COMP 333 — Week 6 Welcome

Welcome to COMP 333 Data Analytics

In Week 6 (this lecture) and Week 7 we will cover Data Wrangling
which is the most time-consuming phase of Data Analytics.

Data Wrangling is the ETL process of data warehouses
applied more generally as part of Data Analytics.

It is very important to clean and organize your data.

Remember GIGO (Garbage-In, Garbage-Out)

This week we cover

- Example of Data Wrangling
  to create an actual product on POI (Points-Of-Interest) for the travel industry;
- Data Wrangling Process
  with an introduction to each step.

In Week 7 we will go into the step of Data Cleaning in depth.

READ the files marked READ.

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

Go back to the video of the POI system development several times
to see each step of Data Wrangling in action.
And again after seeing Week 7 material.

Chapter 7 of the `pandas` book is on Data Wrangling

Chapter 5 of the `pandas` book has a section on *Handling Missing Values*

and Chapter 9 covers more advanced features for *Data Aggregation and Group Operations*

The recap in Week 7 will re-iterate the points covered in the lectures.

All the best, Greg.