Assesment_R_Programming_Basics

Igor Ruiz de los Mozos

Assessment: `biomed_data.csv` Exploration & Analysis

Instructions:
- Create a new R script or R Markdown document named `STUDENT_NAME_biomed_analysis.Rmd`. This could be a Word transformed to PDF but I encourage you to learn markdown.
- For each question below, include (a) the R code you used, (b) any plots or summary tables, and (c) a brief written interpretation (2–4 sentences).
- Knit your R Markdown to HTML or PDF and submit the rendered document.

---

1. Linear Modeling & Interpretation

0. **Prepare enviroment and load dataset**

```
# Set your working directory once per session
getwd()
```

```
[1] "/home/rstudio"
```

```
setwd("/home/rstudio")

# Install required libraries once:
install.packages(c("tidyverse", "viridis", "GGally"))
```

```
Installing packages into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)
```

```r
# Load & packages
library(readr)
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(tidyr)
library(ggplot2)
library(GGally)
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

```r
# Load & Inspect
biomed <- read_csv("/home/rstudio/biomed_data.csv", show_col_types = FALSE)
glimpse(biomed)      # columns & types
```

```
Rows: 200
Columns: 14
$ patient_id <chr> "P1", "P2", "P3", "P4", "P5", "P6", "P7", "P8", "P9", "P10"~
$ sex        <chr> "Female", "Female", "Female", "Female", "Male", "Female", "~
$ age        <dbl> 64, 27, 32, 43, 28, 44, 43, 31, 56, 41, 49, 59, 26, 61, 71,~
$ group      <chr> "Treated", "Treated", "Treated", "Control", "Control", "Con~
$ region     <chr> "North", "North", "North", "South", "South", "North", "Sout~
$ dose       <chr> "Low", "High", "High", "Low", "Medium", "High", "High", "Me~
$ marker_A   <dbl> 5.64, 3.61, 3.87, 3.86, 4.31, 5.17, 5.59, 4.18, 2.14, 5.95,~
$ marker_B   <dbl> 14.07, 11.44, 11.98, 13.23, 13.29, 10.17, 14.07, 8.11, 7.78~
$ marker_C   <dbl> 49.52, 46.27, 51.05, 45.76, 53.96, 50.15, 53.02, 45.03, 48.~
$ marker_D   <dbl> 1.23, 1.81, 1.91, 1.12, 0.99, 1.24, 1.54, 1.31, 0.79, 1.04,~
```

```
$ marker_E   <dbl> 11.38, 11.62, 10.65, 10.73, 9.75, 8.69, 10.27, 10.29, 9.00,~
$ expression <dbl> 8.66, 12.54, 11.99, 11.27, 8.67, 10.80, 8.90, 8.97, 6.92, 1~
$ heart_rate <dbl> 69, 88, 84, 64, 91, 100, 64, 85, 92, 91, 99, 99, 94, 94, 63~
$ RBC_count  <dbl> 4.47, 4.38, 3.99, 4.02, 5.32, 4.45, 4.73, 4.74, 4.95, 4.41,~
```

```
summary(biomed)      # summary stats & NAs
```

```
  patient_id            sex                 age           group
 Length:200         Length:200         Min.   :25.00   Length:200
 Class :character   Class :character   1st Qu.:38.00   Class :character
 Mode  :character   Mode  :character   Median :49.00   Mode  :character
                                       Mean   :50.09
                                       3rd Qu.:61.25
                                       Max.   :75.00

