Noteable™

Analysing Data

Analysing data in R

Resource

Secondary

14-18 years

Contents

Noteable Activities for Schools: Analysing data in R

These resources are a guide for teachers to demonstrate to the whole class or direct individual students as appropriate. The activities below can be directly distributed to pupils.

For instructions on how to install and use Noteable resources, please look at our guides for teachers in GLOW: GLOW guidance for teachers to start using Noteable.

Content and Curriculum links

Level	Context	Indicators
14-18	Data Analysis	Graphics Descriptive Statistics Correlation and Regression

Knowledge	Using bullet point lists to give instructions
Curriculum links (England) Computing KS4	• develop and apply their analytic, problem-solving, design, and computational thinking skills
Scottish Curriculum for Excellence	**Experiences and Outcomes:** • I can evaluate and interpret raw and graphical data using a variety of methods, comment on relationships I observe within the data and communicate my findings to others. **MNU 4-20a** **Benchmark:** ○ Interprets raw and graphical data. ○ Uses statistical language, for example, correlations, to describe identified relationships.
All: Cross-curricular opportunities	English, Science, Social Studies and Geography

Analysing data in R

To read in data from an Excel csv file called excel_data.csv to R Studio and name it mydata, first use the drop-down menus in R Studio Session > Set Working Directory > Choose Directory to indicate the location of excel_data.csv on your computer. The following code will then read the data into R Studio:

```
R                                                    Copy code

mydata <- read.csv("excel_data.csv")

attach(mydata) # this adds the variable names
```

At the end of the analysis, remember to use detach(mydata) to disassociate the variable names.

The list below displays a list of statistical tools and vocabulary which are needed for Higher Application of Maths. The list has been added for your base knowledge. For this set of activities, only the items in bold and marked with an arrow will be used.

a) Graphics

- hist(X, col="yellow", main="Histogram of X (units)"): This produces a histogram of the variable X.
- **plot(X, Y, xlab="x-axis label", ylab="y-axis label", main="Scatterplot of Y on X", pch=21, bg="black"): This produces a scatter plot of Y on X with the required title, axis labels, and black dots.**
- pie(table(X), main="Title"): This gives a simple pie chart of the categories in variable X with the specified title.
- barplot(table(X), main="title", xlab="x-axis label", col="orange"): This gives a bar chart of the categories in the variable X with the required title, axis labels, and color.
- boxplot(X): This produces a boxplot of the numerical variable X.

b) Descriptive Statistics

- mean(X): Computes the mean of X.
- sd(X): Computes the standard deviation of X.
- summary(X): Computes the mean, median, minimum, maximum, and upper and lower quartiles.
- table(X): Computes the number of observations in each level of the categorical variable X.

- prop.table(table(X)): Returns the proportion of observations in each level of the categorical variable X.
- prop.table(table(X))*100: Returns the percentage of observations in each level of the categorical variable X.
- table(X,Y): Produces a cross-tabulation between the two categorical variables X and Y.

c) Correlation and Regression

- cor.test(X,Y): Computes the correlation between X and Y and performs a test of the null hypothesis of zero correlation.
- lm(Y~X): Fits a linear regression.

d) Hypothesis Testing

t.test(X,Y) — performs a two sample t-test between X and Y

t.test(X,Y,paired=TRUE) — performs a paired t-test between X and Y

prop.test(x = c(a, b), n = c(n1, n2)) — performs a 2-sample test for equality of proportions with continuity correction

Activity 1

NAAS, has invested in an online business called 'EUNAAS Tour', which promotes touristic tours through Europe. He wants to promote his business and one of his ideas is to advertise countries in Europe with highest quality of life in comparison with environmental elements such as pollution and climate. He did his research, but some data needs to be tidied up. Can you use R to write an algorithm for his case?

To help NAAS, you will need to refer to the spreadsheet file 'EU_quality_of_life.csv'.

You will then need to:

