Group Projects 2024

1 Overview

For the Data Analysis Group Project you will be placed into groups and be tasked with writing a concise report summarising the results of a statistical analysis that you have conducted. Each group will create a GitHub repository to store and share all code used throughout the analysis. The group report is worth 20% of your final grade for the course. Group reports must be worked on collaboratively and should be written using quarto.

- all R output, including figures and tables, are appropriately labelled and presented;
- R code should **not** be included in the body of the report; and
- the report should be no longer than 6 pages.

Your group report should include:

- An appropriate **Title**;
- An **Introduction** section detailing the data set and question of interest;
- An **Exploratory Analysis** of the data;
- A **Formal Analysis** of the data; and
- Finish with your **Conclusions**.

1.1 Project details

Section 2 describes the projects from which each group has been allocated one to work on for this assignment (the group number corresponds to the dataset allocated). Please ensure you are working with the correct dataset and are answering your assigned question of interest.

> **Note**
>
> - The data sets are available to download from Moodle through the **Group Project Data Sets** link**.**
> - The **Group projects allocations** file contains the information of the classmates you will partner up for the project. Please contact the classmates in your group.
> - Read the **Group Project Data Set Allocations** file to find out which project your group has been assigned to. The group number corresponds to the dataset your team has been allocated to, e.g., *Group 3* will be analyzing *Project 3 - Obesity prevalence*.

1.2 Project Goals

1. **Analyse the allocated data set using GitHub to share and store code**

   Analyse the data in a collaborative fashion in a file called `Group_##_Analysis.qmd` in GitHub by creating a repository on GitHub named `DAS-Group-##` where `##` corresponds to the number of your group.

   > **Important**
   >
   > Only one person will need to create the repository, then add everyone else. Please also add Jafet and Isa to your repository using the GitHub usernames **JBelmont89** and **isammarques**. Ensure your repository is set to public. The deadline for creating the repository is specified in Section 1.4 .

   Notes:

   - When using your repository for analysis, make sure to leave detailed comments on any commits you make to the main branch.
   - Make sure you create at least one branch for at least one piece of work. Assign members of your group to review the changes, edit them if necessary, and commit the new changes. When all reviews are complete, merge the branch into the main branch, and delete the branch.

- If you choose to save any relevant outputs on your repository, please put these in a folder with an appropriate title (e.g. Plots).
- The `.qmd` file which you share on Github must be able to render to a `.pdf/html` file which contains the commented code used in the analysis and appropriate report output.

2. **Produce a report on your analysis**

   Your report should summarise the statistical analysis you conducted and a thorough interpretation of your model in relation to the data analysed. You should elect one member of your group to be responsible for submitting the group report as a **.pdf/html** file via the **Group Project Submission** link in section **Group project** on the Data Analysis Moodle page.

   You should submit your group report as `Group_##_Analysis.pdf/html`. Even though one person will be elected to submit the report, all members of the group should be able to see what has been submitted on Moodle.

   Make sure that in your report:

   - all R output, including figures and tables, are appropriately labelled and presented;
   - R code should **not** be included in the body of the report; and
   - the report should be no longer than 6 pages.

   Your group report should include:

   - An appropriate **Title**;
   - An **Introduction** section detailing the data set and question of interest;
   - An **Exploratory Analysis** of the data;
   - A **Formal Analysis** of the data; and
   - Finish with your **Conclusions**.

1.3 Marking scheme

This assessment will be marked out of **16 MARKS** with the report worthing a total of **15 MARKS**, which are broken down as follows:

