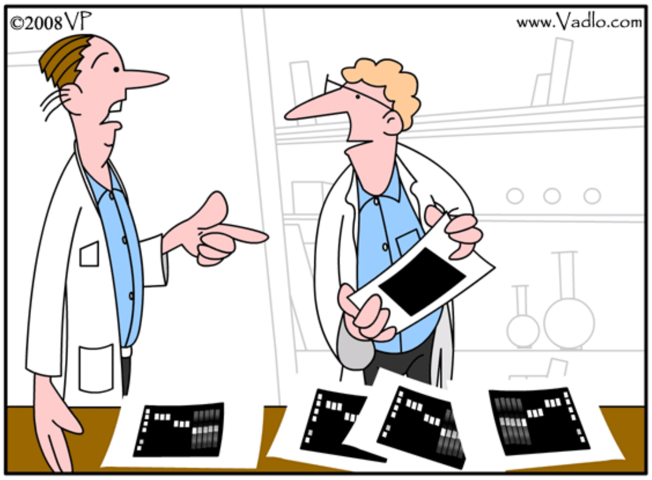
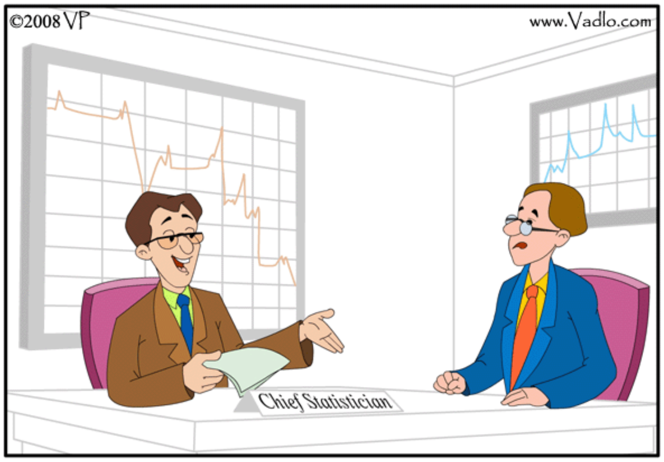
D3M

Data Driven Decision Making (UB 54, Spring 2017)

T/R 14:00-15:15 UC-19



Data don't make any sense, we will have to resort to ***statistics****.*



I can prove it or disprove it, what do you want me to do?

Instructor	Contact	Office hours
Xiao Liu	914 Tisch Hall 212-992-6873 Email: **xliu@stern.nyu.edu**	Friday 15:00 to 17:00 or set an appointment

Course Information

"Every two days we now create as much information as we did from the dawn of civilization up until 2003"
Eric Schmidt, 2009
"Data are widely available; what is scarce is the ability to extract wisdom from them"
Hal Varian (UC Berkeley and Chief Economist, Google)

1. Motivation

The two quotes above summarize the main theme of this course. In every aspect of our daily lives, from the way we work, shop, communicate, or socialize; we are both consuming and creating vast amounts of information. More often than not, these daily activities create a trail of digitized data that is being stored, mined, and analyzed by entities in the private (e.g. Google, Wal-Mart) as well as the public and non-profit sectors (e.g academia, government). The general goals of these data driven initiatives is the hope of generating valuable intelligence that is pertinent to business decisions or public policies. For example, customer transaction databases provide vast amounts of high-quality data that can allow firms to understand customer behavior, and customize business tactics to increasingly fine segments or even segments of one. However, much of the promise of such data-driven policies has largely failed to materialize; primarily due to the difficulty of translating data into actionable strategies.

The objectives of this course are to fill this gap by training you with the tools and techniques needed to analyze large databases, expose you to a wide variety of issues in an empirical context, and instilling an intuition for D3M, i.e. how to generate insights from the volumes of data ('detect signal from noise').

