SCMA 632. Statistical Analysis and Modelling. 3 Credit Hours

Start & End Date: June 3, 2024, & July 28, 2024.

Course Material is uploaded on the CANVAS course page.

SYLLABUS

Class timings:

Theory & R program implementations: Dr. Lalith Achoth, lalithuas@gmail.com
Wednesday and Thursday 7:00-9:15 PM (15 minutes break)
Note: For Theory and R programming related doubts students must contact: Mr. Manojkumar Patil

Python Implementation Lab: Dr. K.B. Vedamurthy, vedandri@gmail.com

Tuesday 7:00-9:15 PM (15 minutes break)
Saturday 2:00-4:15 PM (15 minutes break)
Note: For Python-related doubts, students must contact Dr. K.B. Vedamurthy

Teaching Assistant: Mr. Manojkumar Patil, scma632@gmail.com
Students should submit assignments to Mr. Manojkumar Patil through Canvas before the deadline and approach him for any queries regarding marks.

GitHub Profile link: https://github.com/scma-632

Platform: Please find the Teams meeting links for the following days:

- Tuesday, Wednesday, and Thursday: TU,W,TH Link
- Saturday: Sat Link
- Sunday: Sun Link

Student Representative: Adarsh Bhardwaj (should email and coordinate with instructors)

Course Description:

This will be a hands-on course using live/Real datasets, and it focuses on automation using R & Python, employing contemporary methods of analysis with equal emphasis on interpreting the results and testing the hypothesis. Topics covered will have an applied focus, including data handling, cleaning, anomaly detection, visualization, and data summarization & description. Further, the students will be exposed to analytical techniques such as regression analysis, logistic regression, categorical data analysis, and Time series analysis. Determining the optimum model selection, including using machine learning algorithms to model data like the random forest, market basket analysis, etc. The analysis of data will be conducted using business-oriented problems. Emphasis will be on interpretations of the results and making recommendations.

Learning Outcomes:

At the end of the course, students are expected to be proficient in accessing and cleaning data, describing its distribution, estimating its parameters, and choosing and estimating the appropriate models to analyze data. While doing so, students will learn to fit the best model for the given data & the purpose at hand; students build a model using the training dataset and, validate the model using the validation datasets and try to automate repetitive calculations and interpret the results. It focuses on all three descriptive, predictive, and prescriptive statistics domains. Students learn to analyze data using analytics tools and gather business insights. Students will understand the process of data-driven decision-making to make more informed business decisions as well as, test whether the analyses confirm the hypothesis. Students will apply ML techniques to solve business problems.

Course Instructions:

1. Download the required packages onto your local machine:
    a. Download R and R studio https://posit.co/download/rstudio-desktop/
    b. Download Anaconda https://www.anaconda.com/download/success
2. You have to solve assignments both in R and Python. For each assignment, you will be provided with a dataset. You will have to perform the following tasks:
    a) Run the codes on your desktop in R studio and in Python.
    b) The submitted assignments should have a cover page containing the following: name of the institution (font size 20 and in caps) - course name (font size 20 and sentence case) - assignment number – topic (font size 16 and in caps) - name and id number (font size 15 and in caps) - date of submission (font size 15 at the bottom of the page) - the second page must be the content page - the body of the report should contain the introduction, results, interpretation, and recommendation, as well as the codes - in the introduction, students should write the exercise's objective and its business significance - then the students should present the results and interpret them - ultimately, students should write the implications and list a few recommendations - students should submit the codes - the report must be in pdf format - please note that you should save both the report's results and codes. Upload the assignment to Canvas Course Page. And push codes to GitHub Repo.
    - Refer this link to see submission guideline https://github.com/SCMA-632/assignment-submission-guidline
    - Submit Code in Notebook format (. ipynb)
    c) Students can clarify their doubts during the coaching and doubt-solving sessions. In addition, students can email the questions to scma632@gmail.com. However, the questions will be answered in the doubt-solving sessions.

Topics Covered:

1. **Preliminary preparation and analysis of data- Descriptive statistics**

Manipulation of data, data transformation, 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 outliers, and data visualization. Summarizing data, testing of hypothesis.

