An Introduction to EDA

Berk Orbay, PhD

Essentials of Data Analytics - BDA503
MEF University

September 26, 2017

Briefly About Me

- BS in Industrial Engineering, METU.
- MS in Operations Research, METU.
- PhD in Industrial Engineering, Bogazici University. (Thesis title: Empirical Assessment and Model Selection in Option Pricing)
- Taught/teaching: Computational finance, business analytics, introduction to probability and R.
- Worked on government, education, elections, financial, energy and similar data sets.
- R user since 2011.
- Co-founder of Algopoly, a data science company.

Truth is rarely pure and never simple.

- Oscar Wilde, *The Importance of Being Earnest*

Some Data Problem Types

- Labeling data. (e.g. Heart attack risk calculation)
- Regression data. (e.g. house prices)
- Time series data. (e.g. USD/TRY, energy demand)
- Text data. (e.g. sentiment analysis)
- Image data. (e.g. handwriting OCR)
- Graph data. (e.g. mail leaks, bitcoin wallet activity tracking)

A Simplified Data Process

1. Data cleaning. Manipulating the raw data to the desired format.
2. Insight gathering. Preliminary analyses, feature engineering.
3. Analysis. Modeling and inference.
4. Results. Analysis outputs (e.g. error rates).
5. Reporting. Storytelling.

The course will be more about 2 and 3 (to some extent). But, we will be doing all the parts.

Data Cleaning

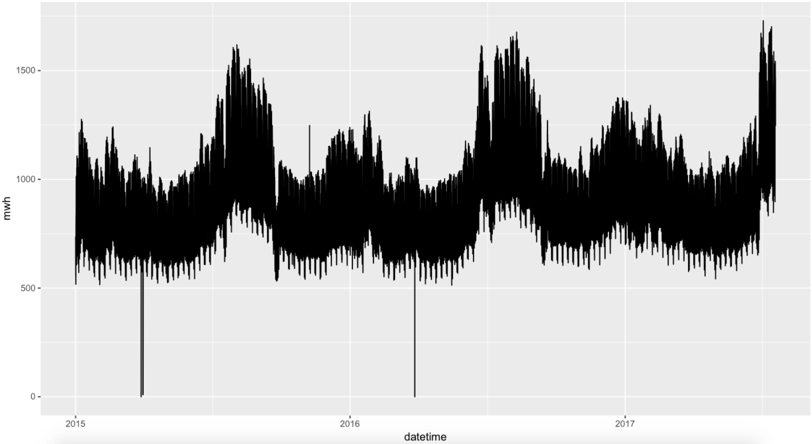
Horrible part (imagine you have to combine with data gathering).
80% of the job (in terms of time) is about data cleaning.
Ultimately necessary.

- Missing data.
- Date/time inconsistencies (time zones).
- Wrong inputs.
- Suspicious outliers.
- Different formats/types of data.
- Encoding issues (UTF-8).

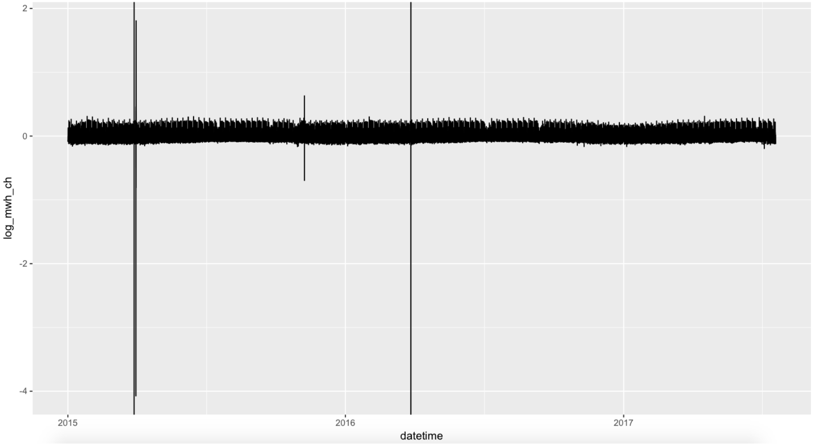
Plus many other unpleasant surprises. (Government data is the worst.)

