WQD7003 DATA ANALYTICS
Assignment

Maximum Marks Available   :   20%
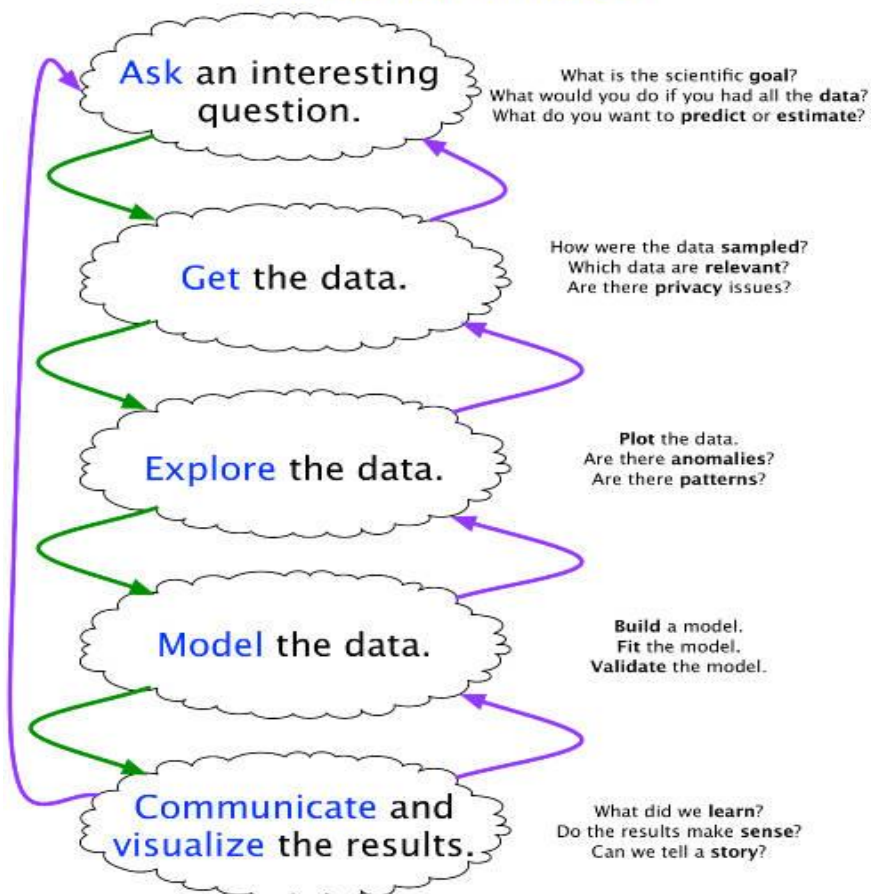
Hand out date   :   10th March 2020
Hand in date   :   12th May 2020

Project

You are required to apply your skills to the real world by choosing a project you care about!. Go through the complete data science process from start to finish.

The Data Science Process

Ask an interesting question.

What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

Get the data.

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Explore the data.

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and **visualize** the results.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course http://cs109.org/.

Suggested Topics:
- Domestic violence
- Drug Abuse
- Dengue
- Breast Cancer
- Parkinson Disease
- Dermatology

- Food Environment Atlas
- Poverty
- Migrant and Refugees
- Aging Society
- Learning Analytics
- Heart Disease

What to expect?
- Initial Questions
  • What questions are you trying to answer?
- Overview and Motivation
  • Consider that this will be read by people who did not see your project proposal.
- Related Work
  • Anything that inspired you, such as a paper, a web site
- Data Cleaning
- Exploratory Data Analysis
  • Visualizations, statistical methods you considered? Justify the decisions you made, and show any major changes to your ideas. How did you reach these conclusions?
- Models used and comparative analysis between the models
- Evaluation
- Final Analysis
  • What did you learn about the data? How did you answer the questions? How can you justify your answers?
- Presentation

(Note: This is a group assignment. Please work in a group of 4 people)

Project Submission

• IPython notebook
  • important part of your project
  • standalone document that **fully** describes your project in detail
  • overview and motivation
  • related Work
  • initial Questions
  • Data Cleaning
  • Visualizations
  • Modelling
  • Evaluation and Analysis
• Turnitin Report

Presentation:-
  • 10 minutes presentation + 5 minutes question

- Problem: Motivations, question
- Data Cleaning
- Modeling
- Evaluation
- Results, can be: predictions, better understanding of a problem
- Future works
- No implementation details (unless a specific technical point matters)

Plagiarism

Your attention is drawn to the no-cheating policy. This covers cheating, attempts to cheat, plagiarism, collision and any other attempts to gain unfair advantage in assessment. The work you submit must conform to this policy. If cheating is discovered, any parties involved will take equal blame (get zero).