Project Goal

To prepare the student for SOA Exam PA in the context of regression and time series analysis. Learning objectives from SOA Exam PA that this project supports are listed below and the grading rubric is taken/adapted from the SOA Exam PA rubric.

1b Predictive Analytics Problems and Tools: Write and execute basic commands in R using RStudio.

2a Problem Definition: Translate a vague question into one that can be analyzed with statistics and predictive analytics to solve a business problem.

3a Data Visualization: Understand the key principles of constructing graphs.
3b Data Visualization: Create a variety of graphs using the ggplot2 package.

4b Data Types and Exploration: Identify the types of variable and terminology used in predictive modeling.
4e Data Types and Exploration: Apply univariate and bivariate data exploration techniques.

5b Data Issues and Resolutions: Identify opportunities to create features from the basic data that may add value. See discussion of "features".
5c Data Issues and Resolutions: Identify outliers and other data issues.
5d Data Issues and Resolution: Handle non-linear relationships via transformations.

6a Generalized Linear Models: Implement ordinary least squares regression in R and understand model assumptions
6d Generalized Linear Models: Interpret model coefficients, interaction terms, offsets, and weights.
6f (Optional) Generalized Linear Models: Explain the concepts of bias, variance, model complexity, and the bias-variance trade-off.

7a (Optional) Decision Trees: Understand the basic motivation behind decision trees.
7b (Optional) Decision Trees: Construct regression and classification trees.

9a Communication: Develop and justify a recommended analytics solution.
9b Communication: Communicate in a clear and straightforward manner using common language that is appropriate for the intended audience.
9c Communication: Structure a report in an effective manner.
9d Communication: Follow standards of practice for actuarial communication.

Project

1. Select a data set from the list provided (other data sets may be acceptable if approved in advance). Identify an appropriate audience/client for your research project. Determine a reasonable/appropriate research question that a regression/time series analysis on your data could answer. Submit as a Word document in Canvas. (5 pt check point)
2. Make any revisions to your problem statement necessary and add a project plan with deadlines to your document. (5 pt check point)
3. Perform the analysis, document your analysis process/results and your work progress, create notes on how to perform the given analysis. You will submit updates on the work accomplished including a self-evaluation of the quality of your work and your ability to manage the project. See Canvas for a discussion of these updates. (5 pt each)
4. Write a report (see specifications below and in rubric). (275 pts)
5. Prepare a 10 to 12 minute oral presentation (you will only present one of the two projects). (75 pts)

Project Deliverables

- Deliverables for SOA Exam PA:
  - Prepare a report that presents and supports your solution. The report will target non-technical decision-makers with an appendix targeting content experts outlining and summarizing techniques tried
    * The written report should be organized with the following sections (failure to do so will incur significant penalty): Executive summary, Introduction, Data exploration/preparation/cleaning/feature selection, Model selection/interpretation and validation, Findings/summary/conclusion, Appendices
    * More information on what should go into each section is given in RM Ch 20 as well as an example of a written report.
  - SOA deliverables will include your report and your R code plus any other information or files you believe will support your solution.
    * R code should not include everything you tried, but be a clean presentation of the pertinent pieces of analysis.
    * R code should be organized in such a way that it can be run and easily mapped to the work presented in the reports.
- Additional Written Deliverables:
  - Work log - you will create a project/analysis plan (with deadlines) and document when work is done each week that accomplishes your goals (and if you meet your deadlines)
    * This is intended to help you learn to log work that may be billed to a client as well as to set and meet deadlines to manage long term projects effectively.
    * Organize the work log according to the sections of the written report (you should also include a writing component for each section). Include dates for each time work as well as total time spent for each section and the project in total.
    * This is probably most efficiently done in a table/Excel file. Complete sentences and such are not expected.
  - How-to log - documents how you go about performing the analysis, how to run things in R, how to interpret the results, what to watch out for to avoid problems/mistakes, etc.
    * This can be developed and organized any way you want (typed, handwritten, Word, OneNote, etc.). However there should be a clear organizational structure that is easy to follow and minimal use of white space (if typed, don't double space, don't have excessive indentations, etc.). Don't be vague, add detail and commentary that will be easy to follow in the future.
- Oral Deliverable:
  - Oral presentation - for one (not both) project, 10 minutes in length, summarizing your research question and your findings/solution.
    * Slides or visual aids of some kind are required. However, DO NOT READ THEM to us.
    * Your oral presentation grade will be based solely on the quality/clarity of the presentation you give (not the quality of your data analysis).

Grading Rubric Exam PA Written Submission $\sim$ 250 Points for Regression Project, 150 Points for Time Series Project

Communication (35%)

- Executive summary – clearly and concisely written summary that is appropriate for someone who reads nothing else
- Problem statement – clearly defines the problem and its business context
- Use of tables and graphs – clearly constructed, labeled, and referenced
- Interpretation of model results – relates the results of the modeling process to the problem statement
- Audience – sections tailored to the audience as described in the project statement
- Code – easy to follow, using intuitive variable names and sufficient comments

Data Exploration and Feature Selection (20%)

- Description of the data – summary statistics and graphs with interpretation
- Identification of issues and corrective steps – includes handling missing data and possible transformations
- Selection of features for use in the model – includes creating new features through transformations, clustering, or principal component analysis as appropriate.
- Code – successfully runs and produces output presented in the report

Model Selection and Construction (45%)

- Selection and justification of model type – relates model choice to the business problem and the available data
- Estimation of model parameters, with explanation – calibrates the selected model, including selecting features from the list previously established
- Validation of the selected model – documents that an appropriate validation method was used and provides an estimate of model accuracy using previously unseen data
- Description of selected model – describes the model in appropriate terms for the stated audience
- Code – successfully runs and produces output presented in the report

Grading Rubric Additional Written Submission $\sim$ 75 Points possible for each project

Worklog (40%)

- Log is well organized/easy to read.
- Log contains all required elements
  - Include sections for Executive summary, Introduction, Data exploration/preparation/cleaning/feature selection, Model selection/interpretation and validation, Findings/summary/conclusion, Appendices as well as Writing the sections
  - Include dates for each time work (mapped to the appropriate section) as well as total time spent for each section of the report and the project in total

How-to Log (60%)

- Clear organizational structure that is easy to follow, minimal use of white space, effective formatting
- Specific detail and commentary that will be easy to follow to reproduce a regression/time series analysis
- Contains all elements of a thorough analysis/validation, including notes on how to run stuff in R, interpret results, etc.

Grading Rubric Oral Presentation ∼ 75 Points possible for one project

Oral Presentation

- Slides should be uploaded to Canvas prior to the presentation.
- Professional personal presentation
- Slides/visual aids are used effectively
  - Neat, easy to follow, employ good choices regarding coloring/formatting/spacing to promote your message
  - Slides are used to complement the presentation (they are NOT read)
  - Your interaction with the slides is comfortable and effective
- Tone of presentation is comfortable and conversational (not tense/awkward); interactions (nonverbal possibly also verbal) with audience exist throughout the presentation and are comfortable
- Efficient/effective/appropriate use of time to communicate clearly
- Clearly evidence that you understand what you are talking about, respond comfortably to any questions.