Contents

**[Q2-Aims and objectives:](#_Toc36834312)** [2](#_Toc36834312)

[**Brief description about the general aims of project:** 2](#_Toc36834313)

[**Objectives of above Aim:** 2](#_Toc36834314)

[**Q3-The Source Data:** 2](#_Toc36834315)

[**Source of data:** 2](#_Toc36834316)

[**Content of data:** 3](#_Toc36834317)

[**Data Quality:** 3](#_Toc36834318)

[**Data Quality Table:** 3](#_Toc36834319)

[**Variable classification I:** 3](#_Toc36834320)

[**Q4-Research Question** 4](#_Toc36834321)

[**Question:** 4](#_Toc36834322)

[**Independent Variables:** 4](#_Toc36834323)

[**Dependent variables:** 5](#_Toc36834324)

[**Q5-Analysis and Findings:** 5](#_Toc36834325)

[**1.** **Types of measure in the variables:** 5](#_Toc36834326)

[**2.** **Critical interpretation and conclusions about those observed correlations:** 5](#_Toc36834327)

[**3.** **Tabular summaries of the data in the form of crosstabs or pivot tables:** 5](#_Toc36834328)

[**4.** **Critical interpretation of above table:** 6](#_Toc36834329)

[**5.** **Relevant Visualization:** 6](#_Toc36834330)

[**6.** **Final answer to the research question:** 8](#_Toc36834331)

[**Analyzing Another Data Set:** 10](#_Toc36834332)

[**7.** **Critical comment on your conclusion** 13](#_Toc36834333)

[**Q6-Project Description** 13](#_Toc36834334)

[**Q7–Reflection:** 14](#_Toc36834335)

[**Q8- References:** 15](#_Toc36834336)

[Cleaning Dirty Data with Pandas & Python 15](#_Toc36834337)

**Executive summary**

In this project we have been assigned three data sets of HESA, the Higher Education Statistics Agency, the experts in UK higher education data, and the designated data body for England. HESA published these details of staff employment at UK higher education (HE) providers on 1st December 2018.

The data sets contains all the details of staff related to staff country, about their service contract, details of numbers of females or males working on a specific zero-hour contract. It also contains the details for their job description whether they are permanent or not whether the staff is working as part time or full time and information about their responsibilities. We explored only one data set for our question and get results how job of staff members depends upon gender of people. We analyzed and discussed the data set in detail.

# **Q2-Aims and objectives:**

## **Brief description about the general aims of project:**

* Learn by practicing how to use data sets, feature engineering (features or columns are very important for our question) and to analyze data set.
* Learn how to use Jupiter notebook for analyzing and working with csv files in analysis by using pandas and NumPy libraries.
* Learn how to use plotting libraries such as matplotlib and seaborn to justify our answer.

## **Objectives of above Aim:**

* Formulate a research question and investigate it and write a report of what I have found.
* To investigate the relationships between a selected independent variable and a selected dependent variable.
* Find the relationship between some variables.
* Use of a proper measure of correlation to show the relationships between our selected variables and appropriate visualizations using different libraries.

# **Q3-The Source Data:**

## **Source of data:**

All the data sets were taken from official website of <https://data.gov.uk/>.

## **Content of data:**

The data sets contains all the detailed information related to higher education staff of United Kingdom according to their people’s country, employer type, their activities and gender.

## **Data Quality:**

After getting data first we clean it by using python script (code) in Jupiter notebook.

First of all we wrote code to checked up for missing and null values, after it we wrote code to check duplicated data and hence we could not found any duplicated and missing data (values) in all three data sets. After it we use python code and identified some outliers in two data sets. 35 outliers were found in ‘Number’ feature of all staff equality data set and 25 outliers were found in ‘Number’ feature of all staff academic dataset. We have wrote code of it in Jupiter notebook as well.

After it we found some unformal data such as Sex column should not have any values ‘All’. Same as we found ‘All’ in zero hour contract and Academic contract marker and category columns should not require total attributes. Same as we don’t require ‘All’ attribute in Mode of employment column and Terms of employment don’t require ‘All’ attributes.

All the python code has been placed in Jupiter notebook to show these dirty data. The dirty data was important to remove otherwise it can compromise our model.

## **Data Quality Table:**

|  |  |
| --- | --- |
| 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. |

## **Variable classification I:**

On the behalf of viewing datasets we are showing all three datasets where these are dependent or independent according to the nature of their attributes.

|  |  |
| --- | --- |
| All Staff Academic | Dependently |
| Academic contract marker | Independent |
| Terms of employment | Dependent |
| Zero hours contract | Dependent |
| Mode of employment | Independent |
| Sex | Independent |
| Hourly paid marker | Dependent |
| Number | Dependent |

|  |  |
| --- | --- |
| All Staff Equality | Dependently |
| Category Marker | Independent |
| Category | Independent |
| Country of HE Provider | Dependent |
| Academic contract marker | Dependent |
| Activity standard occupational classification | Dependent |
| Number | Dependent |

|  |  |
| --- | --- |
| Staff By Mode | Dependently |
| Category Marker | Independent |
| Category | Independent |
| Sex | Independent |
| Percentage | Dependent |
| Number | Dependent |

# **Q4-Research Question**

## **Question:**

Visualize the “Fixed-term contract female employees working on a [on a zero hour’s] contract”.

