UNIVERSITY OF STRATHCLYDE
DEPARTMENT OF MATHEMATICS & STATISTICS

MM104: Statistics and Data Presentation Semester 2
MM107: Statistics and Data Presentation

Project 7: Mystery Project

Overview

This week's project is worth 20 % of your overall mark for the class and thus requires slightly more effort than previous projects. Each group will work with a data set on a different topic; you will be required to decide for yourselves, the appropriate statistical techniques to use and the appropriate results to present as part of a presentation and report.

The presentation and report should be handed in by 10am on Friday Week 9.

The Data Set

Group 1 - Life Expectancy in Scotland

The dataset to be used is `Life_Expectancy.csv`. It contains the life expectancy for men and women living in various small areas of Scotland called intermediate zones. The percentage of working age adults who are employment deprived is also available as well as the health board the intermediate zone is located in. The data come from the Scottish Public Health Observatory.

- What is the association between life expectancy and employment deprivation?
- Is the association the same in the two largest health boards?
- Does life expectancy vary between Greater Glasgow and Lothian?

Variable	Value	Description
Area_Code	Text	Label of the intermediate zone.
Emp.Deprived	Numeric	Percentage of the working age population who are employment deprived.
LE.Males	Numeric	Life expectancy for men living in the intermediate zone, years
LE.Females	Numeric	Life expectancy for women living in the intermediate zone, years
HB.Name	Text	Health Board

Table 1: Group 1 - Life Expectancy in Scotland: Variable names and descriptions.

Group 2 - Instagram and Twitter followers

The dataset to be used is `insta_and_twitter`. It contains the number of Instagram followers and Twitter followers for a number of celebrities. Each celebrity is from a particular category. The data was collected in January 2021 and was correct at that time.

- Is there an association between Twitter and Instagram followers?
- Is the number of followers the same for Instagram and Twitter?
- Is there any association between category of celebrity and followers?

Variable	Value	Description
Name	Text	Name of the "celebrity"
Category	0: Other 1: Musician 2: Actor 3: Sports personality 4: Reality star	Category of the "celebrity"
Instagram followers	Continuous value (millions)	Number of Instagram followers to the nearest million
Twitter followers	Continuous value (millions)	Number of Instagram followers to the nearest million

Table 2: Group 2 - Instagram and Twitter: Variable names and descriptions.

Group 3 - Virus

The dataset to be used is `virus_handsanitiser.csv`. It contains the number of hand sanitisers purchased from selected Boots stores in Glasgow and surrounding area over a 14 day period in January 2020.

- Are the purchasing behaviours of Scottish people controlled ?
- Is there any difference in the number of hand sanitisers purchased in different Scottish regions?
- Are there any trends in the data?

Variable	Value	Description
Day	Discrete value	The day in the study
Area	Text	Area in Scotland
Hand Sanitisers	Discrete value	Approximate number of hand sanitisers sold each day per store.

Table 3: Group 3 - Virus: Variable names and descriptions.

Group 4 - Pollution

The dataset to be used is `pollution.csv`. It contains the number plastic bags that were collected by two environmental agencies at 24 randomly selected sites for each of four ocean areas.

- Is there any difference in the number of plastic bags collected by each agency?
- Is there a difference in number of plastic bags in each ocean ?

Variable	Value	Description
Ocean	Categorical	The name of ocean for which the sea region was located.
Sea	Categorical	The name of the sea region for which pollution was collected from.
Agency	Categorical	The name of the agencies involved in collecting the plastic bags.
Plastic bags	Discrete Value	The number of plastic bags collected.

Table 4: Group 4 - Pollution: Variable names and descriptions.

Group 5 - Obesity, Deprivation and Birth Weight

The dataset to be used is `Birth_Weight.csv`.It contains the information about obesity in mothers, birthweight of babies and children living in low income families in various small areas of Scotland called intermediate zones. The health board the intermediate zone is located in is also available. The data comes from the Scottish Public Health Observatory and was collected in 2019.

- Is there an association between obesity in mothers' and low birth weight of babies?
- Is there an association with deprivation and if so, what is it?
- Is Obesity in mothers the same in Greater Glasgow compared to Lothian?

Variable	Value	Description
Area_Code	Text	Label of the intermediate zone.
Mothers.Obese	Numerical	Percentage of mothers in the intermediate zone classed as obese.
Babies.HBW	Numeric	Percentage of babies born to mothers in the intermediate zone who are a healthy birth weight.
Children.LIF	Numeric	Percentage of children in the intermediate zone living in low income families.
HB.Name	Text	Health Board

Table 5: Group 5 - Obesity, Deprivation and Birth Weight: Variable names and descriptions.

Report and Presentation

Your Word report should be at least 3 to 4 pages, and at least 1000 words (excluding title page and references). Your report should have sections which correspond to the sections in your PowerPoint presentation as well as a references section. Your PowerPoint presentation should be 6-7 minutes.

Remember: the first or second page should contain a declaration section. This should list the names of all group members and sentence outlining each person's contribution to the project. Each person should sign this sheet electronically.