In Vino veritas

Finding truth in wine with Data Science

Evgeniy Paskin

October 12, 2018

LIFE IS TOO SHORT TO DRINK BAD WINE !!!

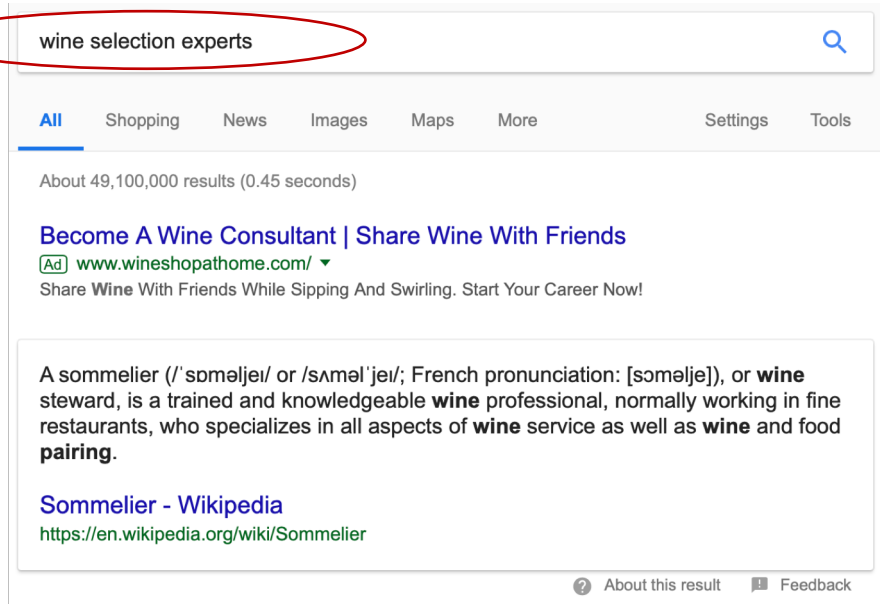
BUT HOW TO CHOOSE ?

- If you are not an expert in wines selection is actually very hard
- Looking at the bottle in the store gives you very little information:
  - Price
  - Age
  - Country (and some times *appellation* - a certain region within the country)

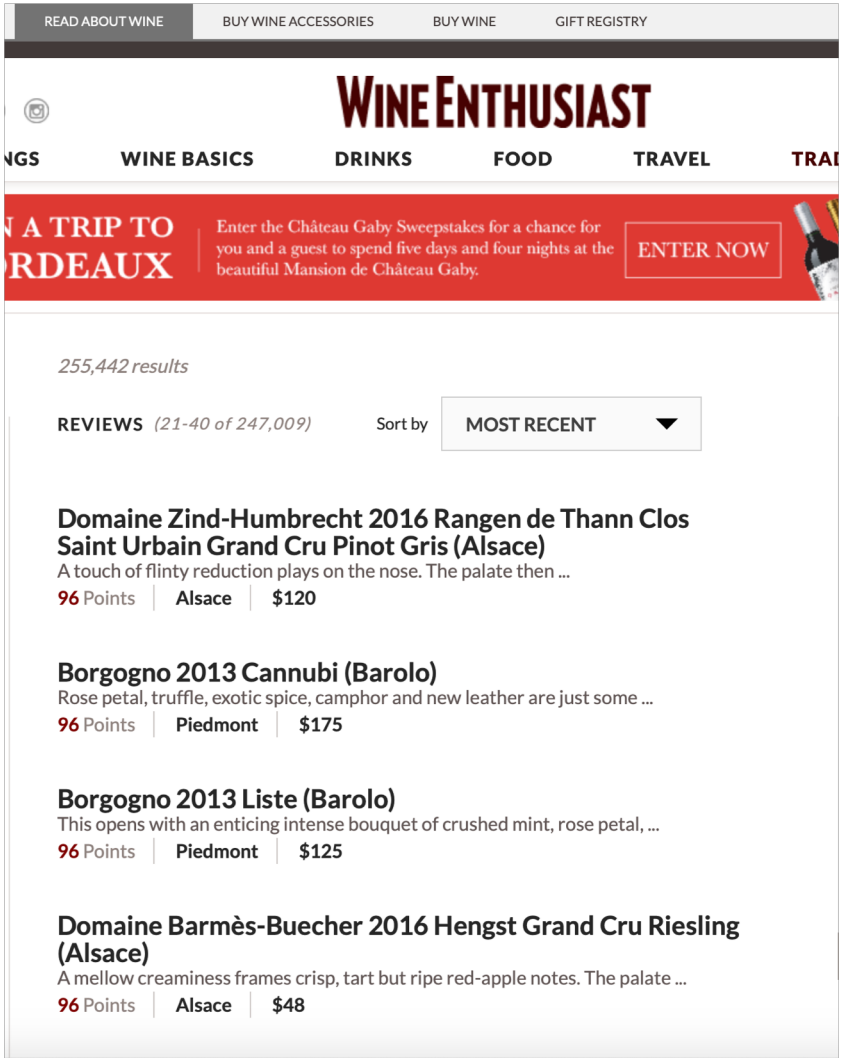
CAN WE IMPROVE WINES SELECTION EXPERIENCE WITH DATA SCIENCE?

