# TM351 Data management and analysis

## Spring 2020 TMA01 & Project TMA Template (Solution document)

|  |  |
| --- | --- |
| **Please complete the information below** | |
| **Branch:** |  |
| **Student Name:** |  |
| **Student ID:** |  |
| **Group number:** |  |
| **Instructor:** |  |

**Question 1 (30 marks)**

Complete and submit All the specified Jupyter notebooks in the form of a "solved" .rar or .zip file:

**Question 2 (10 marks)**

Place all you coding work for this question in the lab notebook template and your 300-word summary here.

For creating our report, we are going to import all Data Frames from our dataset.

**Summary of understandings of the Higher Education Staff Statistics UK, 2018/19 datasets:**

* + **The contents of the data set:**

Data has been collected HESA, the Higher Education Statistics Agency, are the experts in UK higher education data, and the designated data body for England.

The HESA Staff record provides data in respect of the characteristics of staff employed under a contract of employment at a reporting higher education provider (HEP) in the UK.

**The data set contained are:**

1. **Figure 4 - All staff (excluding atypical) by academic contract marker, mode of employment and hourly paid marker 2018/19:**
   * Description: Details about type of contract, mode of employment, terms of employment, zero-hour contract, sex, payment type and number of staff members in each category
2. **Figure 5 - Staff by mode of employment, academic contract marker and sex 2018/19:**
   * Description: Statistics about category of employment in each gender and the percentage of this category.
3. **Figure 6 - All staff (excluding atypical) by equality characteristics 2018/19:**
   * Description: Characteristics about each HE country provider and each contract type with the activities of each category and the number of staff in each category.

**Each data set contain a number of defined attributes as following:**

1. **Figure 4 - All staff:**
   1. Academic contact marker (Academic contract - Non-academic contract)
   2. Mode of employment (Part Time – All – Full time)
   3. Terms of employment (Fixed term contract – Open ended/permanent – All)
   4. Zero hours contact (Not on a zero hours contract - On a zero hours contract – All)
   5. Sex (Female – Male – Other – All)
   6. Hourly paid marker (Salaried - Hourly paid – Total – Other)
   7. Number (Numbers)
2. **Figure 5 - Staff by mode of employment, academic contract marker and sex:**
   1. Category marker (Academic contract marker - Mode of employment)
   2. Category (Full time – Part time – Academic – Non academic)
   3. Sex (Female – Male – Other)
   4. Number (numbers)
   5. Percentage (Percentages)
3. **Figure 6 - All staff (excluding atypical) by equality characteristics 2018/19:**
   1. Category marker (Age Group - Disability status – Ethnicity – Nationality – Sex – Total)
   2. Category (Different Categories)
   3. Country of HE provider (England – Scotland- All – Wales - Northern Ireland)
   4. Academic contact marker (Academic contract - Non-academic contract)
   5. Activity standard occupational classification ( Categories of occupation)
   6. Number (Numbers)
   * **Assessment the Quality of each data set:**
4. **Figure 4 - All staff:**

|  |  |  |
| --- | --- | --- |
| **Type of quality** | **Description** | **Assessment** |
| Validity | Validity refers to how accurately a method measures what it is intended to measure | In Academic contract marker attribute, there are 3 main categories (Academic contract – Non academic contact – Total)  Total category is not valid category type |
| Amount of dirtiness | There is invalid data in Academic contract marker attribute.  Evidence (Python code in the lab notebook template) | |
| Accuracy | Accuracy is determined by how close a measurement comes to an existing value that has been measured | The sex attribute contains 2 data types which are not accurate (ALL – Other) |
| Amount of dirtiness | Unknow sex category  Evidence (Python code in the lab notebook template) | |
| Completeness | Completeness is defined as expected comprehensiveness. Data can be complete even if optional data is missing. As long as the data meets the expectations then the data is considered complete. | There are no incomplete data in the data set |
| Amount of dirtiness | All Data set are complete | |
| Consistency | consistency is typically a measure based on the correlations between different items on the same test (or the same subscale on a larger test). It measures whether several items that propose to measure the same general construct produce similar scores. | There are no inconsistent data in the data set |
| Amount of dirtiness | All data is consistent | |
| Uniformity | Uniformity implies consistency in lack of variation between the items being compared, over a long period and across a wide range. See also comparability analysis. | All Data set and its variables are uniform it’s all depending on text basis |
| Amount of dirtiness | All data are uniform | |

