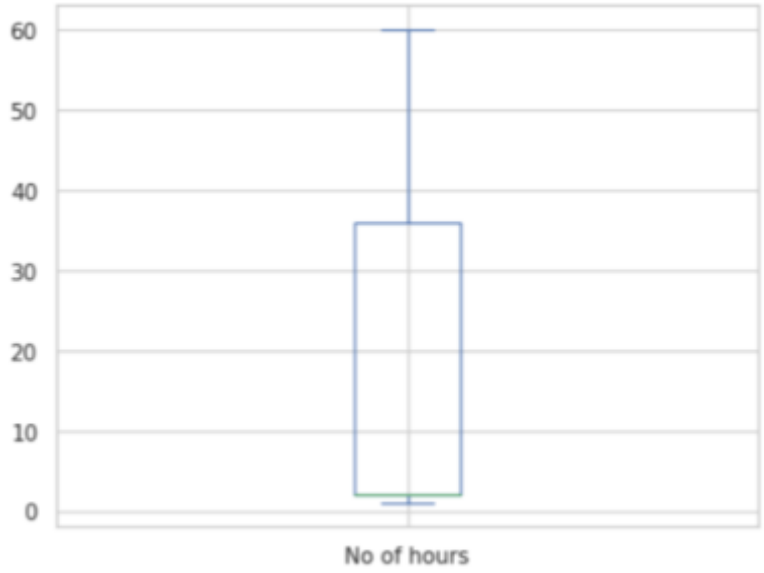


Figure 4: 'No of hours' box plot

Figure 4 provides a visual summary of the previous findings.

- Visual Interpretation: The graph visualizes the statistics mentioned in the analysis of the 'No of hours' attribute shown above.
- Data Distribution: The graph illustrates that the majority of individuals work less than 36 hours, evident from the box's positioning closer to the minimum value, indicating a substantial gap between it and the maximum value.

Visualization 2: Stacked Bar Chart

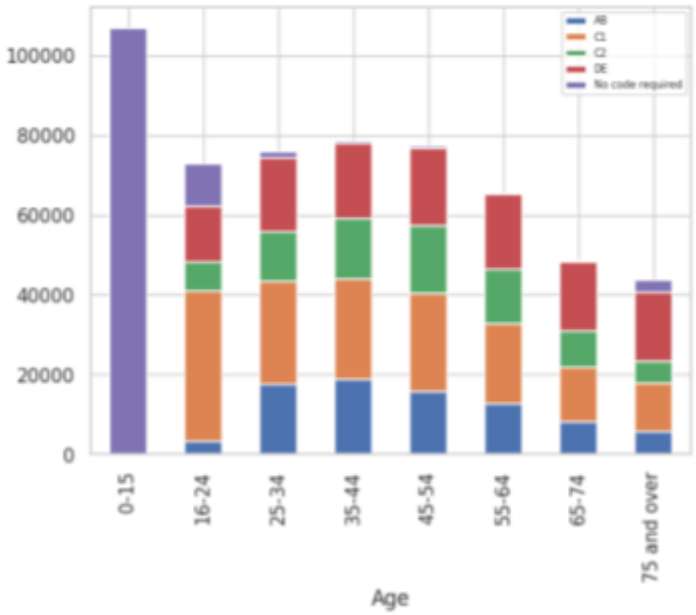


Figure 5: Age and Approximated Social Grade

Figure 5 unveils insights into the Age and Approximated Social Grade relationship:

- Social Grade Distribution: C1 is the most common for individuals aged 16 and above, indicating widespread representation.
- Age Group Insights: Ages 25-54 are predominantly under social grade AB, the highest grade.
- Grade C2: Demonstrates the least representation, mainly among individuals aged 35-64.
- Grade DE: The second most prevalent social grade among individuals aged 16 and over.
- Overall Observation: The chart highlights a significant majority in grades C1 and below, with limited presence in the AB grade.

Visualization 3:

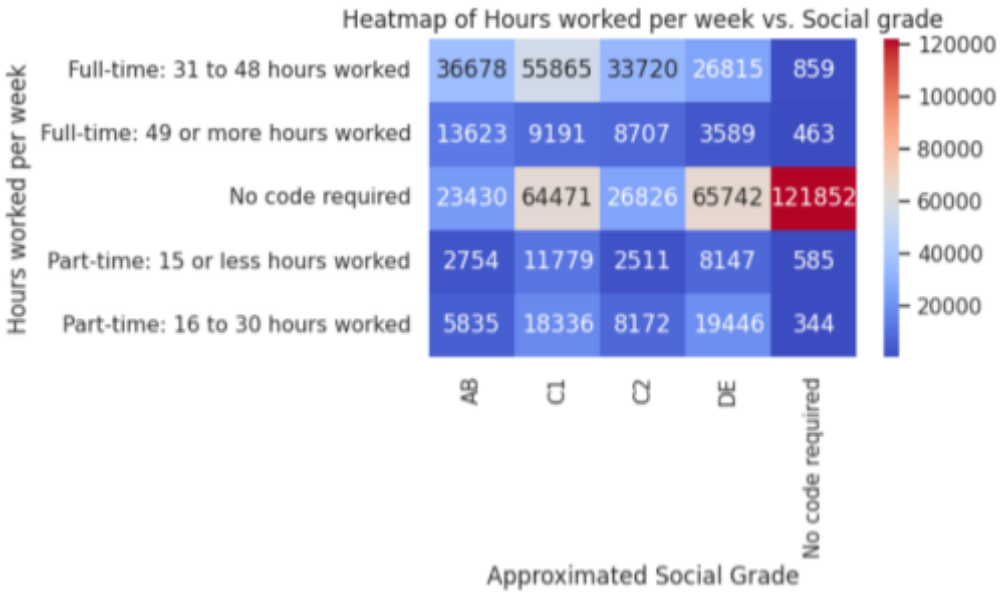


Figure 6: Hours worked per week and Approximated social grade

Figure 6 reveals the link between working hours and social grades:

- Full-time workers (31-48 hours) are mostly in social grade C1.
- Those working longer hours (more than 49 hours) are often categorized as social grade AB.
- Part-time workers (15 hours or less) are mainly in social grade C1.
- Part-time workers (16-30 hours) display a relatively even distribution between social grades C1 and DE.
- The majority across all social grades work full-time for 31-48 hours, reflecting the high number of records in this category.