    region              dose              marker_A        marker_B
 Length:200         Length:200         Min.   :2.140   Min.   : 6.72
 Class :character   Class :character   1st Qu.:4.300   1st Qu.:11.10
 Mode  :character   Mode  :character   Median :4.900   Median :12.58
                                       Mean   :4.896   Mean   :12.43
                                       3rd Qu.:5.525   3rd Qu.:13.89
                                       Max.   :7.350   Max.   :18.09
                                       NA's   :5
    marker_C        marker_D         marker_E        expression
 Min.   :35.35   Min.   :0.1400   Min.   : 5.850   Min.   : 4.000
 1st Qu.:46.63   1st Qu.:0.8875   1st Qu.: 9.265   1st Qu.: 7.280
 Median :49.93   Median :1.1200   Median :10.325   Median : 8.680
 Mean   :49.89   Mean   :1.1838   Mean   :10.183   Mean   : 8.876
 3rd Qu.:52.98   3rd Qu.:1.3550   3rd Qu.:11.110   3rd Qu.:10.340
 Max.   :62.26   Max.   :2.9400   Max.   :15.040   Max.   :16.640
 NA's   :5                                         NA's   :5
   heart_rate       RBC_count
 Min.   : 57.00   Min.   :3.550
 1st Qu.: 67.00   1st Qu.:4.265
 Median : 77.00   Median :4.550
 Mean   : 78.89   Mean   :4.554
 3rd Qu.: 91.00   3rd Qu.:4.860
 Max.   :103.00   Max.   :5.490
                  NA's   :5
```

1. **Fit a multiple linear regression**

```r
lm_fit <- lm(expression ~ marker_B + dose + group, data = biomed)
summary(lm_fit)
```

```
Call:
lm(formula = expression ~ marker_B + dose + group, data = biomed)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7909 -1.3004  0.0164  1.4084  6.0033

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.94424    0.87530  10.218  < 2e-16 ***
marker_B      0.04227    0.06624   0.638 0.524086
doseLow      -1.50361    0.35446  -4.242 3.46e-05 ***
doseMedium   -2.10358    0.35634  -5.903 1.61e-08 ***
groupTreated  1.15258    0.29319   3.931 0.000118 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.039 on 190 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.2251,    Adjusted R-squared:  0.2088
F-statistic:  13.8 on 4 and 190 DF,  p-value: 6.719e-10
```

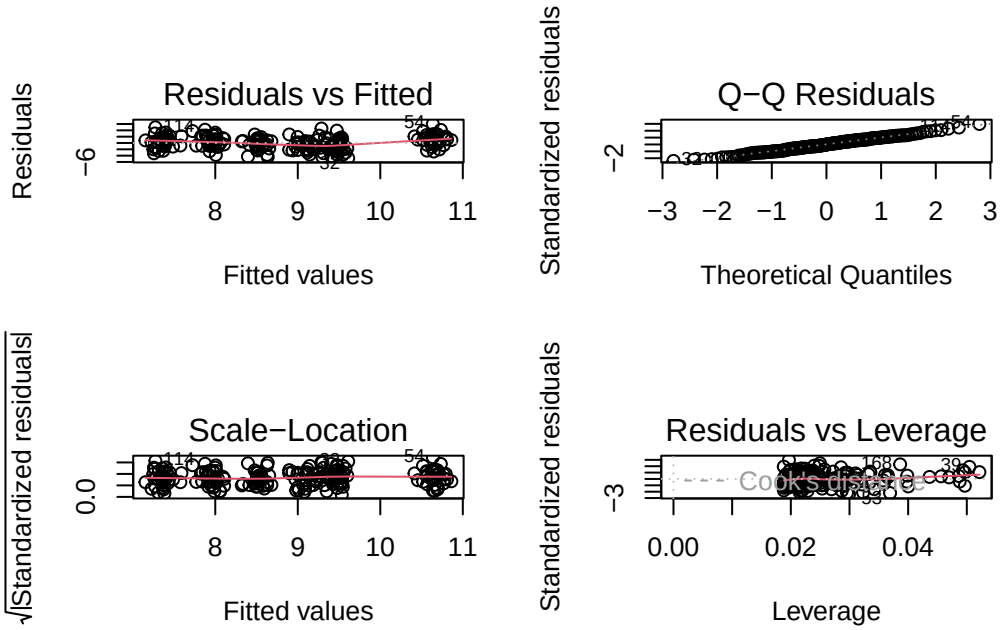
Report the estimated coefficient for marker_B, doseMedium, doseHigh, and groupTreated.

Interpret each coefficient in the context of how expression changes per unit increase in marker_B and between dose/group categories (holding other variables constant).

Assess model assumptions

Produce the 4 base R diagnostic plots:

```r
par(mfrow = c(2,2))
plot(lm_fit)
```



Identify any potential outliers or non-constant variance. Which plot indicates this, and what would you do next?

2. Two-Way ANOVA & Interaction

Fit a two-way ANOVA