1. **Title** and **Introduction** `[1 MARK]`
    - Appropriate title
    - Description of background to problem and questions of interest.
2. **Exploratory data analysis** `[5 MARKS]`
    - Appropriate exploration of the allocated dataset using multiple exploratory techniques and identification of possible patterns/anomalies in the data.
    - Appropriate comments relating to the summary statistics / plots and the questions of interest.
3. **Formal data analysis** `[6 MARKS]`
    - Appropriate statistical methods have been correctly applied that answers the questions of interest and produce appropriate results.
    - Clear and proper explanations of analysis and presentation of key results.
    - Appropriate interpretation of the model and how it answers the questions of interest.
4. **Conclusions** `[2 MARKS]`
    - Discussion of key results and validity of conclusions.
    - Discussion of future work/extensions
5. **General layout** `[1 MARK]`
    - Appropriate layout with nicely displayed output and appropriate use of tables and/or figures with sensible labels and cross-referencing/hyperlinks.
    - Overall visual appeal of the report/readability
    - Overall organisation and structure of report is clear

In addition to the final report, you will also be assessed on your use of GitHub to work in a collaborative environment:

6. **Collaborative Coding** `[1 MARK]`
    - At least one branch created to look at a section of analysis. This branch should be then merged into the main branch. At least one member of the group should review the changes, comment and - potentially after several reviews -, merge the branch into main and delete it.

- Demonstrate good practice of using GitHub. For example, provide meaningful commit messages, review others' commits to the repo, keep repository organised and tidy, provide a meaningful README file. Consider using .gitignore to ignore files you don't want to track.

1.4 Submission and deadlines

Everything that you are required to submit for your group project should be submitted by **11:00 am 22nd March 2024** via Moodle. You must decide on one member of each group to be responsible for submitting. Details of each submission are given below.

- Setup GitHub Repository – **Week 8** by **11:00 am Friday 1st March 2024**
- Report submission – **Week 11** by **11:00 am 22nd March 2024**
- Contribution evaluations and Declaration of Originality – **Week 11** by **11:00 am 22nd March 2024**

> **!** Important
>
> Note that the GitHub Repository must be created during **Week 8** so you can start working on your analysis.

1.4.1 Contribution Form

Alongside the group report, each member will be need to separately submit a **Contribution Form** as a **.pdf** file via the **Group Project Contribution Form Submission** link in section **Group project** on the Data Analysis Moodle page. You should submit your contribution form as **GroupNumber_MatriculationNumber_ContributionForm.pdf**. Here, you will provide information on how much you believe each member of your group contributed to the group project. Contribution of each member may consist of the following:

- attending arranged group meetings;
- analysis of the data set via writing `R` code and/or interpretation of the results;
- writing the groups findings and sections of the report in quarto; and
- being cooperative and supportive throughout the project.

If a meaningful discrepancy in the contributions of group members is observed then that may result in different grades being awarded to different group members. For example, no grade will be awarded to a group member who does not contribute anything to the group project.

1.4.2 Declaration of Originality Form

Each member should also submit a **Declaration of Originality Form** as a **.pdf** file via the **Group Project Declaration of Originality Form Submission** link in section **Group project** on the Data Analysis Moodle page. You should submit your Declaration of Originality form as **GroupNumber_MatriculationNumber_DeclarationForm.pdf**.

2 Group Project descriptions

> **!** Important
>
> Ensure your group downloads and analyses the correct data set you have been assigned otherwise your report will be void and you will not receive a grade.

2.1 Modelling weight/obesity in Scotland

The weight and categorisation of weight (namely obesity) in Scotland has been monitored since the introduction of the Scottish Health Survey, which is designed to monitor the health of the Scottish population living in private households. The main aim of the survey is to keep an eye on health trends in Scotland. The Scottish Heath Survey data will be used to explore trends in weight/obesity in Scotland.

Project 1 - Obesity prevalence

Data are available on socio-economic and lifestyle factors from the 2008 - 2012 Scottish Health Surveys. The data are stored in `DAProject1.csv` and contain the following columns.

- `Age` - Age of individual
- `Sex` - Sex of individual (Male / Female)
- `Education` - Highest educational qualification of individual
- `Veg` - Consume recommended daily vegetable intake (Yes / No)
- `Fruit` - Consume recommended daily fruit intake (Yes / No)
- `Year` - Year of the Scottish Health Survey
- `Obese` - Indicator of individuals obesity classification (Yes / No)

Questions of interest

The main questions of interest are:

- Has the prevalence of obesity in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in obesity by age, gender, socio-economic status or lifestyle factors?