2. Course Philosophy

Extracting useful insights from the vast amount of information involves a combination of analytical skills and intuition. It is both art & science. The pedagogic philosophy in this course embraces the principle of learning-by-doing. Each concept that we cover has a software implementation and a problem/case whose resolution can be enhanced through the use of data.

Statistical tools covered in the class will range from simple data analysis and visualization, to advanced methods such as non-linear regressions, multivariate statistics, and mining of 'unstructured' data. Our emphasis will be on applications and interpretation of the results for making business/policy decisions. Beyond what is necessary, we will focus less on the mathematical and statistical properties of the techniques used to produce these results.

Since this is primarily a Marketing course, emphasis will be given to quantitative aspects of marketing decision making such as segmentation, forecasting demand, designing/positioning new products, customer relationship management (CRM) and evaluation of policies (Can I price better? Did my x-million $ spent on adverting do anything?).

3. Objectives

Regardless of your chosen field or major, it is virtually impossible to survive in the professional world without a working knowledge of basic data analysis and use of some statistical software. The course is designed to expose & train you in a wide spectrum of problems that you are likely to encounter in your workplace. Some of the quantitative methods and concepts are fairly advanced and may seem intimidating at the beginning. It is important to note that **n**o one understands the full scope of every method/technique that exists. Regardless of your prior background, focus of this course should be on continuous improvement by benchmarking your own progress. In particular, you will get most of this class by focusing on (a) removing your fears (if any) of data analysis, (b) enhancing your toolkits, and (c) (most importantly) internalizing the broad **a**nalytical intuition.

Prerequisites: An introductory class in statistics/regression and working knowledge of MS Excel. Experience in any form of computer programming is always a plus but not required. However, the single most important prerequisite for the class is a positive attitude towards learning.

4. Course Organization

Textbook:

There are no required text books. We will rely on:

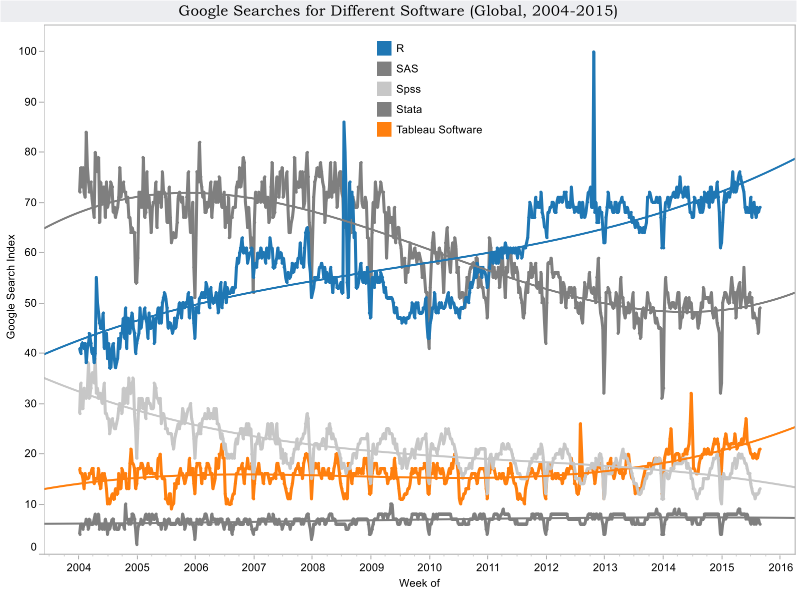
(1) Your old statistics/regression text (if you still have it) ✚
(2) Free online resources (there are infinite-- I will point out where to find for each topic) +
(3) Extensive notes (Combination of ppt/pdf documents that I will provide)

Course Website:

All relevant material related to the course will be posted on NYU Classes. As needed, additional reading materials, and class notes will be made available.

