Lecture 13 - The Data Analysis Process

Lecture outline:

- What is Data Analysis?

- The three dimensions of Data: Population, Variables, and Time

- Determining the ideal data cube for your question/problem

- Further useful online ressources

In [1]:

```
%load_ext rmagic
```

What is Data Analysis?

Selected parts of the coursera Data Analysis Lecture (https://class.coursera.org/dataanalysis-002/class/index) by Jeff Leek

- The landscape of data analysis (http://prezi.com/fhumwa8tb3fs/the-landscape-of-data-analysis/)
- Types of Data Analysis Questions (https://class.coursera.org/dataanalysis-002/lecture/55)
- Sources of Data Sets (https://class.coursera.org/dataanalysis-002/lecture/57)

> **Data Analysis** is a **collection of methods** allowing us to **answer questions** or **solve problems** in any given field from **data** relevant to the questions or problems.

The **fields** concerned by the questions/problems at hand may be **very varied**:

- biology
- physics
- chemistry
- medecine
- computing
- economy
- sociology
- politics
- buisness
- law
- urbanism
- etc. (and even nowdays, mathematics!)

A **first step** in data analyis is to **identify the right data set** that may **contain the solution** to our question/problem.

For that, one needs

- an **understanding of the field** that the question address (**field expertise**), so that we can gather **relevant** data,
- an **understanding of the type of data** our question/problem requires.

Statistics gives an abstract framework to identify the **type of data** or **form of the data** required by the question/problem. Field expertise and intuition is very important though, since, even though the type of data may be the right one to feed the type of analysis required by our problem, the actual data *content* may not be relevant, or may lack important variables relevant to the problem.

Field expertise (or sometimes good judgement) may help you identify the relevant **population** to study and the **relevant variables** to consider, as well as the **relevant time-scale** for the observations, as we will see below.

In the next section, we will give concrete examples and more details on what is meant by **type of data**.

Data Analysis: Question/Problem $\longrightarrow$ type of data needed

Field expertise: type of data needed $\longrightarrow$ right data set for the problem

The **second** step is to **infer** from **the question/problem** and from the **relevant data set** which **type of data analysis** needs to be performed to solve the problem.

Only when

- the **question/problem** has been defined
- the **ideal data set** has been determined
- the **right data analysis method** has been specified

the concrete data analyis process can began.

Identifying this three components (question/problem, ideal data set, right analysis type) will be the bulk of your class project prospectus.

Remark: Even if the question/data set/data analysis first considered may evolve some upon **exploration** of the **actual gathered data**, you should have a clear understanding of these three points, before starting the analysis.

Randomly taking a data set and analysing it without concrete purpose or angle provided by a thought out question is a recipe for disaster.

The **prospectus** is there to gard you from this danger.

The actual **data anlysis process** goes as follows:

- **data gathering:** Collect **raw data** from various source data sources containing the information necessary to build your ideal data set (or the best approximation you can)
- **data exploration:** Look at all the data collected, check for **value integrity**, look for **defective/aberant values**, isolate the good data, and disgard the rest
- **data cleaning:** From the **raw data** collected build the **clean data set** (as close as possible to the ideal one) on which you'll perform data anlysis
- **data analysis:** Apply the right data analyisis on your clean data set to answer your question, or solve your problem
- **result delivery:** From you analysis create a **data visualization** that answers your question if possible, or write up **an algorithm** that solves your problem.
- **data report delivery**: Write a report restating clearly your question/problem, giving you data sources and analysis methods, and comments on the accuracy of your results.

In the rest of this lecture, we learn how to recognize

- the **ideal data set**
- the **right analysis type**

to answer your question, or solve your problem.

Doing so will also be your main goal, while writing your class project prospectus.