## **Independent Variables:**

In this question independent variables are those by whom we found dependent variables which are:

* On a Zero-hour contact,
* Not on a zero hour contact
* Open-ended/permanent
* Fixed-term contract

## **Dependent variables:**

We want to investigate dependent variables:

* Female,
* Other
* Male

# **Q5-Analysis and Findings:**

1. **Types of measure in the variables:**  To find correlation between two variable we need to have at least one variable based on scale. But in our data set we have nominal variables so we cannot find correlation. But we can do frequency calculation as in histograms created on Jupiter notebook and down in part 4 show a graph. From the figure we can say those females who are on a zero hour and on fix term is equal or same in number with females with not on a zero hour and permanent because both graphs shows equal number of females.
2. **Critical interpretation and conclusions about those observed correlations:** The thing we have observed so far has brought us conclusion that our output variable Sex is equal distributed among all other independent variables we have, and no variation exists. Moreover more than 50% of our data values in Sex column are all or other which hinder our analysis and to find any difference.

## **Tabular summaries of the data in the form of crosstabs or pivot tables:**

As already we stated that our data is nominal, we can’t find mean etc. using describe function. But use the method to describe all details whose code is written in jupyter notebook. The picture given down displays details regarding only three columns of data set ‘All\_staff\_Academic’. When this code was written on that time our data was full cleaned and we don’t have any ‘All’ or ‘other’ attributes which made our data so messy.

A screenshot of a cell phone

Description automatically generated

In Jupiter we also wrote this code to describe all staff Academic data set at initial stage without cleansing but it was so big so we can’t make screenshot to paste here and that code contains a lot of all attributes which could change results that’s why we drop that like rows.

## **Critical interpretation of above table:**

Well as stated above our data set is so nominal. Secondly our data set was so consistent. One can see that it’s not possible to find difference because the number of dependent female and male are equal. Same as independent variables zero-hour contact and not on a zero-hour contact are same in size also open-ended/permanent is equal in total number of fixed terms. So it was impossible to show any pattern of difference or correlation.

## **Relevant Visualization:**

Well Using on a zero-hour Contract and Fixed-term contract as independed to find relation with dependent Sex column whose attributes are male, female and others. We display code for all three possibilities to find any difference but only once there is difference of other attributes with perspective to female and male which is shown down by graph in detail. However, in all three cases we found same number of female and male. Description is down in detail with graphs.

1. **Using on a Zero hour And Fixed-term contract As Independed to find relation with dependent Sex Column**:

A screenshot of a cell phone

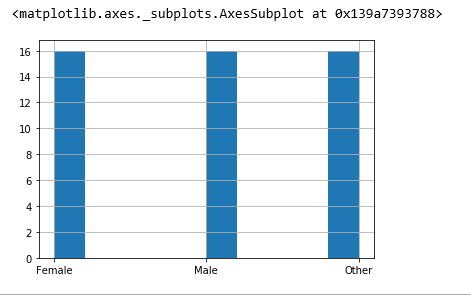
Description automatically generated

**Conclusion:**

In this graph one can see that total number of male and female are equal in numbers which are 16 when there are two independent variables name as zero-hour contact and fixed term which was our project problem to solve.

However, there is so much consistency in data of female and male we can’t find any supposition that independent variables impact highly for males or females. One can say that it is because of nominal data.

1. **Using Non-Zero hour And Fixed-term contract As independed to find relation with dependent Sex Colum**:



**Conclusion:**

In this graph again one can see that total number of male and female are equal in numbers which are 16 when there are two independent variables name as Non zero hour contact and fixed term contact.

So again, by passing two other independent variables we got same consistency between male and female. we could find any result that independent variables depend more on female variable or male variable which are dependent as we mentioned in our project problem. that is because it is nominal data.

1. **Using Non-Zero-hour contract And Open-ended/permanent contract as independed to find relation with dependent Sex Column:**

**A picture containing drawing

Description automatically generated**

**Conclusion:**

In this third graph one can see that total number of male and female are equal in numbers which are 16 when there are two independent variables name as non-zero-hour contact and open-ended/permanent contact.

By third graph we want to show how independent depend on dependent variables but could find any advantage because total number of male and females are same as all the independent variables are same in number as well.