Software:

A number of statistical packages have built in capacities to execute statistical procedures we need (and do so efficiently on large datasets): SAS, SPSS, STATA, R, Minitab, and so forth. In my own research I tend to use SAS (for data cleaning), R (for analysis/graphics), and MATLAB (for advanced models—not required for most business applications you will encounter). The two most widely used in the business world currently are SAS and SPSS, although the momentum is with R. This is primarily for 2 reasons: (1) ℝ is open source software that is absolutely free & (2) has the most dynamic/ innovative global community comprising of academics as well as professionals (wiki model of advanced statistical computing).



Downside of R is steep learning curve particularly if you don't have programming experience. Software, which has existed for a while, but has recently made significant strides in capabilities, is JMP. It has a simple menu driven interface, with advanced modeling features and great visual capabilities. The full professional version of JMP is available to NYU students for $50 (only cost in this class). We will be using a combination JMP Pro 12 + Tableau + R. Tableau is probably the best BI tool available with terrific visualizations, and is quite useful to understand the patterns in your data—a prerequisite before moving to advanced modeling techniques. Tableau offers free student/academic desktop licenses. Note that if you are already comfortable with any particular statistical software, feel free to use that. I can provide help with SAS/R/Stata/SPSS/JMP/Tableau. The key is to pick up and master at least one statistical language (Excel alone would not do). This involves developing skills to load and manipulate large data files, run various statistical methods, and convey your message in a concise/visually appealing way.

Basis for Final Grade:

There are 3 main components to the final grade (**100%** in total):

1. Assignments **40%** (4 Assignments, 10% each)
2. In-class Exams **30%**
   a. Exam-1 **15%**
   b. Exam-2 **15%**
3. Marketing Project **30%**
   a. Mid-term Presentation **10%**
   b. Final Presentation **10%**
   c. Final Report **10%**

All exercises for the assignments/cases/final are application-oriented and incorporate extensive use of software. They are drawn from three primary sources:

- Proprietary data from the business world (mostly used in my research),
- Publically available data from various businesses (e.g. Google), government (e.g. Census, BLS), international agencies (e.g. UN/World Bank)
- Published research in the top academic journals from various fields. Many (good) journals require authors to publish the data used in their research. We will replicate the findings from some of the most influential papers from various fields for which data are available.

All assignments/cases/final are open book, open notes, open internet. We will mimic the "real" world where your objective is to solve the problem at hand and make recommendations based on your analysis. I will NOT penalize you if you can find a solution to any exercise on the internet (you can copy/paste if you do). On the contrary, I will point out (and provide links to) the original sources of data for all assignments/cases.

Assignments: There will be 4 quantitative assignments during the semester. The objective of the assignments is to provide you with a working knowledge of the tools and techniques commonly used in the industry. We will learn how to summarize and visualize data; execute advanced statistical models; and interpret how the output can be used for decision making. We should think of these assignments as learning a new language-- they form the backbone for longer case studies and project.

Case studies: These go a step beyond simply executing models. Instead, they are designed to challenge you to (1) Understand the problem at an intuitive level, (2) Use simple data analysis and visualization to verify (or falsify) your intuition, (3) Use appropriate statistical analysis to present your arguments. In order to imitate the real life challenges, the case studies are fairly open-ended and provide little step-by-step instructions.

Group Work: All assignments and exams are done individually. The project will be done in groups. Please formulate groups yourself, with a minimum of 1 student and a maximum of 4 students per group. It is important to note that the material covered in the class can only be understood by hands-on work. Hence, working in a group should not imply division of labor-- between say (analysis/writing).

Although groups in the class are self-selected, membership is 'non-binding'. If at some point you feel that learning can be enhanced by working alone or group dynamics are not working out, feel free to 'divorce' the group. In summary, feel free to pick & choose group vs. individual work based on what enhances your learning.

Due Date, Submission, and Grading: All assignments are to be turned in electronically via NYU Classes. **Please find the due dates in section 6 on page 9. The due time is 8:00 am on the due date.** All assignments should be submitted as a MS Word document converted to a PDF document. We will use minimal writing in the class, beyond bullet point arguments. Our job will be to illustrate the points using clean tables and graphs—let data talk.