Garbage In Garbage Out

Example: Energy Demand Data



Example: Energy Demand Data



Example: Energy Demand Data

What went wrong?

- Huge blackout on March 31,2015.
- Winter-time / summer-time adjustments.
- No winter time adjustment for Turkey in 2016!
- Some duplicate consumptions (different day, same hourly consumption for each hour)

What are the possible consequences if irregularities are not handled? What are the possible solutions?

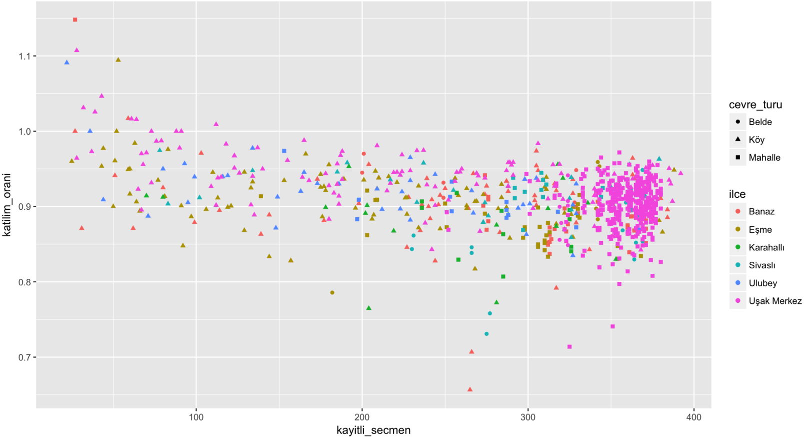
Insight Gathering

Humans are incredible pattern recognizers and they are good with visualizations.

- Attribute (covariate) interactions. (e.g. correlations)
- Exploratory plotting.
- Outlier detection.
- Data transformations (e.g. log, $1/$x, x^2).
- Feature extraction (to enhance your models).

Do not get carried away. Beware of wishful thinking and overfitting.

Example: Elections Data



Model Building

There are many models out there

- Sampling (stratified), experiment design
- Model choice
- Feature engineering
- Fitting
- Out of sample predictions
- Cross validation

Your introduction to models will be limited in this course.

Results and Reporting

Now that you did all your homework. It is time to check the results.

- Is the objective achieved? (e.g. minimize absolute error)
- What do different models tell about?
- How do you interpret the outcome?
- Are your predictions reliable? How confident are you?
- Reproducible research

Conclusion

- EDA is the starting point. If not done properly, it will affect your whole process (GIGO). Take data cleaning seriously.
- Come up with a clear objective.
- Analyze your results based on that objective.
- Convey your findings with good storytelling.
- Automate if possible.
- Bonus: Organize in a way that others can follow your steps to reproduce your findings by themselves from the raw data. (reproducible research)

You will be fine.

Course Progress

Lectures

- 6 lectures + 1 presentation week. A lecture every 2 weeks.
- There will be minimal lecturing (except first two lectures).
- Most of the lecture hours will be about hands-on coding, discussions and analysis.
- Block hours (intermission as required).
- Different learning speeds are expected. More material will be provided with progress.

Grading

- Homeworks (15%). Nothing too hard.
- Quizzes (15%). Mostly to check your progress on Udacity and other material.
- Group Project (35%). Most important part. This is where you show your knowledge and skill.
- Final (35%).

Group Projects

Crux of the course.

- Groups of 3-4 students.
- You should work on interesting data sets. I suggest Turkey-related data sets.
- No secret data (no exception). No data set related to work (exceptions apply).
- You will display all your data, work, code etc.
- Reproducible research.
- Peer review.
- There will be a guideline.

Communications, Materials and Colloboration

- All materials will be on `http://mef-bda503.github.io/`. Perhaps on Blackboard too.
- HW/project submissions on (wait for it, probably GitHub Classroom). Most work should be done in RMarkdown.
- Colloboration is highly encouraged (except for individual in-class assessment i.e. quiz, final).
- E-mail me at `orbayb@mef.edu.tr`