### 1. Data preparation and understanding

#### 1.4 Explanation of code preparation & understanding

**Overview:**

In this work, I am using a dataset from the authors for HESA that is used to show the performance of students on issues such as employment, salary, etc., after they have been educated in different courses at the university. The dataset consists of a large number of entries, and the downloaded zip file contains more than ten csv files, and on its open source website, the dataset has the following official description:

The discover uni webset provides comparable sets of information about full- and part-time undergraduate courses. it is run by the Office for Students and It is run by the Office for Students and is designed to meet the information needs of prospective students.

Due to the extremely large number of csv files contained in this dataset, I was not able to process all the entries in this work. Therefore, after screening, I chose to use employment.csv as the dataset for my work. This file is described in the corresponding readme as follows:

EMPLOYMENT.csv.

Include employment statistics entity, Contains information relating to student employment outcomes.

Contains information relating to student employment outcomes

This file was chosen because the main focus of this dataset is to share information about student employment after graduation, and EMPLOYMENT has a comprehensive and informative description of employment, including a large number of course-specific student employment rates, making it the best choice for exploring employment. Most of the other documents contain more marginal and trivial information, so this document will suffice as a source for exploring student employment.

**Preparation:**

After deciding which document to explore, the first task that needed to be undertaken was: to find out what information and data entries this document contains. Therefore, I went to hesa's official introduction page for this dataset. I was able to gather information about the headers of all the data columns in the employment file, which are listed below:

The table header columns contained in the entire csv file totaled 18 items, which are:

1.UKPRN: UK provider reference number, which is the unique identifier allocated to providers by the UK Register of Learning Providers (UKRLP)

2.PUBUKPRN Publication UK provider reference number for where the course is primarily taught

3.KISCOURSEID An identifier which uniquely identifies a course within a provider

4.KISMODE The mode of the KIS course (full-time, part-time, both) (where 1 is full-time, 2 is part-time, 3 is both)

5.EMPUNAVAILREASON Indicator of the reason why data for a course may not be available

6.EMPPOP Number of students in the population from which the employment data is derived for the course

7.EMPRESPONSE Number of respondents to the survey

8.EMPSAMPLE Number of students in the survey population

9.EMPRESP\_RATE Response rate of students in the population from which the employment data is derived for the course

10.EMPAGG Aggregation level applied to the employment data for the course

11.EMPSBJ CAH Level subject code

12.WORKSTUDY Proportion of students in work and/or study 15 months after their course ended

13.STUDY Proportion of students who are undertaking further study 15 months after their course ended

14.UNEMP Proportion of students unemployed 15 months after their course ended

15.PREVWORKSTUD Proportion of students previously in work or study 15 months after their course ended

16.BOTH Proportion of students in work and study 15 months after their course ended

17.NOAVAIL Proportion of students who are not available for work or study 15 months after their course ended

18.WORK Proportion of students in work 15 months after their course ended

The eighteen header columns can be broadly categorized into three groups: firstly, the various ids used to identify the data information, including the number of the UK educational institution, the course number, the subject code, etc. Secondly, the ids associated with the respondent/student, the course number, and the subject code. The second is information relating to the respondent/course, which includes how many people participated in the survey, whether the course was full-time or part-time, etc. The third is the most important information in this dataset. The third is the most important uh part of the information in this dataset, which is the relevant employment status. For these employment situations, this dataset distinguishes between a number of different scenarios. These include the proportion who chose to participate in work/study after the program, the proportion who were unemployed, and the proportion who had participated in work. One of the more intuitive pieces of this information is the entry UNEMP: Proportion of unemployed students. This data entry will also be used as the main object of study in the subsequent understanding of the data.

After understanding the meaning of each data entry, the next step to be taken is to determine which of these data are meaningful to this work, and if a particular piece of data is deemed to be meaningless to both this task and the application in cw2, then it should be considered for deletion. After analyzing the data, I think that the meaningless data are mainly used to identify the information, such as "UKPRN", "EMPUNAVA", "EMPUNAVA" and "EMPUNAVA".

"EMPUNAVAILREASON", "EMPAGG", "EMPSBJ" and so on. However, after consideration, UKPRN may have a role in cw2, as a data visualization page presented online may need to fully display the information of government agencies behind the data presented. Therefore, I kept UKPRN and PUBUKPRN, and deleted "EMPUNAVAILREASON", "EMPAGG", "EMPSBJ" columns that do not help us to understand the data.

After processing the columns, we should consider the rows. Row-related data cleanup work is mainly to remove null values. If a row contains null data, the row should be deleted. Therefore, I have performed this operation. After completing this operation, the total number of rows in this dataset was reduced from over 40,000 to about 30,000.

At this point, the prepare for this data is complete and I have stored the processed data as a new csv file.

**Understand**

In fact, in the preparation section, I have already demonstrated a partial understanding of this dataset, which consists mainly of me fully understanding what information each data item represents. What I need to do next is to explore the specifics of each data item.

In this in-depth exploration, I accomplished the following three main parts of the work:

--Plot the distribution of data on each valid data item (specifically: student-related situation, employment-related situation) to enable an observation of the distribution of the employment-related data. This helped me to get a basic understanding of the data as a whole.

