Foundations of Data Science

Group 7 - Wine Insights (...or *"I heard it through the grapevine"*)

Project choice: wine insights

- A well-chosen wine collection can be hugely profitable
- Insights may have implications for portfolio optimisation
- Clear commercial potential - trade and forecasting
- Predicting “good” or “bad” vintages for investment

Can we discover insights about fine French wine using information that is available before bottling?

Problem domain

- Many factors (some psychological) affecting “quality” of wine;
    - Region
    - Climate and weather
    - Soil profile
    - Chemical profile (tannins, phenols, other aromatic compounds)
    - Price[1]
    - Reputation of vineyard/producer
    - Respected critics’ reviews/ratings

[1] *“Marketing actions can modulate neural representations of experienced pleasantness”* - Plassman, et al.

Tooling (and libraries)

- Data Acquisition
  - Python
  - Unix tools
  - MongoDB
  - MLab - MongoDB SaaS

- Exploratory Analysis
  - R (ggplot2)
  - Python (sklearn, seaborn)
  - Tableau

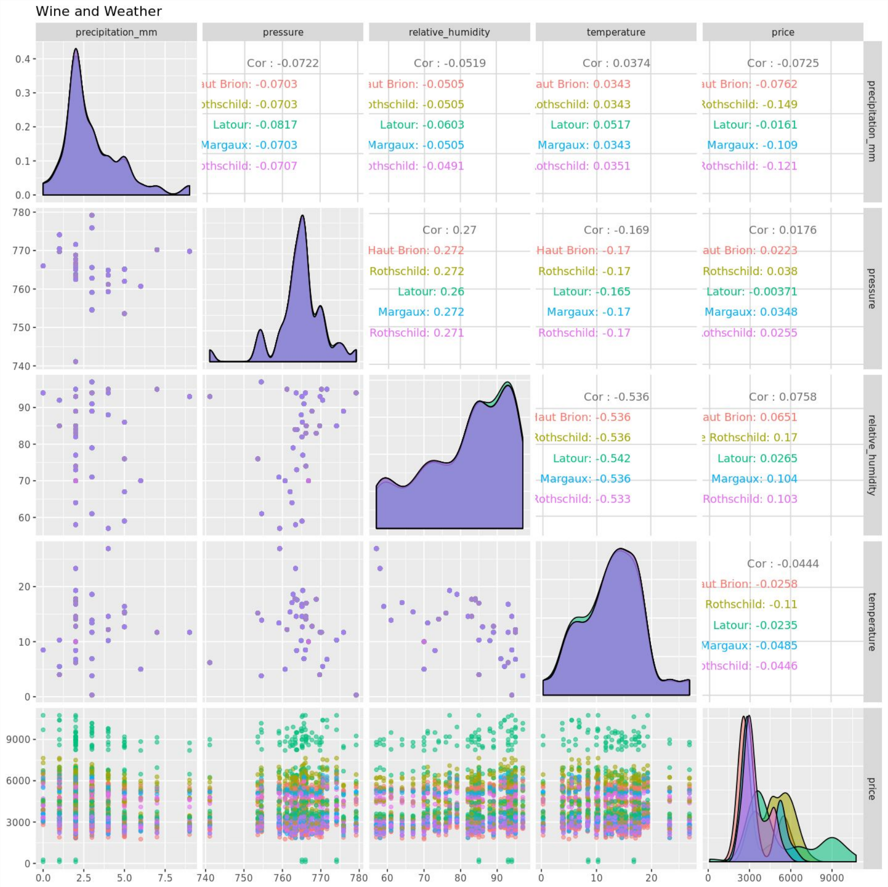
- Web Dashboard
  - NVD3, D3, D3-Cloud
  - Bootstrap
  - Python (tornado, pymongo)
  - RESTful API