Assignment A1a: Using the provided data, you must create an Excel file with the state assigned to you. Name it and then import it into Excel. Subset the variables assigned to you and perform the following operations using the software. You must discuss your results.

- Check if there are any missing values in the data, identify them, and if there are, replace them with the mean of the variable
- Check for outliers, describe your test's outcome, and make suitable amendments.
- Rename the districts and sectors, viz., rural and urban.
- Summarize the critical variables in the data set region-wise and district-wise and indicate the top and bottom three districts of consumption.
- Test whether the differences in the means are significant or not.

* Use the dataset [data "NSSO68.csv"]

Assignment A1b: Using the files pertaining to the IPL, you are required to

- Extract the files in R/Python
- Arrange the data IPL round-wise and batsman, ball, runs, and wickets per player per match. Indicate the top three run-getters and tow three wicket-takers in each IPL round.
- Fit the most appropriate distribution for runs scored and wickets taken by the top three batsmen and bowlers in the lost three IPL tournaments.
- Find the relationship between a player's performance and the salary he gets in your data.
- Last three-year performance with latest salary 2024
- Significant Difference Between the Salaries of the Top 10 Batsmen and Top Wicket-Taking Bowlers Over the Last Three Years

* Use the data sets [data "Cricket_data.csv" & "Salary 2024.csv"]

2. **Inferential statistics, Testing Population, Sample, and estimation**

Data: Types of data, most relevant visual representation of data, measures of central tendency and dispersion, correlation among variables. Distribution parameters hypotheses: Construct confidence intervals to convey the reliability of estimates; State and test the hypothesis on whether the parameters differ from one sample to the other. Null and Alternate Hypothesis, steps in TOH, alpha, and p values, Type – I and Type- II errors, One-sample z test, one-sample t-test, two-sample test, F test, ANOVA, and chi-square test. Regression - Notation and Assumptions, Interpretation of Multiple Regression Equation, Partial Regression Coefficients; MLR - Estimation, Hypothesis Testing of coefficients,

*Use data sets [data "NSSO68.csv"] & [data "Cricket_data.csv"]

3. **Regression - Predictive Analytics**

Regression - Notation and Assumptions, Goodness of Fit - Concept of R2 - Multiple coefficients of correlation and determination, Adjusted R square; Panel data regression.

Assignment A2: Using the data created as part of assignment 1.

- Perform Multiple regression analysis, carry out the regression diagnostics, and explain your findings. Correct them and revisit your results and explain the significant differences you observe. [data "NSSO68.csv"]

- Using IPL data, establish the relationship between the player's performance and payment he receives and discuss your findings. * Use the data sets [data “Cricket_data.csv”]
- Analysing the Relationship Between Salary and Performance Over the Last Three Years (Regression Analysis)

4. Limited dependent variable Models - Classification Analysis

Logistic Regression - Assumptions of logistic regression - How to fit logistic regression models in R - How to interpret logistic regression models - How to assess the fit of logistic regression models - Testing for overfitting and numerical stability: Reporting on a logistic regression analysis. Use Logistic regression to find factors that influence limited dependent variable; Introduction to Probit / Tobit regression and inverse Mills ratio to handle zero values in data; Decision Trees

Assignment A3: Using the data created as part of Assignment 1

- You are provided a file for logistic regression. Perform logistic regression, check the assumptions' validity, evaluate the model’s performance using a confusion matrix, and draw a ROC curve. Interpret the results and the model’s efficacy in predicting the event under study. [data “heart.csv”] & [data “Campaign Data.csv”]
- Fit a probit regression to identify non-vegetarians in your sample. Discuss your results and the characteristics of a probit model. [data “NSSO68.csv”]
- Employ decision tree analysis for the data in part a) of this assignment and compare the results of logistic regression and decision tree. * Use the data set [data “heart.csv”]
- Tobit Regression

5. Multivariate Analysis and Business Analytics Applications

Principal component analysis, Factor Analysis, Cluster Analysis - Segmentation for Business Analytics, Multi- dimensional scaling.