--Find the highest data in each column and find the kiscourseid that corresponds to that data.The main purpose of this exercise is to allow us to explore which course is actually responsible for those particular data. For example, which is the course with the highest unemployment rate? Which course is the one with the best employment data? This allows us to see which courses are "special."

--Observe whether there is a direct relationship between the unemployment rate data and full-time/part-time programs. In other words, to see which full-time courses or part-time courses are likely to lead to higher unemployment rates. The reason why I chose to carry out this work is that there is a specific reference to information about full-time and part-time in the introductory part of this dataset, and therefore, I believe that this work is of particular significance to the understanding of this data and.

The first part of the work is presented below:

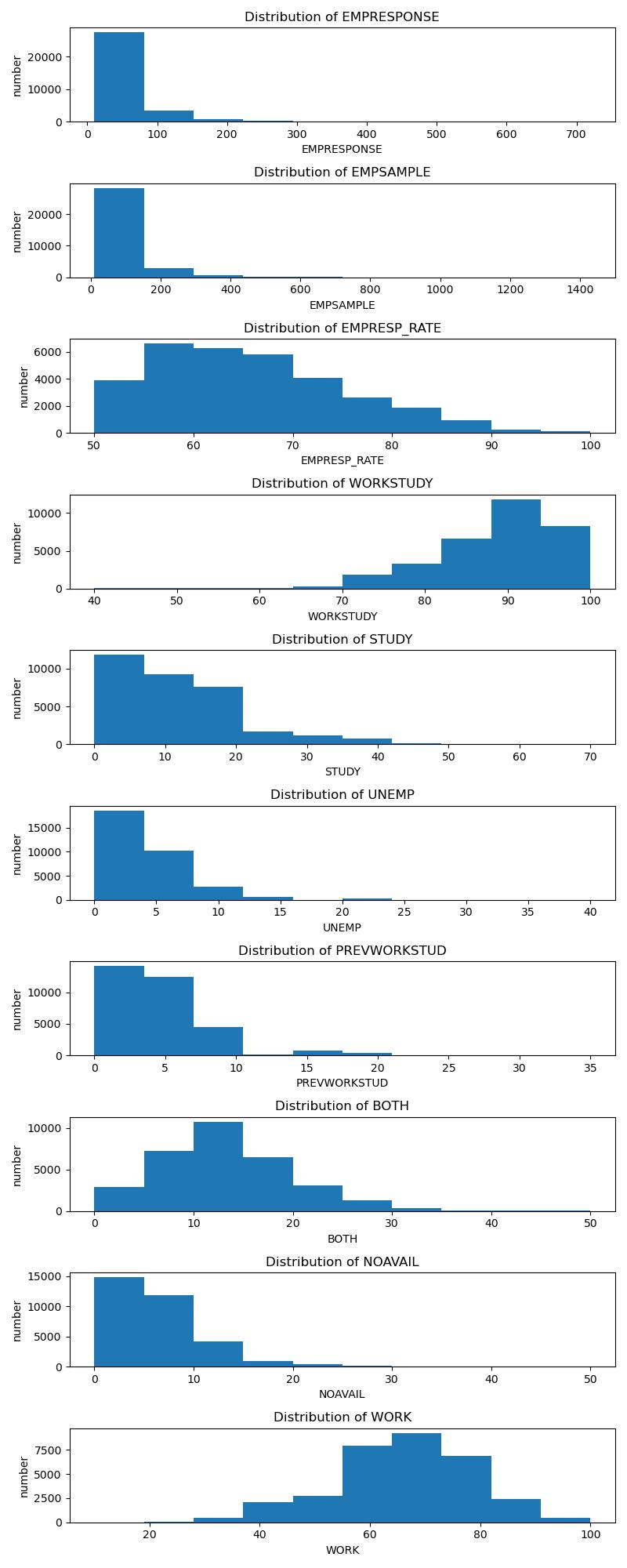


Figure 1

The above figure shows the bar chart of the distribution of all the valid data. The reader can get the distribution of all the data from it. A few of the data are very noteworthy:

As we can see in the distribution of this data of empresponse, most of the courses have no more than one hundred students interviewed.

In the workstudy data we can see that most of the courses still give students the opportunity to have a job or to continue their education after graduation. This is also reflected in the data on unemployed, which is a minority in terms of distribution.

The chart below shows the second part of the work:

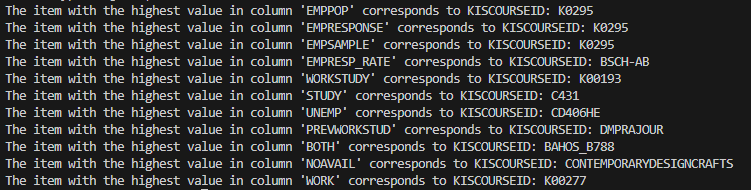


Figure 2

The obvious message to take away from this graph is that if you are a student who has a high expectation of employment after graduation, then please do not take the course with the kiscourseid CD406HE. Instead, take the course K00193 whenever possible.

