Foundations of Data Science with R

Jameson Watts, Ph.D.

Agenda

1. Course Overview and Expectations
2. Example Analysis of Wine Prices

Course Overview and Expectations

About Me

- Background
  - BS in Computer Science from UC, Boulder
  - MBA from Willamette
  - Ph.D. in Marketing from U of A (minor in computational linguistics)
  - ~10 years programming professionally + ~10 years programming for research
- Contact
  - Website jamesonwatts.github.io
  - Email: jwatts@willamette.edu
  - Office Hours: after class
  - Appointments: jamesonwatts.youcanbook.me (Skype or phone call)

Class Materials

- Base R
- RStudio 1.2
- R for Data Science
- DataCamp Classroom

Other resources:

- https://twitter.com/r4dscommunity
- https://bookdown.org/yihui/rmarkdown
- http://google.com

Reading the Course Outline

- From the syllabus
- Class Topics
  - Subjects I plan to cover during that day’s lecture
- Reading and Assignments
  - DCC: assignments in the DataCamp Classroom
  - R4ds: chapters to read in the online textbook

Assignments

- DataCamp homework assignments (25%)
- Midterm exams (50%)
- Final Presentations and Report (25%)

Course Policies and Expectations

- Name tents
- Collaboration
- Late work
- Effort
    - 2-4 hours outside of class each week
    - struggle, Google, StackExchange, struggle, Google, doh!
    - start with the basics… ramp up very fast

Analysis of wine prices

Overview of Data

- Grabbed from Kaggle here
- Scrape of wine reviews, scores, and prices from Wine Enthusiast during week of 6/15/2017
- Includes region, taster’s name, variety and winery
- 130k observations
- Some background reading here

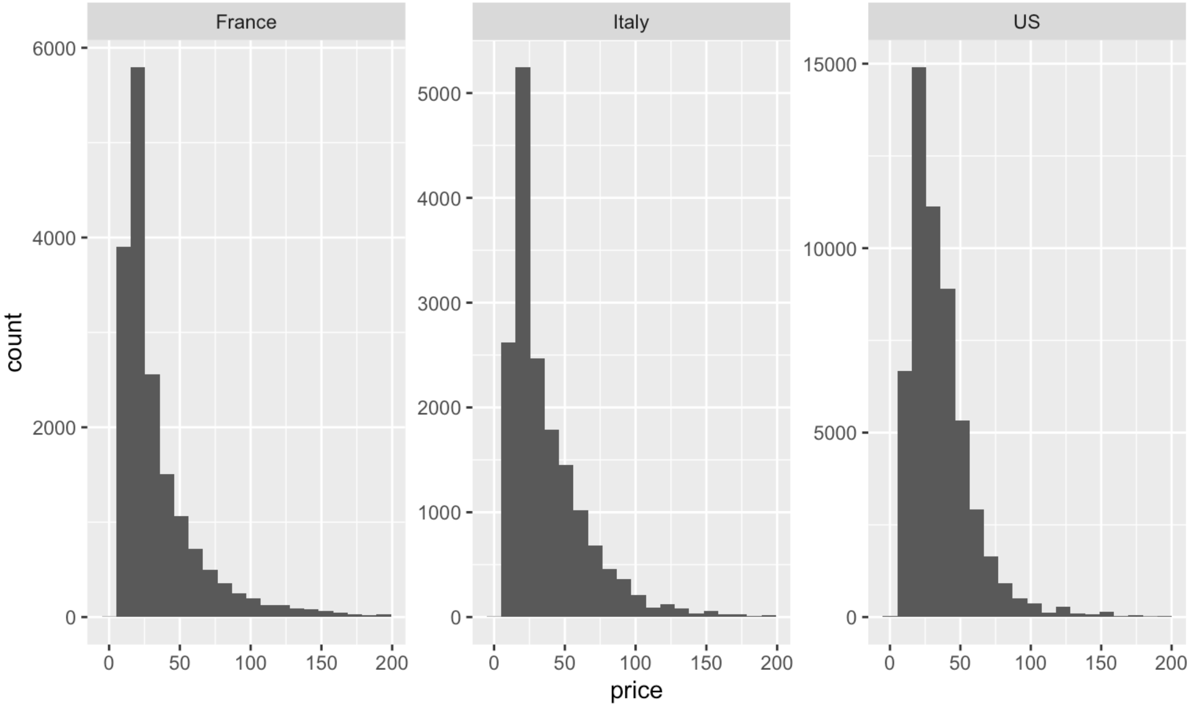
Summarize Dataset

```
## Observations: 129,971
## Variables: 14
## $ id                    <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, …
## $ country               <chr> "Italy", "Portugal", "US", "US", "US", "Sp…
## $ description           <chr> "Aromas include tropical fruit, broom, bri…
## $ designation           <chr> "Vulkà Bianco", "Avidagos", NA, "Reserve L…
## $ points                <dbl> 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, 87…
## $ price                 <dbl> NA, 15, 14, 13, 65, 15, 16, 24, 12, 27, 19…
## $ province              <chr> "Sicily & Sardinia", "Douro", "Oregon", "M…
## $ region_1              <chr> "Etna", NA, "Willamette Valley", "Lake Mic…
## $ region_2              <chr> NA, NA, "Willamette Valley", NA, "Willamet…
## $ taster_name           <chr> "Kerin O’Keefe", "Roger Voss", "Paul Gregu…
## $ taster_twitter_handle <chr> "@kerinokeefe", "@vossroger", "@paulgwine …
## $ title                 <chr> "Nicosia 2013 Vulkà Bianco  (Etna)", "Quin…
## $ variety               <chr> "White Blend", "Portuguese Red", "Pinot Gr…
## $ winery                <chr> "Nicosia", "Quinta dos Avidagos", "Rainsto…
```