- There are a lot of wine selection experts and different ratings available
- **Can we gather this data to predict wine rating based on the information available to you in the store?**

Let's try!



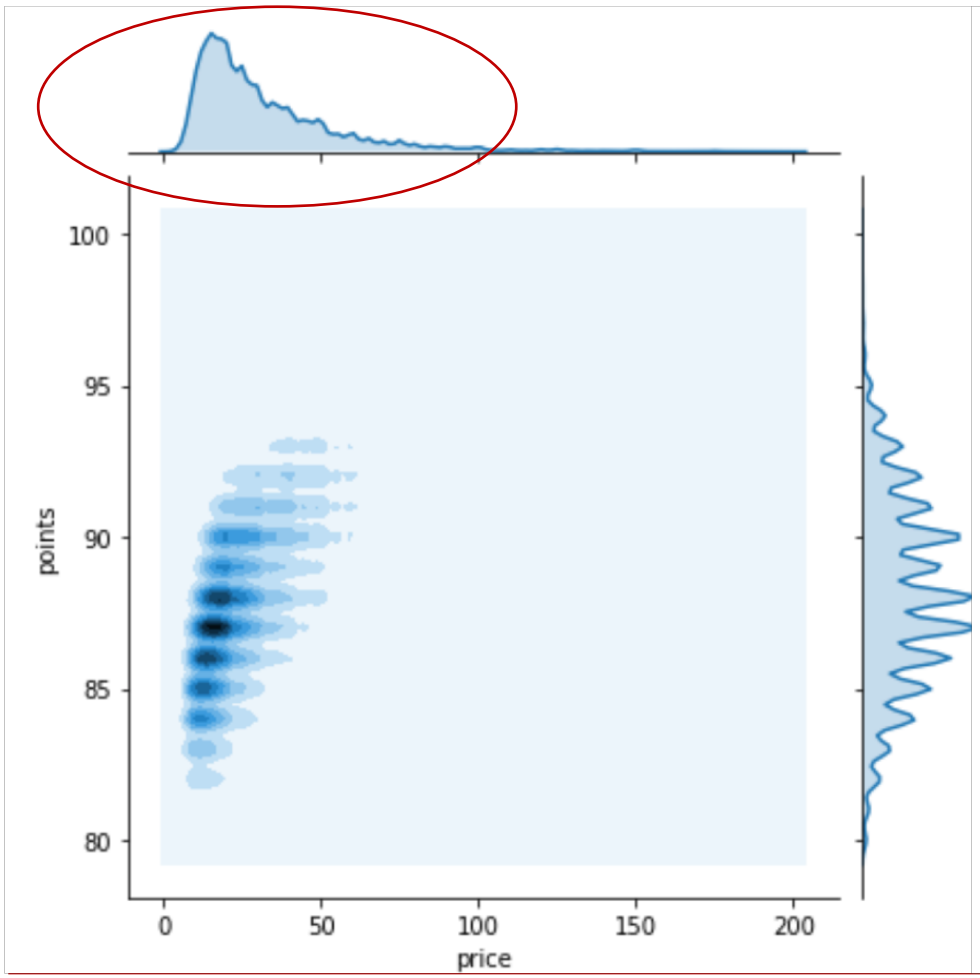
STEP #1: SCRAPING WINE RATINGS DATA

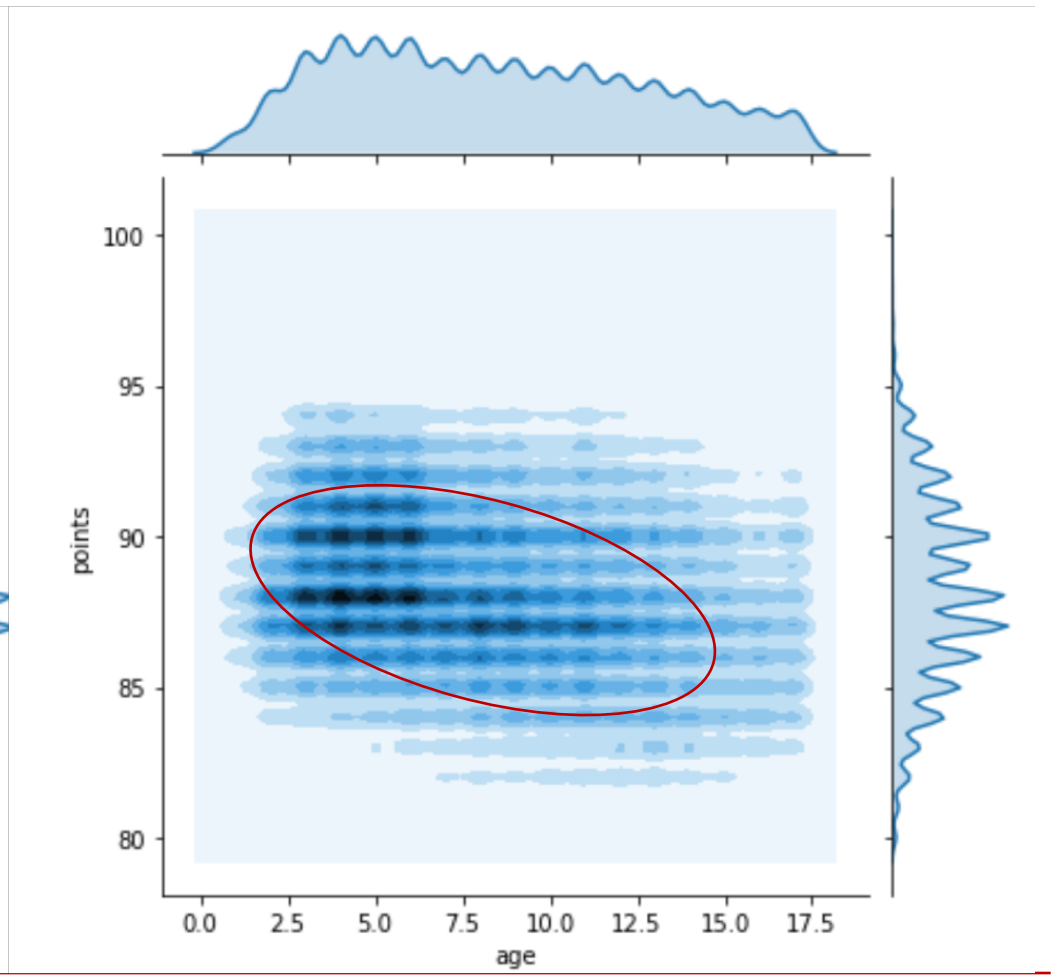


- Target: WineEnthusiast.com
- Data:
  - 20 wine descriptions per page x 12k pages
  - Rating + Price + Appellation + Short Description
- Obstacles:
  - Site protection
  - Slow speed
  - No info about country of production

STEP #2: CLEANING DATA & EDA

- Data cleaning: dropping outliers ($>200, age > 18 years), NAs, etc. Final dataset 197k rows
- Ranking from 80 to 100 (so actually 20 points). Price distribution looks like log-normal





