Big Data Analytics, Spring 2018

GPIM 452, Section ID 931863

Professor Roger Bohn, UC San Diego

Syllabus v 1.05 April 4, 2018

The latest syllabus and assignments are at BDA2020.wordpress.com, as an evolving document.

Rbohn@ucsd.edu

Contact information:

Class meets Monday Wednesday at 11. Initially in Gardner Room. Later in 3203

Professor Roger Bohn
Office = RBC 1315. Phone and text (858) 381-2015.
Email Rbohn@ucsd.edu. *Put #BDA18 in the subject line of all emails.*

Office Hours

Wednesday 6:30 to 7:30 PM in Peet's Coffee.

Other times to be determined.

TA Feiyang.Chen@rady.ucsd.edu Office hours to be agreed on

Web site: A blog will provide announcements, updates of assignments, etc. Class handouts will also be available. **https://bda2020.wordpress.com.** These are supplements to the textbook, and discuss various problems and solutions in data analytics. Be sure to subscribe at least to the blog page.

Visit this page for the latest versions of syllabus, assignments, lecture notes, and other written material.

https://bda2020.wordpress.com/2018/04/04/latest-syllabus-assignments-and-notes/

Assignments should be submitted and returned through Tritoned: tritoned.ucsd.edu GPIM 452

A Google folder for the course, containing files to download, is at:

https://drive.google.com/drive/u/1/folders/1K54L2Y-AFPEZoUGofYFeVOPlOJ6CLvtU

Other syllabus material: The formal syllabus will be on the web site, as a pdf document. Lecture notes and homework will be provided in the same way.

What is Big Data Analytics?

The Big Data Analytics course teaches how to analyze large data sets, using non-traditional tools. This topic has vaulted into prominence in the last five years because it is able to uncover new insights, and business value, that were previously very difficult to get. Companies and organizations that master it can gain considerable advantage over laggards. In fifteen years, the techniques we learn in this course will be routine, and surviving companies will all use them in some form. But by then, new BDA techniques will have been developed, which will create new competitive advantages.

There is actually no agreed on definition of BDA. Closely related concepts that are sometimes interchanged include Data Mining, Artificial Intelligence, Business Analytics, Data Science, and several others. This course will use Big Data Analytics as an umbrella term for data-intensive, decision focused, analysis.

BDA starts from the same base of statistical methods as other quantitative methods courses at GPS. Understanding OLS Regression, to the level of QM2, is a prerequisite. (Speak with me if you are in doubt.) However, its methods and its philosophy are quite different from

statistical analysis for academic research. Some of its characteristics are:

- It uses large data sets, collected in the course of normal business. Millions of records and thousands of variables are common. Often the original data is so large because it is extremely disaggregated, such as every transaction of a customer at a web site.
- It relies primarily on natural experiments rather than controlled experiments. The structure of the data usually makes it difficult or impossible to infer causality, but its advantage is that it is copious and cheap.
- It is designed to *make decisions*, rather than to *prove hypotheses*. Is a patient sick, and therefore must be treated? Is a credit charge fraudulent, and therefore it should be turned down?
- It accepts that all decisions have some chance of error. There is no arbitrary cutoff such as "95 percent confidence" that the null hypothesis is wrong; BDA does not use the idea of a null hypothesis. Instead, the cost structure of decisions drives how to think about the chances of different kinds of errors.
- It is a mixture of *art* and *science*. The textbooks teach the science (formal techniques), but becoming a good analyst and decision maker require also learning the art. In the BDA course, the art issues are approached via a major open-ended project which you design and conduct yourself. (The same mixture of art and science applies to good academic research.)

Techniques: BDA, as we will do it in this course,

- Is a mixture of statistics, computer science, and domain knowledge.
- Uses the open source programming language R, and some of its many add-ins, rather than proprietary statistical packages.
- Uses a variety of statistical models, only a few of which are based on least squares regression. Many of the others are known as *non-parametric models*, i.e. they make no assumptions about underlying statistical distributions of the data such as being Gaussian (Normal).