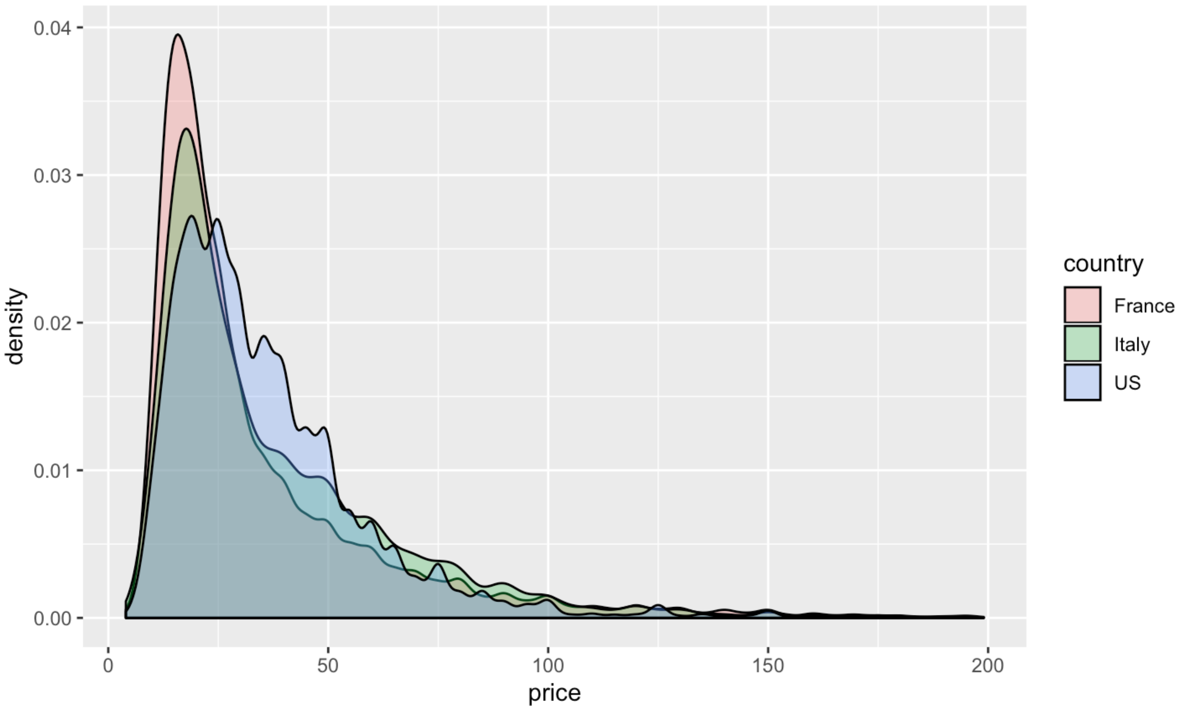
Possible Research Questions

- What is the mean/median rating and cost of a bottle of red wine?
- Is wine from the Willamette Valley more or less expensive than wine from elsewhere?
    - Against which regions do we have a comparative advantage?
    - Where are we at a disadvantage?
- Do the most prolific tasters have a preference for a certain region or type of wine?
- What is the relationship between rating and price? Are there confounds?
- Are there certain words always associated with the highest rated wine?
- Which wines are a ‘good’ deal?

