Hi All,

For this project I will be using data from Crunchbase database (<https://data.crunchbase.com/docs/daily-csv-export)>. The Crunchbase data mainly describes about the Startups and other companies , investments and funding for the companies, Acquisitions by other companies and the funding rounds for the companies for the years 2010 to 2015.

I am planning to do an analysis to apply Sequential Pattern discovery on how to the funding in different regions influence the New Startups to set up in that region. I will use various graphs to show the Funding in different regions and the trend in the count of startups. In particular, I would consider California Vs Other Regions in USA.

Thanks,

Geetha Akkineni

Zhiwei Qu

I am interested in data mining, more specific, using data to analyze customer's behavior.

In my final project, I found an algorithm called apriori, which describes support, confidence and lift of different elements in a data set. The mathematical way of the calculation called association rules, and the formulas are shown as below:

support(A ⇒ B) confidence(A ⇒ B)

lift(A ⇒ B)

= P(A∪B) = P(B|A) = P(A∪B)

P (A)  
= confidence(A ⇒ B)

P(B) = P(A∪B)

P (A)P (B)

To apply the data, I found a library called arules, which contains this algorithm and the syntax is pretty straight forward. Also, a related library called arulesViz is found to be an interesting tool for visualization.

Sukesh Reddy Kotha

My project is about reading the csv files and showing the data in a graphical view using graphs and charts with Shiny in R studio.

Like all projects, R programs expressly report the means of your investigation and make it simple to imitate or overhaul examination, which implies you can rapidly attempt numerous thoughts and right issues.

I think the most intriguing point is information representation. R Programming offers an acceptable arrangement of inbuilt capacity and libraries, (for example, ggplot2, pamphlet, cross section) to manufacture representations and present information.

Regards,

Sukesh

Wei He

My project is about pairs trading in financial topics, which is very interest to me. I will try to explore some stock dataset and futures dataset by using R. Before learning R language, I read some stock information and futures information in Yahoo.financial or on NASDAQ stock market web, they do have nice graphs and tables but I am just lack of tools to make analysis by my own. From this class, I figure out that R is exactly the best fitful statistic tool to analysis data like stock and futures for me. It is not only as powerful as SQL to process table data but also provide more functional visualised graphs.

In my project, I will first try some web data processing such as “Quantmod” package to process Yahoo or Google financial page dataset or “Quandl” package to process Quandl dataset. In order to plot these dataset I process from webpage into nice graphs, time series package such like zoo and xts, index calculation package like TTR, visualization package ggplot2 will be also used in data processing.The final goal I want to achieve is to get the visual output of trading signals which can lead the trading. if it works, it might also help in real life, which means a lot to me.

Last but not the least, I want to mention that I have very less machine learning experience before but this class definitely drive my interest in digging more about this area.

Priyanka Das

Hello Everyone,

I have prepared a presentation on my Final Project in R Language. I am attaching it to this discussion. Please view it whenever you get a chance. It is basically the Family Income Analysis. It discusses about the different classification algorithms.

[Final\_Project\_Presentation\_Priyanka\_Das.pptx](https://s3.amazonaws.com/itu.etpo.production/attachments/000/316/737/original/Final_Project_Presentation_Priyanka_Das.pptx?1477890064)

Please let me know in case any more details is required in any of my topics discussed.

Thanks,

Priyanka Das

Karan Gupta

In this project I am using a data set provided by the Kaggle competition: leaf classification.

The train.cvs dataset consists approximately 1,584 images of leaf specimens (16 samples each of 99 species) which have been converted to binary black leaves against white backgrounds. Three sets of features are also provided per image: a shape contiguous descriptor, an interior texture histogram, and a ﬁne-scale margin histogram. For each feature, a 64-attribute vector is given per leaf sample.

My research is to utilize and analyze data sets of leaves, containing binary leaf images and extracted features, including shape, margin & texture, to accurately identify 99 species of plants.

There are two papers interesting on this topic:

T. Beghin, J. Cope, P. Remagnino, and S. Barman.   
Shape and texture based plant leaf classification. In Advanced Concepts for Intelligent Vision Systems,   
pages 353. Springer, 2010.

Charles Mallah, James Cope, James Orwell. Plant Leaf Classification Using Probabilistic Integration of Shape, Texture and Margin Features. Signal Processing, Pattern Recognition and Applications, in press. 2013.

I think it's a good idea to use the K nearest neighbor algorithm as the classifier, and manipulate K values and using data preparation method like normalization. These are the tasks I've been focusing on. And very good accuracy can be reached.

And this is a good post that summarized eight different classification methods:

http://machinelearningmastery.com/non-linear-classification-in-r/

Rui Huang

Which R topics are of interest to you? Are you interested in using R to explore a data set? Do you want to perform statistical methods and techniques to understand a specific data set?

Through this course, I have better idea about R language and even machine learning. I like explore a data set using R.

R is the leading tool for statistics, data analysis, and machine learning. It is more than a statistical package; it’s a programming language, so you can create your own objects, functions, and packages.Speaking of packages, there are over 2,000 cutting-edge, user-contributed packages available on CRAN.