## **Final answer to the research question:**

As in project problem we said to find females of on a zero-hour contact with fixed term contract. We applied techniques to find out independent variable depend on females or male but there were major issues like first data was nominal secondly all the data where it is dependent(male, female) or Independent(on a zero hour, not on a zero hour or fixed term, open-ended) all are same in numbers with each other for e.g. zero hour contact is equal in number of non-zero hour. It was fully consistence data so we could not define independent data is more dependent on male or female. But with figure we can say that there are total 16 females/males who are on a zero-hour contact and with fixed term.

A screenshot of a cell phone

Description automatically generated

## **Analyzing Another Data Set:**

Moreover, this problem is coming with another dataset in All staff Equality as well.

I just wrote some code at the end of Jupyter book. Where there is not finding any distinguish while we let some independent/dependent variables to find relation between them. Because all the data is very nominal.

The list is as below of following results.

we are taking column[Country of HE provider] as dependent variables and will check relation with two other column variables.

* 1- Showing the diagram status of all depend variables.

Which is very clear in figure almost every bar is at same peak mean they are same in number so can’t we distinguish that on which variable independent variables more relay.

A picture containing drawing

Description automatically generated.

* 2- Relation of Dependent Column with independent Variables name as ‘25 and under’ from column of category.

A screenshot of a cell phone

Description automatically generated

One can see clearly that all five columns are giving same result we can’t find difference whether is more dependent upon which country from all of 5.

* 3- When independent variable is ’26-35 ’ from column category.

A screenshot of a social media post

Description automatically generated

* 4- When independent variable is ‘white’ from category column. Which also can’t distinguish between any dependent variable same as before.

A screenshot of a cell phone

Description automatically generated

* Let we change another Independent column and check for result upon dependent Column ‘country of HE Provider’.

Where one can observe there is distinguish between any dependent variable.

A screenshot of a social media post

Description automatically generated

* And at last again I am showing Relation between independent variable ‘Disability Status’ of column name Category maker with Dependent column ( Country of HE Provider) variables.

A screenshot of a social media post

Description automatically generated

Let me add some words in end the data set we downloaded contains similarity in data it is so consistent specially our dependent variables and hence it was not possible to find strong relation between dependent and independent variables.

## **Critical comment on your conclusion**

Number of male and female are equal distributed among all variable values as shown above in histograms figure.

Same as Number of different countries [ England, Northern, Scotland, Wales] are equal distributed among all variables which can’t define relations between independent/ dependent variables.

# **Q6-Project Description**

* **1- Planned our project:**
* I have read the TMA project question document carefully.
* For each question I had some notes and search in the web for some on things that I don’t understand
* I made a timeframe for starting and ending the project.
* **2-Acquiring our data:**
* I have found all data from HESA website whose link is:

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

* **3-Preparation:**
* I downloaded all the three datasets from given link.
* Downloaded Anaconda
* Start coding on jupyter notebook
* **4-Analyzing:**
* Looking for the datasets and the variables inside
* Checking the lookup table to find the relation between variables
* Checking the dependent and independent variables
* Starting to use python trying to analyzes the data using what I have learned
* **5- Report of our findings:**

Loaded the dataset. Remove unnecessary columns. Consider rows with combination of independent and dependent variables. Other Important things in this project

* Exploratory Data Analysis (EDA) with Visualization
* Data Pre-processing
* Data Analysis

# **Q7–Reflection:**

* Our experience with the project:

The experience regarding this project was quite amazing. I have a good experience and I’ve got learned many things

* Learned:

Got learned many things about python and analysis

* what we went well:

Everything was good except Nominal data.

* what went wrong:

I have tried to use some dataframe functions and got many errors, then I have fixed them.

* and how can you benefit from this experience in future projects

I've learned things on Data Frame and histograms most importantly. So, it will benefit me afterwards.

# **Q8- References:**

|  |  |  |
| --- | --- | --- |
| Author | Type of search | Link |
| [**AnkitRai01**](https://auth.geeksforgeeks.org/user/AnkitRai01/articles) | Selecting rows in pandas dataframe on conditions.  I use stack overflow and other sites to search this as well. | <https://www.geeksforgeeks.org/selecting-rows-in-pandas-dataframe-based-on-conditions/> |
| **Brad Solomon** | Histograms | <https://realpython.com/python-histograms/> |
| Nadia Rivera | Cleaning Dirty Data with Pandas & Python | <https://www.developintelligence.com/blog/2017/08/data-cleaning-pandas-python/> |
| Not seen on site | Removing Rows | <https://thispointer.com/pandas-skip-rows-while-reading-csv-file-to-a-dataframe-using-read_csv-in-python/> |

Well there was a lot of search beside it but not due to specific website I found many information specially from stack overflow to find error answers and many more.