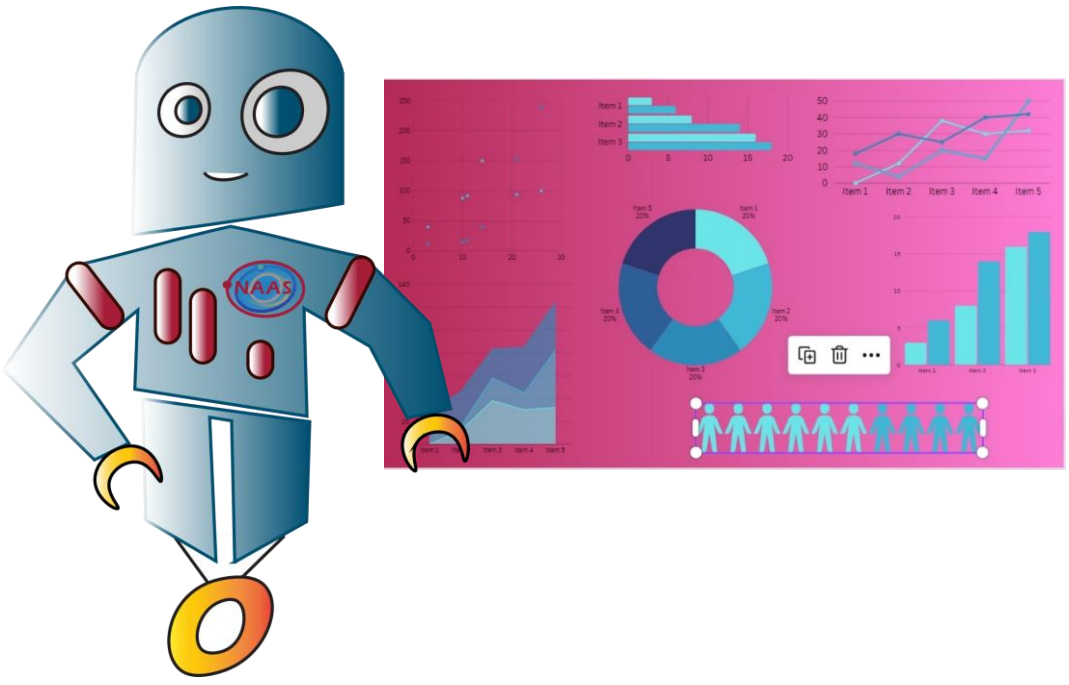
a) construct a scatter plot comparing the quality of life and pollution for each of the 36 European countries in the csv file,
b) construct a scatter plot comparing the pollution and climate for each of the 36 European countries in the csv file.

Before you get started, let's understand what each of the variables mean:

c) **Quality of Life Index**: This index represents the overall quality of life in each country. Higher values indicate better quality of life.

d) **Pollution Index**: This index measures the level of pollution in each country. Lower values indicate less pollution.

e) **Climate Index**: This index reflects the climate conditions in each country. It considers factors like temperature, precipitation, and overall climate comfort. Higher values indicate more favourable climates.



Activity 1 – Answer (for teachers)

The codes below create two scatter plots; one for comparing quality of life and pollution and the other for comparing pollution and climate for each of the 36 European countries in the csv file.

R Copy code

```
# Load the data
data <- read.csv("EU_quality_of_life.csv")


# Create a scatter plot comparing quality of life and pollution
plot(data$Pollution.Index, data$Quality.of.Life.Index,
    xlab = "Pollution Index",
    ylab = "Quality of Life Index",
    main = "Scatterplot of Quality of Life on Pollution",
    pch = 21, bg = "black")
```

R Copy code

```
# Create a scatter plot comparing pollution and climate
plot(data$Pollution.Index, data$Climate.Index,
    xlab = "Pollution Index",
    ylab = "Climate Index",
    main = "Scatterplot of Climate on Pollution",
    pch = 21, bg = "blue")
```

In these plots:

- xlab specifies the label for the X-axis.
- ylab specifies the label for the Y-axis.
- main provides the title for the plot.
- pch = 21 sets the point shape to a filled circle.
- bg = "black" fills the points with black color.

Activity 2 – A step forward

NAAS also wants to find out:

a) the correlation coefficient between the quality of life and the climate and
b) the correlation coefficient between the quality of life and the pollution.

1. Can you write the code for each of the tasks above and interpret the correlation coefficient for each of them?
2. Are the two results similar?
3. What conclusion will NAAS come to when he compares climate and pollution coefficients?

Activity 2 – Answer (for teachers)

To implement the algorithm for tracking and updating the average ratings for NAAS's juices, you can use a similar approach as before. Here's a basic example:

R Copy code

```
# Load the data
data <- read.csv("EU_quality_of_life.csv")


# Calculate the correlation coefficients
climate_cor <- cor.test(data$Quality.of.Life.Index, data$Climate.Index)
```

R Copy code

```
# Load the data
data <- read.csv("EU_quality_of_life.csv")


# Calculate the correlation coefficients
pollution_cor <- cor.test(data$Quality.of.Life.Index, data$Pollution.Index)
```

Interpretation:

1. **Correlation Coefficient between Quality of Life and Climate**:
    - The correlation coefficient between quality of life and climate is approximately cor_quality_climate.
    - A positive coefficient indicates that as the climate index increases, the quality of life tends to improve. However, the magnitude of the coefficient tells us how strong this relationship is.
    - Since the coefficient is positive, we can infer that better climate conditions are associated with higher quality of life.

2. **Correlation Coefficient between Quality of Life and Pollution**:

- The correlation coefficient between quality of life and pollution is approximately cor_quality_pollution.
- A negative coefficient suggests that as pollution levels increase, the quality of life tends to decrease.
- The magnitude of the coefficient indicates the strength of this relationship.

