Regression and Analysis of Variance STAT 3340 / MATH 3340
Fall 2020
Final Project

Weight (% of final grade)**:** 25%
Due Date: 11:59 ADT December 11th, 2020

Motivation:
Upon successful completion of this project, students will possess a working knowledge of Github and R Markdown, and mastery in the practice of regression analysis. These skills are highly valued globally by employers in search of data scientists.

Project Description:
Each group will be assigned a dataset. Collectively, group members are to perform a complete regression analysis of their data, details of which must be presented on **Github** (https://github.com) using **R Markdown** (https://rmarkdown.rstudio.com/articles_intro.html).

The following sections must be included:
Abstract (150 words or less)
Introduction (must contain a thorough description of the questions of interest)
Data Description (must contain data visualizations that are properly labelled and explained)
Methods (must contain a complete description of all analysis tools used)
Results (all figures should be properly labelled and discussed)
Conclusion (must contain a concise discussion of what has been learned from the analysis)
Appendix (must include all data and R Markdown files for reproducibility)

Data:
Datasets are found at https://lionbridge.ai/datasets/10-open-datasets-for-linear-regression/. Groups 1-5 are to analyse Dataset 1 (Cancer), Groups 6-10 are to analyse Dataset 2 (CDC) , Groups 11-15 are to analyse Dataset 3 (Fish Market), Groups 16-20 are to analyse Dataset 4 (Medical Insurance), Groups 21-25 are to analyse Dataset 5 (New York Stock Exchange), Groups 26-30 are to analyse Dataset 7 (Real Estate), Groups 31-35 are to analyse Dataset 8 (Red Wine), Groups 36-40 are to analyse Dataset 9 (Vehicle), Groups 41-45 are to analyse Dataset 10 (WHO).

Note: Before commencing your analysis, **you must introduce one new additional data point** into your assigned dataset. A description of this unique data point must be included in your *Data Description* section along with some rationale for the values chosen.

Grading Scheme:
6 Overall presentation and organization of materials
3 Quality of data visualizations
6 Correctness of analysis
4 Quality and selection of relevant figures
6 Interpretation of results
--
25