Exams: The exams are in-class for 75 minutes. The exams will cover short exercises pertaining to each topic covered and will be similar in spirit to the case studies/exercises covered in the class. The exams are open book, open notes, open internet.

Class Participation: A substantial part of the benefit that you will derive from this course is a function of your willingness to expose your viewpoints and conclusions to the critical judgment of the class, as well as your ability to build upon and critically evaluate the judgments of your classmates. You are strongly encouraged to share articles/videos on any topic that you find interesting and voice your opinions. We will use this as an opportunity to implement the scientific approach to decision making on contemporary issues in the media.

Assignment Discussions: There are 4 classes for assignment discussions. You will be assigned to discuss one problem in one of the five assignments. The matching will be based on the alphabetical order of your last name. Please find the date and problem for you to discuss in section 7 on page 10.

Laptop Use: Majority of the topics/methods will require use of laptop computers during the class sessions. Please be courteous to people around you and the educational institute by refraining from text messaging, FB, etc.

Feedback: Some of the material covered in the class is fairly advanced. Regardless of your current comfort level with data/technology/statistics, it is my objective to make sure that every student gets a good grasp of the concepts, methods, and implementation (literally 'no child left behind'). If at any time you feel falling behind in class, please contact me. I am happy to work with you individually or in a group to catch up. However, please note that it is your responsibility to seek help.

It is my goal is to make this an excellent course. I encourage you to provide feedback on any issue that can enhance your learning and progress.

Guest Speakers:

We will have two guest speakers coming to present their real-world applications of how to make decisions with data (what we learnt in the course). They are Data Scientists in leading Marketing Analytics companies.

Project:

Please see section 8 on page 11 for details.

5. Methodological Topics

PART 1: Fundamentals of Analytics

Topic 1: An Introduction to Data-driven Decision Making

We will begin the course with a general introduction on what we mean by data driven strategy and why it is important. We will use several examples and mini-case studies to illustrate the role of statistical analysis in decision making. These lectures will provide an intuition for 'data linguistics'.

Topic 2: Basic Data Analysis &

In this session we will discuss various types of data that are commonly collected by firms. What methods to use and what inferences/insights can be obtained depend on the type of data that are available (stated versus revealed preference, level of aggregation, cross-sectional, time series, panel data and so forth). We will cover some of the nuts and bolts of preparing data for analysis, and use several mini-cases to review some basic yet extremely useful techniques such as frequency distributions, mean comparisons, and cross tabulation. Statistical inferences using chi-square, t-test and ANOVA will be discussed.

Topic 3, 4, 5: Intro to R/JMP/Tableau and Data Visualization

We will first look at the basics of the R/JMP/Tableau then start with data visualization. A graph is worth a thousand words. It's important to use graphs to tell the stories behind the data. We'll introduce how to visualize univariate information, relationships (bivariate, mapping and networks) and text. We will use dashboards and animation in presentations to make communications dynamic and interactive. We will show the comparative advantages of each software, Tableau, JMP and R, in three lectures.

Topic 6: Experimental Design and Natural Experiments

Experimental designs are often regarded as the "gold standard" for making causal or cause-effect inferences. We will discuss the issues of design of experiments and internal and external validity. Several case studies in marketing, economics, and medicine that range from controlled lab and field experiments, A-B testing, and circumstances that provide us with "natural" experiments will be discussed using hands-on implementation.

PART 2: Prediction Tools

Topic 7: Regression Analysis

In this topic we will turn our attention to the relationships among variables. Regression is by far the most useful tool for analyzing relationships between a phenomenon of interest (independent variable) and one or more predictor variables. We will spend a fair amount of time on regression and its applications. Emphasis will be on use of regression output in forecasting, elasticity analysis, and various applications such as promotional planning and optimal pricing.