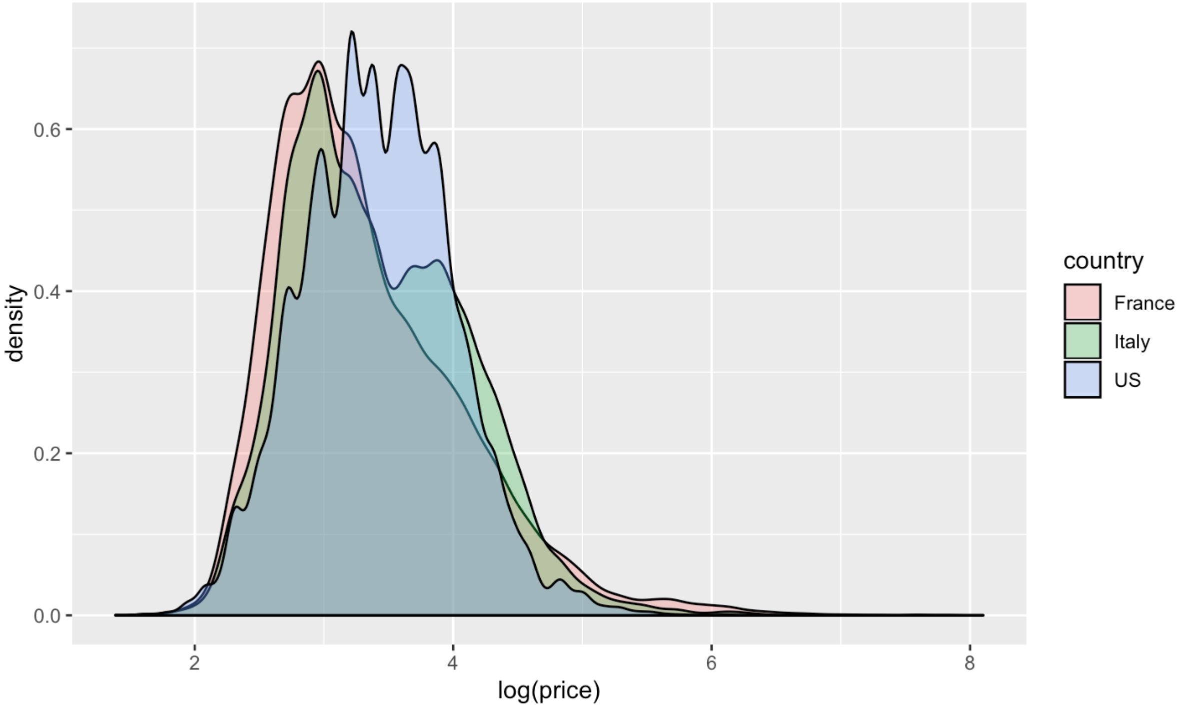
Wine Prices (< $200) Histogram



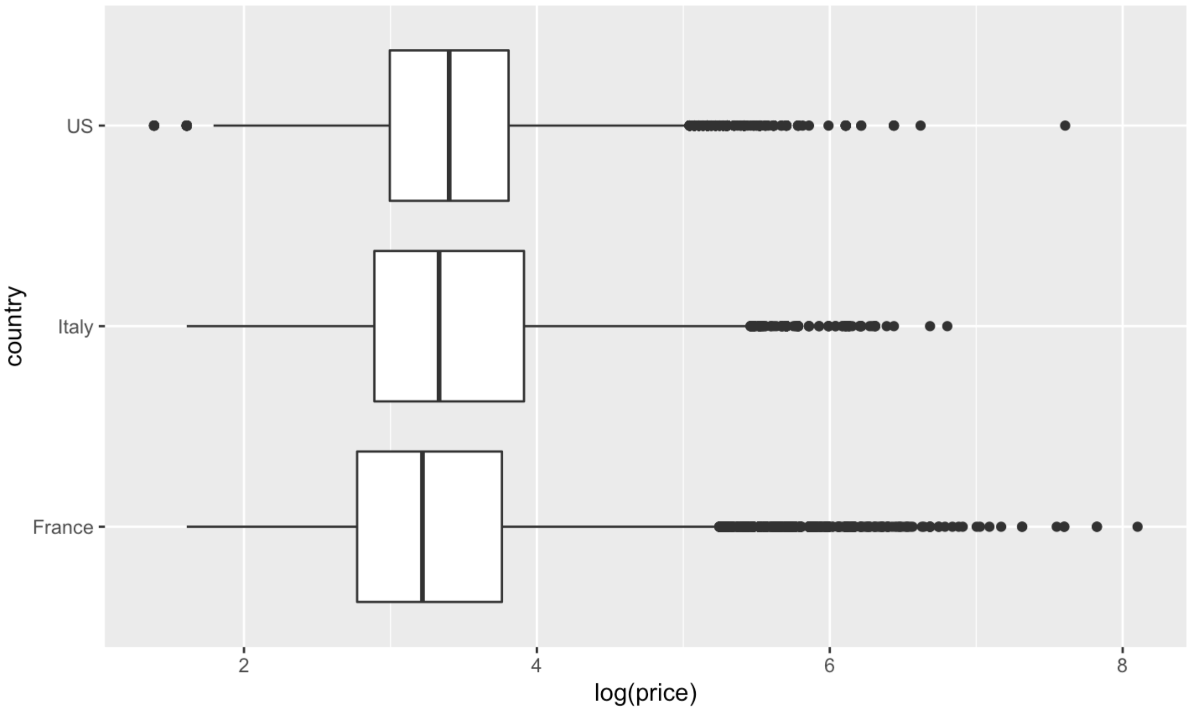
Wine Prices (< $200) Density