```
aov_fit <- aov(expression ~ group * dose, data = biomed)
summary(aov_fit)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
group         1   68.2   68.20   19.68 1.55e-05 ***
dose          2  159.5   79.74   23.01 1.14e-09 ***
group:dose    2  136.1   68.07   19.64 1.78e-08 ***
Residuals   189  655.1    3.47
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
5 observations deleted due to missingness
```

Report the F-value and p-value for the group:dose interaction term.

Interpret whether there is a statistically significant interaction between treatment group and dose on expression.

Post-hoc comparisons (optional, bonus)

Use TukeyHSD(aov_fit) to examine pairwise differences. Which combinations of group × dose differ significantly?

3. Missing-Value Imputation & Re-visualisation

Impute marker_C with its median

```
med_C <- median(biomed$marker_C, na.rm = TRUE)
biomed_imputed <- biomed %>%
  mutate(marker_C = ifelse(is.na(marker_C), med_C, marker_C))
```

Re-plot the histogram of marker_C before and after imputation side by side (use gridExtra::grid.arrange or patchwork).

Comment on how the distribution changed.

4. New Factor & Comparative Boxplots

Create an age_group factor

```
biomed2 <- biomed %>%
  mutate(age_group = cut(age,
                         breaks = c(-Inf,35,60,Inf),
                         labels = c("<35","35-60",">60")))
```

Boxplot heart_rate by age_group with points overlaid (geom_jitter).

Describe any trends you observe (e.g., does heart rate vary by age group?).

5. Tidying & Faceted Violin Plots (Challenge)

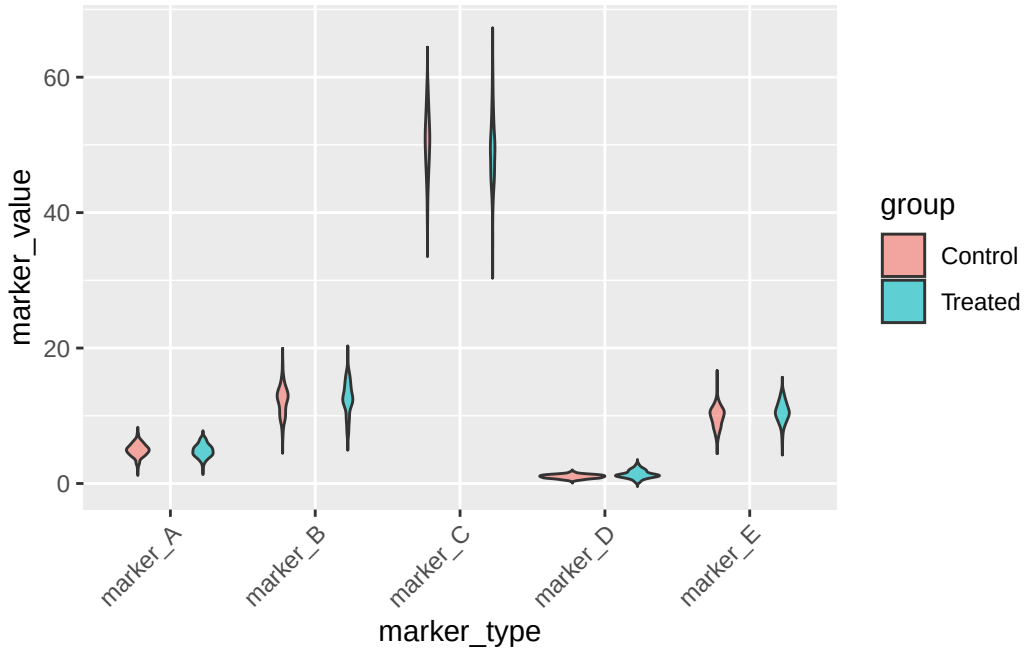
Pivot all marker_ columns longer

```
biomed_long <- biomed %>%
  pivot_longer(cols = starts_with("marker_"),
               names_to = "marker_type",
               values_to = "marker_value")
```

Create violin plot of marker_value by marker_type, colored by group.