Project 2 - BMI distribution

Data are available on socio-economic and lifestyle factors from the 2008 - 2012 Scottish Health Surveys. The data are stored in `DAProject2.csv` and contain the following columns.

- `Age` - Age of individual
- `Sex` - Sex of individual (Male / Female)
- `Education` - Highest educational qualification of individual
- `Veg` - Consume recommended daily vegetable intake (Yes / No)
- `Fruit` - Consume recommended daily fruit intake (Yes / No)
- `Year` - Year of the Scottish Health Survey
- `BMI` - Body Mass Index of individual

Questions of interest

The main questions of interest are:

- Has the body mass index (BMI) in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in the BMI distribution by age, gender, socio-economic status or lifestyle factors?

Project 3 - Obesity prevalence

Data are available on socio-economic and lifestyle factors from the 2013 - 2016 Scottish Health Surveys. The data are stored in `DAProject3.csv` and contain the following columns.

- `AgeGroup` - Age range of individual
- `Sex` - Sex of individual (Male / Female)
- `Employment` - Employment status of individual
- `Veg` - Consume recommended daily vegetable intake (Yes / No)
- `Fruit` - Consume recommended daily fruit intake (Yes / No)
- `Year` - Year of the Scottish Health Survey
- `Obese` - Indicator of individuals obesity classification (Yes / No)

Questions of interest

The main questions of interest are:

- Has the prevalence of obesity in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in obesity by age, gender, socio-economic status or lifestyle factors?

Project 4 - BMI distribution

Data are available on socio-economic and lifestyle factors from the 2013 - 2016 Scottish Health Surveys. The data are stored in `DAProject4.csv` and contain the following columns.

- `AgeGroup` - Age range of individual
- `Sex` - Sex of individual (Male / Female)
- `Employment` - Employment status of individual
- `Veg` - Consume recommended daily vegetable intake (Yes / No)
- `Fruit` - Consume recommended daily fruit intake (Yes / No)
- `Year` - Year of the Scottish Health Survey
- `BMI` - Body Mass Index of individual

Questions of interest

The main questions of interest are:

- Has the body mass index (BMI) in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in the BMI distribution by age, gender, socio-economic status or lifestyle factors?

Project 5 - Obesity prevalence

Data are available on socio-economic and lifestyle factors from the 2013 - 2016 Scottish Health Surveys. The data are stored in `DAProject5.csv` and contain the following columns.

- `Age` - Age of individual
- `Sex` - Sex of individual (Male / Female)
- `Education` - Highest educational qualification of individual
- `Veg` - Consume recommended daily vegetable intake (Yes / No)
- `Fruit` - Consume recommended daily fruit intake (Yes / No)
- `Year` - Year of the Scottish Health Survey
- `Obese` - Indicator of individuals obesity classification (Yes / No)

Questions of interest

The main questions of interest are:

- Has the prevalence of obesity in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in obesity by age, gender, socio-economic status or lifestyle factors?

Project 6 - BMI distribution

Data are available on socio-economic and lifestyle factors from the 2013 - 2016 Scottish Health Surveys. The data are stored in `DAProject6.csv` and contain the following columns.

- `Age` - Age of individual
- `Sex` - Sex of individual (Male / Female)
- `Education` - Highest educational qualification of individual
- `Veg` - Consume recommended daily vegetable intake (Yes / No)
- `Fruit` - Consume recommended daily fruit intake (Yes / No)
- `Year` - Year of the Scottish Health Survey
- `BMI` - Body Mass Index of individual

Questions of interest

The main questions of interest are:

- Has the body mass index (BMI) in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in the BMI distribution by age, gender, socio-economic status or lifestyle factors?

Project 7 - Obesity prevalence

Data are available on socio-economic and lifestyle factors from the 2008 - 2012 Scottish Health Surveys. The data are stored in `DAProject7.csv` and contain the following columns.