Here is the third part of the work:

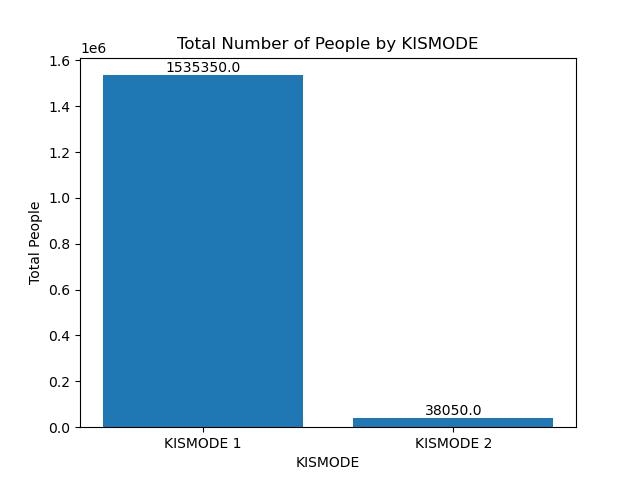


Figure 3

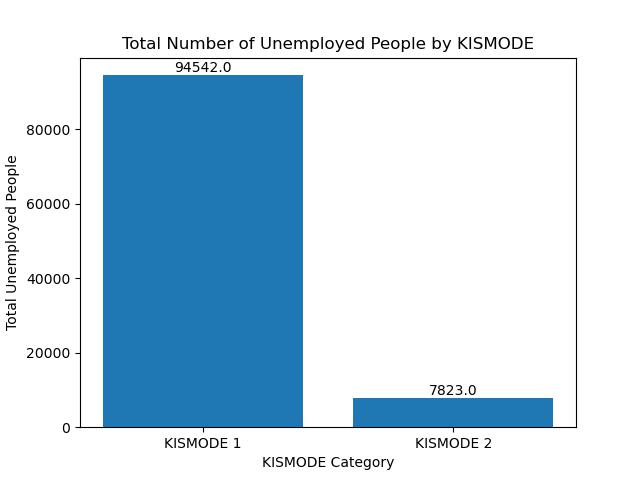


Figure 4

The two charts above show the respective total numbers of students enrolled in full-time and part-time courses, as well as the number of students enrolled in full-time and part-time courses who were unemployed after graduation. As can be seen from the charts, the total number of students enrolled in full-time courses is 1535 350. The number of students enrolled in part-time courses is 38,050, which is one-fortieth of the number of students enrolled in full-time courses. As for the number of unemployed, only 94,542 full-time students were unemployed, less than one-tenth of the total number of students enrolled, while 7,823 students enrolled in part-time courses were unemployed. This is more than one-fifth of the total number of students.

This also means that students enrolled in full-time programs have better employment rates, while part-time education has room for improvement in terms of securing employment for students. This information is of great interest to students who are choosing their course of study.

At this point, all the work for the data preparation and comprehension section is complete.

### 2. Product and project definition

#### 2.1 Problem statement

#### 2.2 Product overview

1.The first app I will be developing in coursework2 is a rest api.Rest api is a design guide for an api that allows developers to access a service in a simple way. And this rest api is a specific api for accessing the dataset which is pre-processed by me and authored by HESA under the name Discover Uni dataset (formally Unistats). developers can access the employment.csv sub dataset of this dataset in a simple way.

2. the second app that i will be developing in coursework2 is a data visualization app. this app is a web app and the user can access this app by visiting the online link. the function of this app is: to take the pre-processed dataset that i have made, authored by HESA, with the name of Discover Uni dataset (formally Unistats) that I have preprocessed and authored by HESA under the name Discover Uni (formally Unistats) for visualization. This app allows users to visualize more about the employment rate after graduation of users who have received various undergraduate courses, allowing students to better choose the direction of their courses.

#### 2.3 Persona

Role: Students who are about to decide their direction of further education

Name: Tom

Background and Needs:

Tom is a high school student from an ordinary family in the UK. Neither of his parents had a university education and Tom's own grades were average, so his family was not able to provide him with a good plan for his life, and he did not have a clear idea of what kind of course he could take in the future, whether he should go to school full-time or part-time, and which course would be better.

His needs include:

Tom would like to know what courses would give him a better chance of getting a job.

Tom wants to know the pros and cons of each potential course option in a clearer way.

How this app meets the user's needs:

This app preprocesses the Discover Uni dataset and visualizes that dataset. It will allow tom to have sufficient knowledge about how good or bad each course is in terms of employment situation.

This app visualizes the data and can allow tom to draw conclusions in a clearer and more intuitive way, without the need to search through huge amounts of data.

#### 2.4 Project goal & objectives / Questions

### 3. Tools & techniques

#### 3.1 Source code control

Here is the github repository I use for code control:

https://github.com/Harry460731/comp0035CW1

#### 3.2 Linting

#### 3.3 Project planning and tracking

#### 3.4 Use of AI

I did not use any AI related tools in this work of mine.

### 4. Methodology

#### 4.1 Methodology selection

### 5. References