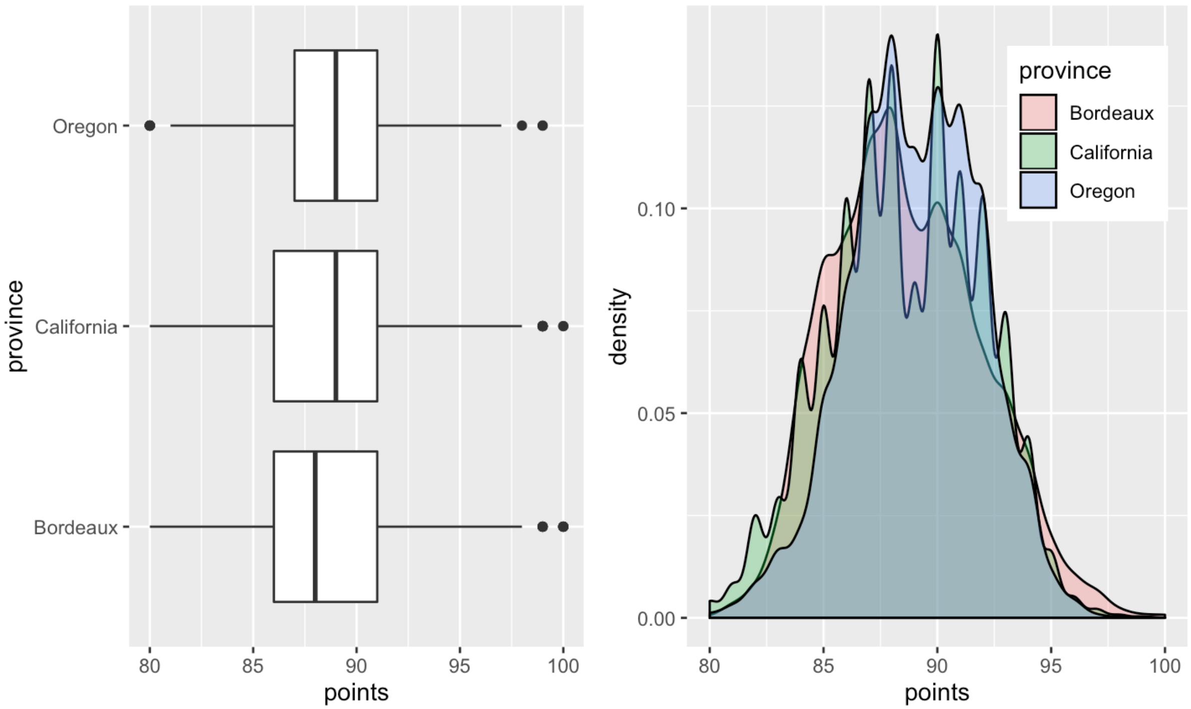
Wine log(Prices) Density



Means and Medians

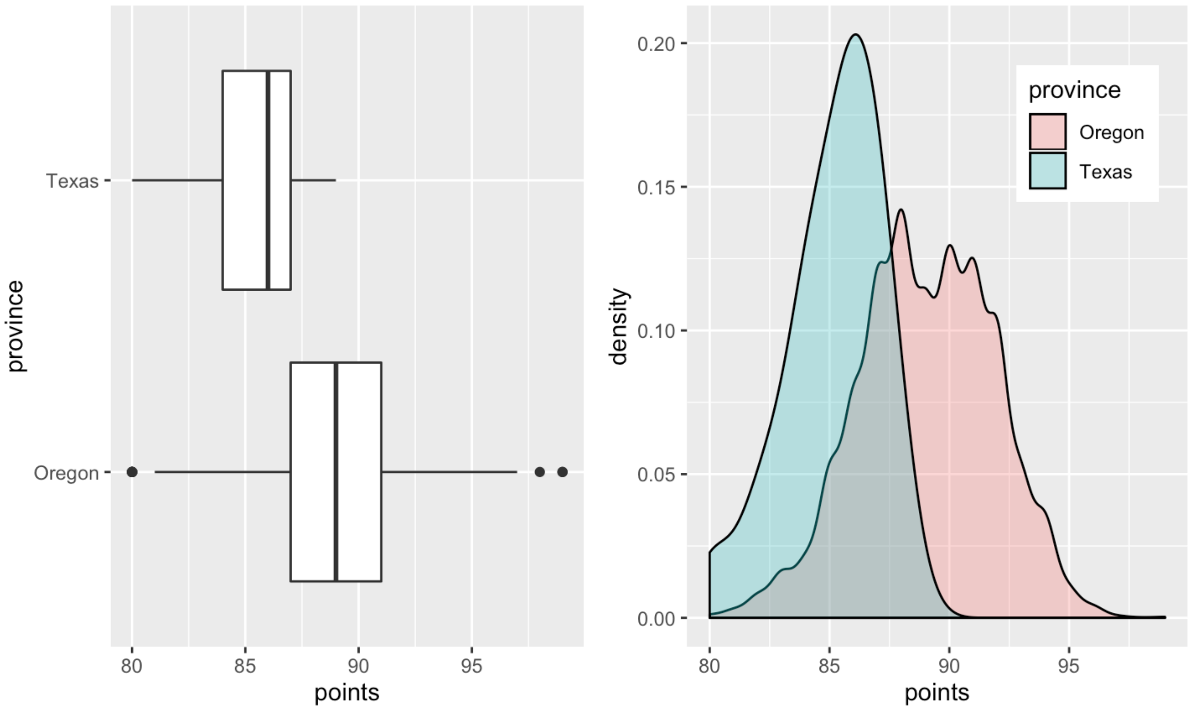


