COMP 333 — Week 2 Recap

In Week 2 (this lecture) we will cover three examples of Exploratory Data Analysis
as we continue our high-level overview of Data Analytics.

Think of these as introductions, that you can come back to time and time again,
to see Data Analytics in action as you learn more about the individual steps and techniques.

These examples are often used to illustrate EDA.

They range from simple to very complex.

They deal with domains, such as serving in a restaurant, high-school education, and the story of the Titanic that are familiar to all of you.

- Example 1 is a small dataset about tipping in restaurants.
- Example 2 is an official dataset from OECD on educational performance of 15-year olds.
- Example 3 is an investigation of survival on the Titanic when it sank.

You are not expected to follow every detail of an example at this stage.

Use the examples to orient yourself as to

What is Data Analytics?

READ the files marked READ.

For these examples, it is very worthwhile to also read/watch the supplementary material.

Data Collection None of these examples addresses *data collection* other than to fetch a `csv` file.

In the real world, a lot of effort is devoted to this task:

- crawling the web and scrapping web pages
- fetching entries, mainly text, from social media
- running surveys or robo-call polls
- designing experiments to generate your data
- merging data sources, such as in data warehouses

The course *COMP 479 Information Retrieval and Web Search* covers data collection and handling text data, so we do not cover that in this course.

Data Wrangling This topic is covered very briefly in Example 3 when it talks about missing values. But data wrangling is not covered in the other Examples.

Descriptive Data Analysis The major focus of Examples 2 and 3 is DDA. They do an excellent job of introducing *descriptive statistics* and *types of data*, and then cover the appropriate ways to plot univariate and bivariate relationships.

Modeling Example 1 briefly introduces EDA through a simple linear regression model. And then goes on to show what other insights might be found through the exploration steps of EDA.

Example 1 does not discuss modeling.

Example 3 touches on modeling, but only briefly discusses the steps producing the models, evaluating their performance, and comparing them that is behind the table comparing 8 models performance.

Nor does Example 3 go further and discuss how one *validates* a model.

Modeling is important as the model is a major deliverable for data analytics.
Validation of the model is vital to its successful deployment.

Feature Engineering Example 3 gives a good first introduction to feature engineering.