Who should take this course?

This course is for students who want to understand how to use new analytic concepts, often referred to as "data mining," "big data," or "machine learning," to understand and solve problems. The case studies will mainly be from business, with some government examples such as deploying police forces. The course covers the entire cycle of problem solving, from selecting the problem to presenting the results in an understandable way, and putting them into action. Business problems discussed include web analytics, marketing, operations, credit scoring, product recommendations, etc.

We will use the R language, a computer language designed for statistical computing. Applying data mining methods, understanding their strengths/weaknesses, and using the results will be emphasized over statistical theory. A background in statistics is required,

especially understanding of classical least squares regression. (OLS) Therefore IRCO 454 or equivalent are a prerequisite.

Some students will also consider taking QM3, advanced regression methods. Taking both simultaneously is very difficult due to the workloads. The two courses aim at different employers. QM3 emphasizes regression methods for research, and it emphasizes hypothesis testing. BDA covers big data methods for managers, and it emphasizes decision making. In fact, there is a lot of overlap between the two sets of tools. Scientists are using more and more big data techniques for hypothesis testing, such as text mining, and regression methods are part of the BDA toolkit.

Using R; Special R Certification

The R language is widely used for data analytics, but it is far from the only language available. In fact, no single programming language has a 50% market share, and most of the concepts are language-independent. Therefore, you are not required to use R for your course project. 80% of the work can be done in Stata or Excel if you are already expert in one of them. The remaining 20% is hard to do without R, but you are welcome to try.

For 2018, Big Data Analytics will offer an optional R module that goes beyond what is essential for analysis. Choosing the module will not directly affect your course grade. It has four requirements.

- From Week 3 onward, all homework must be done using R.
- Attend the weekly TA tutorials, which will cover both R and data analytics topics. These tutorials are recommended for all students.
- Take a special evening R test in week 8. The test is open book, open computer.
- Submit all code for your final project as a loadable R workspace.

With the exception of the special test, all of these are worth doing as part of the regular course. All students who meet these requirements at an acceptable level will get a special certificate from GPS, signed by the Associate Dean and the BDA professor.

Major project

A major part of the Big Data Analytics course is a project. Projects use real-world data and analyze it using the tools of the course. Depending on your topic and your data set, you will need to learn specific tools that are not covered in the main class. For example to capture data from the web, you will learn web scraping methods,. There are numerous books and online references available for each specific technique.

A good project should be something that 1) you can get data for, 2) you are curious and can formulate interesting questions about it, 3) is feasible to solve, or at least make major progress, in about 8 weeks. Condition 1 is not as hard as it sounds, as there are many existing data sets created specifically for various data mining activities. The easiest projects are ones where there is a published article documenting a similar problem, or the data, or both. Ask parents, friends, and business connections if they have a pile of non-confidential data sitting around, that has not been thoroughly examined and may have some hidden insights.

You are encouraged to use data and a problem domain that you are already familiar with. However you must use new analytic methods, completely different than QM3 or QM4. Thus if you use familiar data, you must approach it with new questions and analyze it in new ways. In keeping with the theme of the course, your work must have a practical orientation, rather than social science. You may do many kinds of analyses, but *every project must include making a prediction*.

How much data you use will depend on the topic and what is available. Generally you should have between 20,000 and 1 million observations. More than 1 million is feasible, but you will probably end up using a subset of the data. The data should have a *minimum* of 10 underlying variables, and 50 or more variables is much better. Projects with 20 or fewer variables in the original data set will require close consultation with faculty.

Projects can be done in teams of two. Diverse partners are usually better: different nationality, academic interests, gender, etc.

This year for the first time, we are providing some outside data sets from specific companies that you can choose to work on. These are informal consulting projects; the companies are interested in what you find.