In the coming lectures, we will discuss in details every steps of the actual data analysis process, using R. For now, here are some videos concerning the other steps of the data anlysis process, if you can not wait:

Gathering Data:

- Getting Data (Part 1) (https://class.coursera.org/dataanalysis-002/lecture/73)
- Getting Data (Part 2) (https://class.coursera.org/dataanalysis-002/lecture/75)
- Data Ressources (https://class.coursera.org/dataanalysis-002/lecture/77)

Exploring Data:

- Summarizing Data (https://class.coursera.org/dataanalysis-002/lecture/79)
- Exploratory Graphs (Part 1) (https://class.coursera.org/dataanalysis-002/lecture/85)
- Exploratory Graphs (Part 2) (https://class.coursera.org/dataanalysis-002/lecture/87)

Cleaning Data:

- Data Munging Basics (https://class.coursera.org/dataanalysis-002/lecture/81)

Analysing Data:

- Hierachical Clustering (https://class.coursera.org/dataanalysis-002/lecture/91)
- K-Mean Clustering (https://class.coursera.org/dataanalysis-002/lecture/93)
- Dimension Reduction (https://class.coursera.org/dataanalysis-002/lecture/95)
- Clustering Example (https://class.coursera.org/dataanalysis-002/lecture/97)
- Basic Least Squares (https://class.coursera.org/dataanalysis-002/lecture/99)
- Inference Basics (https://class.coursera.org/dataanalysis-002/lecture/101)
- P-values (https://class.coursera.org/dataanalysis-002/lecture/105)
- Regression with Factor Variables (https://class.coursera.org/dataanalysis-002/lecture/107)
- Multiple Variable Regression (https://class.coursera.org/dataanalysis-002/lecture/103)
- Regression in the Real World (https://class.coursera.org/dataanalysis-002/lecture/109)

The 3 dimensions of Data: Population, Variables, and Time

The population and the variable dimensions

So far, we saw that **statistical analysis** is based on **two main objects**:

- A set Ω representing the **population** under study

- A subset F of functions on Ω representing the **population characteristics** we are interested in.

Depending on the field both O and F have different names:

Example: A class professor can decide to perform a data anlysis on its class students. In this case, the population under study Ω will most probably be the set of all the students in the class. Depending on the questions the professor is asking himself about his class, different **characteristics** or **variables** may be considered.

Try to find out for yourself, what variables may be relevant in order to answer the following questions:

- Are the quality of the different discussion sections affecting my student performances?

- Can the poor exam performance of certain of my student be attributed to a common weak background?

- Can I predict the final score of a given student if I know its total homework score and its midterm score?

- Can I identify the students that have cheated at the midterm, if any?

Now, what if the professor decides to investigate the following question:

- Can I identify which of my students have been improving in the class, and which have been regressing?

The temporal dimension in data analysis

There is a **third aspect** that we didn't talk about so far: **TIME**

Namely, all **processeses take place in time**:

- populations **evolves** in time (for instance: individual can die, be born, etc.)
- their **characteristics vary** in time

Example:

- The **price of a given stock** at the stock exchange varies in time.
- The **temperature at a given spot** on earth varies in time.
- The score of a given student at the **weekly** quizzes and homework assignments varies from week to week.

We will forget here that the set Ω itself can change over time.

Time gives a **third dimension** to data analysis (corresdonding to the notion of `axis` in Pandas).

```
dimension 1 = population individuals (data frame rows)
dimension 2 = population variables (data frame columns)
dimension 3 = observation times (third dimension!)
```

To encode time in our analysis, we can think of the **population variables as depending on time**:

$$X : O \times \mathbb{R} \longrightarrow A, \quad (\omega,\, t) \longrightarrow X_t(\omega) \in A,$$

where A is the set of possible values of the variable X, and where

$X_t(\omega)$ is the **value** of the **variable** X for **individual** ω **observed at time** t.

Example: Consider again the professor question that we saw above:

- Can I identify which of my students have improved in class, and which have regressed?

So here the **right population** to consider is again the population Ω of all the class students.

The **relevant variables** that may help us answer this question must somehow depend on time. Here are two:

- The **weekly** quiz scores Q
- The **weekly** homework scores H

Both of these variables have a value depending on which week we are observing the quiz or homework score.