Visualization 4: Stacked bar chart

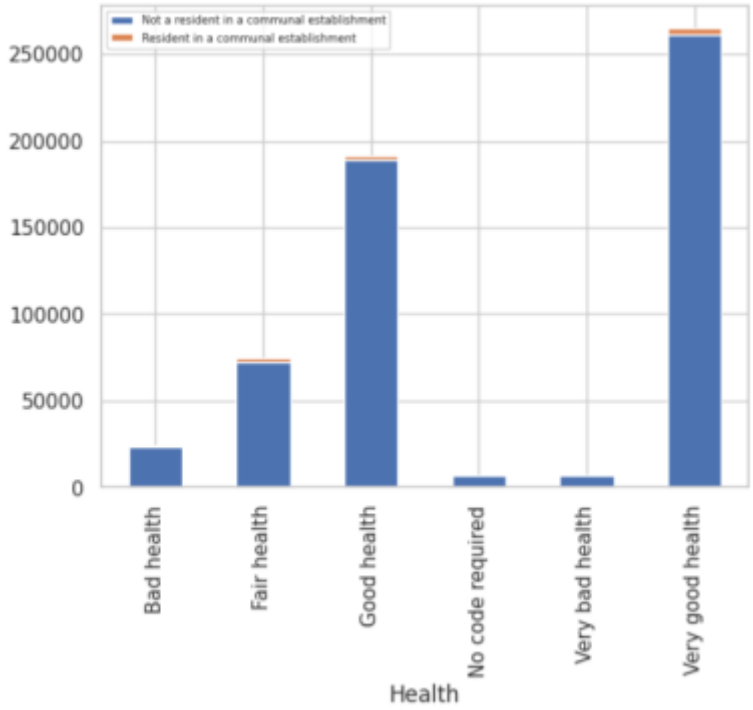


Figure 7: Health and Residence type

Figure 7 illustrates the correlation between health and residence type.

- Contradictory Findings: Surprisingly, most individuals in communal establishments report fair health or higher, contradicting the expectation that such establishments, including hospitals, would house individuals with poorer health. Individuals with bad health are not found in communal establishments based on the visualization.
- Overall Health: The dataset suggests that the majority of individuals enjoy good health or better, indicating a positive trend.
- Exceptional Health: A noteworthy portion of individuals in communal establishments reports very good health. This might be explained by the inclusion of boarding schools and student halls in the communal establishment category.

Visualization 5: Countplot(Sahoo et al., 2019)

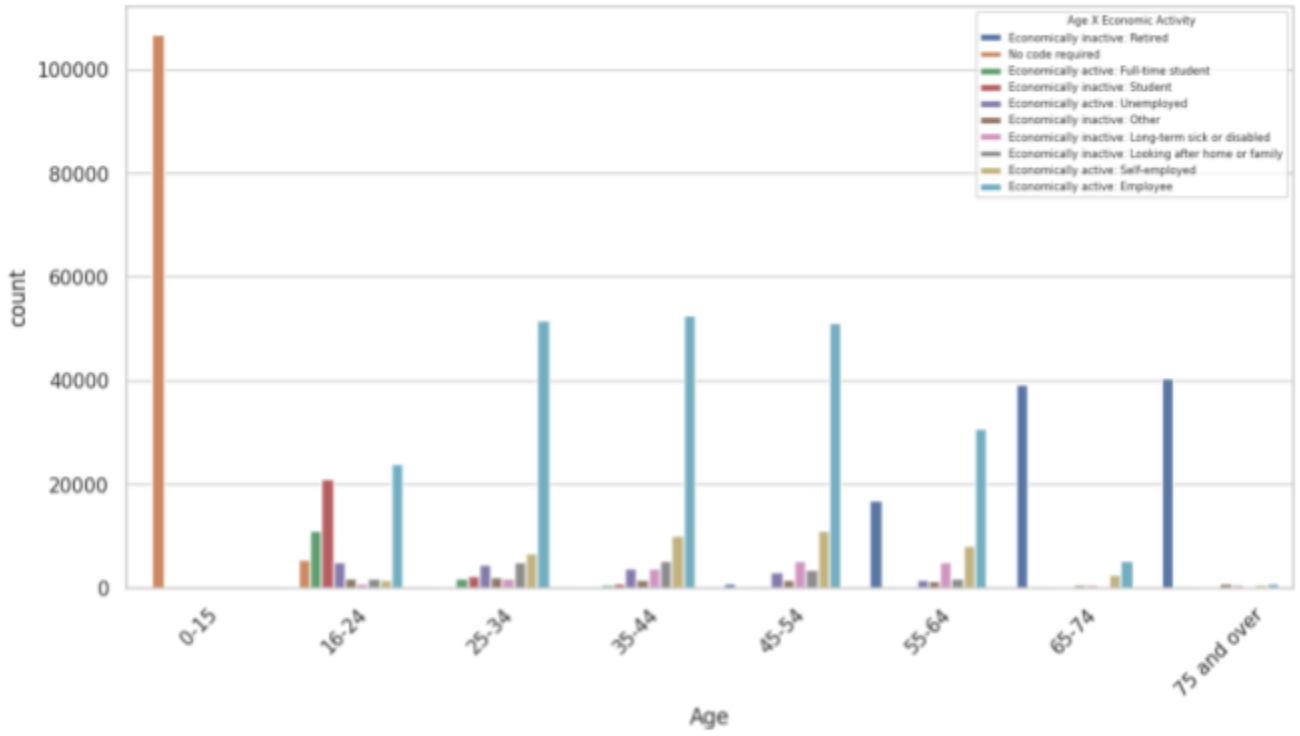


Figure 8: Age and Economic Activity

Figure 8 illustrates the relationship between age and economic activity.

- Individuals aged 16-64 are predominantly economically active employees, making it the most common group within that age range.
- Individuals aged 65 and above are primarily in the retired category.
- The majority of self-employed individuals fall within the age range of 25-54, suggesting a higher inclination toward entrepreneurship or self-employment in this age group.
- The age group of 16-24 shows the highest variance in economic activity, reflecting a diverse range of employment statuses within this specific age category.

