SCMA 632. Statistical Analysis and Modeling. 3 Credit Hours.
Start & End Date: June 28, 2021 & August 14, 2021.
Material is uploaded on the CANVAS course page https://canvas.instructure.com/courses/2967284

Part I: SYLLABUS

Class timings:

Theory & R program implementations: Dr. Lalith Achoth, lalithuas@gmail.com
Monday 2.00-4.15PM (15minutes break)
Friday 2 PM-4.15 PM (15 minutes break)
Note: For Theory and R programming related doubts students must contact: Dr. Lalith Achoth

Python Implementation Lab: Dr. K.B. Vedamurthy, vedandri@gmail.com
Thursday 5.00-7.15PM (15minutes break)
Saturday 2.00-4.15PM (15 minutes break)
Note: For Python-related doubts, students must contact Dr. K.B. Vedamurthy

Teaching Assistants: Ms. Sanjana Athiyedath, scma632@gmail.com
Students should submit assignments to Ms. Sanjana before the deadline and approach her for any queries regarding marks.

Platform: WebEx
Student Representatives: Shivani Kolluri (Pl email and coordinate with instructors – need to (i) create a WhatsApp group; (ii) setup WebEx; (iii) help students in installing latest R Studio/ Python Anaconda as well as resolve GitHub issues ; and coordinate any changes in class timings due to exigencies)

Course Description:

This will be a **hands-on course using live/Real datasets** and it focuses **on automation using R & python** using the **contemporary method** and **equal emphasis on the interpretation** of the results. Topics covered have an applied focus and may include logistic regression, bootstrap estimation, permutation tests, categorical data analysis, model selection, sparse methods, and Bayesian methods. Statistical analysis of data will be conducted using business-oriented computational software. While all the statistical analysis is done using R package, Python labs will run parallel. In Python lab, students besides learning python, will implement handpicked statistical analysis covered in this course.

Learning Outcomes:

At the end of the course, students are expected to access and clean data, describe its distribution, estimate its parameters, and estimate the strength of the linear relationship between two or more variables. While doing so, students will learn to fit the best model for the given data & the purpose at hand; students learn how to build a model using the training dataset and validate the model using the validation datasets. Students also become familiar with codes in R, Python and working with GitHub workflow

Topics Covered:

1. **Data cleaning and manipulation**
   Treatment of missing values; Indexing data; finding the most appropriate empirical distribution to the given dataset, computing descriptive statistics using R & Python; detecting and treating outlier and data visualization.

2. **Inferential statistics, Testing Population, Sample, and estimation**
   State and test distribution parameters hypotheses; Construct confidence intervals to convey the reliability of estimates; State and test hypothesis on whether the parameters differ from one sample to the other.

3. **Applications of Regression Models**
   Perform Regression analysis to identify factors influencing an independent variable; diagnostic checking of the regression like outlier detection in Y and X's and; estimate the price and income elasticity of demand using regression coefficients.

4. **Introduction to Panel data regressions and their use.**
   Fit a regression model to a panel data set; interpret panel data regression results

5. **Applications of Limited dependent variable models logistic, probit, and Tobit regression.**
   Use Logistic regression to find factors that influence limited dependent variable; use Probit/Tobit regression and inverse Mills ratio to handle zero values in the data

6. **Conjoint analysis**
   Identification of critical attributes and defining their levels; developing concept cards and collecting data; implementing conjoint analysis in R/Python and interpreting the results.
7. **Time Series Analysis**
   Estimating trend and seasonality in the time series data; fitting ARIMA process and forecasting. Exploring other models such as Artificial Neural Network(ANN) and Facebook profit; Estimating interrelationships between two or more time-series data sets using transfer function; Estimating risk using Value at Risk, ARCH/GARCH. Explore applications of vector error correction, Granger causality, and the impulse response function.

Required Textbook:

Business Analytics: The Science of Data-Driven Decision Making by U Dinesh Kumar, Wiley.

Grading System:

Final grades will be based on performance evaluation, assignments, and the final exam.
Weights will be applied as follows:

Weekly Assignment	70%
Final Exam	30%

The following grading scale shall be used:

Marks Obtained (%)	Grade
≥ 90%	A
≥ 80% and < 90%	B
≥ 70% and < 80%	C

Weekly Assignments

The instructor will demonstrate how to perform analysis on a database in R. Students should perform the same analysis for the data set assigned to them and submit it before the deadline.

Examinations

The final exam will cover both theory and problems. Problems will include coding in R and Python

Attendance Expectations

Attendance Compulsory

Software

The course requires R Studio and Python anaconda packages as well as working with GitHub. Students should have a laptop during every class.

VCU Honor System: Plagiarism and Academic Integrity

Submit your work. Do not submit anyone else's work. You may be randomly called to run the code and explain how you did the analysis.