Like all programs, R programs explicitly document the steps of your analysis and make it easy to reproduce or update analysis, which means you can quickly try many ideas and correct issues.

I think the most interesting topic is data visualization. R Programming offers a satisfactory set of inbuilt function and libraries (such as ggplot2, leaflet, lattice) to build visualizations and present data.

I hope I can use R more skillful and explore more. Very enjoy this course.

Rui Huang

Taneema Chowdhury

Hi all,

Today for my Final project I will be using data from World Bank data bank. My data will comprise of different types of data for Bangladesh. It will have the GDP, Export/Import details, Population and Gross capital for a period of 6 years from 2010 to 2015.

I am planning to do a study on all these above details of Bangladesh's economic stand using different methods in R that was taught in this class. I will try to use graphs and interpret what those graphs and plots mean with the data provide.

Thanks

Taneema.

Jia Wang

In this project, the main topic is to find the suggested buy and sell price according to the backtrack data. It is a very interesting topic because this is related to finance. If the result looks great or better than manually manipulate, it is worth trying in the real world with real money. The result we expect can potentially help to find some trends about the company and give a better picture of the company as well. There are some research work on this trading algorithm, however it is worth to run the result by myself using R. And even though the algo is guided to use 50 and 200 days to compute and find the trends, it is still interesting to see whether there will be a better comparison days of moving average or even just to approve they are the best number to use.

Wei Xu

I am especially interested in exploring data set and generating plots using R becauseit is actually greatly related to my internship. My final project uses many R functions to explore the words Shakespeare used in his works. In the project, I used the following packages:

Google’s BigQuery  
dplyr  
magrittr  
tidyr  
ggplot2

This is great practice for me to use almost everything we have learned so far in the R class in order to get meaningful results from a big data set.

Paramesh Palanisamy

I am interested in Functions and Interactive Graphs. Functions makes easy the procedural problems. Interactive graphs are helpful to analyze the data in the different angle. The user interactive way makes our work easy and comfortable.

Yes. I am interested in exploring data set using R especially data sets like iris and mtcars gives the ideas to get explore in R language.

Statistical Methods and techniques helps me to understand the data sets . Since we can adopt to different approaches we can perform unique and best solutions based on the data set.

The data set mtcars . This data set helps me to practice more about the plots and functions.

I was using Matlab to analyze the data set which we are getting for my project. Now, I am writing the R code to implement all the Matlab functionality. In implementation, wise, R programming looks very easy and user convenient. I am interested to explore in medical data set for which I am working now. I am working on sensor which provide the input parameters to deduct the heart rate. We are using Matlab algorithms for detecting the Heart Rate. I am migrating to R little by little from Matlab.

Cheng Jui Tu

Previously , I have experience in writing javascript data visualization. during those time , I use library d3.js and highchart .js

Both of the library is really useful to create web application chart.

I think the data visualization is very interesting. I think data is useful when it became a graph, and this is more understandable for people.

I think if I work more on R, I will try to reproduce this chart in R.

The following two link is good to take a look in d3 and highchart:

[HighChart : http://www.highcharts.com/](http://www.highcharts.com/)

[D3.js :https://d3js.org/](https://d3js.org/)

Raffael Da Silva Nagel

In the past, have you used another language to code? Was there code that you wrote to perform a task that was fun for you? Would you like to rewrite that code in the R language?

By the time I started to code in R I already knew languages like Java, PHP, C/C++, etc. Although R has a very different approach, this previous knowledge helped me in many levels. Besides, I began to compare R to advanced languages like Java to see which features the R language needed in order to be on the same level as the second language. After a lot of research I started to develop functions and methods to fullfil R. This work resulted in a new metodology that allow R programmers to create full enterprise applications, that run outside the R development environment with graphical users interface, installation packages, and all the features that a real app need.

What types of problems intrigue you? Are you interested in text mining, web data processing, or bioinformatics? Is there something you are working on that could be developed into a project for this class?

I'm really interested in developing GUIs for R apps. In my opinion the R greatest problem is that it's not possible to someone without knowledge in coding or in statistics to use it.

Bo Wang

My point of interest is statistical analysis and exploration of data, more specifically, my interest is to extract data from a set of measurement results which I have no idea how accurate they are, and to get information such as if the measurement results are valid, if the parts been measured are OK to use and if the production line is running normal.

Such an activity is called quality and process engineering. These two manufacturing specific fields are developed under a sad fact that there is nothing perfect--no matter how accurate the parts are, error still presents. One simple example is you may think the edge of a knife is smooth and accurate, however if look under microscope or SEM it became a street with pitholes all over the place. Therefore, the concept of "spec" is generated, which means a range is given and as long as the part falls into this range it can be considered OK. And process engineering is built upon the idea to make OK parts, quality engineering, on the other hand, works to making sure that a part does look OK, and they are really OK, then the challenge comes.