Visualization 6: Countplot

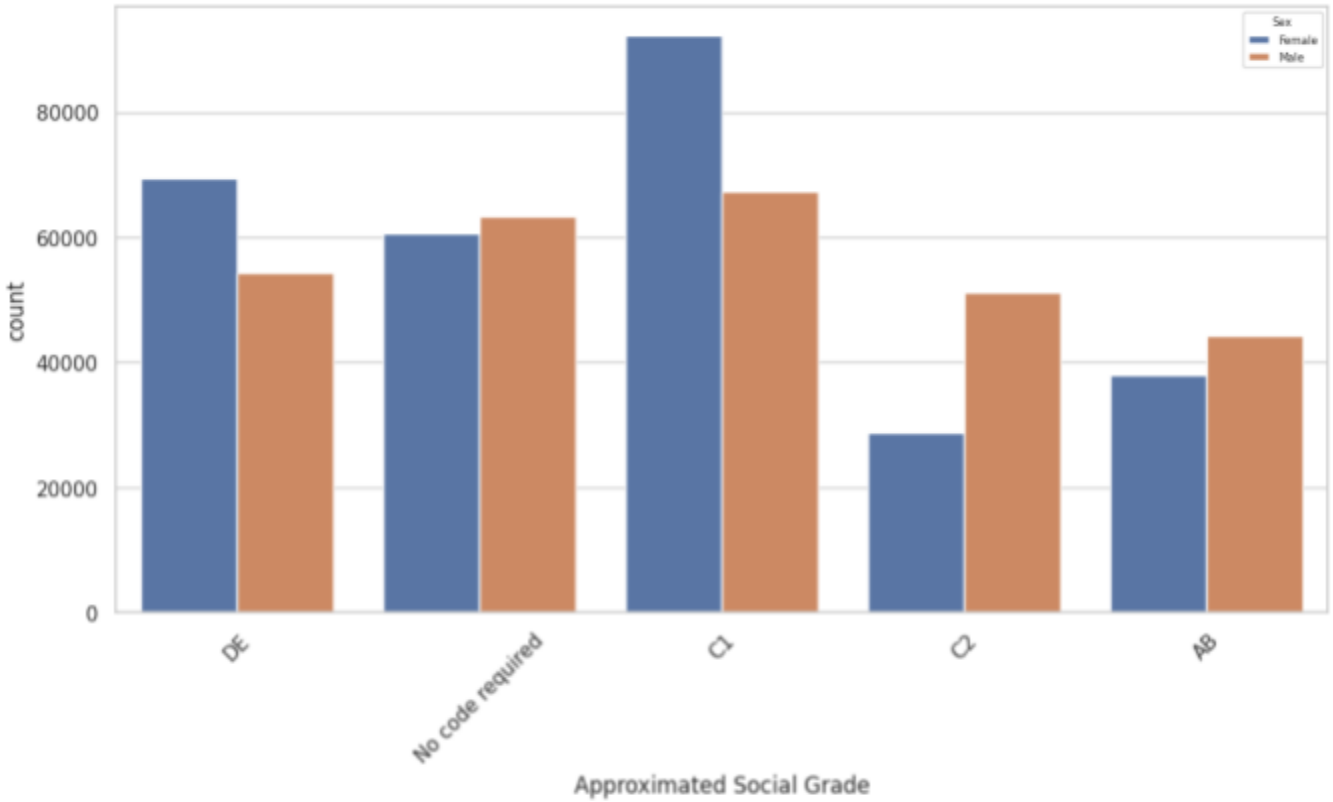


Figure 9: Approximated social grade and Sex

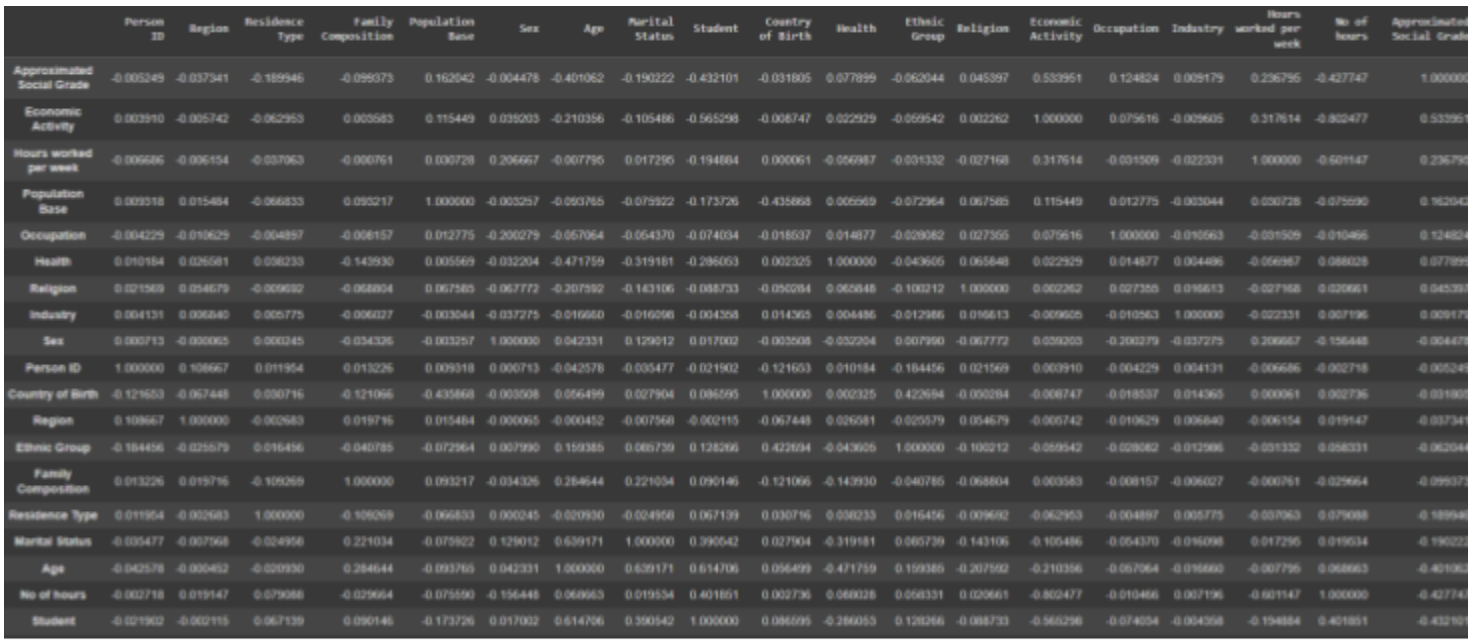
Figure 9 depicts the relationship between 'Sex' and 'Approximated Social Grade' in the dataset.

- The highest social grade, AB, contains a higher proportion of men, while the lowest grade, DE, has more women.
- Social grade C1 is predominantly composed of women, whereas social grade C2 is predominantly men.
- Social grade C1 emerges as the most prevalent category in the dataset.

Task 2: Classification

Data preparation steps:

- Data Transformation: Attributes with -9 values were converted to meaningful categories, and then those attributes along with the rest of the nominal ones were transformed into numerical representations using Label Encoder, effectively handling -9 values.
- Handling Missing Values: 'No of hours' attribute with numerous missing values was filled using the backward fill method, ensuring dataset completeness for analysis after numerous experiments as mentioned above.
- Data Normalization: Crucial for fair attribute comparisons, it ensures uniformity across different attributes. This is particularly important because certain algorithms are sensitive to the scale of input data. Without normalization, larger-scale attributes could disproportionately influence the classification results, leading to suboptimal outcomes.
- Dimensionality Reduction: Pearson correlation analysis along with objective understanding guided the selection of attributes for classification. Figure 10's last column displays attribute correlations in ascending order. Notably, 'Age', 'Student', 'Economic Activity', 'Occupation', 'Industry', and 'Hours worked per week' were chosen.