Topic 8: Advanced Regression Models (the dreaded log makes an appearance)

This session covers some important aspects of regression modeling including measures to control for seasonality and trend, interactions, and use of appropriate functional forms (semi-log, log-log).

PART 3: Advanced Decision Tools

Topic 9: Multivariate Analysis (Unsupervised learning)

Cluster Analysis: Hosts of algorithms that allow "grouping a set of objects in such a way that objects in the same group is more similar (in some metric) to each other than to those in other groups. **Factor analysis** is a "method used to describe variability among correlated variables in terms of a potentially lower number of latent **factors.** Factor analysis originated in psychometrics, and is used in behavioral sciences, social sciences, marketing, product management, operations research, and other applied sciences that deal with large quantities of data" (from Wikipedia)

Topic 10: Machine Learning

It is often thought that the value of a firm can be computed using the metric of life time value of its customer base. This topic will cover the important and growing area of CRM and customer equity. We will discuss various tools in database/direct marketing used to model customer acquisition and retention. We will also cover few extremely useful techniques for data mining/reduction such as **Decision tress, Elastic Nets and Random forests.**

Topic 11: Text Mining and Sentiment Analysis

Text analytics refers to the process of deriving insights from text. Perhaps the most widely known is Google's n-gram (frequency of words and phrases) found in over 5.2 million books digitized by Google Inc. Most of the actual analysis of text (e.g. sentiment analysis, topic modeling) requires using R or Python. However, we will cover the general idea and use (processed) data from Amazon reviews to see what such text analysis can teach us.

6. Timeline and Due Dates

Date	Wk	Day	Lecture	Part	Topic	Out	**Due**
24-Jan	1	T	1	PART 1: Fundamentals of Analytics	Topic 1: Introduction to Data-Driven Decision Making	HW1	
26-Jan	1	R	2		Topic 2: Basic Data Analysis	Project	
31-Jan	2	T	3		Topic 3: Data Visualization: Tableau		
2-Feb	2	R	4		Topic 4: Data Visualization: JMP		
7-Feb	3	T	5		Topic 5: Data Visualization: R		
9-Feb	3	R	6		HW1 Discussion		HW1
14-Feb	4	T	7	PART 2: Prediction Tools	Topic 6: Experimental Design and Natural Experiments	HW2	
16-Feb	4	R	8		Topic 7: Regression Analysis		
21-Feb	5	T	9		Topic 8: Advanced Regression Models (log)		
23-Feb	5	R	10		Case 1: Progresso (Pricing)		
28-Feb	6	T	11		HW2 Discussion + Midterm Review		HW2
2-Mar	7	R	12		Project Midterm Presentation		
7-Mar	7	T	13		Project Midterm Presentation		
9-Mar	7	R	14		Exam 1		
14-Mar	8	T			Spring Recess No Class		
16-Mar	8	R			Spring Recess No Class		
21-Mar	9	T	15	PART 3: Advanced Decision Tools	Topic 9: Multivariate Analysis (Cluster/Factor Analysis)	HW3	
23-Mar	9	R	16		Case 2: Brand Valuation		
28-Mar	10	T	17		Guest 1: Robert Moakler (Facebook)		
30-Mar	10	R	18		HW3 Discussion		HW3
4-Apr	11	T	19		Topic 10: Machine Learning: JMP	HW4	
6-Apr	11	R	20		Topic 10: Machine Learning: R		
11-Apr	12	T	21		Case 3: Yelp Predict Rating		
13-Apr	12	R	22		Guest 2: Xiaohan Zhang (Integral Ad Sciences)		
18-Apr	13	T	23		Topic 11: Text Mining and Sentiment Analysis: R		
20-Apr	13	R	24		Case 3: Yelp Predict Review Stars		
25-Apr	14	T	25		HW4 Discussion + Final Review		HW4
27-Apr	14	R	26		Exam 2		
2-May	15	T	27		Project Presentation		
4-May	15	R	28		Project Presentation		
8-May	16	M					Report

