**Visual Analytics Course Work**

**Abstract**

This report examines the economic activity status of students in various regions of England and Wales and its related influencing factors. Correlation analysis and dimensionality reduction PCA reveal the connections behind the influencing factors, using data such as correlation heat maps, geographical distribution maps, and bar charts. The display method, combined with the tableau dashboard, effectively helps users identify areas of high and low student economic activity and presents key data. After studying relevant data analysis theories, the display and evaluation are performed, providing data analysis insights for target users.

**1. Introduction**

By analysing the latest census data for England and Wales, this report examines a variety of factors that may influence student economic behaviour, such as population density, educational attainment and household composition. This research is aimed at policy makers, educational administrators and social science researchers, who need to have an accurate picture of the current state of student economic activity and its changing trends in order to develop effective support measures and services. Of course, as a student doing a postgraduate degree at Bristol, I think most of my fellow students are interested in the state of student economic activity, as well. Currently, the student body is experiencing significant changes as a result of economic volatility, particularly notable in terms of the impact on their economic participation and employment opportunities. The continuing rise in the cost of education and the instability of the job market make it particularly urgent to gain a deeper understanding of the economic status of students and the factors that influence it. Data on student economic activity is available from the Department of Statistics website, and it is obvious that we have a series of charts that provide a visual perspective, but this is not enough.

There are a number of reasons why student economic activity shows different results across regions, such as the population density of the region, which is the most basic and parsimonious value; also for example, the composition of the household, where single-occupancy family units tend to have some correlation with student economic activity; and there are also some migration contexts to the student question, for example, some places with a high level of migration tend to have high levels of student economic activity. active, so some data on the country of origin of immigrants, year of arrival in the UK, etc. are also needed as correlates of impact. In addition, the educational characteristics of the population is also an important factor to consider, higher education level rate and student rate, in general, higher education level rate indicates that the area has more highly educated people, which may be associated with higher level of economic development, better employment opportunities, and higher quality of life; while the level of student rate can indicate the proportion of young population in an area as well as the level of participation in education. Therefore, we cite the above influencing factors to do some data analysis and visualisation of the economic activity of students, so that readers interested in this topic can more deeply explore the scenarios behind the data that they are interested in.

**2. Data preparation and extraction**

The data used for this study are derived from the 2011 and 2021 Censuses for England and Wales. The data was accessed via the official census website and covers socio-economic indicators for all regions across the UK.

The data is summarized below:

*geography code: Region identifier (e.g., "E06000001").*

*Household composition: One person household; measures: Value: Count of one-person households.*

*Population Density: Persons per square kilometre; measures: Value: Population density.*

*higher\_education\_ratio: Ratio of the population with higher education.*

*student\_ratio: Ratio of students in the population.*

*Country of birth: Europe; measures: Value: Count of individuals born in Europe.*

*Country of birth: Middle East and Asia; measures: Value: Count of individuals born in Middle East and Asia.*

*Country of birth: The Americas and the Caribbean; measures: Value: Count of individuals born in the Americas and Caribbean.*

*Economic activity status: Economically active (excluding full-time students): Count of economically active individuals (excluding full-time students).*

*Economic activity status: Economically active and a full-time student: Count of economically active full-time students.*

*Economic activity status: Economically inactive: Count of economically inactive individuals.*

During the data cleaning process, we first identified and processed the missing values in the dataset. The aim was to prioritise filling these missing values with the median of each variable, taking into account the completeness of the data and the needs of the analysis. , but in the actual exploration, due to the relatively large geocoding scale we used (starting with E06), there were no missing values in the 2021 target and test impact data.

In order to analyse student economic activity more effectively, we carried out feature engineering to extract and calculate the proportion of students and the proportion of higher education as key indicators from the raw data. In addition, the data were correlated and analysed according to the needs of the analysis.

1-Population Density: Persons per square kilometre; measures: Value - has a low correlation with most household types and is suitable to be retained in order to study the independent effect of population density on economic activity.