Oregon vs. California vs. Bordeaux (Ratings)



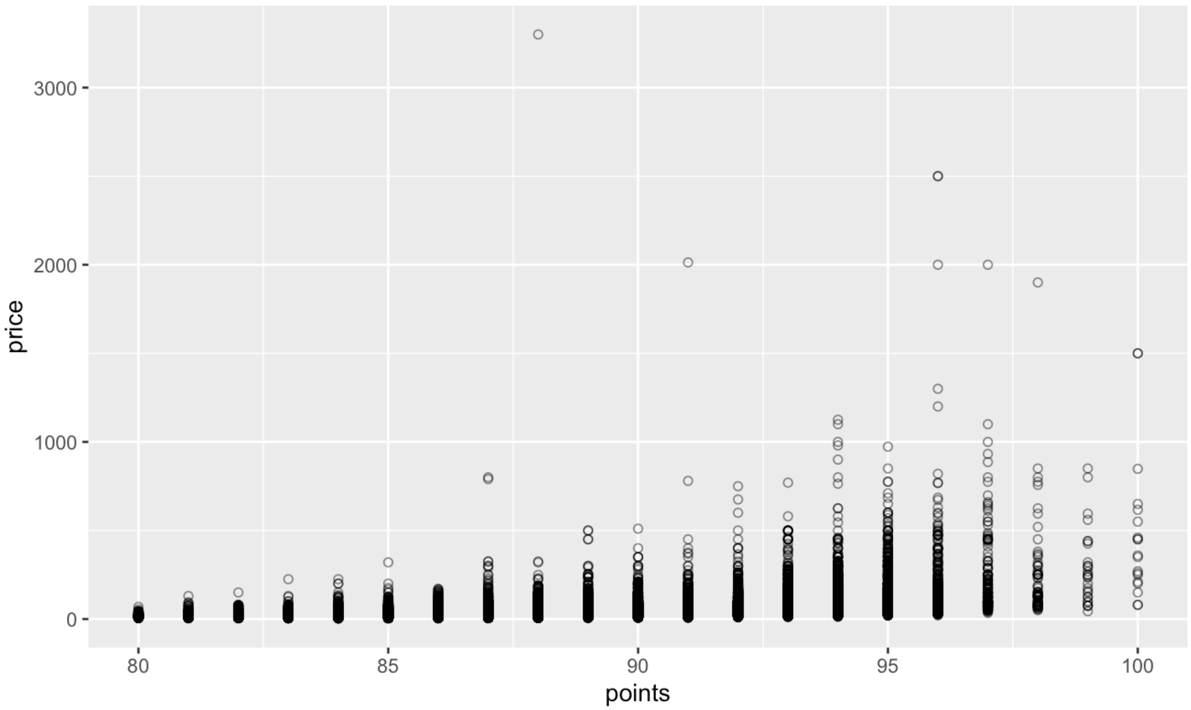
Ok, all pretty quality. How do we compare with Texas?

