STAT 230 Short Report 2

Due November 8th at 10pm (no penalty late submission at 2am on 11/9)

There are two options for this short report: understanding wine quality and predicting forest fire size. As with the first short report, you will need to do some background reading and here you will cite at least **three** peer-reviewed sources to give context to your study. The report will be graded with higher expectations of both statistical analysis **and** coherence/writing quality than the first report.

Wine Dataset

The goal of this analysis is to **understand factors that improve the evaluated quality of wine samples** based on known information about the wine. Input variables (listed below, based on physicochemical tests) will be used to help describe the output variable *quality* (score between 0 and 10). The score was the median of at least 3 evaluations made by wine experts. Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

- Fixed acidity (g(tartaric acid)/dm^3)
- Volatile acidity (g(acetic acid)/dm^3)
- Citric acid (g/dm^3)
- Residual sugar (g/dm^3)
- Chlorides (g(sodium chloride)/dm^3)
- Free sulfur dioxide (mg/dm^3)
- Total sulfur dioxide (mg/dm^3)
- Density (g/dm^3)
- pH
- Sulphates (g(potassium sulphate)/dm^3)
- Alcohol (vol.%)

You can read in the dataset with the following code:

```
wine_dat <- read.csv("https://www.math.carleton.edu/ckelling/data/wine_project.csv")
```

Forest Fire Dataset

For this dataset, you will develop an analysis to **predict the size of forest fires in the northeast region of Portugal based on environmental factors**. The response variable in this analysis is **area** - the burned area of the forest (in hectares). The list of potential explanatory variables present in this analysis are as follows:

- X - x-axis spatial coordinate within the Montesinho park map: 1 to 9
- Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9
- month - month of the year: 'jan' to 'dec'
- day - day of the week: 'mon' to 'sun'
- FFMC - Fine Fuel Moisture Code index from the FWI system: 18.7 to 96.20
- DMC - Duff Moisture Code index from the FWI system: 1.1 to 291.3
- DC - Drought Code index from the FWI system: 7.9 to 860.6
- ISI - Initial Spread Index index from the FWI system: 0.0 to 56.10
- temp - temperature in Celsius degrees: 2.2 to 33.30
- RH - relative humidity in
- wind - wind speed in km/h: 0.40 to 9.40

- rain - outside rain in mm/m2 : 0.0 to 6.4

Hint: If you want to log() transform the area (response) variable, try adding 1 to all responses first. This transformation should be well-documented if used. **Note** that this kind of transformation is not appropriate in all cases (e.g. when transforming percentages).

```
fire_dat <- read.csv("https://www.math.carleton.edu/ckelling/data/forestfires.csv")
```

Overall tips

When analyzing these datasets, be sure to do the following:

- Keep in mind the goals of the analysis (understanding vs prediction)
- Consider consolidating variables
- Check for interactions and the need for transformations
- To reach satisfactory depth of analysis, your final model should include at least three explanatory variables, unless fully justified that it is not necessary/appropriate. This includes quantitative and categorical variables.
- Check model assumptions often
- Check for collinearity and adjust models accordingly
- Check for outliers/influential points and only remove points for a justifiable reason
- Compare models
- Fully justify the final dataset and model using tools discussed in class

Report Guidelines

The main text should be **under 4 pages**, including figures. Your report should be organized and clearly written. It should contain accurate and precise language and justifiable interpretations of statistical evidence, and should have the components listed below. The report should be a summary of your analysis, for an informed audience more interested in conclusions than in technical details. This text should include the following clearly labeled parts. *Clearly labeled* means that each section should have a bold-font section title that corresponds to the labels given below. The percentages listed below indicate how the report will be graded (**see also attached Rubric**). You should include the RMarkdown appendices (code and knitted pdf) as separately attached documents.

- **Title and Author(s)** (5%): Your title should be an informative and engaging description of your report (i.e., not "Short Report").
- **Introduction** (15%): A brief introduction to the questions addressed. Also include any relevant information about the data and how they were collected, and background information as appropriate. Avoid making broad and unsubstantiated claims, and cite evidence for claims that are not common knowledge if applicable. Give enough background in the intro so that a reader doesn't need to refer to source material, but take care not to plagiarize your source material above. Cite sources as needed. You will need to cite *at least three peer-reviewed papers* to provide context and motivate for your analysis! Be sure to connect these papers to your current research questions.
- **Results** (50%): Both exploratory and formal statistical analyses. Include the following: (1) at least one relevant graphical exploratory display of your data (labeled appropriated, and described in text, *not involving a model*), (2) estimates and uncertainty quantification from your model, and (3) relevant interpretations in context. Before jumping into the model building process, you should start your Results section by introducing your dataset and performing some EDA to motivate your analysis. Technical language should be used correctly. This section should focus on conveying the *results* of your analysis to your reader and should not be a step-by-step description of what you did. For example, there is no need to tell the reader how many different transformations you tried before finally finding a linear association. Do not include screenshots of R tables- these should be formatted nicely in your report as tables.