We'll also check if this joke makes sense from data science point of view

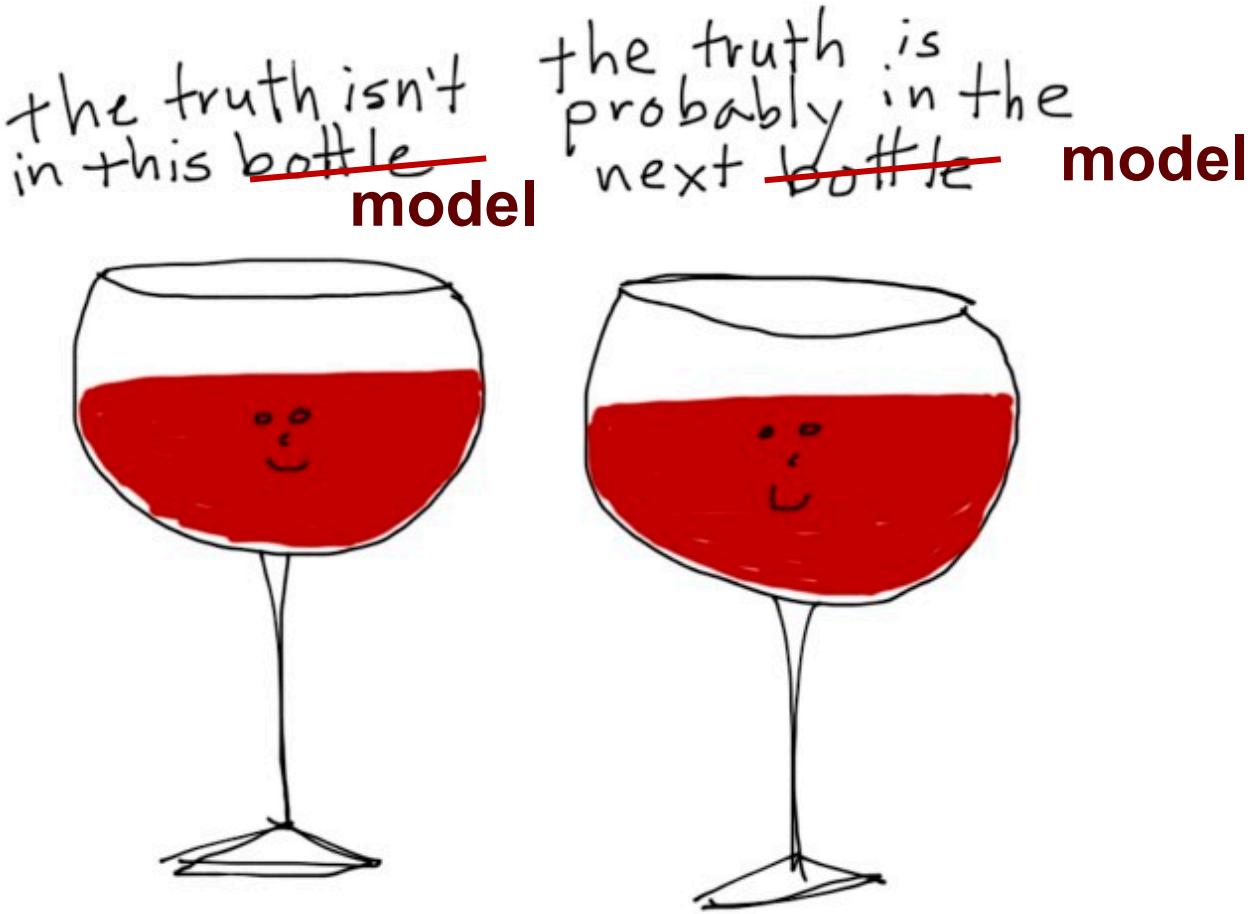
STEP #3: ITERATIVE MODELLING

- Initial model: the simplest linear regression with no pre-processing based on age and price only **scored R^2 of 0.32**
- Maximum scored model:
  - All data (179 features, including country & appellation)
  - StandardScaler preprocessing
  - 1st order polynomial
  - ElasticNetCV (50/50) for Alpha selection
  - **Score 0.43**
- Final model:
  - Same process as for maximum scored model, but without appellation data to reduce number of features (I can not remember 120+ appellations) and improve model interpretability
  - **Score 0.41, Intercept = 88 points**

WHAT “WINE TRUTH” DID WE FIND WITH THE MODEL?

- **Knowing price, age and region of production is not enough to reasonably predict wine rating ☹**
- **Age of wine actually doesn’t matter – feature weight zero (other models showed even negative relation of age to wine rating points)**
- **Don’t go for expensive wine! ☺**
- **New World wines are doing pretty good**

	beta	feature	beta_abs
2	1.892870	Argentina	1.892870
1	-0.458386	log_price	0.458386
5	0.230886	Chile	0.230886
12	0.175240	South Africa	0.175240
8	0.128138	Italy	0.128138
14	-0.088784	US	0.088784
3	-0.083716	Australia	0.083716
4	0.077025	Austria	0.077025
11	-0.074111	Portugal	0.074111
6	-0.071321	France	0.071321
7	0.062947	Germany	0.062947
13	0.054980	Spain	0.054980
10	0.023415	Other	0.023415
0	0.000000	age	0.000000
9	0.000000	New Zealand	0.000000


the truth isn't in this ~~bottle~~
model
the truth is probably in the next ~~bottle~~
model

NEXT STEPS & LESSONS LEARNED

Next steps:

- Explore model results in more details - some betas were unexpected
- Add new data to improve models. Want to try:
  - Historical weather data for region & year (e.g. count of sunny days, average temperature)
  - NLP of wines descriptions
  - Application of other models, including classification algorithms

Lessons learned:

- Start EDA & Modelling earlier with less data while scraper working, add more later
- For me it actually took **4x** time than expected (not 2x!)