1. **Figure 5 - Staff by mode of employment, academic contract marker and sex:**

|  |  |  |
| --- | --- | --- |
| **Type of quality** | **Description** | **Assessment** |
| Validity | Validity refers to how accurately a method measures what it is intended to measure | All of the data variables are valid and reliable |
| Amount of dirtiness | No invalid data is found all data are classified with specific values. | |
| Accuracy | Accuracy is determined by how close a measurement comes to an existing value that has been measured | The sex attribute contains a data types which are not accurate (Other) |
| Amount of dirtiness | Unknow sex category  Evidence (Python code in the lab notebook template) | |
| Completeness | Completeness is defined as expected comprehensiveness. Data can be complete even if optional data is missing. As long as the data meets the expectations then the data is considered complete. | There are no incomplete data in the data set |
| Amount of dirtiness | All Data set are complete | |
| Consistency | consistency is typically a measure based on the correlations between different items on the same test (or the same subscale on a larger test). It measures whether several items that propose to measure the same general construct produce similar scores. | There are no inconsistent data in the data set |
| Amount of dirtiness | All data is consistent | |
| Uniformity | Uniformity implies consistency in lack of variation between the items being compared, over a long period and across a wide range. See also comparability analysis. | All Data set and its variables are uniform it’s all depending on text basis |
| Amount of dirtiness | All data are uniform | |

1. **Figure 6 - All staff (excluding atypical) by equality characteristics 2018/19:**

|  |  |  |
| --- | --- | --- |
| **Type of quality** | **Description** | **Assessment** |
| Validity | Validity refers to how accurately a method measures what it is intended to measure | All of the data variables are valid and reliable |
| Amount of dirtiness | No invalid data is found all data are classified with specific values. | |
| Accuracy | Accuracy is determined by how close a measurement comes to an existing value that has been measured | The Category Marker attribute contains data types which are unknown (Total)  The Category attribute contains data types which are unknown (Total- Other - Black) |
| Amount of dirtiness | Unknow category marker and category (Total)  Evidence (Python code in the lab notebook template) | |
| Completeness | Completeness is defined as expected comprehensiveness. Data can be complete even if optional data is missing. As long as the data meets the expectations then the data is considered complete. | There are no incomplete data in the data set |
| Amount of dirtiness | All Data set are complete | |
| Consistency | consistency is typically a measure based on the correlations between different items on the same test (or the same subscale on a larger test). It measures whether several items that propose to measure the same general construct produce similar scores. | There are no inconsistent data in the data set |
| Amount of dirtiness | All data is consistent | |
| Uniformity | Uniformity implies consistency in lack of variation between the items being compared, over a long period and across a wide range. See also comparability analysis. | All Data set and its variables are uniform it’s all depending on text basis |
| Amount of dirtiness | All data are uniform | |

**Question 3 - Project (60 marks)**

Place all you coding work for this question in the lab notebook template and your complete project report including copies of all visualizations here.

***Project Summary***

Our project will discuss the 3 data sets Fig4, Fig5 and Fig6. Our investigation will discuss the relationships between a selected independent variable and a selected dependent variable.

We will use a proper measure of correlation to show the relationships between selected variables using appropriate visualizations with Pandas data frame.

We will produce our conclusion according to 10 different question on the 3 data sets.

***Aims and objectives:***

* + 1. Aims:
       1. Showing different correlation between columns.
       2. Visualizing the relation between variable.
       3. Specifying criteria based on Academic contract marker, Terms of employment, Zero hours contract, Sex, Category and Category Marker.
       4. Showing some statistics in each data sets.
    2. Objectives:
       1. Using ANACONDA frame work to import CSV files into PANDAs data frame.
       2. Importing required libraries into ANACONDA.
       3. Using different visualization models to show the correlations between variables in specific conditions or criteria.
       4. Using SQL language and Pandas data frame

***Data Source***

1. **Source:**

The main data source is the HESA, the Higher Education Statistics Agency, are the experts in UK higher education data, and the designated data body for England, they work with HE providers in each of the four nations, collaborating with them to collect and curate one of the world’s leading HE data sources. Using this knowledge and expertise we provide a rich, open source of HE information for data users.[[1]](#footnote-1)

1. **Contents:**

Inclusion of an individual in the Staff record depends upon the existence of one or more contracts of employment between the HEP and the individual and/or the liability of the HEP to pay Class 1 National Insurance contributions for that individual.

An individual's employment with a higher education provider (HEP) will be governed by one or more legally binding contracts. The Staff record collects attributes of the HE provider, person, contract of employment, and activities within each contract. Each person will have one or more contracts; each contract will have one or more activities. A member of staff may be employed under a single contract of employment or a number of separate contracts: this will depend on the HEP's employment practices.

The range of data required about an individual and the contract(s) that they hold will depend on the nature of those contracts and also the classification of the activity for which the contract exists[[2]](#footnote-2).

1. **Quality of Data:**

We discussed in Question 3 the data quality according to its validity, accuracy, completeness, consistency and uniformity.