	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	Hours worked per week	No of hours	Approximated Social Grade
Approximated Social Grade	-0.005249	-0.037341	-0.189946	-0.099373	0.162042	-0.004478	-0.401062	-0.190222	-0.432101	-0.031805	0.077899	-0.062044	0.045397	0.533951	0.124824	0.009179	0.236795	-0.427747	1.000000
Economic Activity	0.003910	-0.005742	-0.062953	0.003583	0.115449	0.039203	-0.210356	-0.105486	-0.565298	-0.008747	0.022929	-0.069542	0.002262	1.000000	0.075616	-0.009605	0.317614	-0.802477	0.533951
Hours worked per week	-0.006686	-0.006154	-0.037063	-0.000761	0.030728	0.206667	-0.007795	0.017295	-0.194884	0.000061	-0.056987	-0.031332	-0.027168	0.317614	-0.031509	-0.022331	1.000000	-0.601147	0.236795
Population Base	0.009318	0.015484	-0.066833	0.093217	1.000000	-0.003257	-0.093765	-0.075922	-0.173726	-0.435868	0.006969	-0.072964	0.067585	0.115449	0.012775	-0.003044	0.030728	-0.075690	0.162042
Occupation	-0.004229	-0.010629	-0.004897	-0.008157	0.012775	-0.200279	-0.057064	-0.054370	-0.074034	-0.018537	0.014877	-0.028082	0.027355	0.075616	1.000000	-0.010563	-0.031509	-0.010466	0.124824
Health	0.010184	0.026581	0.038233	-0.143930	0.005569	-0.032204	-0.471759	-0.319181	-0.286053	0.002325	1.000000	-0.043605	0.065848	0.022929	0.014877	0.004486	-0.056987	0.088028	0.077899
Religion	0.021569	0.054679	-0.009692	-0.068804	0.067585	-0.067772	-0.207592	-0.143106	-0.088733	-0.050284	0.065848	-0.100212	1.000000	0.002262	0.027355	0.016613	-0.027168	0.020661	0.045397
Industry	0.004131	0.006840	0.005775	-0.006027	-0.003044	-0.037275	-0.016660	-0.016098	-0.004358	0.014365	0.004486	-0.012986	0.016613	-0.009605	-0.010563	1.000000	-0.022331	0.007196	0.009179
Sex	0.000713	-0.000065	0.000245	-0.034326	-0.003257	1.000000	0.042331	0.129012	0.017002	-0.003508	-0.032204	0.007990	-0.067772	0.039203	-0.200279	-0.037275	0.206667	-0.156448	-0.004478
Person ID	1.000000	0.108667	0.011954	0.013226	0.009318	0.000713	-0.042578	-0.035477	-0.021902	-0.121653	0.010184	-0.184456	0.021569	0.003910	-0.004229	0.004131	-0.006686	-0.002718	-0.005249
Country of Birth	-0.121653	-0.067448	0.030716	-0.121066	-0.435868	-0.003508	0.056499	0.027904	0.086595	1.000000	0.002325	0.422694	-0.050284	-0.008747	-0.018537	0.014365	0.000061	0.002736	-0.031805
Region	0.108667	1.000000	-0.002683	0.019716	0.015484	-0.000065	-0.000452	-0.007568	-0.002115	-0.067448	0.026581	-0.025579	0.054679	-0.005742	-0.010629	0.006840	-0.006154	0.019147	-0.037341
Ethnic Group	-0.184456	-0.025579	0.016456	-0.040785	-0.072964	0.007990	0.159385	0.085739	0.128266	0.422694	-0.043605	1.000000	-0.100212	-0.069542	-0.028082	-0.012986	-0.031332	0.058331	-0.062044
Family Composition	0.013226	0.019716	-0.109269	1.000000	0.093217	-0.034326	0.284644	0.221034	0.090146	-0.121066	-0.143930	-0.040785	-0.068804	0.003583	-0.008157	-0.006027	-0.000761	-0.029664	-0.099373
Residence Type	0.011954	-0.002683	1.000000	-0.109269	-0.066833	0.000245	-0.020930	-0.024958	0.067139	0.030716	0.038233	0.016456	-0.009692	-0.062953	-0.004897	0.005775	-0.037063	0.079088	-0.189946
Marital Status	-0.035477	-0.007568	-0.024958	0.221034	-0.075922	0.129012	0.639171	1.000000	0.390542	0.027904	-0.319181	0.085739	-0.143106	-0.105486	-0.054370	-0.016098	0.017295	0.019534	-0.190222
Age	-0.042578	-0.000452	-0.020930	0.284644	-0.093765	0.042331	1.000000	0.639171	0.614706	0.056499	-0.471759	0.159385	-0.207592	-0.210356	-0.057064	-0.016660	-0.007795	0.068663	-0.401062
No of hours	-0.002718	0.019147	0.079088	-0.029664	-0.075690	-0.156448	0.068663	0.019534	0.401851	0.002736	0.088028	0.058331	0.020661	-0.802477	-0.010466	0.007196	-0.601147	1.000000	-0.427747
Student	-0.021902	-0.002115	0.067139	0.090146	-0.173726	0.017002	0.614706	0.390542	1.000000	0.086595	-0.286053	0.128266	-0.088733	-0.565298	-0.074034	-0.004358	-0.194884	0.401851	-0.432101

Figure 10: Pearson Correlation

Algorithms Comparison:

	Algorithms	Precision	Recall	F1 Score	Accuracy Score
0	MLPPartitioning	0.806721	0.798933	0.799453	0.798933
1	MLPCrossValidation	0.805256	0.797697	0.798660	0.797697
2	KNNPartitioning	0.779006	0.770219	0.772774	0.770219
3	KNNCrossValidation	0.774062	0.766753	0.769274	0.766753
4	SVMPartitioning	0.742573	0.710053	0.715983	0.710053
5	SVMCrossValidation	0.738885	0.707446	0.713860	0.707446
6	LRCrossValidation	0.609680	0.613525	0.602104	0.613525
7	LRPartitioning	0.601340	0.608748	0.595716	0.608748

Figure 11: Classification Algorithms Comparison

Figure 11 displays the algorithm performance metrics, where higher values indicate superior performance. All models underwent training with both cross-validation and partitioning for accuracy comparison. However, only cross-validation results will be considered during confusion matrix analysis,because despite the apparent superiority of partitioning algorithms, cross-validation offers more reliable accuracy by leveraging multiple folds.

Multi-Layer Perceptron (MLP) Models:

MLP, chosen for its ability to model complex, non-linear relationships, outperformed other models with the highest accuracy, precision, recall, and F1 score. Figure 12 visually confirms its strong performance, with the confusion matrix showing predominant alignment between predictions and true labels, particularly excelling in classes 1, 3, and 4.

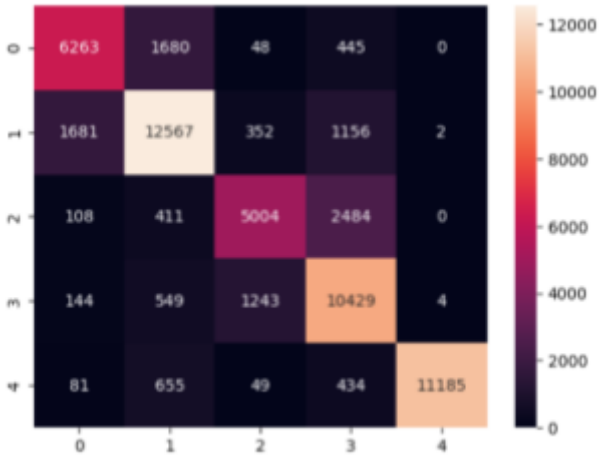


Figure 12: MLP Confusion Matrix

K-Nearest Neighbors (KNN):

KNN, chosen for its simplicity and flexibility, underwent experiments to optimize the number of k-neighbors (see Figure 13). KNNPartitioning and KNNCrossValidation ranked competitively, 3rd and 5th, respectively. Known for simplicity and effectiveness, KNN models performed well, as evidenced by metrics in Figure 11 and the confusion matrix in Figure 14, showcasing accurate classification across categories, especially on classes 1, 3, and 4 as the previous modes. This model demonstrates adequate performance for this task.