The challenges raised to quality engineers are 1) how can we check and make sure that the part is good? 2) how can we tell that our check is good? Ideally we would like a CMM machine to measure everything to an accuracy of 1 micron, but if so then a normal lug nut on the wheels of a car may worth $2,000. Therefore, a variety of tools are developed: there are tools to make sure that a measurement system works well, there are tools to actively monitor trending and abnormalities of data and attempts to warn people before problem occurs, and there are tools to determine if a production line is capable of performing the job...but all these handy tools need a tremendous amount of data been analyzed. Therefore I would like to write some code and make these tools easier to use and read.

Jiangli Shi

**Reproducible research is a hot topic. Have you read an article that contains references to a specific data set?**

It is an interesting topic. Reproducible research is key to new discoveries in a lot of areas. The paper below which is published in Nature shows that 47 out of 53 medical research papers focused on cancer research were irreproducible.

Begley, CG (2013). "Reproducibility: six flags for suspect work". *Nature*. **497**: 433–434.

Yiling Liu

What types of problems intrigue you? Are you interested in text mining, web data processing, or bioinformatics? Is there something you are working on that could be developed into a project for this class?

I am interested in exploring data set, generating plots using R because this allows me to draw to interesting conclusions. That's why my final project is about exploring imdb movie data set using the techniques I learnt during the semester. For example, I used knit, ggplot, qplot, seperation and aggregation of data to analyze the dataset. I am also interested in making predictions using statistical models and applying machine learning algorithms. R is also a powerful tool for that. So in the future I want to learn more about doing data mining and machine learning projects using R.

Nikhil Chitlur Navakiran

Which data sets capture your attention?   
A successful company makes data driven decisions, I like to work with data that impact the decisions that the company or the product takes.

What are you interested in analyzing?   
I am interested in analyzing monetization trends of users and analyzing interaction with my product.

Do you like sports and want to work with sports data. Does the company you are working at have some questions that could be explored using R?   
The company creates games and i want to explore game data, some of the questions that can be answerd using R are player engagement with a new feature, predict monetization based on A/B test data for features yet to be released.

Would you like to download the twitter feed and analyze that data?  
No

What about looking at yelp data?   
No

Are you interested in exploring medical data or census data?  
No

Sarah Huangting Jauw Gunawan

My Final Project and research question is focused on generating a code for research and experimental studies that would be used in analysis. The idea is to use a data set generated by an instrument that runs Complete Blood Count (CBC) on blood samples in a laboratory setting and translate that into valuable data that can be visualized and quantified. It must be specified that the code if meant to be used as a research and development tool for analysis and is not meant to be used in a clinical setting in running and analyzing clinical patient data. The data source is mainly data, mined from experimental research in a biotechnology company (proprietary data) of CBC (complete blood count) test which consists of different parameters. With a wide data set and parameters consisting of (WBC, RBC, HGB, PLT, MCV, MCHC, MCH, RDW, MPV, %LYM, %BASO, %NEU), the code manipulates and analyze the data to present variables and bias on the measurands from different instruments. The data will be visualized with ggplot and ggvis. Particularly with a measurand that fluctuates over time Mean Platelet Volume (MPV) will be presented in an interactive and dynamic manner via ggvis.

The code is conceptualized to act as a template that can be used repeatedly for different experimental studies that allows one to analyze the same parameters and measurands of CBC while allowing the operator the ease and capability of manipulating different variables and output for analysis throughout the course of the experimentation. This research code has large value as it translates data values from instrument (source) to a comparison measurement that is quantifiable allows ease for analysis. The analysis allows for manipulating data from the dataset that translates data into graphs, charts and interactive diagrams via ggvis. As the tool is meant for research only and provides ease in focusing on different variables in CBC, it allows the operator to visualize and compute dataset. As such, no accurate dataset is needed to be referred to. Furthermore, I will be tagging a code that analyzes another dataset that calculates florescence stain index and intensity of antibodies that have been tagged on cells and are measured, for TBNK (T-cell, B-cell, Natural Killer cell subset).

Additional bioinformatics is a great way to investigate and develop my code.

Aleksandar Milenov Gyorev

I'm interested in exploring R for image classification and face recognition.

I believe that now with the emerging of augmented reality this field will become more and more dominant, and what better way to learn about it than doing a final project about it with one of the best data modelling languages there are.

Given the space and the complexity of the problem it is amazing how easy it is to do it with just a couple of matrix multiplications and decompositions, using R.

In my past I've used Python to do a small image scanner for rectangular objects on a plane, but due to the nature of the language, everything was way to unstructured and I found it quite difficult to debug. Writing my final project in R was very straightforward and I didn't have any problem with using the language.

The dataset that caught my attention for my final project is the very popular AT&T faces dataset -- [http://www.cl.cam.ac.uk/research/dtg/attarchive/fa...](http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html)

I will be running the Eigenfaces algorithm, encapsulating different energies and plotting the accuracy for each one respectively.

To put a small twist to the problem, I will also be using my model to match a famous celebrity face image to the closest one in the dataset, so that we can see to which "normal" person does this celebrity look like :D

Mingshu Xin