- `AgeGroup` - Age range of individual
- `Sex` - Sex of individual (Male / Female)
- `Employment` - Employment status of individual
- `Veg` - Consume recommended daily vegetable intake (Yes / No)
- `Fruit` - Consume recommended daily fruit intake (Yes / No)
- `Year` - Year of the Scottish Health Survey
- `Obese` - Indicator of individuals obesity classification (Yes / No)

Questions of interest

The main questions of interest are:

- Has the prevalence of obesity in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in obesity by age, gender, socio-economic status or lifestyle factors?

Project 8 - BMI distribution

Data are available on socio-economic and lifestyle factors from the 2008 - 2012 Scottish Health Surveys. The data are stored in `DAProject8.csv` and contain the following columns.

- `AgeGroup` - Age range of individual
- `Sex` - Sex of individual (Male / Female)
- `Employment` - Employment status of individual
- `Veg` - Consume recommended daily vegetable intake (Yes / No)
- `Fruit` - Consume recommended daily fruit intake (Yes / No)
- `Year` - Year of the Scottish Health Survey
- `BMI` - Body Mass Index of individual

Questions of interest

The main questions of interest are:

- Has the body mass index (BMI) in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in the BMI distribution by age, gender, socio-economic status or lifestyle factors?

Project 9 - Obesity prevalence

Data are available on socio-economic and lifestyle factors from the 2008 - 2012 Scottish Health Surveys. The data are stored in `DAProject9.csv` and contain the following columns.

- `AgeGroup` - Age range of individual
- `Sex` - Sex of individual (Male / Female)
- `Employment` - Employment status of individual
- `Year` - Year of the Scottish Health Survey
- `BMIgroup` - Indicator of individuals weight classification group

Questions of interest

The main questions of interest are:

- Has the prevalence of obesity in Scotland changed over the given years of the Scottish Health Survey?

- Are there any differences in obesity by age, gender, socio-economic status or lifestyle factors?

> Hint
>
> You will need to create a new binary response variable for obesity classification from the `BMIgroup` variable.

Project 10 - Obesity prevalence

Data are available on socio-economic and lifestyle factors from the 2008 - 2012 Scottish Health Surveys. The data are stored in `DAProject10.csv` and contain the following columns.

- `Age` - Age of individual
- `Sex` - Sex of individual (Male / Female)
- `Education` - Highest educational qualification of individual
- `Year` - Year of the Scottish Health Survey
- `BMIgroup` - Indicator of individuals weight classification group

Questions of interest

The main questions of interest are:

- Has the prevalence of obesity in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in obesity by age, gender, socio-economic status or lifestyle factors?

> Hint
>
> You will need to create a new binary response variable for obesity classification from the `BMIgroup` variable.

Project 11 - Obesity prevalence

Data are available on socio-economic and lifestyle factors from the 2013 - 2016 Scottish Health Surveys. The data are stored in `DAProject11.csv` and contain the following columns.

- `AgeGroup` - Age range of individual
- `Sex` - Sex of individual (Male / Female)
- `Employment` - Employment status of individual
- `Year` - Year of the Scottish Health Survey

- `BMIgroup` - Indicator of individuals weight classification group

Questions of interest

The main questions of interest are:

- Has the prevalence of obesity in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in obesity by age, gender, socio-economic status or lifestyle factors?

> Hint
>
> You will need to create a new binary response variable for obesity classification from the `BMIgroup` variable.

Project 12 - Obesity prevalence

Data are available on socio-economic and lifestyle factors from the 2013 - 2016 Scottish Health Surveys. The data are stored in `DAProject12.csv` and contain the following columns.

- `Age` - Age of individual
- `Sex` - Sex of individual (Male / Female)
- `Education` - Highest educational qualification of individual
- `Year` - Year of the Scottish Health Survey
- `BMIgroup` - Indicator of individuals weight classification group

Questions of interest

The main questions of interest are:

- Has the prevalence of obesity in Scotland changed over the given years of the Scottish Health Survey?
- Are there any differences in obesity by age, gender, socio-economic status or lifestyle factors?