There are a series of intermediate deadlines, some of which have formal deliverables. These are, roughly:

1. What data set or sets will you analyze? What questions will you ask? Discuss articles that have done something similar.
3. Get the data cleaned, merged, and loaded. Descriptive statistics.
4. Exploratory data analysis. What are some patterns that seem interesting?
5. Do a "quick and dirty" prototype of an analysis, using real data.
6. Create a set of new variables and choose issues to investigate in more detail.
7. Run your analysis on a subset of the data. Refine your analytic methods.
8. Produce some preliminary results. Submit an initial report and discuss it with your supervisor (in this case, your professor).
9. Refine your preliminary results. Discuss the managerial significance. Prove to yourself and your supervisor that you did the analysis correctly.
10. Give a presentation of your problem and results.
11. Turn in a professional quality final report. It should be something you could show to a potential employer.

A **one page progress memo** is due from each team, most weeks. Write it as a business memorandum, with attached exhibits. More information about project management and project

deliverables will be provided.

Warning: **Treat the intermediate deliverables on your projects very seriously**. Having problems in the early stages is normal, but paralysis due to confusion or fear is not acceptable, and can be fatal. If you are not progressing, talk with faculty sooner, rather than later. Learning how to execute a significant programming/research project is one of the arts to learn in the BDA course.

The rules for **outside assistance** on your project are simple: *you may consult anybody about anything*, provided that you cite them. That includes the TA, professor, other faculty, friends you went to college with, web pages, other people's term papers, etc. Of course, *you must fully footnote contributions from outside*, including code sources, advice, and assistance from UCSD faculty and other students. Presenting work as your own when it is based on outside material is a serious offense, and the penalty can range up to failing the course. Be safe: keep track of outside sources you use during your work, and cite them in the final report. The converse is also true: If you did a substantial amount of R programming for some purpose, include the R code in the report and make clear how much of it is your own work. In this course you get credit for both effective effort, and results.

Week/ Class	**Date**	**Topic**	**Readings**	**Project Assignments**
Wk 1, Class 1	April 2, 2018	Introduction. Examples of applications		
1, 2	April 4	Applications. R software	DMBAR Ch 1 + 2	
2, 3	April 9	Intro to Rattle; Classification Trees	DMRR See homework assignment	Talk with faculty and potential partners. Find a diverse partner.
2, 4	April 11	Classification tree Toyota case Measuring performance : Confusion/error matrix, lift curves, Validation samples	See C 4 CART Toyota	Preliminary ideas. URLs / descriptions of data set you might use. One proposal per person.
3	April 16	Linear Models. Logistic regression. Introduction to R and Rstudio.		Formal proposal. URLs / descriptions of the data set you intend to use. Outside literature on the topic.

4		Text Mining		Load and clean a sample of the data. Provide descriptive statistics.
5		Random forests and boosting.		Exploratory data analysis. Final project proposal, including specific questions, discussion of variables, relevant literature
6		Regularization, LASSO		Initial data mining results
7		Case Studies		Alternative techniques, refining
8		Learning by controlled experiments (A/B trials)		Draft paper **Optional R exam**
9		Problems with big data. Loose ends.		Detailed analytic results
10		Presentations and final drafts.		Hand in view-graph version
11		Final projects due. No final exam.		Final papers DUE.

MEMORANDUM – Syllabus Supplement

To: Students in Big Data Analytics and other courses

Subject: General procedures for homework assignments

From: Prof. Roger Bohn Rbohn@ucsd.edu
Date: April 4, 2018 (version 1.3.0)
Note: this memo is a supplement to the syllabus.

This course requires learning by doing, not by memorizing or studying in the conventional sense. Most weeks will have 1 long problem set and 1 short problem set. Not all of these will be turned in, but be prepared to present your solutions in class.

Graduate students learn more from each other than from the faculty. Therefore *you may do all assignments in teams of 2.* I actually recommend teaming in preference to solving problems by yourself. Working in pairs is faster because you don't get stuck as much. From my observation, teaming also leads to better work quality. I

don't know why for sure, but I suspect it is because your partner helps you to avoid analytic dead ends as well as to be more creative.