We have 12 weeks in a semester, which gives us a kind of discrete time:

$T = \{1, 2, 3, \ldots, 12\},$

where the number i denotes the i^{th} week of instruction.

So both variables take a time $t \in T$ and student $\text{Bob} \in \Omega$ and yield back two values:

$$Q_t(\text{Bob}) \qquad \text{and} \qquad H_t(\text{Bob})$$

which are respectively, the score of Bob at the quiz and homework assignment of week t.

Answering the professor question will thus involve a kind of temporal analysis of this two variables for example.

We will come back to that late. For now, just keep in mind that **your question or problem may involve a temporal aspect**, whose analysis will necessite particular methods. You need to be aware of that and to be able to recognize any temporal aspect in your clas project question or problem. This will determine the type of data you'll need to gather and the type of analysis you'll need to perform on it.

Determining the ideal data set for your problem

Suppose you have a certain question you'll wish to answer using data anlysis techniques;.

Your **first order of business** is to **identify the tree dimensions that your problems involves**.

You'll need to dertermine:

- What is the **population** Ω relevant to the problem?
- What are the **population variables** $\{X_1, \ldots, X_n\}$ relevant to the problem?
- What is the **time frame** $T \subset \mathbb{R}$ relevant to your problem?

Example: Suppose you are environemental scientist, and you'd like to settle the question about the earth global temperature increasing. As rough simplification, you may take your population Ω to be all the points at the surface of the earth: i.e. all the points (θ, γ) describing a location on the earth by giving a longitude θ and latitude γ. The variable of interest here is the temperature $T_t(\theta, \gamma)$ at a given location (θ, γ) at a given time t.

A field expert may advice us to enlarge our population by considering all the points not only on the surface of the earth, but also those in a small layer of the earth atmosphere. This expert for good reason, but unknown to us, may also suggest other variables to take into consideration. This is exactly at this point of the data analysis that field expertise is crucial.

Another aspect that we must consider here is the **time-frame** relevant to our problem: Are we considering a period of 10 years, 50 year, 500 years, 500000 years? Again, the field expert will tell us, but we need to be aware that this issue must be addressed.

Remark: Your problem may or may not involve all of the three dimensions, but you need to be very clear about what dimensions are involved at the beginning of your project. Problems involving **all the 3 dimensions** will be most probably be **much more challenging** than those, where only one or two dimensions are relevant.

This increased complexity has two reasons:

- **Practical reason**: it's harder (if not impossible in certain cases) to collect data along all the three axes
- **Theoretical reason**: you may need to **perform data analysis for each of the axis**: a **population analysis**, a **variable analysis**, and a **temporal analysis**.

So in your **class project**, it is strongly encouraged that you **avoid questions involving all the three dimensions of Data!** It won't be managable.

In practice, we don't have access to the whole population, nor the whole time-frame, and even maybe not all the values for our ideal set of variables. We will need to content ourself probably only with

- a **finite sample** $S = \{s_1, \dots, s_m\} \subset \Omega$ of our ideal population (even if m can be very large)
- a **finite choice of variables** $F = \{X_1, \dots, X_n\}$ of our ideal variable collection (which may be infinite...)
- a **finite number of observation times** $I = \{t_1, \dots, t_k\}$

The **collection of values** corresponding to

- the **best possible finite sample** S
- the **best possible finite variable collection** F
- the **best possible finite observation times** I

is our **IDEAL DATA SET**.

Observe that the **ideal data set comes geometrically in the form of cube** with one axis for the population, one axis for the variables, and one axis for the times.

We will call it the **IDEAL DATA CUBE associated to our initial question**.

As already noted, a lot of problems do not involve **the three data dimensions**.

Depending on the initial question, the **ideal data set** may only correspond to **slices of the full data cube**.

For instance, the relevant data to solve the intial problem may be

- a **time series**: one idividual and one variable observed at different times (ex: single stock prices)
- a **time frame**: several individual and one variable at different times (ex: portfolio prices)
- a **data frame**: several individual and several variables constant in time

On the other hand, **only a certain subsets or slices of the ideal data cube may also be available to you**...