Residence type: Lives in a household; measures: Value and Household composition: Total; measures: Value - These two variables show very high correlations with many other variables, especially those related to other household compositions. The various household type variables (e.g. mono nuclear, lone parent, married or civil partner, etc.) also show high correlations with each other. Therefore the variable Residence type was excluded and only Population Density and Household composition were retained.

图表, 树状图

描述已自动生成

2- In analysing the variables country/region of birth and year of arrival in the UK, the years of arrival from 2014 to 2021 showed very high correlations with each other (0.86 to 1.00), suggesting redundancy and that using all of these variables in a single model could lead to problems of multicollinearity. I therefore merged the data for the arrival years 2014-2021 (which in practice are separate data in the statistical returns). The results of the correlation analyses show a high correlation between country of birth and year of arrival. The total number of permanent residents is highly correlated (0.70 and above) with the breakdown of year of arrival, suggesting that recent arrivals can be closely predicted from the total number of residents. Moderate to high correlation between different birth regions: Regions such as Europe, the Middle East and Asia show moderate correlation (0.58 to 0.59) with the Americas and the Caribbean, suggesting that while there is some overlap, each region also provides unique information. This means that taking country-of-birth migrants from Europe and overlaying the information from analysing year-of-arrival data produces a highly correlated profile, whereas for regions such as the Middle East and Asia versus the Americas and the Caribbean data, the country-of-birth data for Europe is still present and necessary. I have therefore excluded the year of arrival variable and retained only the three variables for country of birth.

图片包含 图表

描述已自动生成

3-. The correlation analysis graph shows that the correlation coefficient between the higher\_education\_ratio and the student\_ratio is 0.26. This indicates that there is a slight positive correlation between the two, but the correlation is not particularly strong.

图表, 树状图

描述已自动生成

In order to better visualise the functionality of the data, the data was subjected to dimensionality reduction. I used two methods for dimensionality reduction analysis respectively. The first option is the linear dimensionality reduction method of PCA, and the results are as follows. Since the dimensionality reduction method of PCA is a dimensionality reduction method implemented using linear combination, we have analysed the loading scores and factor loading matrices for the PCA results.

PCA Loading Matrix:

|  |  |  |
| --- | --- | --- |
| **Variables** | **PC1** | **PC2** |
| Household composition: One person household | 0.376714 | 0.377875 |
| Population Density: Persons per square kilometre | 0.32126 | -0.332549 |
| higher\_education\_ratio | 0.184267 | -0.363528 |
| student\_ratio | 0.304778 | -0.133342 |
| Country of birth: Europe | 0.361683 | 0.418572 |
| Country of birth: Middle East and Asia | 0.409864 | -0.109297 |
| Country of birth: The Americas and the Caribbean | 0.366241 | -0.20403 |
| PCA1 | 0.365362 | 0.412346 |
| PCA2 | 0.248213 | -0.44484 |

It is clear that for principal component 1, the highest loading scores are found for the variable associated with "Country of birth: Middle East and Asia; measures: Value", which suggests that this variable is the most important in explaining the variation in the dataset in the PC1 direction. This suggests that this variable is the most important in the PC1 direction for explaining variation in the dataset. This is unexpected, and although this value does not exceed 41%, it is clearly the most important component. The other important variables are the single family variable, and the values for the other two countries of origin of the immigrants. The single family variable is well understood, as students statistically account for a significant portion of the single family numerical situation, which was the original purpose in examining this variable, while the other two immigrant source countries show a considerable importance, which suggests that there is currently some connection between the economic activity of the students and their immigrant situation.

For the PCA algorithm, the factor loading matrix occupies a very important position to describe the relationship between the dependent original variables and the principal components, which helps to reveal key and hidden information about the data. Factor loadings are usually defined by the correlation coefficients between the original variables and the principal components. They represent the weights of the original variables on each principal component.