For the same reasons, you are permitted to study and confer with classmates (and others outside the class). However, *you must list anyone* who you worked with at the end of your memo.

Mechanical details:

- Submit written assignments through TritonEd.
- Follow the memorandum on how to write and submit a memo.
- If you are working with a partner, BOTH students should submit the IDENTICAL file, with both of your names on it.
- Almost all work should be submitted as a single PDF file.
    - The only exception is when you are submitting R code or other technical information.
    - If your memo has a supplemental file, be sure to list it by file name and topic in the main memo.
- Unless otherwise noted, assignments are due at 11PM the day before class (Sunday night or Tuesday night).
    - We will attempt to make these deadlines clear on Ted. Occasionally, a date might get mixed up.
    - Late assignments are discouraged, but Ted will be set to accept them.
- **DMRR** is the supplemental textbook, *Data Mining with Rattle and R*
- **DMBAR** is the main textbook, *Data Mining for Business Analytics with R*

Files from the textbook are available from a shared Google Folder.

Early Schedule

Week/ Class	**Date**	**Topic**	**Readings**	**Project Assignments**
Wk 1, Class 1	April 2, 2018	Introduction. Examples of applications		
1, 2	April 4	Applications. R software	DMBAR Ch 1 + 2	

2, 3	April 9	Intro to Rattle; Classification Trees	DMRR See homework assignment	Talk with faculty and potential partners. Find a diverse partner.
2, 4	April 11	Classification tree Toyota case Measuring performance : Confusion/error matrix, lift curves, Validation samples	See C 4 CART Toyota	Preliminary ideas. URLs / descriptions of data set you might use. One proposal per person.
Week 3, Day 5	April 16	Logistic regression. Introduction to R and Rstudio.	See homework assignment	Formal proposal. URLs / descriptions of the data set you intend to use. Outside literature on the topic.

MEMORANDUM - Syllabus Supplement

To: Students in Big Data Analytics and other courses
From: Prof. Roger Bohn Rbohn@ucsd.edu

Subject: Format for homework memos and other submissions

Date: April 4, 2018 (version 1.3.0)
Note: this memo is a supplement to the syllabus.

Homework Format recommendation:

Make your work look professional, as if you were in a company or organization. Write your results *as a memo*, from you (one or both authors) to an appropriate person who wants to know your results. Memo format includes: To, From, Subject, Date. Include email addresses for replies. If you are cowriting something, always include your colleague's email in all correspondence, so that Reply All will reach both of you.

The most important part of your homework is the text discussion up front. It should state your conclusions and summarize the reasoning, with references to specific figures, tables, and printouts. Just "dumping" figures and tables at the end of the memo has little value, especially if there are a lot of them. Highlight, circle, or otherwise indicate which numbers are important in the exhibit. The computer creates the raw table; you add value by pointing out what is important.

In the text of a memo, cite *specific* numbers. Instead of saying "As figure 1 shows, product Z should sell very well," be specific, such as "Figure 1 predicts that product Z will have a demand demand of 110,000 units per year, which is 38% of total forecast sales. This is good, because it will be the third-largest product."

Submission format:

All memo assignments should be turned in as a **single PDF file**. Although Microsoft Word is often used as a *lingua franca*, there is no guarantee that a file in MS Word will look the same on someone else's computer. Appendices and exhibits should be part of the same PDF. Any decent PDF program on the Mac, including *Preview*, will allow you to merge PDFs from two sources, and I assume there are counterparts on Windows. Occasionally a single PDF file will be too limiting, for example if you are trying to submit actual data. In those cases you can link to a Google document that you have stashed somewhere.

Emails:
Email is a key medium of communication in business, but good email protocols are still emerging. If you send an email, include the tag #BDA18 somewhere in the subject line. (Other courses should use a different tag, obviously.) This enables automated filtering. Second, make your subject line useful. A subject such as "question" is too short. "Question about confusion matrices," or "Confused about homework for 4/15" are better. Third, set up the email so that all I have to do is hit "Reply All." For example, cc your partner if you are working with someone.