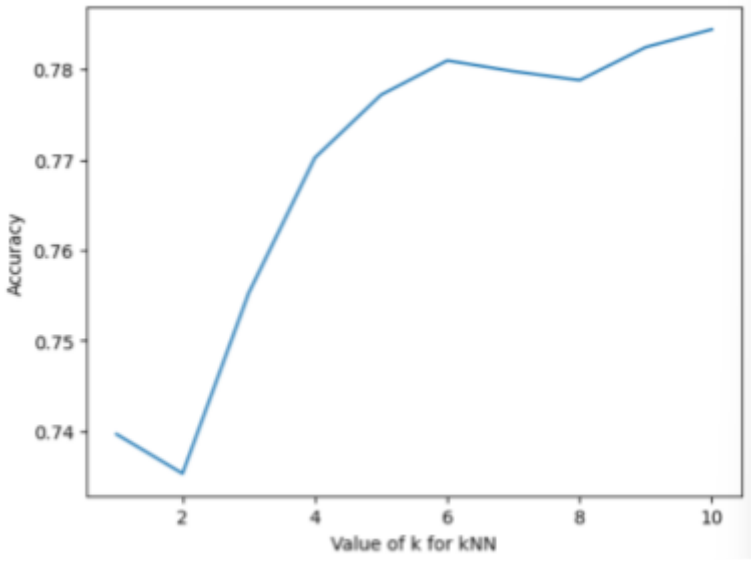


Figure 13: KNN performance with different K

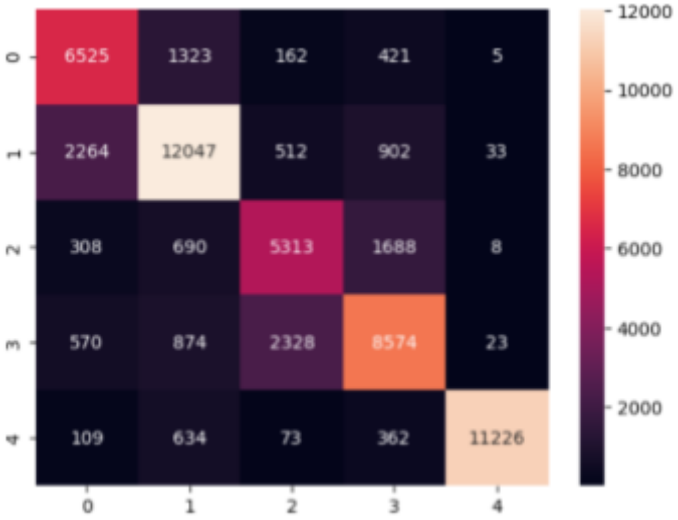


Figure 14: KNN Confusion Matrix

Support Vector Machine (SVM) Models:

This algorithm was used due to its ability to utilize kernels and ability to perform well on complex tasks and would be a great comparison with the MLP algorithm.The poly kernel yielded the highest accuracy after experimentation. SVMCrossValidation and SVMPartitioning showed similar accuracy trends. Like the previous algorithms, superior classification was observed for classes 1, 3, and 4, evident in Figure 15. Overall, considering all metrics, the model's performance is acceptable and suitable for the task

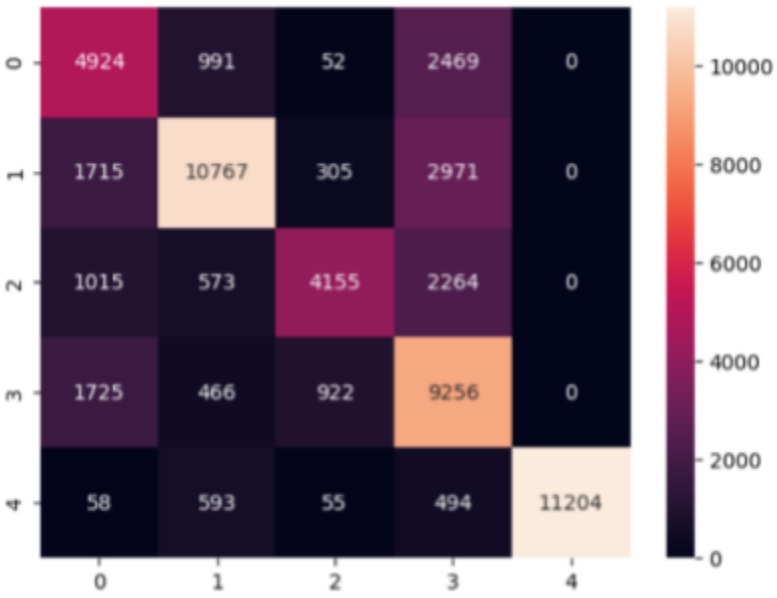


Figure 15: SVM Confusion Matrix

Logistic Regression (LR) Models(LaValley, 2008):

LR, chosen for simplicity and efficiency on large datasets, underperformed compared to other models. LRPartitioning and LRCrossValidation exhibited lower performance, consistent with LR's simplicity. Figure 16's confusion matrix reveals decent performance on classes 1,3 and 4 as the previous algorithms but very poor on the other classes, rendering the model unsuitable for this classification.

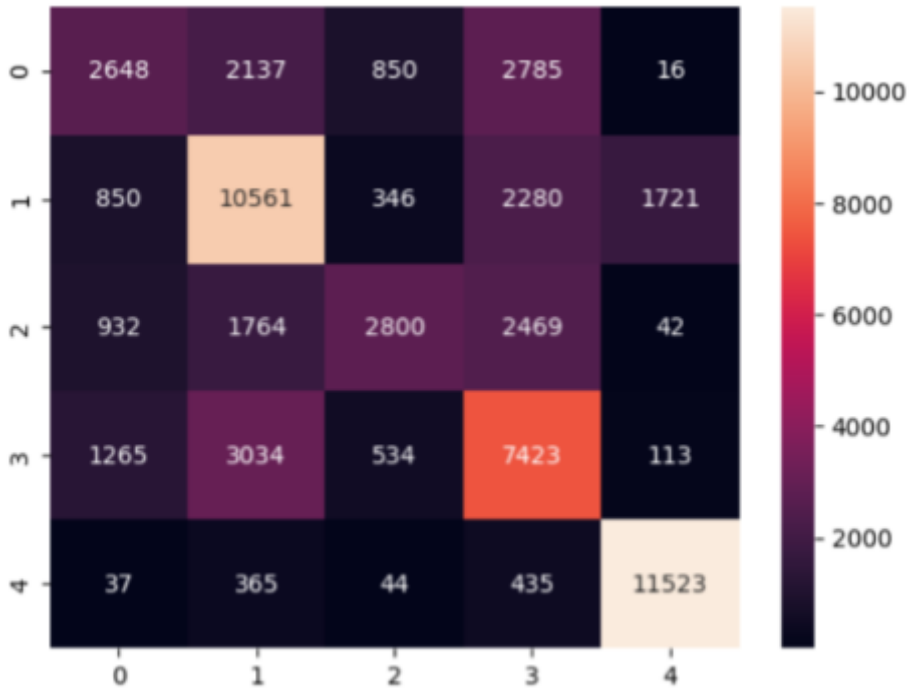


Figure 16: LR Confusion Matrix

Task 3: Regression

Data preparation steps:

For regression, the process mirrored classification, except for the Dimensionality reduction part. Attributes ('Industry,' 'Student,' 'Occupation,' and 'Approximated Social Grade') were chosen based on the correlation table (Figure 17) but also the new objective understanding due to the different target attribute.