**Formular:**

**Where:**

|  |  |  |
| --- | --- | --- |
| **Variables** | **PC1** | **PC2** |
| Household composition: One person household | 0.806902 | 0.566436 |
| Population Density: Persons per square kilometre | 0.688121 | -0.498493 |
| higher\_education\_ratio | 0.394691 | -0.544931 |
| student\_ratio | 0.652818 | -0.19988 |
| Country of birth: Europe | 0.774706 | 0.627442 |
| Country of birth: Middle East and Asia | 0.877906 | -0.163837 |
| Country of birth: The Americas and the Caribbean | 0.784469 | -0.305843 |
| PCA1 | 0.782586 | 0.61811 |
| PCA2 | 0.53166 | -0.666818 |

It is worth noting that the variables associated with "Country of birth: Middle East and Asia; measures: Value" and "Country of birth: Europe; measures: Value" have higher positive loadings on PC1, while "higher\_education\_ratio" has the highest negative loadings on PC2. Value" have high positive loadings on PC1, while "higher\_education\_ratio" has the highest negative loading on PC2, showing the influence of different demographic characteristics on economic activity characteristics.

In addition to this I applied another downscaling method, a non-linear downscaling method, t-SNE downscaling method, which is a non-linear downscaling means to demonstrate the downscaling features, which also achieved good results. Unlike the PCA downscaling approach, this shape varies a lot but can still be analysed and studied using clustering.

图表, 散点图

描述已自动生成

**4. task definition using Munzner's taxonomy**

Task 1: Analysing the factors influencing the economic activity status of students

|  |  |
| --- | --- |
| **task1: Analyze Factors Influencing Students' Economic Activity Status** | **Content** |
| **What (Object)** | Indicators of students' economic activity and influencing factors |
| **Items (Projects)** | Various regions of England and Wales, including comprehensive analysis of major regions |
| **Attributes (Attributes)** | Relationship between students' economic activity and factors such as population density, proportion of single-person households, population from different birthplaces, education level |
| **How (Method)** | Correlation analysis (Pearson correlation coefficient), dimensionality reduction analysis (PCA), heatmaps, scatter plots, bubble charts |
| **Analyze (Analyze)** | Data aggregation, use of stacked bar charts and scatter plots to show the relationship between economic activity status and influencing factors |
| **Explore (Explore)** | Identify regions with high and low student economic activity, use heatmaps to show regional differences |
| **Compare (Compare)** | Use ANOVA and t-tests to compare the economic activity of students with other population groups across different regions |
| **Why (Purpose)** | Reveal regions with high and low economic activity status, analyze possible reasons |
| **Discover (Discover)** | Uncover patterns and characteristics of students' economic activity in different regions |
| **Explain (Explain)** | Explain significant factors affecting students' economic activity and explore their mechanisms and possible policy suggestions |

Task II: Ten years of changes in student economic activity

|  |  |
| --- | --- |
| **Task 2: Ten-Year Changes in Students' Economic Activity** | **Content** |
| **What (Object)** | Students' economic data |
| **Items (Projects)** | Various regions |
| **Attributes (Attributes)** | Relationship between students' economic activity status and factors such as population density, proportion of single-person households, population from different birthplaces, education level |
| **How (Method)** | Statistical methods, time series analysis, regression analysis |
| **Analyze (Analyze)** | Use of time series charts, trend charts, and stacked area charts to display changes |
| **Summarize (Summarize)** | Comprehensive analysis of the correlation between students' economic activity and the aforementioned factors |
| **Why (Purpose)** | Explore changes in time series data to understand long-term trends and driving factors of students' economic activity |

**5. Visual Argumentation**

In order to analyse the state of students' economic activity and the factors influencing it, I chose a variety of visualisation techniques to demonstrate data relationships and patterns.