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. Market Basket Analysis. Portfolio Optimization. * Use the dataset - [data “Groceries.csv”] & [data “Bill_all.csv”]

Assignment A4: Using the data set provided to you

- Do principal component analysis and factor analysis and identify the dimensions in the data,
- Conduct cluster analysis and characterize the respondents based on their background variables.
- Do multidimensional scaling and interpret the results.
- Conjoint analysis

* Use the data sets [data “Survey.xlsx”] & [data “icecream.csv”]

6. Visualization - Perceptual Mapping for Business

Different types and scales of data (ratio, interval, nominal, and ordinal); Data summarization and visualization methods; Tables, Graphs, Charts, Histograms, Frequency distributions, Relative frequency measures of central tendency and dispersion; Box Plot; Chebyshev’s Inequality. Data visualization and storytelling with data. Introduction to Tableau for Dashboard Building

* Use the data set [data “NSSO68.csv”]

Assignment A5:

- Draw a histogram of the data in Exercise 1 to indicate the consumption district-wise.
- Depict the consumption on the state map, showing consumption in each district.

* Use the data set - [data “NSSO68.csv”]

7. Time Series Analysis

Estimating trend and seasonality in the time series data; fitting ARIMA process and forecasting. Exploring machine learning models such as Neural Networks and Tree based models and other models Facebook prophet; Estimating interrelationships between two or more time-series data sets using the transfer function. Estimating risk using Value at Risk, ARCH/GARCH. Explore applications of vector error correction, Granger causality, and the impulse response function.
* Use the data set – [data "stock prices"] [data "commodity prices"] (e.g. Oil, Gold, Silver, Wheat)

Assignment A6a: Download the data form www.investing.com or Yahoo finance

- Clean the data, check for outliers and missing values, interpolate the data if there are any missing values, and plot a line graph of the data neatly named. Create a test and train data set out of this data.
- Convert the data to monthly and decompose time series into the components using additive and multiplicative models.
- Fit a Holt Winters model to the data and forecast for the next year.
- Fit an ARIMA model to the daily data and do a diagnostic check validity of the model. See whether a seasonal ARIMA fits the data better and comment on your results.
- Forecast the series for the next three months.
- Fit the ARIMA to the monthly series.
- LSTM
- NN
- Tree based models
- Using Indexes Like NASDAQ and Indicators Such as RSI and Bollinger Bands to Forecast and Capture Market Sentiment: A Machine Learning-Based Approach

Assignment A6b: Download the data form www.investing.com or Yahoo finance

- Check for ARCH /GARCH effects, fit an ARCH/GARCH model, and forecast the three-month variability.
- VAR, VECM
- [data "commodity prices"] for ex: Oil, Sugar, Gold, Silver, Wheat and Soyabean

8. Applications Development

This module introduces students to application development, including managing virtual environments, building web applications using Streamlit, and deploying these applications on Streamlit Cloud. By the end of this module, students will gain hands-on experience in creating, managing, and deploying web applications.

Assignment A7: Building 2 Apps

1. "Hello World!" App
   - Create a simple "Hello World!" application using Streamlit
   - Learn the basics of Streamlit components and layout
2. Exploratory Data Analysis (EDA) App
   - Develop an application that performs exploratory data analysis on a given dataset
   - Utilize various Streamlit features to display data insights and visualizations
   - Deploy the EDA application on Streamlit Cloud

References:

- Business Analytics: The Science of Data-Driven Decision Making by U Dinesh Kumar, Wiley.
- Basic Econometrics by Damodar N. Gujarati
- Multivariate Data Analysis by Joseph F. Hair, William C. Black, Barry J. Babin, and Rolph E. Anderson
- Applied Econometric Time Series by Walter Enders
- Time Series Analysis by James D. Hamilton

Grading System:

Final grades will be based on assignments, viva and exams. In case of assignments, the instructor will demonstrate how to analyze a database in R or Python. In addition, Mr. Manojkumar Patil will solve a sample assignment and clarify the doubts, if any. Students should perform the same analysis for the assigned data set and submit it before the deadline. On the other hand, examinations cover specific questions based on the results abstained in various analysis. In Viva Students will be asked questions on any of the assignments to check they have understood the concepts and the results of the analysis. Weights will be applied as follows:

Assignment	A1a	5%
	A1b	2%
	A2	10%
	A3	10%
	A4	10%
	A5	8%
	A6a	5%
	A6b	5%
	A7	5%
Exams	Exam 1	15%
	Exam 2	15%
	Viva	10%
Total		100%

The following grading scale shall be used:

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

Attendance Expectations

Attendance Compulsory

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.