Determining the right type of analysis for your problem

A **data cube** can be analysed in essentially 3 ways, corresponding to the 3 axis:

- **Variable Analysis:** i.e., analysis by **columns**
- **Population Analysis:** i.e., analysis by **rows**
- **Temporal Analysis:** i.e., analysis along the **time axis**

Once the

- the **problem** has been chosen,
- the **ideal data cube** identified,

it is necessary to identifiy the **right type of analysis** to solve your problem. (The form in which you want your solution, will also prescribe certain sub-types of analysis.)

Variable analysis

> The main problems that **Variable analysis** addresses may be split into the following three:
>
> - **describing** the **repartition of values** among the population
> - **groupping** the variables that have a **(co)dependence relationship**
> - **predicting** the value of a **target variable** from the values of the other variables

Variable analysis is a **column analysis**.

These are the three main type of variable analysis *we will need for this class*.

Remark: Be aware that there are **other types of variable analysis**. Two other very important types, **factor analysis** and **principal component analysis**, aim at **reducing** the number of variables **while retaining most of the information**; In a sense, these two anlaysis attempt at finding a smaller set of other variables on which our larger set of initial variables depends upon. You most probably won't need this type of analysis for your class project.

Example1 (Prediction): Suppose you are real-estate-agent, who buys and sells houses in the Bay Area. One important aspect of your buisness is then to determine the best prices at which you will be able to sell the house in your house pool. So here is your problem:

At what price should I put this particular house back on the market right now?

For this particular data analyis problem:

- the **ideal population** Ω would be the set of **all houses in the Bay Area**
- the **ideal variable collection** would be **all the possible house characteristics** (location, number of rooms, date of construction, current state, size, number of floor, buying price, and what not...)
- the **ideal time frame** can be taken as now, which means that you may disregard the temporal dimension for this problem althougehter (it is a simplification of course, since housing prices evolve in time...)

Realistically, you'll need to settle down for the houses you can get information on (for instance, by collecting data from the website trulia (http://www.trulia.com/). As for the variables, a good simple first start would be to consider two variables:

- the variable X giving you the **house size** in feet
- the variable Y giving you the **house price** in thousands of dollars

So the data cube you need to construct for this problem is a **data frame** (no time dimension), and the analysis required by your problem is a **variable analysis** (along the column axis). You want to **predict** the value of the variable Y (called the **target**) frome the values of the variable X.

Example 2 (Correlation): Suppose your are a professor, and you'd like to know if the homework sets you gave to your students helped them do well in the final exam. Even better, you'd like to know, if at all possible, which homework sets were the most effective.

Here the **population** Ω is again the class students, while the **variables** may be taken to be each of the homework scores, say H_1 to H_{14}, along with the final score F. Again, the data cube for this problem is actually a **data frame** with rows corresponding to the students, and columns corresponding to the variables above.

The question here involves again a **variable** or **column** analysis, but it is not a **prediction** problem in this case, but a **correlation** problem. You want to understand correlation between the homework scores and final scores.

Beware that this type of **causal of analyis** between variables should be taken with much caution, since **causality relations** of the type "a good grade for homework 8 **causes** a good grade for the final exam are very **hard to determine for sure**.

Example 3 (Description): Our class professor has just given an exam to his students, and would like to know if the exam was too easy, too hard, or just right. Here we again have Ω to be the class students, and we have a single variable X, the exam score. So our data cube is in fact a **data line**.

Solving this problems amounts to compute the distribution of grades among the student, and interpret it.

Observe that this analysis involves a **implicit assumption** about the **probability distribution of grades**: If our class is a typical class, one may assume that the student skills are distributed in a **normal way** with a lot of average students, a fewer number of weak students, and a fewer number of very talented students: the underlying assumption is that talent is shared according to a **normal distribution** among a typical class of students.

From this assumption, the professor will be able to interpret the exam by observing deviations from this normal distribution pattern. Again, here many things may be going on, and conclusions must be drawn with care.