	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	Hours worked per week	No of hours	Approximated Social Grade
No of hours	-0.002718	0.019147	0.079088	-0.029664	-0.075590	-0.156448	0.068663	0.019534	0.401851	0.002736	0.088028	0.058331	0.020661	-0.802477	-0.010466	0.007196	-0.601147	1.000000	-0.427747
Student	-0.021902	-0.002115	0.067139	0.090146	-0.173726	0.017002	0.614706	0.390542	1.000000	0.086595	-0.286053	0.128266	-0.088733	-0.565298	-0.074034	-0.004358	-0.194884	0.401851	-0.432101
Health	0.010184	0.026581	0.038233	-0.143930	0.005069	-0.032204	-0.471759	-0.319181	-0.286053	0.002325	1.000000	-0.043605	0.065848	0.022929	0.014877	0.004486	-0.056987	0.088028	0.077899
Residence Type	0.011954	-0.002683	1.000000	-0.109269	-0.066833	0.000245	-0.020930	-0.024958	0.067139	0.030716	0.038233	0.016456	-0.009692	-0.062953	-0.004897	0.005775	-0.037063	0.079088	-0.189946
Age	-0.042578	-0.000452	-0.020930	0.284644	-0.093765	0.042331	1.000000	0.639171	0.614706	0.056499	-0.471759	0.159385	-0.207592	-0.210356	-0.057064	-0.016660	-0.007795	0.068663	-0.401062
Ethnic Group	-0.184456	-0.025579	0.016456	-0.040785	-0.072964	0.007990	0.159385	0.085739	0.128266	0.422694	-0.043605	1.000000	-0.100212	-0.059542	-0.028082	-0.012986	-0.031332	0.058331	-0.062044
Religion	0.021569	0.054679	-0.009692	-0.068804	0.067585	-0.067772	-0.207592	-0.143106	-0.088733	-0.050284	0.065848	-0.100212	1.000000	0.002262	0.027355	0.016613	-0.027168	0.020661	0.045397
Marital Status	-0.035477	-0.007568	-0.024958	0.221034	-0.079922	0.129012	0.639171	1.000000	0.390542	0.027904	-0.319181	0.085739	-0.143106	-0.105486	-0.054370	-0.016098	0.017295	0.019534	-0.190222
Region	0.106667	1.000000	-0.002683	0.019716	0.015484	-0.000065	-0.000452	-0.007568	-0.002115	-0.067448	0.026581	-0.025579	0.054679	-0.005742	-0.010629	0.006840	-0.006154	0.019147	-0.037341
Industry	0.004131	0.006840	0.005775	-0.006027	-0.003044	-0.037275	-0.016660	-0.016098	-0.004358	0.014365	0.004486	-0.012986	0.016613	-0.009605	-0.010563	1.000000	-0.022331	0.007196	0.009179
Country of Birth	-0.121653	-0.067448	0.030716	-0.121066	-0.435868	-0.003508	0.056499	0.027904	0.086595	1.000000	0.002325	0.422694	-0.050284	-0.006747	-0.018537	0.014365	0.000061	0.002736	-0.031805
Person ID	1.000000	0.106667	0.011954	0.013226	0.009318	0.000713	-0.042578	-0.035477	-0.021902	-0.121653	0.010184	-0.184456	0.021569	0.003910	-0.004229	0.004131	-0.006686	-0.002718	-0.005249
Occupation	-0.004229	-0.010629	-0.004897	-0.008157	0.012775	-0.200279	-0.057064	-0.054370	-0.074034	-0.018537	0.014877	-0.028082	0.027355	0.075616	1.000000	-0.010563	-0.031509	-0.010466	0.124824
Family Composition	0.013226	0.019716	-0.109269	1.000000	0.093217	-0.034326	0.284644	0.221034	0.090146	-0.121066	-0.143930	-0.040785	-0.068804	0.003583	-0.008157	-0.006027	-0.000761	-0.029664	-0.099373
Population Base	0.009318	0.015484	-0.066833	0.093217	1.000000	-0.003257	-0.093765	-0.079922	-0.173726	-0.435868	0.005069	-0.072964	0.067585	0.115449	0.012775	-0.003044	0.030728	-0.075590	0.162042
Sex	0.000713	-0.000065	0.000245	-0.034326	-0.003257	1.000000	0.042331	0.129012	0.017002	-0.003508	-0.032204	0.007990	-0.067772	0.039203	-0.200279	-0.037275	0.206667	-0.156448	-0.004478
Approximated Social Grade	-0.005249	-0.037341	-0.189946	-0.099373	0.162042	-0.004478	-0.401062	-0.190222	-0.432101	-0.031805	0.077899	-0.062044	0.045397	0.533951	0.124824	0.009179	0.236795	-0.427747	1.000000
Hours worked per week	-0.006686	-0.006154	-0.037063	-0.000761	0.030728	0.206667	-0.007795	0.017295	-0.194884	0.000061	-0.056987	-0.031332	-0.027168	0.317614	-0.031509	-0.022331	1.000000	-0.601147	0.236795
Economic Activity	0.003910	-0.005742	-0.062953	0.003583	0.115449	0.039203	-0.210356	-0.105486	-0.565298	-0.006747	0.022929	-0.059542	0.002262	1.000000	0.075616	-0.009605	0.317614	-0.802477	0.533951

Figure 17: Pearson correlation for 'No of hours'

Algorithms Comparison:

	Algorithms	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	R2_Score	Adjusted R2_Score
0	RegressionTreeCrossValidation	12.216958	252.949783	15.904395	0.303486	0.303437
1	RegressionTreePartitioning	12.308050	255.670481	15.989699	0.295101	0.294853
2	LinearRegressionCrossValidation	13.889294	274.141145	16.557208	0.245134	0.245081
3	LinearRegressionPartitioning	13.928690	275.024666	16.583868	0.241740	0.241474
4	SVRPartitioning	12.150630	280.799894	16.757085	0.225817	0.225545
5	SVRCrossValidation	12.140411	283.302992	16.831607	0.219907	0.219852

Figure 18: Regression Algorithms Comparison

Figure 18 illustrates the comparative performance of all algorithms. Higher R2 and Adjusted R2 scores indicate better performance, while lower values for other metrics suggest better performance.

Regression Tree:

The Regression tree algorithm excels in capturing complex data relationships, outperforming both SVR and Linear Regression by having the highest R2 and Adjusted R2 scores while having also the lowest value of the other error metrics.

Support Vector Regression(Awad et al., 2015):

SVR also handles nonlinear relationships well, but the big performance difference in comparison with RT is attributed to the Regression tree's superior ability to capture complexity.