The first is a correlation heat map analysis, which is used to show a matrix of correlation coefficients between multiple variables, using shades of colour to indicate the strength of the correlation. The correlation between student economic activity status and factors such as population density, the proportion of one-person households, the number of people in different birth regions, and education level can be visualised to help quickly identify significant relationships. A portion of statistical methods are also utilised, dimensionality reduction methods such as Principal Component Analysis, PCA is used to simplify the dataset and highlight the main trends of change. Multi-dimensional data can be simplified into 2D or 3D graphs to help identify the main factors affecting the state of students' economic activity and to reduce data complexity while retaining the amount of information. Based on Munzner's theory [1], the choice of visualisation techniques should be in line with the basic principles of human perception and cognition. Techniques such as correlation heatmaps and PCA 3D visualisation can effectively demonstrate multivariate relationships, in line with human recognition advantages of colour and shape.

In the display of the dashboard, I made full use of the staggered arrangement to ensure the aesthetics of the visualisation without losing the effective information, and cleverly utilised the rule-based and data-driven approaches, which is in line with Zhou's theory of colour mapping techniques [2].

In the process of map visualisation, I used interleaved small and large area displays, inspired by Heer et al.'s design of design space analysis of architectural and interface issues [3], which maximises the total-subdivision design and facilitates a quick understanding of the information while at the same time having an understanding of the two variables.

**6. Evaluation**

For the evaluation metrics system, I used a number of Munzuner-based rating system metrics, Domain Problem Validation, Task Abstraction Validation, Coding and Interactive Techniques Validation and Data and View Validation.

The purpose of domain question validation is to understand the patterns and characteristics of student economic activity and influencing factors in different regions with reference to several key indicators such as population density, proportion of single-person households, proportion of tertiary education, and proportion of students, which help to understand regional differences in the state of student economic activity and can explain correlations and potential influencing factors. Attention needs to be paid to the needs of the target users in this evaluation criterion, with a different focus for student populations, educators and social researchers. The task abstraction validation focuses on whether the intended objectives are met, with a focus on explaining the relationships between influencing factors and helping to visualise them. Attention was given to whether the adopted graphical presentation met the needs and clearly demonstrated the data characteristics. Coding and Interactive Techniques validates that the data has been processed in a way that makes it easier to understand, and focuses on whether the dashboard functionality provides an exploratory experience, and whether the readability and colour contrast meets the needs. Data and view validation looks at whether the presentation is logical and effective, with maps showing the distribution of students by region and key statistics, and bar charts and tables providing detailed comparisons and analysis of values. Data is judged to be handled appropriately, with views presented that accurately reflect key trends and relationships in the data.

In the intra-group mutual assessment, 4/5 students agreed that the domain questions evaluated the indicators as excellent, with one student pointing out that regional variations needed to be compared with proportional data. This view is helpful, but I have thought about this before, and the reason why I did not use proportional data for the presentation was because in terms of metric granularity, due to the overabundance of refined regions, sorting (colour shades) tends to be more effective in providing a viewing intuition, and proportional data is too small in terms of decimal places. Task abstraction validation was rated by 5/5 students as having clear objectives and excellent metric selection. Coding and Interactive Technical Validation had 3/5 students agreeing that the code was technically competent and the dashboards were highly explored, one student pointed out that the correlation analysis was not displayed in the dashboards, this was accepted as a criticism as I really hadn't thought about how to display the correlation analysis in the dashboards, perhaps this point is a different advantage of using python for data analysis and using tableau for data analysis Point. Data and view validation 5/5 students agreed that the views presented accurately reflected the main trends and relationships in the data and were excellent.

**Reference**

[1] Munzner T. Visualization analysis and design[M]. CRC press, 2014.

[2] Zhou L, Hansen C D. A survey of colormaps in visualization[J]. IEEE transactions on visualization and computer graphics, 2015, 22(8): 2051-2069.

[3] Heer J, Mackinlay J, Stolte C, et al. Graphical histories for visualization: Supporting analysis, communication, and evaluation[J]. IEEE transactions on visualization and computer graphics, 2008, 14(6): 1189-1196.