Population Analysis

> **Population analysis** seeks to identify
>
> - **identify** natural **groups** or **clusters** in a population
> - **explain** these clusters in terms of **extra (or hidden) variables**
> - **detect anomalies**, **defects**, or **outliers** in the population

Population analysis is a **row analysis**.

Example 1: (Clustering) Suppose a class professor wants to identify different groups among its students, according to their learning style (for instance: hard-working students having difficulties with the class material, a group lazy but talented students, a group of perfectly balenced students for instance).

For this purpose, the professor decides to have weekly challenging quizzes requiring no preparation at all, straightforward but voluminous weekly homework assignements, and a fair but comprehensive final exam.

So here the population Ω is the class students, the set of variables is the total quiz grade Q, the total homework grade H, and the final grade F. Suppose that the variable value ranges are all between 0 and 10.

The question involves a **population analysis**, namely clustering students into different groups, and interpreting these groups, if at all possible, as reflecting different learning styles.

So for each student $\omega \in \Omega$, one can associate a point

$$x_\omega = \big(Q(\omega), H(\omega), F(\omega)\big) \in \mathbb{R}^3$$

Now the whole class can be represented as a **cloud of points** of such points lying in a three dimensional cube.

We can now look for groups of points that are close to each other, forming independent **clusters**.

For instance, we may interprete a cluster of points accumulating around the the cube vertex $(0, 10, 0)$ as being the hard-working students having difficulty with the material, or a cluster of points accumulating around the cube vertex $(10, 0, 0)$ as corresponding to a cluster of student with talent, but not involved in the class.

It may also be that not cluster at all appear and that all the points accumulate around the cube center $(5, 5, 5)$ indicating a normal distribution centered at score 5 of talent, work, and class performance.

Although cluster may be present in the data, their interpretation should also taken with appropriate care.

Example 2 (Anomaly Detection): Suppose a class professor wants to detect who may have cheated at a given exam. The grading for this exam was negative, and points were removed as new mistakes were appearing during the grading.

Again, the popluation Ω for this problem consists the class students. Now a relevant set of variables for the problem would be the a list of categorical variables with to values corresponding to each of the mistakes encountered in the exam.

For instance, the variable X_i would corresponding to mistake number i and $X_i(\omega)$ would have value 0 if the student ω didn't commit this mistake and value 1 otherwise.

For each student, one obtain a vector of mistakes:

$$x_\omega = (X_1(\omega), \dots, X_n(\omega))$$

where n is the total number of mistakes for the exam.

The problem is a **population analysis** problem where one needs to detect in our student population the individuals having abnormally close mistake vectors, indication potential cheating cases. To be sure, one then need to look at the actual exam copies, and sit placement to confirm the indication.

Time Analysis

> **Time analysis** seeks to

> **Time analysis** seeks to
>
> - **describe trends** updownward or downard in a series of time observations
> - **identify recuring patterns** such as **cycles** or **constant dependence to past**
> - **predict future** observation values from **past** ones

Time analysis is an analysis of values along the time dimension.

Example 1 (Dependence and Prediction): The price of a given stock at the stock exchange is a very difficult thing to model or to predict. If one considers only the time series of prices for a given stock at the stock exchange as our unique source of information, a first subproblem is to identify whether past prices will influence future prices, and if so, how many price observations should we take into account to try to predict tomorrow price?

Intuitively, it seems clear that the price of a stock tomorow will be relatively weakly influenced by the stock prices 50 years ago. But what about last year, or last month? Should we only take into account the prices of the past week?

Here the problem is clearly a **time analysis** problem. The population is formed by a single individual, the stock we are looking at, and there is only one variable, the stock price. So our data cube is actually a **time series**.

The problem here to determine the level of correlation (or **auto-correlation**) of the price at a given date, and the prices at past dates.

Of course, this is a very rough analysis, and should be taken with extreme care since the **time scale** we are taking may also change also in time... according to other variables still to be determined.

