* Question
  + Characterists of students who pass projects?
  + How to stock store?
  + How long to submit pages?
  + How do students who pass their projects differ from those who don’t?
  + Be curious!
  + Think of at least 5 questions and share them on the forums:
    1. What is the average number of courses visited per student?
    2. What is the average number of minutes watched per student?
    3. What is the average number of completed lessons per student?
    4. What is the average number of completed projects per student?
    5. What is the relationship between courses visited, minutes visited per course, lessons completed per course and projects delivered per course?
    6. What is the average time of days to cancel?
    7. What is the percentage of current enrollments on the total?
    8. What is the percentage of canceled enrollments on the total?
    9. What is the relationship between the average number of cancellations and the join date?
    10. What is the relationship between the average amount of cancellations and the cancel date?
    11. How many cancellations of is\_udacity enrollments?
    12. How many current of is\_udacity enrollments?
    13. What is the average time of the projects completed?
    14. What is the percentage of projects completed?
    15. What is the percentage of each rating on the total of projects completed?

Caroline’s questions:

* How much time students spend taking classes?
* How time spent relates to lessons/projects completed?
* How engagement changes?
* How many times students submit projects?
* Wrangle

1. Data acquisition
   * Downloading files
     + Data format
   * Accessing an API
   * Scraping a web page
   * Combine data from different formats
2. Data cleaning
   * CSV: Comma Separated Values
     + Like a spreadshit with no Formulas
     + Easy to process with code (unlike .xlsx)
   * Caroline’s questions:
     + 1. More unique students in enrollments than engagement table;
          - Why are students missing from daily\_engagement?
       - Fix:
         * Identify surprising data points (if any)

Any enrollment record with no corresponding engagement data?

* + - * + Print out one or a few surprising data points

Do you see anything strange? No

* + - * + Fix any problem you find

More investigation may be necessary

Or there might no be a problem!

* + - 1. Column named account\_key in two tables and acct in the third;
      * Fix: Change column from acct to account\_key;
        + Rename the acct column to account\_key in the daily\_engagement table;
        + Enter the output of daily\_engagement[0][‘account\_key’] here:
* Explore
  + Build intuition
  + Find patterns
  + Question: How do numbers in the daily engagement table differ for students who pass the first project?
  1. This will include data from after the project submission;
  2. This compares data from different lengths of time;
  3. Includes engagement in courses not related to the first project;
  + Revision: Only look at engagement from first week and exclude students who cancel within a week;
  + Getting started: Created a dictionary of students who either:
    - haven’t canceled yet; (days\_to\_cancel is None)
    - stayed enrolled more than 7 days (days\_to\_cancel > 7)
    - Keys: account\_keys
    - Values: enrollment date
    - Name your dictionary: paid\_students
    - How many students are in the dictionary?
* Draw Conclusions (or make predictions)
  + Predict: which movies users will like
    - Tentative conclusion: students who pass the subway project spend more minutes in the classroom during the first week;
      * But is this a true difference or due to noise in the data?
      * You can check this using statistics
    - Correlation does not imply causation
      * Students who pass the first project are more likely (correlation) to visit the classroom multiple times in their first week;
      * Does visiting the classroom multiple times cause (causation) students to pass their project?
      * Third factors that could cause visiting the classroom and passing projects:
        + Level of interest;
        + Background knowledge;
      * Or this correlation could be because of causation!
        + We just don’t know!
        + To find out run an A/B test;
    - Making predictions
      * Which students are likely to pass their first project?
        + Could take a first pass using heuristics but getting really good prediction this way could be difficult;

Lots of different pieces of information to look at;

These features can interact;

* + - * + Machine learning can make predictions automatically;
  + Conclude: Users are less likely to click certain articles. Usually requires statistics of machine learning
* Communicate
  + Blog Post, paper, email, powerpoint, conversation
  + Data visualization is almost always useful

|  |  |
| --- | --- |
| **What findings are most interesting?** | **How will you present them?** |
| Difference in total minutes | Report average minutes |
| Difference in days visited | Show histograms (polish any visualizations) |

|  |  |  |
| --- | --- | --- |
| **Data Analysis and Related Terms** | | |
| **Data Science** | **Data Engineering** | **Big Data** |
| Similar to Data Analysis | More focused on data wrangling | Fuzzy term for “a lot” of data |
| More focused on building systems | Involves data storage and processing | Data analysis, scientists and engineers can all work with big data |
| May require more experience |  |  |