Data analysis project guide

How to write a clear, convincing data analysis report using R Markdown.

Markdown data analysis report is like a brief research article

- Contains same sections - Introduction, Results, Discussion (optional), Conclusion
- Includes ***text*** and ***code***
  - Code shows steps used for data analysis
  - Text explains the code
    - Why it's there - questions answered and why they matter
    - States and explains the results from running the code
- If you continue in science/bioinformatics, use your Markdowns as starting material for formal presentations, research articles, posters

Sections

- **Introduction**
- **Results** (also sometimes called "Analysis")
- **Discussion** (optional if results discussed at length in **Results**)
- **Conclusion**

What to include in your report

- **Introduction** - explain the topic, questions you will answer
  1. Explain: Why does it matter?
  2. State question(s) you plan to answer. Discuss possible answers. What do you expect, and why?
  3. Introduce data you'll use to answer the question- variables tested, number of replicates, brief description of experiment or link to publication if available.

What to include in your report

- **Results** - R code & prose that describes results
  1. Exploratory data analysis (if relevant)
     - number observations, summary statistics on variables to be used, etc.
  2. Code that answers questions from **Introduction**
     - Write multiple small code chunks, not one long one.
       - Explain why you included each chunk. What's its function?
     - Automate display of results using in-line R code.
     - Write text that explains results from code chunks.
  3. If you use plots:
     - Label axes. Include a title.
     - Include a legend describing the plot, if not obvious.
     - State take-home message from in easy-to-understand prose, e.g., "The plot shows that..."

Example from a Results section (.Rmd file)

states major result from previous plot

```
Visual inspection of the scatter plot suggested that men's weights increased with height faster than women's weights.

To quantify this difference, use linear regression:

```{r}
females=h[h$gender==1,]
model.female=lm(females$weight~females$height)
males=h[h$gender==0,]
model.male=lm(males$weight~males$height)
```

The slope of the men's model was `r model.male$coefficients[2]`. The slope of the women's model was `r model.female$coefficients[2]`. The difference (men - women) rounded to three decimal places was: `r round(model.male$coefficients[2]-model.female$coefficients[2],3)`.
```

states why the next chunk is included in the Markdown

code chunk; only four lines (short, easy-to-understand)

in-line R code; automatically updates if you change previous code

This section of text states the outcome from the previous code chunk. This is a "result", not a conclusion.

Example from a Results section (.Rmd file)

```
Visual inspection of the scatter plot suggested that men's weights increased with height faster than women's weights.

To quantify this difference, use linear regression:

```{r}
females=h[h$gender==1,]
model.female=lm(females$weight~females$height)
males=h[h$gender==0,]
model.male=lm(males$weight~males$height)
```

The slope of the men's model was `r model.male$coefficie
model was `r model.female$coefficients[2]`. The difference (men - women) rounded to three decimal places was: `r round(model.male$coefficients[2]-model.female$coefficients[2],3)`.
```

Note: What if there's a bug? For example, gender indicator variable (1 or 0) could be incorrect. What if 0 is female and 1 is male?

IMPORTANT: By using in-line R code instead of "hard-coding" the literal values, you make your data analysis Markdown more robust. It automatically updates to use the newer values the next time you "knit" your Markdown after changing previous code chunks to fix an error or improve your analysis methodology.

Always explain your figures - an example

Introduce the figure.
Explain why it's included.

```
Is there are relationship between height and weight, and does gender matter? To answer this question, a scatter plot plotting weight (y axis) against height (x axis) and color-coding points by gender to illuminate whether gender is a lurking variable.

```{r, fig.height=5, fig.width=5}
xlab="Height (inches)"
ylab="Weight (lbs)"
main="Height and Weight"
plot(h$height,h$weight, xlab=xlab, ylab=ylab,pch='.',main=main)
points(h$height[h$gender==0],h$weight[h$gender==0],col="lightblue",pch=15)
points(h$height[h$gender==1],h$weight[h$gender==1],col="orange",pch=16)
legend(55,225, pch=c(16,15), col=c("orange","lightblue"),legend=c("females","males"))
fname="ScatterPlot.png"
if (!file.exists(fname)) {
 quartz.save(fname,dpi=300,type="png")
}
```

Visual inspection of the scatter plot suggested that men's weights increased with height faster than women's weights.
```

Format using reasonable sizes.

Label the axes.
Include a title.

Include a legend as needed.

Always state the main result (or results) from the figure.

What to include in your report, con't

- **Discussion** - Discuss biological significance of findings
  - Investigate newness of findings - has anyone observed these results before?
  - Refer to external sources.
- **Conclusion** - Answer the question(s)
  - restate questions & insert answers; pretend reader skips everything else; summarize briefly

Quality - if other people must read it, make sure it's your top work!

- Few or no spelling or grammar errors, formatting errors
- Language
  - Terms used correctly?
  - Clear, easy to understand prose?
- Content (figures, text) relevant & complete?
  - Should be obvious why each figure, text is included
  - Are answers well-supported?

Quality tips

- Knit your Markdown early and often
  - Proof-read it
  - Fix errors
  - Repeat
- Fix typos using the RStudio spell-checker
- Show it to the TA before you turn it in.
  - Get suggestions for improvement
- Ignoring the code, does it still make sense?

More quality tips

- Always turn in your best effort
  - Edit and re-knit multiple times
    - Remember: In "real life," sloppy work gets you fired.
  - Review the slides to check that you followed instructions.
    - You will lose ***many*** points if you missed something obvious.
- Professor Loraine will read your work and give you personalized feedback.
  - This is the real value you're getting from your tuition money - personalized help from a highly experienced scientist.
  - If you turn in your best possible effort, she'll give you advice & feedback to help you improve.
    - Don't waste her time & attention with crap work.