Oregon vs. Texas (Ratings)



…thank goodness. Let’s get back to the relationship between ratings and price.

Ratings and Price



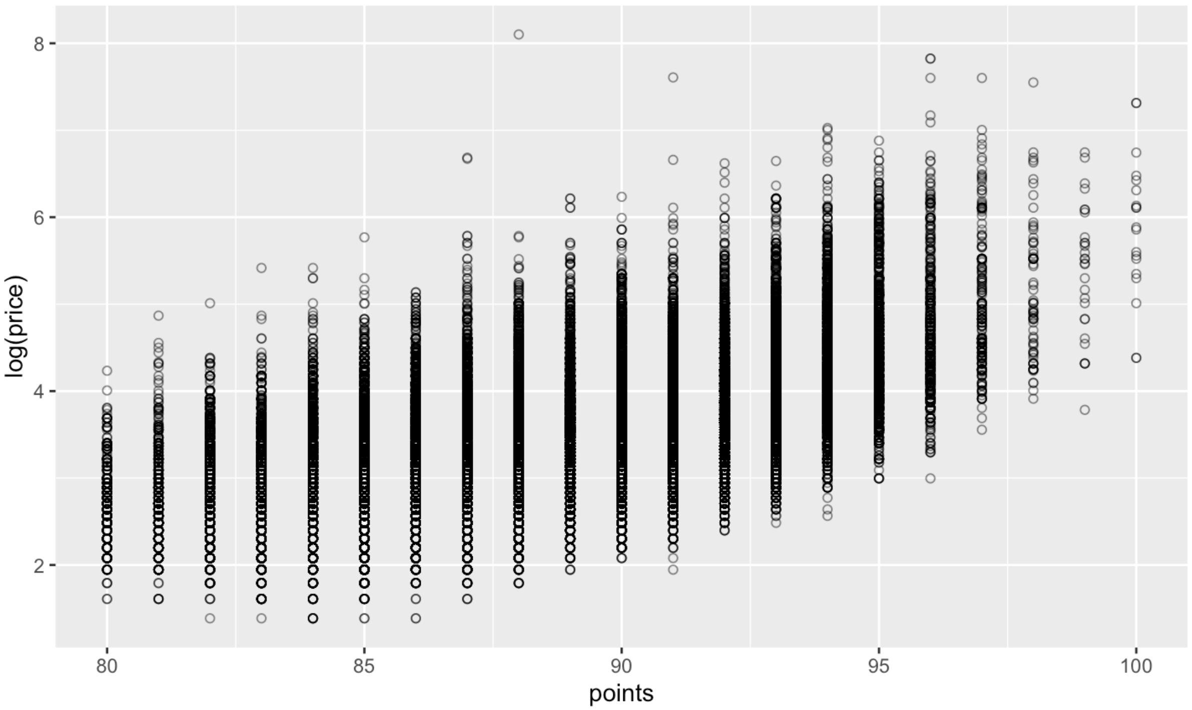
So perhaps we can start to see what is a ‘good’ deal and what isn’t. Let’s look at the crazy outliers.