Core datasets

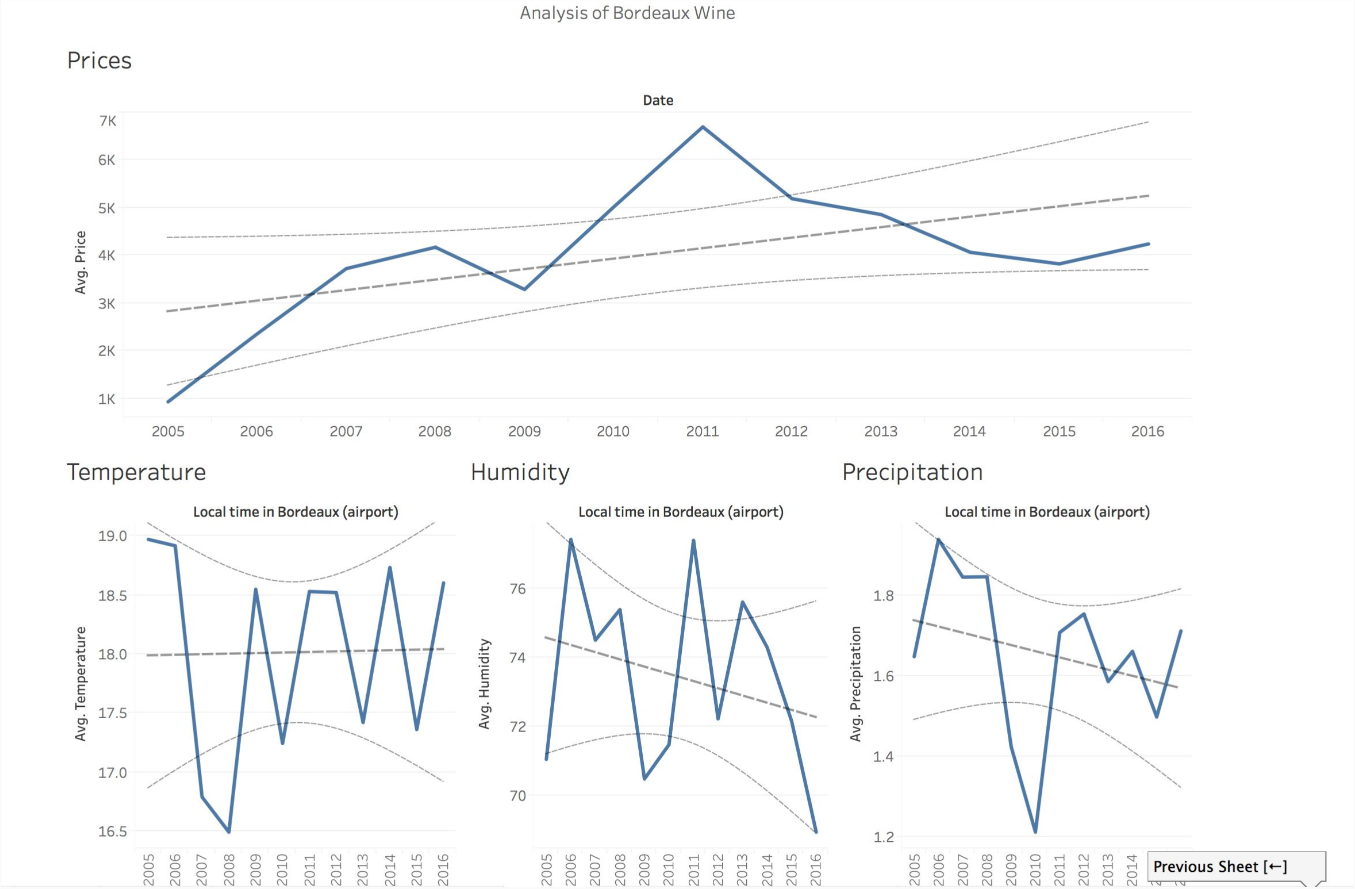
- RP5.am - huge METAR weather data archive
  - Historical archives from French airports which could be downloaded as CSV
- Wine-Searcher - wine review data
  - Information scraped from website using Python and BeautifulSoup
- Cellar-Watch - historical wine prices (auction, market, trade)
  - Top wines for each region from the Liv-Ex Fine Wines Investables index
- MSCI World index - global stock index
  - This was used as a yardstick for global economic performance