Linear Regression:

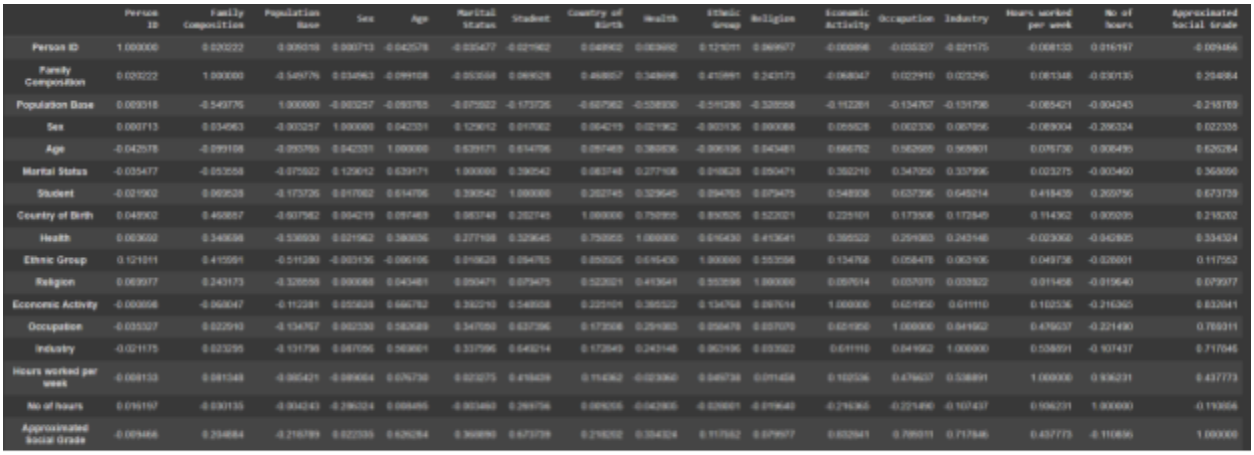
This algorithm performs poorly as it is better suited for linear relationships between predictors and targets.

In conclusion none of the algorithms used were suitable to complete the regression tasks and performed poorly probably due to the low correlation between the predictor and target variables, but by the dataset provided the best model that could be created with the highest accuracy is by far the Regression tree model.

Task 4: Association Rule Mining

Data Preparation steps:

For Association Rule Mining, data was loaded,the less correlated attributes were removed according to Pearson correlation (Figure 19). Numeric attributes were converted to nominal, filled missing values in 'No of hours' with backward fill. Used a sample for Apriori algorithm, while tweaking support confidence and attributes for meaningful rules with high lift.



	Person ID	Family Composition	Population Base	Sex	Age	Marital Status	Student	Country of Birth	Health	Ethnic Group	Religion	Economic Activity	Occupation	Industry	Hours worked per week	No of hours	Approximated Social Grade
Person ID	1.000000	0.020222	0.009318	0.000713	-0.042578	-0.035477	-0.021902	0.048902	0.003692	0.121011	0.069377	-0.000898	-0.035327	-0.021175	-0.008133	0.016197	-0.009466
Family Composition	0.020222	1.000000	-0.549776	0.034963	-0.099108	-0.053058	0.069528	0.468857	0.348698	0.410991	0.243173	-0.068047	0.022910	0.023256	0.081348	-0.030135	0.204984
Population Base	0.009318	-0.549776	1.000000	-0.003257	-0.093765	-0.075022	-0.173726	-0.607562	-0.538300	-0.511280	-0.328558	-0.112281	-0.134767	-0.131798	-0.085421	-0.004243	-0.216789
Sex	0.000713	0.034963	-0.003257	1.000000	0.042331	0.129012	0.017082	0.064219	0.021962	-0.003136	0.000388	0.055828	0.002330	0.007096	-0.089004	-0.206324	0.022358
Age	-0.042578	-0.099108	-0.093765	0.042331	1.000000	0.639171	0.614796	0.097469	0.380836	-0.006196	0.043481	0.666702	0.582695	0.565801	0.076730	0.006495	0.620284
Marital Status	-0.035477	-0.053058	-0.075022	0.129012	0.639171	1.000000	0.390542	0.083748	0.277108	0.010628	0.050471	0.350210	0.347050	0.337996	0.023275	-0.003460	0.369056
Student	-0.021902	0.069528	-0.173726	0.017082	0.614796	0.390542	1.000000	0.282745	0.329645	0.094765	0.079475	0.548308	0.637396	0.645214	0.418439	0.205795	0.673739
Country of Birth	0.048902	0.468857	-0.607562	0.064219	0.097469	0.083748	0.282745	1.000000	0.750955	0.850525	0.522021	0.225101	0.173908	0.172849	0.114362	0.005205	0.219202
Health	0.003692	0.348698	-0.538300	0.021962	0.380836	0.277108	0.329645	0.750955	1.000000	0.616430	0.413641	0.386522	0.291983	0.282148	-0.023060	-0.042805	0.334324
Ethnic Group	0.121011	0.410991	-0.511280	-0.003136	-0.006196	0.010628	0.094765	0.850525	0.616430	1.000000	0.553598	0.134768	0.058478	0.063106	0.049738	-0.026001	0.117552
Religion	0.069377	0.243173	-0.328558	0.000388	0.043481	0.050471	0.079475	0.522021	0.413641	0.553598	1.000000	0.057614	0.037070	0.033822	0.011458	-0.019640	0.079377
Economic Activity	-0.000898	-0.068047	-0.112281	0.055828	0.666782	0.350210	0.548308	0.225101	0.386522	0.134768	0.057614	1.000000	0.651950	0.611110	0.102536	-0.216365	0.832041
Occupation	-0.035327	0.022910	-0.134767	0.002330	0.582695	0.347050	0.637396	0.173908	0.291983	0.058478	0.037070	0.651950	1.000000	0.841952	0.479637	-0.221490	0.789311
Industry	-0.021175	0.023256	-0.131798	0.007096	0.565801	0.337996	0.645214	0.172849	0.282148	0.063106	0.033822	0.611110	0.841952	1.000000	0.538891	-0.107437	0.717846
Hours worked per week	-0.008133	0.081348	-0.085421	-0.089004	0.076730	0.023275	0.418439	0.114362	-0.023060	0.049738	0.011458	0.102536	0.479637	0.538891	1.000000	0.936231	0.437773
No of hours	0.016197	-0.030135	-0.004243	-0.206324	0.006495	-0.003460	0.205795	0.005205	-0.042805	-0.026001	-0.019640	-0.216365	-0.221490	-0.107437	0.936231	1.000000	-0.110856
Approximated Social Grade	-0.009466	0.204984	-0.216789	0.022358	0.620284	0.369056	0.673739	0.219202	0.334324	0.117552	0.079377	0.832041	0.789311	0.717846	0.437773	-0.110856	1.000000

Figure 19: Attributes Pearson Correlation

Rules Interpretation:

Rule 1:

Items	Antecedent	Consequent	Support	Confidence	Lift
{'UK', 'Not a student', 'Not a resident in a communal establishment', 'White', 'Christian'}	{'White', 'Not a resident in a communal establishment', 'Christian'}	{'UK', 'Not a student'}	0.41477970861857116	0.7702086049543676	1.1562261982674658

This rule implies that when an individual is White, not a resident in a communal establishment, and Christian, there is a 77.02% chance that they are from the UK and not a student. The lift of 1.16 indicates a positive correlation, suggesting that the presence of the antecedent increases the likelihood of the consequent.