Who are the crazy outliers? (price > 1000)

```
## # A tibble: 14 x 5
##    points price country province   title                                
##     <dbl> <dbl> <chr>   <chr>      <chr>                                
##  1     88  3300 France  Bordeaux   Château les Ormes Sorbet 2013  Médoc 
##  2     96  2500 France  Bordeaux   Château Pétrus 2014  Pomerol         
##  3     96  2500 France  Burgundy   Domaine du Comte Liger-Belair 2010  La …
##  4     91  2013 US      California Blair 2013 Roger Rose Vineyard Chardonn…
##  5     97  2000 France  Bordeaux   Château Pétrus 2011  Pomerol         
##  6     96  2000 France  Burgundy   Domaine du Comte Liger-Belair 2005  La …
##  7     98  1900 France  Bordeaux   Château Margaux 2009  Margaux        
##  8    100  1500 France  Bordeaux   Château Lafite Rothschild 2010  Pauillac
##  9    100  1500 France  Bordeaux   Château Cheval Blanc 2010  Saint-Émilion
## 10     96  1300 France  Bordeaux   Château Mouton Rothschild 2009  Pauillac
## 11     96  1200 France  Bordeaux   Château Haut-Brion 2009  Pessac-Léognan
## 12     94  1125 France  Burgundy   Domaine du Comte Liger-Belair 2006  La …
## 13     97  1100 France  Bordeaux   Château La Mission Haut-Brion 2009  Pes…
## 14     94  1100 Austria Wachau     Emmerich Knoll 2013 Ried Loibenberg Sma…
```

…so there’s something going on with the French Bordeaux region. We should keep this in mind when we model price. But let’s get back to price/ratings relationship…

Ratings x log(price)



Okay, so the relationship is a bit clearer. But also, there is definitely some variance. Let's first get an estimate of the slope and then see if things are different by region.

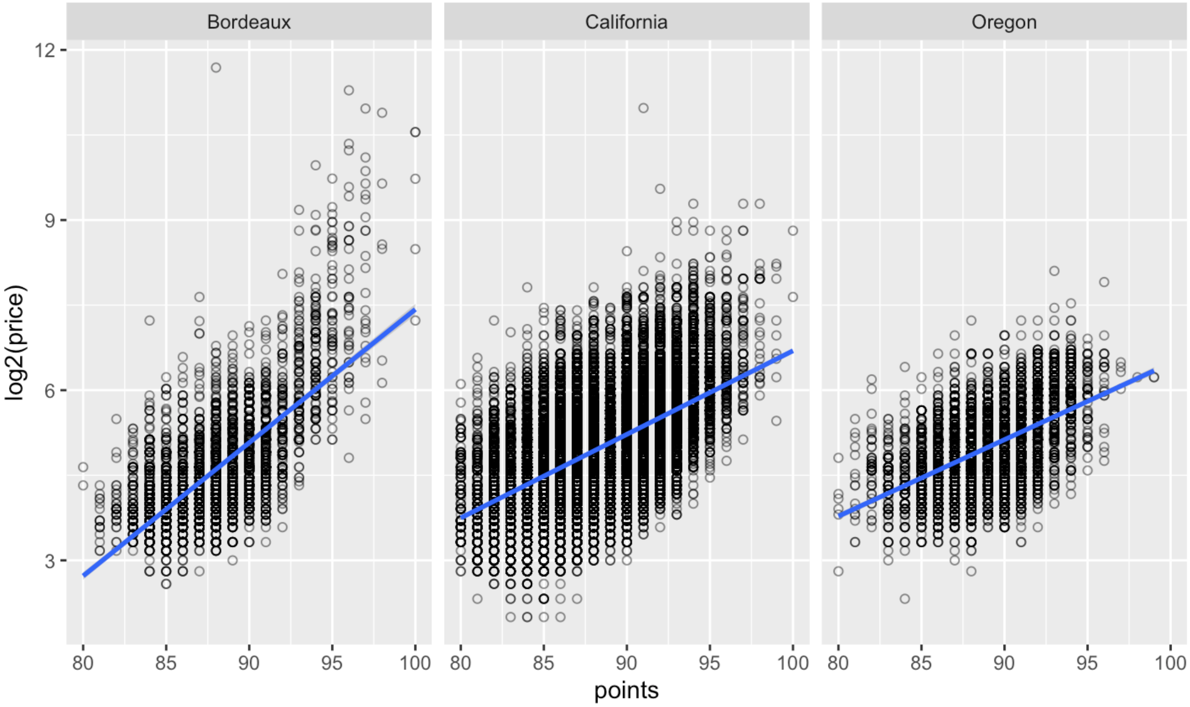
Simple linear model

```
##
## Call:
## lm(formula = lprice ~ points, data = wine %>% mutate(lprice = log(price)))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7076 -0.3688 -0.0405  0.3177  4.8425
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.3076501  0.0432237  -192.2   <2e-16 ***
## points       0.1314413  0.0004885   269.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5173 on 120973 degrees of freedom
##   (8996 observations deleted due to missingness)
## Multiple R-squared:  0.3744, Adjusted R-squared:  0.3744
## F-statistic: 7.239e+04 on 1 and 120973 DF,  p-value: < 2.2e-16
```