Exploratory Analysis

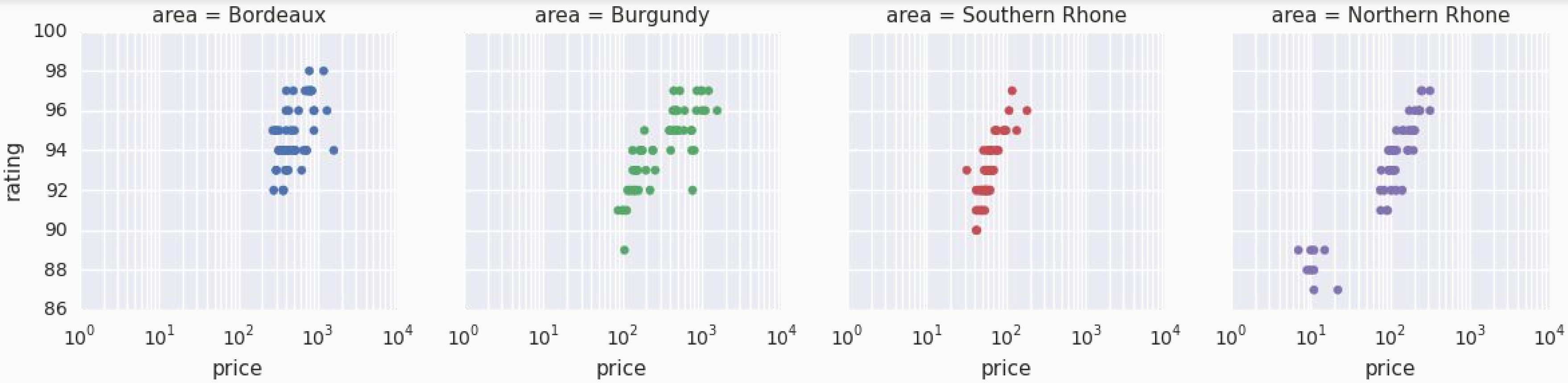
Pairwise plots of weather and price data in R

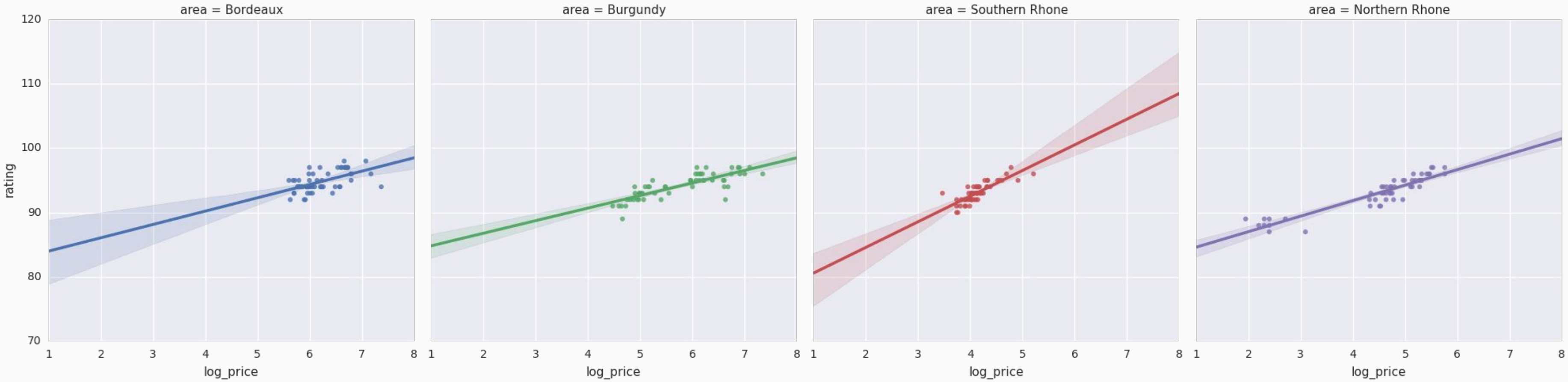


Weather analysis in Tableau



Analysing the relationship between prices and reviews





$$\rho = 0.746$$

Correlation between $\log_e$(price) and average rating.

This suggests a remarkably strong relationship between $\log_e$(price) and review score.

Web Dashboard

Dashboard Charts

- Log-price vs rating (including regressions)
- Price vs. weather trends
  - Humidity
  - Precipitation
- Prices over time (with MSCI stock index)
- Review words (word cloud)
- Parker effect
  - Highlighting the effect of Robert Parker’s review score on prices

Impact

Commercial potential

- Wine portfolio optimisation
  - Only buying “good” vintages means strong long-term portfolio returns
- Prediction of “good” vintages before bottling
  - Wine at the pre-bottling stage is known as “en primeur”
  - En primeur wine is cheaper to buy (at trade prices)
  - Prediction of “good” vintages enables guaranteed returns
- Ensuring that a purchase is at a “reasonable” price
  - Consumers may want to make sure they’re not paying over the odds
  - This data could spin off an online service to value wine

Future work

- More regions and vineyards could be added for worldwide comparison
- Many other features could be analysed;
    - Leaf cover on vineyards (possible with hyperspectral satellite imagery)
    - Chemical profile of wine in barrels
    - Chemical profile of soil (terroir)
    - Reputation of vineyard (could be measured with semantic analysis)
- The relationship between rating and price could be tested for causality
- Price prediction based on wine features