Comparison:

- Comparing the two coefficients:
    - The climate coefficient is positive, indicating a positive relationship between climate and quality of life.
    - The pollution coefficient is negative, suggesting a negative relationship between pollution and quality of life.
    - The results are not similar; they have opposite signs.
    - This implies that better climate conditions are associated with higher quality of life, while higher pollution levels are associated with lower quality of life.

Conclusion for NAAS:

- NAAS should prioritize addressing pollution reduction to improve overall quality of life.
- Efforts to improve air quality may have a more direct impact on well-being than climate-related measures.
- Policies and actions aimed at reducing pollution can lead to better quality of life outcomes for individuals and communities.

Activity 3

NAAS needs a few more results before coming to conclusions. You can help him by:

a) finding the equation of the regression line of quality of life and climate,
b) finding the equation of regression line of quality of life and pollution,
c) find the equation of climate and pollution,
d) From your findings, interpret the slope and intercept parameters,
e) Estimate the quality of life in Luxembourg if pollution was 27.5.

Activity 3 – Answer (for teachers)

First, **Explore the Data**:

- Check the structure of the data using str(data).
- Ensure that the columns for Quality of Life, Climate Index, and Pollution Index are numeric.

```R
# Check the structure of the data
str(data)
```

a) **Linear Regression (Quality of Life vs. Climate Index)**:

- Perform linear regression between the Quality of Life Index (dependent variable) and the Climate Index (independent variable).
- The equation would be of the form:

  $Quality\_of\_Life = \beta 0 + \beta 1 \cdot Climate\_Index$

```R
# Perform linear regression
lm_quality_climate <- lm(Quality_of_Life_Index ~ Climate_Index, data = data)

# Get regression coefficients
summary(lm_quality_climate)
```

b) **Linear Regression (Quality of Life vs. Pollution Index)**:

- Similarly, perform linear regression between the Quality of Life Index and the Pollution Index.
- Equation:

Quality_of_Life=β0+β1·Pollution_Index

```R
# Perform linear regression
lm_quality_pollution <- lm(Quality_of_Life_Index ~ Pollution_Index, data = data)

# Get regression coefficients
summary(lm_quality_pollution)
```

c) **Linear Regression (Climate vs. Pollution Index)**:

- Similarly, perform linear regression between the Climate Index and the Pollution Index.
- Equation:

  Climate=β0+β1·Pollution_Index

```R
# Perform linear regression
climate_pollution_reg <- lm(Climate.Index ~ Pollution.Index, data = data)

# Get regression coefficients
summary(lm_quality_pollution)
```

The calculation for each, a, b and c activities can also be simplified as below:

```R
# Load the data
data <- read.csv("EU_quality_of_life.csv")


# Calculate the regression lines
quality_climate_reg <- lm(Quality.of.Life.Index ~ Climate.Index, data = data)
quality_pollution_reg <- lm(Quality.of.Life.Index ~ Pollution.Index, data = data)
climate_pollution_reg <- lm(Climate.Index ~ Pollution.Index, data = data)
```

d) **Interpretation of Parameters**:

- The coefficients from the regression output will give you the slope (β1) and intercept (β0) for each relationship.
- The slope represents the change in Quality of Life for a one-unit change in the independent variable.
- The intercept represents the Quality of Life when the independent variable is zero.

e) **Estimate Quality of Life in Luxembourg**:

- To estimate Quality of Life in Luxembourg when pollution is 27.5, use the regression equation for Quality of Life and Pollution.
- Plug in the Pollution Index value (27.5) to find the corresponding Quality of Life.

Two-Sample t-Test:

- To perform a two-sample t-test between X and Y, you can use the t.test function as follows:

```R
# Load the data
data <- read.csv("EU_quality_of_life.csv")
# Example: Two-sample t-test
result <- t.test(X, Y)
print(result)
```

Paired t-Test:

- For a paired t-test between X and Y, set the paired argument to TRUE:

```R
# Load the data
data <- read.csv("EU_quality_of_life.csv")
# Example: Paired t-test
result_paired <- t.test(X, Y, paired = TRUE)
print(result_paired)
```

Two-Sample Test for Equality of Proportions:

- To perform a two-sample test for equality of proportions with continuity correction, use the prop.test function:

```R
# Example: Two-sample test for proportions
result_proportions <- prop.test(x = c(a, b), n = c(n1, n2))
print(result_proportions)
```

Cross-curricular opportunities

1. **Geography:**

   Students can explore the geographical locations of the countries in the dataset and compare their climate and pollution levels with other regions of the world.

2. **Science:**

   Students can investigate the causes and effects of climate change and pollution on the environment and human health.

3. **Social Studies:**

   Students can analyse the impact of climate change and pollution on different societies and cultures and explore the policies and actions taken by governments and organizations to address these issues.

4. **Language Arts:**

   Students can write essays, reports, or stories about the relationship between quality of life, climate, and pollution and how they affect people's lives.

These cross-curricular opportunities can help students develop a deeper understanding of the complex issues related to quality of life, climate, and pollution and their interconnections with different subjects and disciplines.

Copyrights