THANK YOU FOR ATTENTION AND THESE TOOLS FOR GREAT JOB

scikit learn

python™

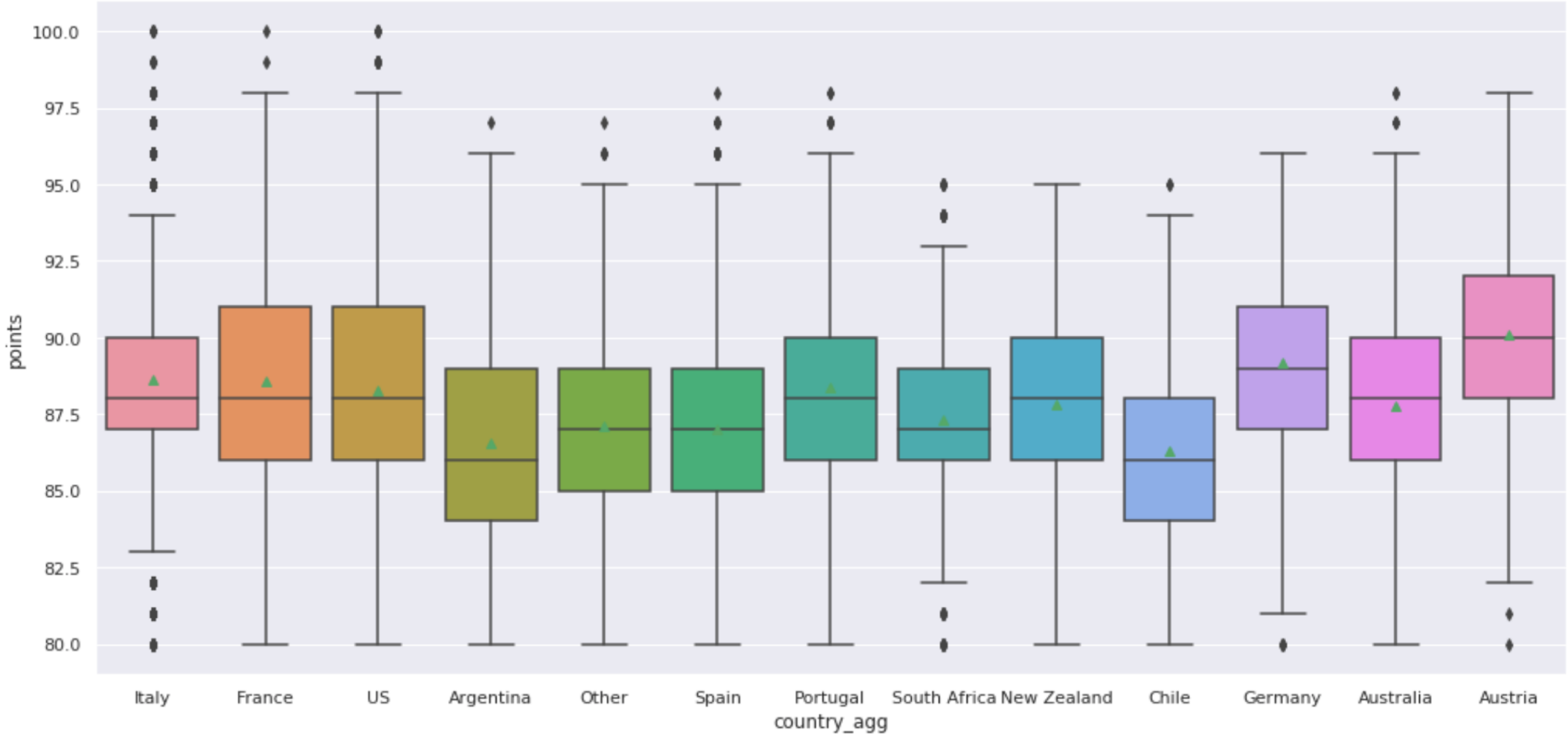
matplotlib

Google Maps Platform

pandas

$y_i t = \beta' x_{it} + \mu_i + \epsilon_{it}$

jupyter

Appendix

STEP #2: CLEANING DATA & EDA