Rule 2:

Items	Antecedent	Consequent	Support	Confidence	Lift
{'UK', 'Not a student', 'Not a resident in a communal establishment', 'White', 'Married or in a registered same-sex civil partnership'}	{'Married or in a registered same-sex civil partnership', 'White'}	{'Not a resident in a communal establishment', 'UK', 'Not a student'}	0.30068457082675093	0.9184986595174263	1.3953831635388738

The above rule implies that individuals who are 'Married or in a registered same-sex civil partnership' and characterized as 'White' exhibit a substantial likelihood (91.85%) of also being 'Not a resident in a communal establishment', residing in the 'UK', and being 'Not a student'. The high lift value indicates that this association is stronger than what would be expected by chance.

Rule 3:

Items	Antecedent	Consequent	Support	Confidence	Lift
{'Student', '(0.999, 2.0]', 'Single (never married or never registered a same-sex civil partnership)'}	{'(0.999, 2.0]', 'Single (never married or never registered a same-sex civil partnership)'}	{'Student'}	0.20098297349482183	0.6550343249427918	2.9406860119772142

The given rule signifies that individuals who work '(0.999, 2.0]' hours worked per week, and are 'Single (never married or never registered a same-sex civil partnership)', are highly likely (65.50%) to be classified as 'Students'. The elevated lift value of 2.94 indicates an incredibly strong association between antecedent and consequent.

Rule 4:

Items	Antecedent	Consequent	Support	Confidence	Lift
{'Full-time: 31 to 48 hours worked', 'Economically active: Employee', 'Not a resident in a communal establishment'}	{'Economically active: Employee'}	{'Full-time: 31 to 48 hours worked', 'Not a resident in a communal establishment'}	0.2339827979638406	0.6148523985239852	2.261339002189247

This association rule reveals that those labeled as 'Economically active: Employee' have a 61.5% likelihood of falling into the categories 'Full-time: 31 to 48 hours worked' and 'Not a resident in a communal establishment.' The considerable lift value of 2.26 emphasizes a robust connection, surpassing what might occur randomly by more than twice.

Rule 5:

Items	Antecedent	Consequent	Support	Confidence	Lift
{'(36.0, 60.0]', 'Not a student', 'Economically active: Employee'}	{'Not a student', 'Economically active: Employee'}	{'(36.0, 60.0]'}	0.20484465508162192	0.5382841328413285	2.171816363170

The rule above suggests that individuals who are 'Not a student' and are 'Economically active: Employee', are highly likely (53.83%) to fall into the category of working '(36.0, 60.0]' hours per week. The substantial lift value of 2.17 indicates a remarkably strong association between the conditions of being 'Not a student' and 'Economically active: Employee' and the outcome of working '(36.0, 60.0]' hours per week.

Task 5:Clustering

Data Preparation steps:

In this step, the process involved handling missing data using the backward fill method, removing uncorrelated features (Person ID, Sex, Population Base) through Pearson correlation analysis. Columns with -9 values were addressed as before. A 10% sample was extracted for clustering, and categorical attributes were transformed and scaled. Clustering was performed using the K-means and Agglomerative algorithms.

Algorithms Comparison:

	Algorithms	Cluster 1 Coverage	Cluster 2 Coverage	Silhouette Score
0	Agglomerative	887	56087	0.46
1	K-Means	41529	15445	0.18

Figure 20: Algorithm Comparison

K-Means:

- Cluster 1 (41,529 instances): This cluster represents a substantial portion of the dataset, suggesting a prevalent category.
- Cluster 2 (15,445 instances): A smaller cluster compared to Cluster 1, indicating a less common group.
- The positive silhouette score suggests a moderate level of cohesion, but less than the Agglomerative method.

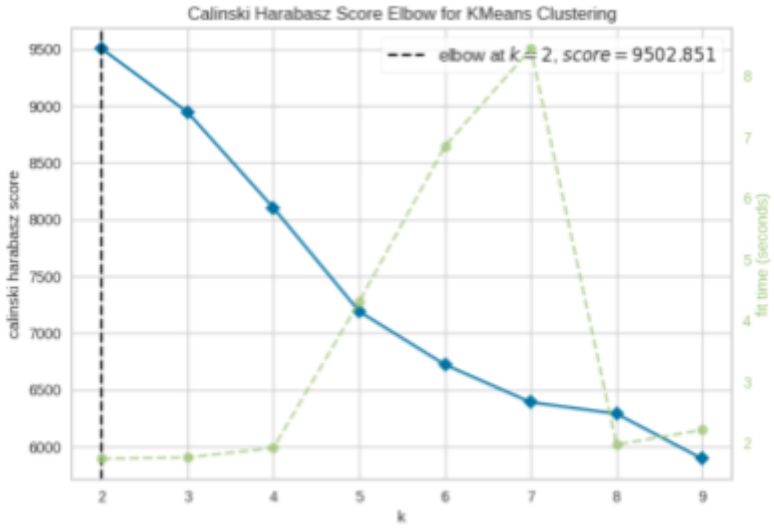


Figure 21: Elbow Score

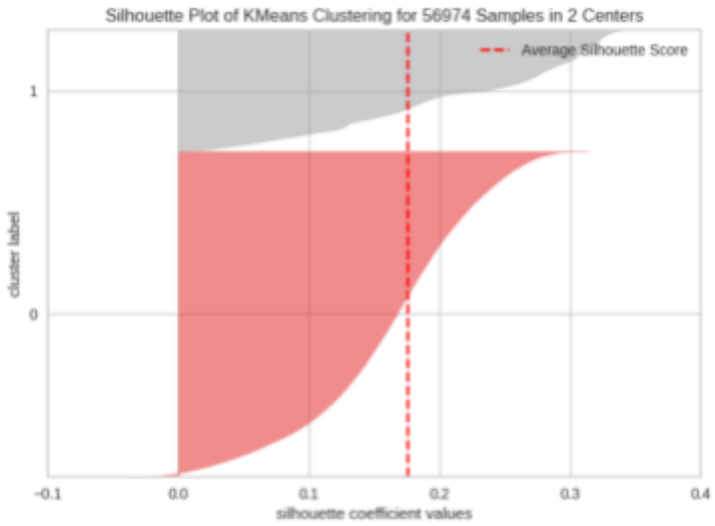


Figure 22: K-means Silhouette Score

Agglomerative:

- Cluster 1 (887 instances): This cluster exhibits a distinct group with relatively low coverage.
- Cluster 2 (56,087 instances): This larger cluster suggests a more widespread category, possibly capturing a diverse set of data points.
- The positive silhouette score indicates a good balance of cohesion within clusters and separation between clusters making this clustering more meaningful.

References:

LaValley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395-2399. https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.106.682658

Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory data analysis using Python. International Journal of Innovative Technology and Exploring Engineering, 8(12), 4727-4735. https://www.researchgate.net/profile/Dr-Subhendu-Pani/publication/337146539_IJITEE/links/5dc70b124585151435fb427e/IJITEE.pdf

Awad, M., Khanna, R., Awad, M., & Khanna, R. (2015). Support vector regression. Efficient learning machines: Theories, concepts, and applications for engineers and system designers, 67-80. https://link.springer.com/chapter/10.1007/978-1-4302-5990-9_