File names: get in good habits now. File names should be *useful to the recipient*, so that they do not have to rename the file. What is in this file, and who is it from? Names should also be useful to you, the sender. Never, send *two files with the identical name and different content*. I use either dates or version numbers in file titles to avoid this. So a file name for this course might be something like: *BDA18 Assign 2016-01-18 Jessica Jones A.pdf* or a short version *BDAjjones@ucsdJan19.pdf*. Subsequent files in the same stream (even if on different days) could have a last letter B, C, etc.

And of course, never send a file with a name like "homework.pdf." Not even you will know what is in it. Yes, I still get these, and I even create them myself by mistake occasionally.

Longer reports: Longer reports benefit from additional practices. Which to use depends on the length, audience, and nature of the report. Here is a list of items that should be included in all reports longer than 10 pages.

- Page numbers. I personally include a header or footer, usually with date.
- Cover page, including all relevant information such as email addresses, and a memorable relevant picture. Suppose someone wants to follow up; how can they reach you post-graduation?
- Executive summary. This is not the same as an introduction. It is not the same as an abstract. It is a self-contained summary.
- Table of contents
- It is acceptable, and far more convenient, to put figures and tables after the main text. But anything you can do to break up the prose, such as inserting figures, makes the report easier to read.
- Use color appropriately, and remember that some readers will print in black and white.
- Use version numbers for drafts, especially if you are part of a team. Dates are not enough. *Never send two files with the same name but different (updated) contents!* Notice the fifth line of this memorandum, for example.

Presenting numerical information: That requires another discussion. But here are two rules:

1. Be clear about the units of measurement. For example data about automobile transportation could be stated in: Cars, drivers, occupants, 1000 cars, or per 100 cars.
2. Show numbers rounded off and with commas, not raw. Writing "the total predicted level is *47130924.22*" is a failure to communicate. Quick: is that more or less than 10 million? The correct way to write that number is approximately: *47.1 million*, or *47,100,000.*

Conclusion: Standardized procedures are a simple but very powerful idea. They reduce errors by 90% or more, and make your work more professional. Develop your own as you gain expertise. This memo includes more than 10 standardized items for semi-formal communication in any professional setting. If you choose to skip any of them, that is your privilege, but do it on purpose rather than by accident.

Early assignments

MEMORANDUM — Assignment
To: Big Data Analytics students
From: Prof. Roger Bohn

Subject: Introducing Rattle and Classification Trees

Date: April 4, 2018. - For Class 3

Part 1: Introducing R

Finish the exercise we started in class on Wednesday. Hand type and run the code on pages 24 to 27 of the DMBA textbook. Play around with subsetting the data (Table 2.3).

Part 2: Introducing Rattle and Trees

This is the first of a two-day sequence on classification trees (CART). On the first day, follow the steps in the DMRR Rattle book, Chapter 11, to do a simple analysis of weather data, using Rattle. Be prepared to explain your steps and your results in class.

- Data frame *weather* is built into Rattle
- Remember to click on "Execute" button
- Do decision tree tutorial, DMRR Section 11.4

There is no formal homework due for this, except be ready to demonstrate in class. But a long assignment will be due Tuesday night for Wednesday's class, So get all the reading done over the weekend.

Classification trees are completely unlike anything you have studied in econometrics. They are a very nonlinear modeling method: they make almost no assumptions about the structure of the problem or the data.

Reading: DMRR Chapter 2, 3, 4.1 on Loading Data - as needed
DMRR Chapter 11 Decision Trees - entire chapter
DMRR Chapter 15.1 to 15.3 Evaluating models using validation data sets.
DMBA chapter *9. (Needed for Wednesday's class.)*

This assignment assumes that we have Rattle working for Mac users. We are reasonably confident. Check the web site for Thursday TA hours to get help with installation.