Tentative Schedule of the course

Date	Time	Instructors	Topics
04-06-2024	7PM-9PM	Vedamurthy	Introduction
05-06-2024	7PM-9PM	Achoth	
06-06-2024	7PM-9PM	Achoth	
08-06-2024	2PM-4PM	Vedamurthy	
09-06-2024	10AM-12AM	Manoj	Assignment A1a & A1b Coaching and Doubt Solving
11-06-2024	7PM-9PM	Vedamurthy	
12-06-2024	7PM-9PM	Achoth	
13-06-2024	7PM-9PM	Achoth	Assignment – A1a & A1b Submission Due 11.59 PM Thursday 13/06/2024
15-06-2024	2PM-4PM	Vedamurthy	Assignment A2 Coaching and Doubt Solving
16-06-2024	10AM-12AM	Manoj	Assignment – A2 Submission Due 11.59 PM Sunday 16/06/2024
18-06-2024	7PM-9PM	Vedamurthy	
19-06-2024	7PM-9PM	Achoth	
20-06-2024	7PM-9PM	Achoth	
22-06-2024	2PM-4PM	Vedamurthy	Assignment A3 Coaching and Doubt Solving
23-06-2024	10AM-12AM	Manoj	Assignment – A3 Submission Due 11.59 PM Sunday 23/06/2024
25-06-2024	7PM-9PM	Vedamurthy	
26-06-2024	7PM-9PM	Achoth	
27-06-2024	7PM-9PM	Achoth	
29-06-2024	2PM-4PM	Vedamurthy	Assignment A4 Coaching and Doubt Solving
30-06-2024	10AM-12AM	Manoj	Assignment – A4 Submission Due 11.59 PM Sunday 30/06/2024
01-07-2024	10AM-12AM		EXAM – I
02-07-2024	7PM-9PM	Vedamurthy	
03-07-2024	7PM-9PM	Achoth	
04-07-2024	7PM-9PM	Achoth	
06-07-2024	2PM-4PM	Vedamurthy	Assignment A5 Coaching and Doubt Solving
07-07-2024	10AM-12AM	Manoj	Assignment – A5 Submission Due 11.59 PM Sunday 07/07/2024
09-07-2024	7PM-9PM	Vedamurthy	
10-07-2024	7PM-9PM	Achoth	
11-07-2024	7PM-9PM	Achoth	
13-07-2024	2PM-4PM	Vedamurthy	Assignment A6a and A6b Coaching and Doubt Solving
14-07-2024	10AM-12AM	Manoj	Assignment – A6a and A6b Submission Due 11.59 PM Sunday 14/07/2024
16-07-2024	7PM-9PM	Vedamurthy	
17-07-2024	7PM-9PM	Manoj	Assignment A7 Coaching and Doubt Solving
18-07-2024	7PM-9PM	Manoj	Assignment – A7 Submission Due 11.59 PM Sunday 18/07/2024
20-07-2024	2PM-4PM	Vedamurthy	
21-07-2024	10AM-12AM	Manoj	
23-07-2024	7PM-9PM	Vedamurthy	Review & Doubt Solving Class

24-07-2024	7PM-9PM	Achoth	
25-07-2024	7PM-9PM	Achoth	
27-07-2024	2PM-4PM	Vedamurthy	
28-07-2024	10AM-12AM	Manoj	
29-07-2024	10-12 AM		EXAM – II and Viva
30-07-2024	10-12 AM		Viva
31-07-2024	10-12 AM		Viva