- **Discussion** (15%): Summarize your findings, and describe any limitations of your analysis as well as questions that you would have been interested in answering but that are outside the scope of this short report. Be sure to include a healthy discussion of any necessary caveats here.
- **References** Include your references in any standard citation style. This does not count towards the 4-page limit.
- **R Markdown Appendices** (5%): Your main text should suppress R commands and output if you write it in RMarkdown. In addition to your report, you will submit your Rmd document and your knitted code as separate documents. R code should be appropriately commented so that the technical details of your analysis are clearly communicated and the code should approximately follow the order of your report.
- (Optional) Appendix: You are welcome to include *non-core* material in an appendix. If you have an appendix, it should be formatted with clarity (eg. complete sentences). Your report should not be overly reliant on this material. It does not count towards the page count.

Finally, 10% of the grade is based on **overall presentation and group evaluation**: is the report structured according to the instructions given above? Is the text neatly presented, broken into readable paragraphs and laid out in a coherent manner? An excellent report will be written so that a reader with basic statistics training can understand your data, your analysis, and your conclusions. Do not simply write a step-by-step description of what you did in R. Don't include unnecessary output. At the conclusion of the project, I will ask you to evaluate your group members. I expect satisfactory peer evaluation showing that you contributed to the assignment.

In describing your conclusions, be sure to pay attention to the study design and be clear about what can be learned from these data (e.g., sampling bias, causality, etc.). You should answer a question that you are interested in, regardless of what the answer ends up being; in particular, a result does not need to be "statistically significant'' for the analysis to be valid and interesting!

Best Practices for Partnered Work

Partners should work on this report *together*. While you are permitted to do some work separately (e.g., revising a section at a time when your schedules don't permit meeting together), I discourage you from completing substantial portions of the work individually. Each partner should contribute substantively to each component of the report, and each partner is assumed to be an equal contributor to the final product. You should be in regular contact with your partner(s) and should arrange at least two meetings to discuss your work. If you are having issues communicating with your partner(s), **please email me as soon as possible.**

My suggested timeline is as follows:

- **Week 1:** Meet as a group to explore data, decide a dataset, solidify research questions, and develop plan for preliminary model-building. Conduct exploratory model building steps.
- **Week 2:** Early in Week 2 of the project, finalize your model and related plots. Begin writing at the latest mid-way through Week 2 to allow for enough time for feedback and editing between group members. The writing of this report should be coherent and smooth and **will be graded with higher expectations than the first report.**

Academic Honesty

Just a reminder, but submitted work is assumed to be by the author(s) unless proper attribution is made. I do not expect you to do significant outside research for this assignment (other than the three required sources for context/motivation), but remember to cite your sources that you use (using any standard citation method). I expect that the primary intellectual contribution to the project is from the authors. One example of permissible outside collaboration is receiving help formatting a graph from the stats lab. One example of impermissible outside collaboration is receiving extensive guidance on how to construct a confidence interval for the quantity you're interested in estimating. Please come to me if you need assistance for this project outside of your assigned group.

Title and Author(s) (5%)

- Has a title and lists the authors
- The title is an informative and engaging description of your report

---

Introduction (15%):

- Introduces and motivates the topic being addressed
- Includes at least three peer-reviewed sources for context/motivation for this study that are well-connected to your questions. Use in-text citations for your peer-reviewed citations (eg. Kelling 2024)
- Explains the questions addressed in the report
- Gives appropriate background information (and avoids making broad and unsubstantiated claims)

Results (50%):

- Focus is on conveying the *results* of the analysis and is not a step-by-step description
- Exploratory graphical summary of data (not involving a model- just EDA)
  - Figure(s) is/are informative about the question addressed in the report
  - Figure(s) is/are labeled appropriately, and described in text
- Statistical inference
  - A complete model statement is given (give the equation!)
  - An appropriate model is chosen
  - Model assumptions are assessed and satisfactory
  - Estimates and uncertainties of model coefficients are given and *both* are interpreted in context
  - The model and estimates are interpreted correctly and in the context of the question being addressed
  - Technical language is used correctly

Discussion (15%)

- Summarizes the main findings and answers the question posed in the introduction
- Describes any limitations of your analysis and any necessary caveats
- Describes any follow-up questions that would be interesting to investigate

---

R Markdown Appendices (separate documents) (5%)

- R code used for the analysis is shown and implemented correctly

- R code is approximately ordered as the results are presented in the report
- Code is commented in a way where results are easy to understand and reproducible

---

Overall presentation (10%)

- A separate references section is included, with any standard citation style
- The report is structured according to the instructions
- The text is neatly presented, broken into readable paragraphs and laid out in a coherent manner
- Peer evaluations were completed for group members and satisfactory from group members