Table of Contents

[**Question 2** 2](#_Toc36832853)

[**Summary which will fulfill:** 2](#_Toc36832854)

[**1.** **Contents of Datasets:** 2](#_Toc36832855)

[**2.** **Data Dirtiness:** 2](#_Toc36832856)

[**3.** **Data Quality** 3](#_Toc36832857)

# 

# 

# **Question 2**

## **Summary which will fulfill:**

### **Contents of Datasets:**

These datasets are taken from official website of data.gov.uk. it is very first release of data from 2018\_2019 HESA record. Well the data contains all the details of all the staff employment at us higher education. We know the importance of data in today’s life. So same here we are going to use previous data to get advantage in near future.

Basically there are three different datasets we were assigned to pick and use for analysis.

1. **All Staff by equality characteristics:**

This data set has information regarding Age Group, Age Category, Country of Higher Education Provider, and Standard Occupation and having information of staff those who are on academic contract.

1. **Staff by mode of employment:**

The data contains important information of Category of employment whether it is full time or part time, gender information and percentage.

1. **All Staff by academic :**

This data has information of mode of employment whether part time/ full time , defines terms of employment, fixed term or permanent, identifies zero hour contract , gender identification and hourly paid marker.

### **Data Dirtiness:**

As we know that data should be clean in data analysis. As garbage data gives as input it will response back the garbage data as output. So we must have to determine dirty data and remove as well.

Well most of the data in all data sets were in non-numeric form but don’t find any missing data, and duplicate data in all the three data sets however 35 outliers were found in all staff equality Data set. We wrote code of it in Jupiter notebook as well.

Secondly the feature ‘Sex’ contains attributes like ‘All’ which can’t be a gender. Similarly we found that ‘total’ attribute in category column which can’t justify any logic and ‘mode of employment’ can’t be ‘All’ in sense. Similarly we have some other non uniform data which are attributes in columns such as ‘All’ And ‘Other’. We will remove all these before analysis in third question. So these are the dirty data in our datasets. We wrote down code in Jupiter notebook to show them.

### **Data Quality**

|  |  |
| --- | --- |
| Term | Tracing and check |
| Validity | Data is perfectly valid. |
| Accuracy | There are some outliers for upper and lower bounds in Number columns of both data sets All staff equality and All staff academic respectively.  We use python code in notebook file to identify these outliers. |
| Completeness | There were no Null or missing values so data set is complete. |
| Consistency | Data is perfectly consistence. |
| Uniformity | There are some unformal data such as ‘All’ attribute in ‘Sex’ column  Similarly in ‘category’ column we have ‘total’ attribute and also ‘mode of employment’ can’t be ‘All’. That’s why these were not uniform data. We determined with help of python code in Jupiter notebook. |

This is the first release of data from the 2018/19 HESA Staff record. It provides details of staff employment at UK higher education (HE) providers on 1 December 2018. It has been produced in partnership with the UK administrations and is released in accordance with the arrangements approved by the UK Statistics Authority.