Examples of Topics for Assignments/Cases/Exams

- Brand Equity: How to quantify the value of a brand?
- Demand Forecast: What's my sales in the next quarter?
- Optimal Pricing: I know! MR=MC. Ok but how do I go about actually doing this?
- Google Trends and Web Traffic: Deep insights from seemingly trivial data
- Customer Lifetime Value (CLV) & Database Marketing: NPV analysis for millions of customers! Cross-selling campaign design!
- Geography: Airbnb Host Location Choice
- Customer Satisfaction
- Sentiment in Amazon Reviews
- Competition: Which search engine provides better information? Google or Bing? Field experiment analysis

7. Guidelines for the Project

Background

You are going to explore Airbnb, a leading sharing economy company that provides a platform for local hosts to rent unique accommodations. Founded in 2008, the company now has over 1,500,000 listings in 34,000 cities and 190 countries. It has significantly disrupted the hospitality industry and the entire housing market.

Multiple decision makers are associated with Airbnb: the Airbnb company itself, hosts and guests. In this project, you and your team are going to choose a role to play and a problem to solve. Then you are going to conduct comprehensive analysis to make data-driven decisions. To make things more relevant, you are going to focus on the New York market.

Potential Problems

Given your role, either an Airbnb manager, a host or a guest, you are going to identify a problem. Below I list some interesting problems. But they are only to get you thinking and please don't be constrained by them. You could have another very cool and important problem in mind!

Airbnb:

Is Airbnb illegal according to the New York State Multiple Dwelling Law which restricts renting out a Class A multiple dwelling for periods of fewer than 30 days (http://www.nydailynews.com/new-york/nyc-airbnb-users-breaking-law-report-article-1.2451722)?

How satisfied are guest consumers with the hosts? In different neighborhoods? At different time of the year? Can we use reviews to predict guest satisfaction?

How much revenue does Airbnb bring to the economy in New York (http://publicpolicy.airbnb.com/tremendous-impact-new-york/)?

Host:

If I want to be a host on Airbnb,

(Product) What features of the listing can attract high demand?

(Place) How many competitors do I have? Which location has high demand and low competition?

(Price) What's my optimal price? Shall I consider seasonality in demand to do dynamic pricing?

Guest:

If I or my family want to have a trip in New York,

(Product): What types of listings are available? How to predict host quality?

(Place) Which neighborhood has the best quality and lowest price Airbnb house or apartment to live?

(Price): What's my expected payment per night?

Data

The data comes from the Murray Cox's website "Inside Airbnb" (http://insideairbnb.com/index.html). Although the original purpose for collecting this data is to "show how Airbnb is being used to compete with the residential housing market", you can use this dataset to help other decision makers.

The data includes four tables. The specific fields in each table are listed below.

1. Listings

id, listing_url, scrape_id, last_scraped, name, summary, space, description, picture_url, host_id, host_url, host_name, host_since, host_location, host_about, host_response_time, host_response_rate, host_acceptance_rate, host_is_superhost, host_picture_url, street, neighbourhood, neighbourhood_cleansed, neighbourhood_group_cleansed, city, state, zipcode, market, country, latitude, longitude, is_location_exact, property_type, room_type, accommodates, bathrooms, bedrooms, beds, bed_type, square_feet, price, weekly_price, monthly_price, guests_included, extra_people, minimum_nights, maximum_nights, calendar_updated, availability_30, availability_60, availability_90, availability_365, calendar_last_scraped, number_of_reviews, first_review, last_review, review_scores_rating, review_scores_accuracy, review_scores_cleanliness, review_scores_checkin, review_scores_communication, review_scores_location, review_scores_value, requires_license, license, jurisdiction_names, calculated_host_listings_count, reviews_per_month

2. Calendar

listing_id, date, available, price

3. Reviews

listing_id, id, date, reviewer_id, reviewer_name, comments

4. Neighborhoods

neighbourhood_group, neighbourhood

Communicating the Project

Your project should include the following parts:

- Part I: Identify the Problem

Who is the decision maker? What problem do you want to solve? Can you break down the problem to sub-problems?