* + - * 1. The data set in Fig 4 were updated on 23 January 2020
        2. The data set in Fig 5 were updated on 23 January 2020
        3. The data set in Fig 6 were updated on 23 January 2020
        4. Validity of data:

The data set in Fig 4 variables are not all valid

The data set in Fig 5 variables are all valid.

The data set in Fig 6 variables are all valid.

* + - * 1. Accuracy:

The data set in Fig 4 variables are not all accurate.

The data set in Fig 5 variables are not all accurate.

The data set in Fig 6 variables are not all accurate.

* + - * 1. Completeness:

The data set in Fig 4 variables are all complete no missing data.

The data set in Fig 5 variables are all complete no missing data

The data set in Fig 6 variables are all complete no missing data.

* + - * 1. Consistence:

The data set in Fig 4 variables are all consistent.

The data set in Fig 5 variables are all consistent.

The data set in Fig 6 variables are all consistent.

* + - * 1. Uniformity:

The data set in Fig 4 variables are uniform.

The data set in Fig 5 variables are uniform.

The data set in Fig 6 variables are uniform.

**4- Variable Classification I:**

**Each column has a specific value defined in the following table for variables.**

|  |  |  |
| --- | --- | --- |
| Figure 4 Variables | Dependencies | Description |
| Academic contract marker | Independent |  |
| Mode of employment | Dependent | This attribute depends on the (Academic contract marker) |
| Terms of employment | Independent |  |
| Zero hours contract | Independent |  |
| Sex | Independent |  |
| Hourly paid marker | Independent |  |
| Number | Independent |  |

|  |  |  |
| --- | --- | --- |
| Figure 5 Variables | Dependencies | Description |
| Category Marker | Independent |  |
| Category | Dependent | This attribute depends on the (Category marker) |
| Sex | Independent |  |
| Number | Independent |  |
| Percentage | Independent |  |

|  |  |  |
| --- | --- | --- |
| Figure 6 Variables | Dependencies | Description |
| Category Marker | Independent |  |
| Category | Dependent | This attribute depends on the (Category marker) |
| Country of HE Provider | Independent |  |
| Academic contract marker | Independent |  |
| Activity standard occupational classification | Independent |  |
| Number | Independent |  |

***Research questions***

The following variables will be under investigation according to specific questions.

**Fig 4 Variable under investigation:**

|  |
| --- |
| Figure 4 Variables |
| Academic contract marker |
| Mode of employment |
| Terms of employment |
| Zero hours contract |
| Sex |

**Fig 5 Variable under investigation:**

|  |
| --- |
| Figure 5 Variables |
| Category |
| Sex |
| Number |

**Fig 6 Variable under investigation:**

|  |
| --- |
| Figure 6 Variables |
| Category Marker |
| Category |
| Country of HE Provider |
| Academic contract marker |

**Questions to investigate the relationship between variables:**

1. How many Females has fixed terms contract? (Fig 4 Data Set)
2. How many male or female in each Terms of employment? (Fig 4 Data Set)
3. Count the number of staff members in each Terms of employment. (Fig 4 Data Set)
4. Maximum salaried staff in each mode of employment. (Fig 4 Data Set)
5. Count the number of staff in Full-time and Part-time mode of employment. (Fig 4 Data Set)
6. Show the hourly paid salary for males in full time contract? (Fig 4 Data Set)
7. How many males and females in each category? (Fig 5 Data Set)
8. How many staff in each country with age category 25 or under? (Fig 6 Data Set)
9. Show the different categories of the Disability status in Northern Ireland. (Fig 6 Data Set)
10. How many academic and non-academic contract in each country. (Fig 6 Data Set)

***Analysis and findings***

Questions analysis:

We used SQL language with panda’s data frame to visualize the statistical measures and matplotlib to show different diagrams for each question.

|  |  |
| --- | --- |
| Question Number | Final Answer  (Coding work and all question are in the lab notebook) |
| 1 | We used the aggregated function to count the number of females in the fixed term contract and we find 689955 females using this type of contact |
| 2 | We grouped each term of contracted with each gender to get a statistical information about the numbers, we find that most of females has Open-ended/permanent contract and most males has also the same type of contact |
| 3 | We got the total staff members in each type of contact the most type was (ALL) and the least contract type was (Fixed term contact) |
| 4 | The Maximum salaried staff in each mode of employment was:   * + 1. ALL: 403120     2. Full Time: 152175     3. Part Time: 63240 |
| 5 | The number of staff in each mode of employment was:   1. Full Time: 4738975 2. Part Time: 2300245 |
| 6 | Showing the hourly paid salary for males in full time contract by relating the (Hourly paid marker) attribute with an aggregated function sum to get the number of staff members and the (Mode of employment) attribute. The result was: 606420 |
| 7 | The number of males and females in each category of employment:   1. Females: 480430 2. Males: 399160 |
| 8 | The number of staff in each country with age category 25 or under:   1. England: 42885 2. Northern Ireland: 440 3. Scotland: 7830 4. Wales: 1985 |
| 9 | Showing different categories of the Disability status in Northern Ireland |
| 10 | The number of academic and non-academic contract in each country was :   1. England: 330 2. Northern Ireland: 315 3. Scotland: 330 4. Wales: 330 |