MEMORANDUM — Assignment
To: Big Data Analytics students
From: Prof. Roger Bohn

Subject: Classification and Regression Trees Assignment for Class 4

Due date: for class 4. Due 11pm April 10.

In this assignment, you will run your first real analysis, of used auto prices. The details of the assignment are in the attached PDF file. The data is on the Wordpress web site, at **ToyotaCorolla_data_binary**

Caution: this assignment will take multiple hours. To make it smoother, we provide some detailed instructions on how to use the software.

Summary: In this exercise, you load data on used Toyota Corolla prices, and try to predict which cars sell for high prices, and which sell for low prices. Everything can be done in Rattle. You can also use raw RStudio, but managing the numerous variables is somewhat tedious.

Load the Toyota used car price dataset, available as an Excel file in the "Downloads" section of the web site. A description and data dictionary are included in the first sheet of the file. Construct several decision tree models *to predict whether the asking price is above 9900 Euros (the median value). Tune the parameters of the decision tree* (section 11.5) to improve the model. One tuning parameter in Rattle is the "minimum bucket size."

Use a 30% validation sample to evaluate performance of each model. (See section 2.7 of DMRR) total error as the error metric (false positive plus false negative, as a percent). Why aren't the best results achieved by building as detailed a model as possible??
Second, re-run the model without the "quarterly road tax" variable, and also drop a number of other variables from the analysis. (In Rattle, just mark them "Ignore.") What is the correlation between road tax and selling price, and why might you not use the tax information when doing a forecast? Compare the two forecasts.

Hand in: A memo documenting your best results and how they were achieved. Explain what they mean. Emphasize their accuracy through the confusion matrix.

Reading: DMRR Chapter 2, 3, 4.1 on Loading Data - as needed
DMRR Chapter 11 Decision Trees - entire chapter
Chapter 15.1 to 15.3 Evaluating models using validation data sets. What does it mean to say results are "good?"

Main textbook, DMBA+R, chapter 9.
You may do this assignment, and almost all others, in teams of 2. Both people submit the same file using your Ted/Tritoned account, with a note giving the name of your partner. See syllabus for more information on format for assignments.

Detailed steps for the assignment are in the attached memo.

MEMORANDUM — Assignment
To: Big Data Analytics students
From: Prof. Roger Bohn

Subject: Building a model using Rattle

Date: April 4, 2018. - Supplementary discussion for Class 4 Toyota problem

Detailed steps for building a Classification Tree using Rattle. This assignment was written originally for Toyota price data, analyzed with CART, but it fits many other situations when using Rattle as the front-end for R. In the future, this level of detail will not be provided unless you have to do something very unusual.

1. Do the readings, so you know what is going on. Read background and tutorial on Rattle, if you have not done so already. Read DMRR chapter 11 about the Classification Tree algorithm. Read DMRR Chapter 15.1 through 15.3 on interpreting results of a model. Read the relevant sections of the main textbook covering the new technique (Chapter 9 on CART).
2. Examine the original data. You may want to use Excel for that purpose, or you can do it after loading it into Rattle. Remember that the data is not in random order. Think about which variables are likely to be very important, somewhat important, or not important to most buyers of these cars, and therefore will probably affect the price.
3. Put the data file in an accessible location on your computer. Start Rattle, and Import the file. A simple rule of thumb is to use CSV file formats when you can. But Rattle can import other formats as well.
4. *Plot* some key variables that you expect to be important. You can also do this in Excel, but you learned enough about plotting to do it directly in RStudio. Or you can experiment with Rattle's plotting capability. Look at some scatter plots and correlation coefficients. Which variables are redundant? Be sure to use the data dictionary (first tab of the Excel spreadsheet with data) to understand the definitions.
5. Create transformed variables as you think appropriate. Rattle handles categorical variables without needing to transform them into dummy variables. (With a few exceptions.) Think about what to do with variables like doors, cylinders, and gears. They are quantitative, but in some ways they are better treated as dummies. (Review any old stats textbook about dummy variables if necessary. They are also discussed in all data mining books, including the supplemental text, *Introduction to Statistical Learning with R*.
6. Go through the first screen in Rattle, assigning roles to different variables. Mainly decide which ones should be *Ignore*d. Your *target* variable is the second one, *Price>Median*. It is a binary variable.
7. Set up your first run of an algorithm, using the *Model* tab in Rattle. Type: Tree.