- Part II: Data-driven Analysis

Please consider using the techniques that you learnt from this course. Here's a couple of tips.

Start with data visualization. Mapping can be very effective for this location data.

When you want to answer a question like "How does A affect B?", consider using tabulation or regression. Pay attention to the data type. Distinguish continuous and discrete variables, especially as dependent variables. They require different methods!!!

If your decision maker has dynamic incentive (like how to calculate consumer lifetime value), don't forget to use dynamic discrete choice models.

If you are doing a prediction, try various machine learning models.

If you have too many variables, consider variable selection methods like PCA or factor analysis). If there are distinct segments (of hosts or guests), clustering will give you better insights.

Try text mining techniques on the reviews. Unstructured data can be used for structured analysis.

- Part III: Recommended Decisions

Give specific recommendations to the decision maker. What, when, where and how to do?

You are going to communicate your findings with the decision makers via two channels: a presentation and a final report.

Team Presentation

You are to take the role of a group of consultants presenting to the decision maker (Airbnb, host or guest). The decision maker may know more about the issue than you do, and will be generally familiar with the firm's situation. Or he/she may have had only a slight opportunity to read the material before a meeting. Leading them through the situation, your analysis and recommendations require considerable skills to hit the right level of detail. Appeal to all these segments. Also make sure that each member of the group gets some "air time." This is difficult to do in practice, but is very effective if the group is well rehearsed and the changeovers are seamless. Be well-prepared and pay extra attention to the substantive content, materials and style of the presentation.

Notes:

(1) Practice your presentation well. Your team will have 15 minutes to present the plan.

(2) If you plan to prepare a PowerPoint presentation, please make sure that you thoroughly test it in the classroom. Too often, presentations fail because of problems related to PowerPoint.

(3) Check out the presentation equipment available in the classroom.

(4) On the day your group is presenting, make sure that all of you are on time.

In the midterm presentation, only Part I and Part II (preliminary analysis) is required.

In the final presentation, Part I, II and III are all required.

Additional issues related to the presentations

1. Bring 5 copies of the overhead slides (2 slides on 1 page) to give to the classmates and me.
2. You can use my laptop for your PowerPoint presentation. In this case, bring your presentation stored on a USB stick to class. You may want to test beforehand whether it works (in the past I have encountered incompatibility issues between MAC and PC, I have a MAC laptop).
3. Teams not presenting will serve as decision makers, will ask questions and will comment on the project. Participating in the decision maker will also count for your participation grade, so all of you should attend the presentations!
4. Rehearse your presentation and time it so that you do not exceed the allotted time!

Final Report

You will prepare the final report for the decision maker. The report will integrate Part I, II and III. You can use the basic structure as outlined for parts I to III. Plan the writing of the report in advance, make a division of tasks and start on time. While working on Parts I, II, and III, consider the consistency among the corresponding sections of the report. Distribute the workload evenly among team members and over time. Include a title page, table of contents, and add a cover to your report. Aim at a length of around 20 pages with a 12-point font plus appendices. When preparing the report pay extra attention to the items listed under 'Grading of Marketing Project.'

Grading of the Project

(30% in total)

1. Final Report (10%)

Clear definition of the problems.
Correct marketing research method. Appropriate data analysis.
Thorough, strong, practical, creative and consistent marketing decision.
Strong basis to support your recommendation.
Practical action programs.
Clarity and conciseness in writing.
Thorough coverage of issues raised during presentation and feedback

2. Presentations (20% in total) - Midterm (10%) and Final (10%)

Message clearly communicated to audience
Use of marketing terminology
Professional and persuasive presentation style
Readability of slides, confident handling of equipment, lighting
Speaking, eye contact, posture, gesture, movements
The result appears as if the team worked well together
Good match of speaker to topic
Good hand-offs