***Project description***

|  |  |
| --- | --- |
| Project planning | 1. Splitting each question among the group members and each question into different parts. 2. Reading each question requirements and gathering the desired data. 3. Installing the required applications to complete the task specially for question 2 and 4. 4. We made a schedule to complete each task in a specific time frame. 5. Combining different answers from the group members into 2 templets one for the coding and the other for the documentation of the project. |
| How data is acquired | Data is downloaded from <https://data.gov.uk/dataset/452fa2dd-72e2-4de3-9e91-25be38dec27d/higher-education-staff-statistics-uk-2018-19>  3 different data sets (Fig4 – fig5 – fig6) |
| Data Preparation | This take a long time from us to prepare the data sets according to our needs.  We used Excel to filter and prepare the final file which is used in our analysis.  Also, we used the virtual machine of the AOU for TM351 to avoid any missing libraries in the anaconda. |
| Data Analysis | Phases of analysis:   1. Identifying variable in each data set. 2. Finding relationship between different data sets 3. Filtering the data to specific date because there are huge number of records. 4. Merging 2 selected dataset into one data set using primary key. 5. Answering different questions using python code to get the desired outcomes. |
| Reporting final findings | هنا على كل طالب كتابه نبذه عن المكتسبات في العمل على تحليل البيانات  النبذه تكون عامه  مثال: we have learned a lot about python coding  وهكذا اي شي تم اكتسابه من مهاره تكتب. |

***Project Reflection***

|  |  |
| --- | --- |
| Project Experience | 1. We have gone through different phases and each phase has its own requirements. 2. Good experience in working as a group 3. We got a good experience in analyzing data sets. 4. Looking for relationships between different variables 5. Filtration of data to minimize the number of records. 6. Combining different data sets according to a specific Primary key. 7. Working in groups and splitting the work among us. 8. Scheduling tasks and committing to the schedule. |
| Learning Outcomes | 1. Python coding 2. Using Anaconda framework 3. Using VM TM351 software. 4. Dealing with Mongo DB 5. Using with Open Refine. |
| What went Well | We finished all tasks |
| What went wrong | Some codes are not running as desired |
| Future benefits | Learning a new programing language. |

References:

Jackson, Joab (Feb 5, 2013). ["Python gets a big data boost from DARPA"](http://www.networkworld.com/article/2163350/application-performance-management/python-gets-a-big-data-boost-from-darpa.html). networkworld. Retrieved October 30, 2014.

Kenny Gorman. ["MongoDB 3.0 WiredTiger Compression and Performance"](http://objectrocket.com/blog/company/mongodb-wiredtiger). Objectrocket.com/. [Archived](https://web.archive.org/web/20170616134542/http:/objectrocket.com/blog/company/mongodb-wiredtiger) from the original on June 16, 2017. Retrieved June 28, 2017.

scalegrid.io. ["Atomicity, isolation & concurrency in MongoDB"](https://scalegrid.io/blog/atomicity-isolation-concurrency-in-mongodb/). scalegrid.io. [Archived](https://web.archive.org/web/20170910125308/https:/scalegrid.io/blog/atomicity-isolation-concurrency-in-mongodb/) from the original on September 10, 2017. Retrieved July 5, 2017.

[*"postgresql-client-10.5p1 – PostgreSQL RDBMS (client)"*](http://ports.su/databases/postgresql,-main). [*OpenBSD ports*](https://en.wikipedia.org/wiki/OpenBSD_ports). October 4, 2018*. Retrieved October 10, 2018*.

 Pihlak, Martin. [*"PostgreSQL @Skype"*](https://wiki.postgresql.org/images/a/a9/Postgresql-at-skype.pdf) *(PDF)*. wiki.postgresql.org*. Retrieved January 16, 2019*.

<https://data.gov.uk/dataset/452fa2dd-72e2-4de3-9e91-25be38dec27d/higher-education-staff-statistics-uk-2018-19>

1. <https://www.hesa.ac.uk/about> [↑](#footnote-ref-1)
2. <https://www.hesa.ac.uk/support/definitions/staff> [↑](#footnote-ref-2)