A related problem is then, once time scale has been determined for our price prediction (i.e. how far in the past are we going to collect prices for our analysis), we may ask ourselve the problem of predicting tomorrow stock price.

A very simple approach (and probably very inefficient) would be to use the same type of **regression techniques** we talked about in variable analysis to predict the value of a given variable, the target, given the values of other variables (here the past prices).

Example 2 (Trend Description): Consider again the professor question that we saw above:

- Can I identify which of my students have improved in class, and which have regressed?

So here the **right population** to consider is again the population Ω of all the class students.

The **relevant variables** that may help us answer this question must somehow depend on time. Here are two:

- The **weekly** quiz scores Q
- The **weekly** homework scores H

Both of these variables have a value depending on which week we are observing the quiz or homework score.

We have 12 weeks in a semester, which gives us a kind of discrete time:

$T = \{1, 2, 3, \dots, 12\},$

where the number i denotes the i^{th} week of instruction.

So both variables take a time $t \in T$ and student $\text{Bob} \in \Omega$ and yield back two values:

$$Q_t(\text{Bob}) \qquad \text{and} \qquad H_t(\text{Bob})$$

which are respectively, the score of Bob at the quiz and homework assignment of week t.

Answering the professor question will thus involve a **temporal analysis** of this two variables and identify **upward** or **downward trends** for any given students.

We will come back to that late. For now, just keep in mind that **your question or problem may involve a temporal aspect**, whose analysis will necessite particular methods.

You need to be aware of that and to be able to recognize any temporal aspect in your clas project question or problem. This will determine the type of data you'll need to gather and the type of analysis you'll need to perform on it.

Further data analysis ressources

Data Analysis with Python

- Python for Data Analysis (http://proquest.safaribooksonline.com/book/programming/python/9781449323592), by Wes McKinney, Springer (2012)
- Wes McKinney's Blog (http://blog.wesmckinney.com/)
- Gitup page with iPython notebooks on data analysis (and more) (https://github.com/ipython/ipython/wiki/A-gallery-of-interesting-IPython-Notebooks)
- Mining the social web (http://nbviewer.ipython.org/urls/raw.github.com/ptwobrussell/Mining-the-Social-Web-2nd-Edition/master/ipynb/Chapter%204%20-%20Mining%20Google+.ipynb)

Data Analysis with R

- Pro Data Visualization using R and JavaScript (http://link.springer.com/book/10.1007/978-1-4302-5807-0/page/1)

- Using R for Data Analysis and Graphics (http://cran.r-project.org/doc/contrib/usingR.pdf) by J. H. Maindonald
- The R Project for Statistical Computing (http://www.r-project.org/)
- Python Computing for Data Science (http://profjsb.github.io/python-seminar/)
- Berkeley Python Bootcamp 2013 (http://www.pythonbootcamp.info/schedule)
- Berkeley Python Bootcamp 2012 (https://sites.google.com/site/pythonbootcamp2012b/agenda)
- Statistical Computing with Python (http://www.astro.cornell.edu/staff/loredo/statpy/)

Coursera Data Science lectures

Directly related to the class

- Data Analysis (R) (https://class.coursera.org/dataanalysis-002/class)
- Computing for Data Analyis (R) (https://class.coursera.org/compdata-003/class)
- Statistics: Making Sense of Data (R) (https://class.coursera.org/introstats-001/class/index)

To further explore if you have liked and finished all of the above...

- Social and Economic Networks: Models and Analysis (https://class.coursera.org/networksonline-001/class)
- Big Data in Education (https://class.coursera.org/bigdata-edu-001/class)
- Web Intelligence and Big Data (Python) (https://class.coursera.org/bigdata-003/wiki/view?page=ProgrammingAssignmentsHW3)
- Machine Learning (Octave) (https://class.coursera.org/ml-003/class)
- Probabilistic Graphical Models (Octave) (https://class.coursera.org/pgm-003/class)

Further upcoming coursera lectures you may want to enroll in (and get a certificate) here (https://www.coursera.org/courses?orderby=upcoming&lngs=en&cats=stats)