```r
ggplot(biomed_long, aes(marker_type, marker_value, fill = group)) +
  geom_violin(alpha = .6, trim = FALSE) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
Warning: Removed 10 rows containing non-finite outside the scale range
(`stat_ydensity()`).
```



Explain which markers show the largest differences between Control and Treated.

6. Extended Exploratory Questions for `biomed_data.csv`

Use these prompts to guide deeper, self-driven investigations of the mock cohort. Pick a handful for your project or discussion. *Choose at least five of these prompts to pursue in your final report. For each, provide code, output (tables/plots), and a brief biological or statistical interpretation.*

1. **Multivariate Correlations**
    - Which pairs of biomarkers (A–E) show the strongest positive or negative correlations? How does this change when stratified by treatment `group` or `dose`?
2. **Predicting Expression**

- Build a multiple regression (or regularized model) predicting `expression` from all five `marker_*` variables plus `age` and `sex`. Which predictors are most important?

3. **Dose–Response Curves**
   - For each `group` (Control vs Treated), plot the mean ± standard error of `marker_D` across `dose` levels. Do dose–response curves diverge between groups?
4. **Age Effects and Interactions**
   - Create an `age_band` (< 35, 35–60, > 60) and test whether the effect of `dose` on `expression` differs across age bands (three-way interaction: `expression ~ group * dose * age_band`).
5. **Regional Variation**
   - Compare biomarker distributions among the four `region`s using box-plots and one-way ANOVA. Which region shows the highest mean `marker_C`, and is it statistically different?
6. **Missing-Data Sensitivity**
   - Impute missing values in `marker_A` and `marker_C` using median, mean, and k-nearest neighbors. Recompute a key summary (e.g., mean of each marker by `group`) and describe how results vary by imputation method.
7. **Patient Clustering**
   - Perform hierarchical clustering or k-means on the numeric biomarkers. How many clusters "make sense"? Do clusters align with `group`, `dose`, or `region`?
8. **Principal Component Analysis (PCA)**
   - Run PCA on the five `marker_*` columns. Plot PC1 vs PC2, coloured by `group` and sized by `dose`. What biological patterns emerge on the principal component axes?
9. **Outlier Investigation**
   - Identify patients whose `expression` lies > 3 SDs from the mean. Examine their full profiles—do they share any common characteristics (age, group, dose, region)?
10. **Time-to-Event Simulation** *(Bonus)*
    - Assume `expression` is a surrogate for time to clinical response. Simulate a binary outcome (`response = expression > threshold`) and fit a logistic regression. Interpret odds ratios for `marker_B` and `dose`.
11. **Model Diagnostics**
    - For your `lm(expression ~ marker_B + dose + group)` model, generate residual vs fitted, QQ, and Cook's distance plots. Which observations drive the model fit?

12. **Interaction Visualization**
    - Use `ggplot2::geom_interaction()` or custom line plots to visualize any significant three-way interactions (e.g., `marker_E ~ group * dose * sex`).
13. **Heatmap of Biomarker Z-Scores**
    - Standardize each `marker_*` to Z-scores and plot a patient-by-marker heatmap (rows = patients, columns = markers). Cluster both dimensions to reveal co-expression modules.
14. **Effect of Sex**
    - Test whether the `group:dose` interaction on `expression` differs between `sex` by fitting separate models or including a three-way interaction: `expression ~ group * dose * sex`. Summarize the findings.
15. **Reproducible Reporting**
    - Package your entire analysis in an R Markdown report with clear section headers, narrative interpretations, and embedded code: set `code-copy: true` in the YAML and ensure every plot is accompanied by a concise caption.
16. **Correlation heatmap**
    - Compute the correlation matrix for all numeric columns (marker_A–marker_E, expression, heart_rate, RBC_count). Plot it as a heatmap (geom_tile) with a diverging color scale. Identify the strongest positive and negative correlations and speculate on their biological meaning.
17. **Group-wise summary table**
    - Produce a summary table showing, for each region, the mean and standard deviation of expression and heart_rate. Display it in R as a nicely formatted tibble, and export to CSV.
18. **Outlier detection**
    - For each numeric variable, compute how many observations lie more than 3 standard deviations from the mean. Discuss whether these outliers should be retained or investigated further.

Submission Checklist

R Markdown file (biomed_analysis.Rmd) with answers, code, and narrative.

Rendered HTML/PDF output with plots and tables.

Any export files (e.g., region_summary.csv) if requested.