You can start with the default value of most coefficients, but change the random number seed to some large number instead of 42.

8. Use the *Evaluate* tab to look at the results. Use *Rules* and *Draw* to look at the solution (model, with coefficients) proposed by Rattle.
9. How good are your results? *Study the confusion matrix until you really understand it.* (See DMRR text chapter 15, and DMBA+R text chapter 5 when it is available.) Do you believe these results — can your model really identify the expensive cars that well?
10. There is a good chance that you got accuracy that is really high, or really low. Either one probably means you made a mistake in step 5 or 6 (picking variables).
11. The variables such as Complexity in the *Model* tab (step 7) will affect how good your results are. So will the choice of variables in step 6.
12. Complete the analysis by running variants of your model and comparing the results.
13. Write up a coherent memorandum, with attached figures and tables, explaining how to classify. Be sure to provide the confusion matrix for your favored model. Length limit 2 pages of text + 3 of exhibits, *maximum*. See the memorandum on how to submit homework assignments. For example, please do not give a chronological recounting of all the things that you tried. Link: Memorandum format for assignments

MEMORANDUM
To: Big Data Analytics students
From: Prof. Roger Bohn
Subject: Project assignment #1, due Tuesday April 18
Date: April 13, 2017

By the end of Tuesday, submit a preliminary project proposal from everyone. 1)What data set, 2)What issue do you plan to investigate? 3)who will you work with, if anyone?

Good topics are in fields that are interesting to you and involve data that you are at least somewhat familiar with, and have limited scope. The best data sets have *event-level data*, not aggregated data. For example for crime, have an entry for every reported crime. For e-commerce, have an entry for every transaction, or every item in the catalog, or every customer. (All 3 would be ideal.) Avoid data that has already been summarized, such as state by state crime statistics.

You may submit 2 different proposals if you have not settled on a final idea. Include them in a single PDF document.

Details:
What is the data set you will use, in as much detail as possible. The original source; the web site where you can get the data (or another place to download it.) If you have 2 people, good projects involve mashing together 2 data sets, or even more. You could mash weather with traffic data to predict the effect of rain and snow. You could mash Twitter streams against news articles, to attempt to predict what news topics will "go viral".

Your data should ideally have 100,000s of observations and between 10 and 200 variables. You may need to merge several different years or locations to get this much data.

Also submit at least one article, preferably an academic paper/working paper, from someone who has either
1) analyzed the same data, perhaps in a completely different way, or
2) Looked at an analogous data set with a similar objective in mind.
For example, if you want to study how Amazon identifies fake reviews, you could submit an article on how Yelp identifies fake reviews. (By the way, identifying fake reviews is an interesting but tough problem, according to a friend at Amazon.)

Submit this article by attaching the PDF to your memo. If you prefer, you may attach just the first page, but include the full URL.

I am not asking for any discussion of what data mining techniques you will use. It's too early to be concerned about that.
More advice is on the Wordpress site, https://irgn452.wordpress.com/projects/ .

Weekly project updates will usually be due on Sundays. You will not always get feedback.
As you get closer to the end of the quarter, write your weekly reports as first drafts of material that will go into the final report. Also, use the updates as a mechanism for *your team* to track what needs to be done. For example, attach a running list of:

- Accomplished last week
- To be done next week
- Issues to think further about.

As you deal with these, cross them out by ~~deliberate strikethrough~~. If you have some way of tracking your work that you prefer, you can use that instead.
==========

MEMORANDUM