> Hint
>
> You will need to create a new binary response variable for obesity classification from the `BMIgroup` variable.

2.2 MMR vaccination in Scotland

The Scottish Childhood Immunisation Record System (SCIRS) holds the individual records of all childhood vaccinations in Scotland. These include measles, mumps, and rubella (MMR) vaccination uptake, which occurs when a child is 12-13 months old and again at 4-5 years of age. Data are available from the SCIRS database between 1998 and 2014. The beginning of this time period was when Wakefield et al. (1998) linked the MMR vaccine with an increased risk of autism, with the media coverage surrounding the article resulting in vaccination rates dropping to around 80% in 2003 in parts of the United Kingdom. These reduced vaccination rates later resulted in large outbreaks of measles in the UK in 2013. The article by Wakefield et al. (2008) was partially retracted in 2004, before being discredited in 2010 after several epidemiological studies failed to find any association with an increased risk in autism.

Project 13 - Measles susceptibility in Glasgow

Data are available on measles susceptibility in pre-school children from the 133 intermediate zones (IZ) comprising Glasgow, which are small geographical units containing, on average, 4000 residents between 1998 and 2014. The data are stored in `DAProject13.csv` and contain the following columns.

- Y - The number of pre-school children susceptible to measles in a given IZ
- N - The total number of pre-school children in a given IZ
- Year - Year the data was collected

Questions of interest

The main questions of interest are:

- Did Glasgow exhibit a change in measles susceptibility following the retraction of the Wakefield article?
- Did the change, if any, in measles susceptibility occur in 2004 alongside the articles' retraction?

> Hint
>
> It is the proportion of pre-school children susceptible to measles that is modelled.

Project 14 - Measles susceptibility in Edinburgh

Data are available on measles susceptibility in pre-school children from the 101 intermediate zones (IZ) comprising Edinburgh, which are small geographical units containing, on average,

4000 residents between 1998 and 2014. The data are stored in `DAProject14.csv` and contain the following columns.

- `Y` - The number of pre-school children susceptible to measles in a given IZ
- `N` - The total number of pre-school children in a given IZ
- `Year` - Year the data was collected

Questions of interest

The main questions of interest are:

- Did Edinburgh exhibit a change in measles susceptibility following the retraction of the Wakefield article?
- Did the change, if any, in measles susceptibility occur in 2004 alongside the articles' retraction?

> Hint
>
> It is the proportion of pre-school children susceptible to measles that is modelled.

Project 15 - Measles susceptibility in Glasgow

Data are available on measles susceptibility in primary school children from the 133 intermediate zones (IZ) comprising Glasgow, which are small geographical units containing, on average, 4000 residents between 1998 and 2012. The data are stored in `DAProject15.csv` and contain the following columns.

- `Y` - The number of pre-school children susceptible to measles in a given IZ
- `N` - The total number of pre-school children in a given IZ
- `Year` - Year the data was collected

Questions of interest

The main questions of interest are:

- Did Glasgow exhibit a change in measles susceptibility following the retraction of the Wakefield article?
- Did the change, if any, in measles susceptibility occur in 2004 alongside the articles' retraction?

> Hint
>
> It is the proportion of primary school children susceptible to measles that is modelled.

Project 16 - Measles susceptibility in Edinburgh

Data are available on measles susceptibility in primary school children from the 101 intermediate zones (IZ) comprising Edinburgh, which are small geographical units containing, on average, 4000 residents between 1998 and 2012. The data are stored in `DAProject16.csv` and contain the following columns.

- `Y` - The number of pre-school children susceptible to measles in a given IZ
- `N` - The total number of pre-school children in a given IZ
- `Year` - Year the data was collected

Questions of interest

The main questions of interest are:

- Did Edinburgh exhibit a change in measles susceptibility following the retraction of the Wakefield article?
- Did the change, if any, in measles susceptibility occur in 2004 alongside the articles' retraction?

> Hint
>
> It is the proportion of primary children susceptible to measles that is modelled.