Data100 Final Project

Introduction

Using any data set(s) that interest you, demonstrate that you've mastered all of the skills in the class!

- Groups of 5
- A single pdf, with code, data, and plot output, submitted to MyLS
- *Strict 10 page limit!* I will not read past the 10th page (not including cover page).
- Use the checklist and structure below to ensure that none of your elements are missed during grading.

Choosing a Data Set

I highly recommend checking out tidytuesday and Kaggle. Look for the following features when choosing a data set:

- There is a **continuous** variable that should clearly act as the target variable (y).
- There are at least 5 features
  - At least one is a **factor** (categorical)
  - At least two are **continuous**
- Some pre-processing would benefit the analysis (especially **regular expressions**).

Code Checklist

All of the following functions/concepts must be used *appropriately*. For example, pivoting something without reason will not check the box.

☐ `pivot_wider()` and/or `pivot_longer()`

☐ `geom_point()`
☐ `geom_histogram()` and/or `geom_density()` and/or `boxplot`, `violin`, or `density_ridges`
☐ `geom_bar()` and/or `geom_col()`
☐ `group_by()` and `summarise()`
☐ `mutate()`
☐ A custom function and/or a `for`/`while` loop
☐ Some sort of regular expression in the data cleaning.
☐ `tidymodels()`, including `workflow()`, `linear_reg()`, and `set_engine()`, and `fit()`
☐ `predict()`
☐ Use a factor variable in a linear model correctly (be very careful!)
☐ Evaluate whether a factor predictor should be included in a linear model
☐ `augment()` and/or `tidy()`

Structure

Following the exact structure below makes grading much easier, and happy graders give higher grades!

1. Introduction
    - Explain the general context of the analysis
2. Goals/Research Question
    - Explain the specific purpose of this analysis
3. Basic Data Cleaning
    - Code should be only what is needed for the modelling
    - Explain the code as if the code isn't there.
    - (**Option 1**) Advanced data cleaning: Do something above and beyond mutate and pivot. Should be as much work as if you did Option 2 below (you only have to do one of the two options).
4. Exploratory plot/table 1
    - Only use the features, *not* target.
    - Either a plot or a table (don't have to do both).
    - Plot should show an interesting relationship among the features that will be relevant to the modeling.
5. Exploratory plot/table 2
    - A different plot/table as above, but same requirements
6. Model Plot 1 (target versus relevant features)

- Plot should show an interesting relationship between the target and at least two of the features (e.g., use the `colour` or `fill` aesthetics).

7. Model Plot 2
    - A different plot as above, but same requirements
8. Exploratory linear model 1
    - Model should be a guess at the final model, based on the plots above.
    - Check the diagnostics!
    - Investigate minor changes, such as adding/removing predictors that you aren't sure about.
    - Do *not* simply check every combination of predictors! Use your knowledge of the context to guide your decision making!
    - Explain your reasoning. This class doesn't go deep into choosing models, so I don't expect perfect answers, but I expect you to put some real thought into it.
9. Exploratory linear model 2
    - Should be a different guess at the final model (e.g., using a different set of predictors that capture similar patterns, as found in the plots above).
    - Same requirements as the first model. Ideally, you'll use the same `workflow()` as the first model.
10. Final linear model, with diagnostics and interpretations
    - Choose between the two models above and investigate other possible relationships.
11. (Option 2) Advanced modelling
    - Spline terms, random forest, `mlp`, etc., with `parsnip`
    - Include diagnostics and interpretations
12. Conclusions
    - Relate the whole analysis back to the context of the data. Simply stating/interpreting the model parameters is not enough; what can you say about the broader context of the study?
13. Limitations
    - Being honest about the limitations in the data/model makes the conclusions more trustworthy.