MBA 655: Turning Big Data into Knowledge

Course Description
In this weekend course we’ll get hands-on experience answering questions with data. Designed for a non-technical audience and using Excel as our primary analysis tool, we will work with a real-world data set and explore modern data science techniques. The course will cover an overview of data science and the kinds of problems that are being answered by these techniques. Then we will dig into some actual data and discuss issues with cleaning, data preparation, data description, and data visualization. The class will finish with group presentations as we practice communicating the story within the data. Students will be evaluated on the presentation as well as a written report due after the weekend.

Readings

- “Analyzing the Analyzers”, Harris, et. al., O'Reilly, June 2013. Copy attached.

Preparatory Exercises
There are two data science exercises that need to be completed before coming to class on Friday (which is one of the reasons the reading list is so modest.)

Data Science Example
Data science is a hot topic in the media these days. Find an example of data being used to make decisions. The jumping-off point could be an article, a company, or a piece of pop culture. For your example, I'd like you to do the following:

- Identify a question that could be answered with data or that has been described;
- Think about the source of the data that would be used to answer the question;
- Estimate the size of that data set;
- Imagine how you'd summarize the data to answer the question; and,
- Describe what sort of analysis you'd want to do to answer the question.

Write down your thoughts about each of these pieces. Feel free to write just a sentence or two for each piece. This isn't meant to take more than 30 minutes once you have your topic. Note, you're not expected to do any data analysis for this exercise. This is just a thought-experiment to begin thinking about these questions.

Craigslist Car Data Set
We're going to go through an exercise where we build a data set and analyze it as part of the class. Before we get into class I'd like you to do some work putting together a data set of your own. Here's what I want you to do:

- Choose a car make and model. I'd recommend picking a car that is popular in the US and has been around for a while. This list is a good starting point. I'd encourage you to pick a car that you're interested in. For instance, you might choose the Honda Fit (one of my personal car crushes).

- Pick a year for your make and model. Choose a year where the car was sold for at least one year on either side, preferably two years. So for the Honda Fit I might pick 2011, since there are used cars being sold on Craigslist from 2009-2013. I wouldn't pick 2014 or 2007 since there isn't data on either side of those years.

- Choose a city within which to search. For instance, Minneapolis. Here’s a list of CL sites. This exercise is easier if your city has a pretty active used car scene, so Seattle is probably better than Moses Lake.

- Find 50 actual used cars from the "For Sale -> Cars/Trucks -> By Owner" section of your city's site, for your chosen make/model, and within two years of the date you chose. Here's the link for Minneapolis. Let's just do cars being sold by Owners, not Dealers, just so we're all working with the same data.

- For each of the 50 cars, I want you to build a spreadsheet that has the following columns: Craigslist city/town, date you pulled the data, car make, car model, car year, the price, and the mileage. If your car has different trim packages (like "plain", SE, SL, etc.), then keep track of that too. If the miles are missing (as you can see some are in my example), just put some marker there. I like "NA".

If it's hard to get 50 cars in your chosen site, feel free to take nearby or similar cities. So if Spokane didn't have enough, feel free to use Moscow, WA. We’re going to work with this data set over the course of the weekend.

**Grading**

<table>
<thead>
<tr>
<th>Item</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Science Example: Written Description</td>
<td>15</td>
</tr>
<tr>
<td>Car Data Set: Preparing Example</td>
<td>15</td>
</tr>
<tr>
<td>Class Participation</td>
<td>50</td>
</tr>
<tr>
<td>Team Analysis Presentation</td>
<td>50</td>
</tr>
<tr>
<td>Individual Written Report</td>
<td>50</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>180</strong></td>
</tr>
</tbody>
</table>

**Tentative Course Schedule**

*These times may change; they will be adapted to the class needs.*
ASSIGNMENTS

Data Science Example
As mentioned in the preparatory exercises, you should find an example of a problem that is amenable to a data scientific solution and prepare a short written report on the question, the source(s) of data, the size of that data set, some thoughts on data summarization, and the analysis you’d do to answer your question. This report can take the form of short bullets addressing each aspect of the assignment.

Although the written report is relatively informal, I do expect excellent communication basics (grammar, spelling, clarity). Imagine you’re sending an email to your boss’s boss and you’ll be in the right mindset. Not only is this an opportunity for you to look good, but it’s an opportunity to make your boss look good.

You should be prepared to give a quick (1 to 2 minute) verbal overview of your example in class on Friday. In that overview I’m really interested in the question description, where you might get the data, and how you’d answer the question.

Car Data Preparation
Again, the exercises section has a fuller description, but I will evaluate the spreadsheet you product on its completeness (did you get at least 50 examples?), accuracy (are you consistent in the recording of your data), and clarity (is the data neatly arranged in a table?). Follow the example that I emailed out and you should crush this assignment.

Group Presentation of Car Data Results
The car data assembled above will form the backbone for the hands-on work in the class. Throughout the class I will lecture about various topics in data science and then you’ll have the opportunity to practice those techniques on the car data you and your group gathered. I will assign you to groups. The end of the class will be an opportunity for your group to present a few slides on what you’ve discovered as well as answering the following question: which car (in your data set) should I buy and why?

I expect your presentation to have a slide that sets the context of the problem, telling us what you collected and any issues in data cleaning and collection. Then you’ll transition to slides on descriptive statistics and visualization. These will be supported by slides on modeling that result in your final, data-supported recommendation.

I want to see evidence that everyone in the group contributed to the presentation and you’ll share presentation responsibilities among the group. Since the presentation time is short, I’m comfortable with some members of the group not speaking to the slides, as long as it’s clear that everyone contributed. Contributions that aren’t speaking include managing the group, formatting slides and figures, ensuring consistency of the written communication, and being the lead person to answer questions about the deck.

The group will get a common grade on the presentation.
Written Report
The written report accompanies your presentation but is done and graded individually and is due one week after the class on Monday, October 13th at 5:00 PM MDT. Although I expect you to prepare your report individually, collaboration with your group is fine. The words should be your own, but the ideas can be shared.

I expect the report to cover the data collection/assembly process, data description through descriptive statistics and visualization, and analysis using the model done in class. If someone wasn’t in class but knew about the project, they should be able to read your report and understand the work you did and the conclusions you reached.

At a bare minimum I expect the report to be a comprehensive recapitulation of the material your group presented, also answering any questions for follow-up identified in class. If you’d like to explore other visualizations or models, this is the opportunity to do so. Feel free to include as many charts and figures as you deem appropriate. (This is why I’m not suggesting a page limit for the report.)