Since we logged the DV, a 1 point ratings increase = `14.05`% increase in price on average. Note:

$$(e^x - 1) * 100$$

Ratings x ln(price) by Region



…so the slopes look different. Let’s actually run a model to see if they are.

Linear models for each province

Bordeaux

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept  -11.1       0.243     -45.9       0  -11.6    -10.7
## 2 points       0.163     0.003      59.1       0    0.157    0.168
```

California

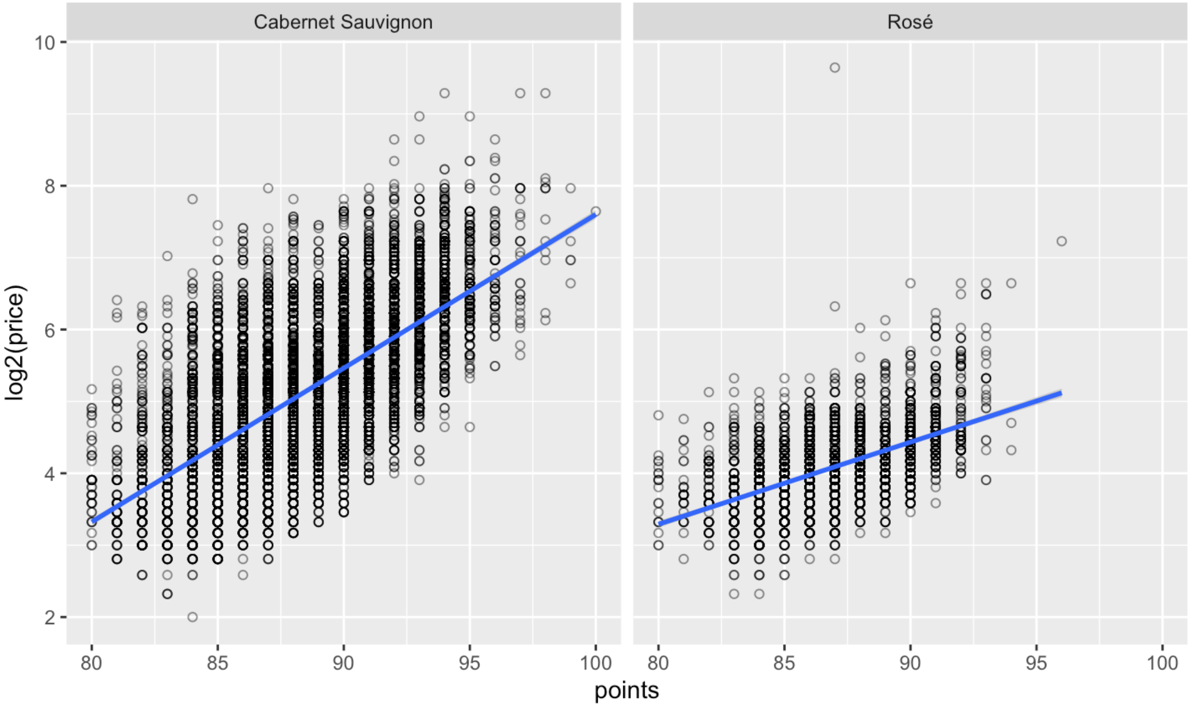
```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   -5.57      0.071     -78.9       0   -5.71    -5.44
## 2 points       0.102     0.001     128.        0    0.101    0.104
```

Oregon

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept   -4.87      0.19      -25.7       0   -5.24    -4.50
## 2 points       0.094     0.002      43.9       0    0.089    0.098
```

What are the percent increases in price for each point by region?

Cabernet or Rose?



Questions?