Lab 2 - Team 3 - Research Proposal

Andrew Abrahamian, Victoria Hollingshead, Heesuk Jang, Hsi-sheng Wei

2022-07-19

Research Proposal

1. Our team will be exploring the following research question: **Are wine ratings influenced by price, grape variety, region, year of vintage?**

- X will be price for a bottle of the wine (metric), grape type (categorical), region(category), and vintage (category).
  - X main concept - Price
  - X sub concepts - Grape variety, region, year of vintage
- Y will be wine rating in an ordinal scale, which is defined in the number of points Wine Enthusiast rated the wine on a scale of 80 -100 (Reviews for wines that score 1 - 79 are not available in the raw data set)

2. The data source will be from Tidy Tuesday: Wine Enthusiast Reviews.There are 129,971 rows and 13 columns. After removing reviews of identical wines and reviewers, the dataset is reduced to 108,290 rows. We are interested in filtering the dataset to the following sample frame, resulting in 26,244 observations.

- Reviews: The sampling frame is exclusively composed of anonymous reviewers. Because we have limited information on reviewer details, we assume all anonymous reviewers are unique, have equal access to all wines in our model, and are pulled from the same distribution.
  - Note on Reviews: Without any filtering, the dataset is overrepresented by 19 named reviewers. Reviews from 19 named reviewers make up over 80% of the overall dataset. We attribute this to the tendency for people who enjoy writing reviews to write more reviews. It is also possible that named reviewers are employed by Wine Enthusiasts, thus incentivizing their voluminous review count. In order to meet the IID requirements, we removed the reviews from these 19 named reviewers. We reason that removing this overrepresented group will allow us to minimize the bias in the sample distribution. Based on this, we assume all unknown reviewers have an equal opportunity to be selected.
- Universe
  - Countries: We will be restricting the sample frame to wines produced in the US. Assuming collinearity between country and region, this restriction allows us to use region as an X concept in our regression models.
  - Grape Varieties: We will be restricting our model to the top 3 grape varieties.
  - Region: We will be restricting our model to the top 3 regions in the US.
  - Vintage: We will include 5-6 vintage year categories.

3. The unit of observation is a unique product review per bottle of wine.

Sush Feedback

1. the research question seems too broad. Can you recognize a primary predictor and build the study on top of it. You may eventually add other covariates to build a better model, but the study should revolve around the primary predictor of interest.

- Points ~ Price

- Options:
  - Points as a primary predictor. Requires Price to be treated as our Y variable
    - treat Price as our outcome

Price (Y) ~ Points (X) - grape variety could be more defensible (more valuable than country/region?) - vintage as a rule of thumb for predicting price ~ useful as a control variable

2. do you think it is a good idea to use an ordinal variable for the response. You could use the usual OLS regression, but your study would then have its limitations. Within this context, can you suggest a different response variable that is metric??

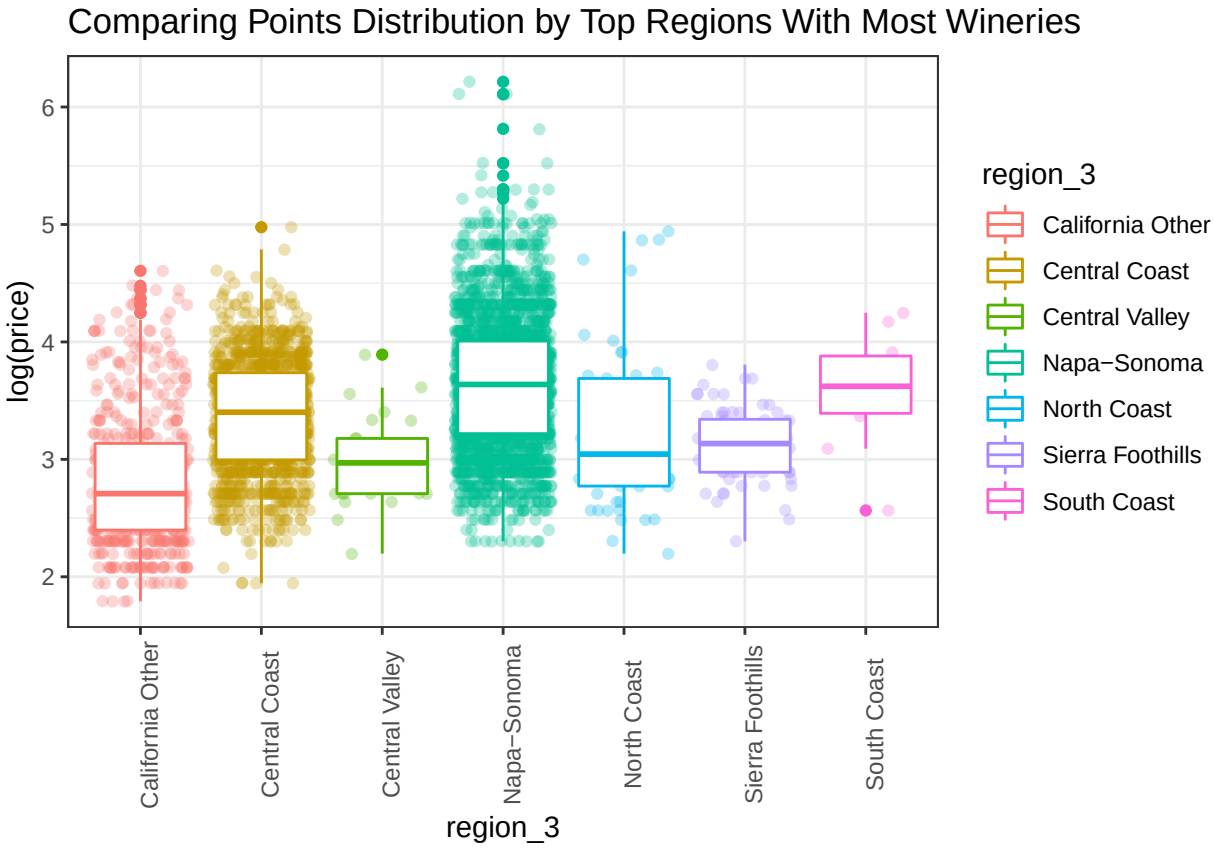
- Evaluate Price as our Y variable

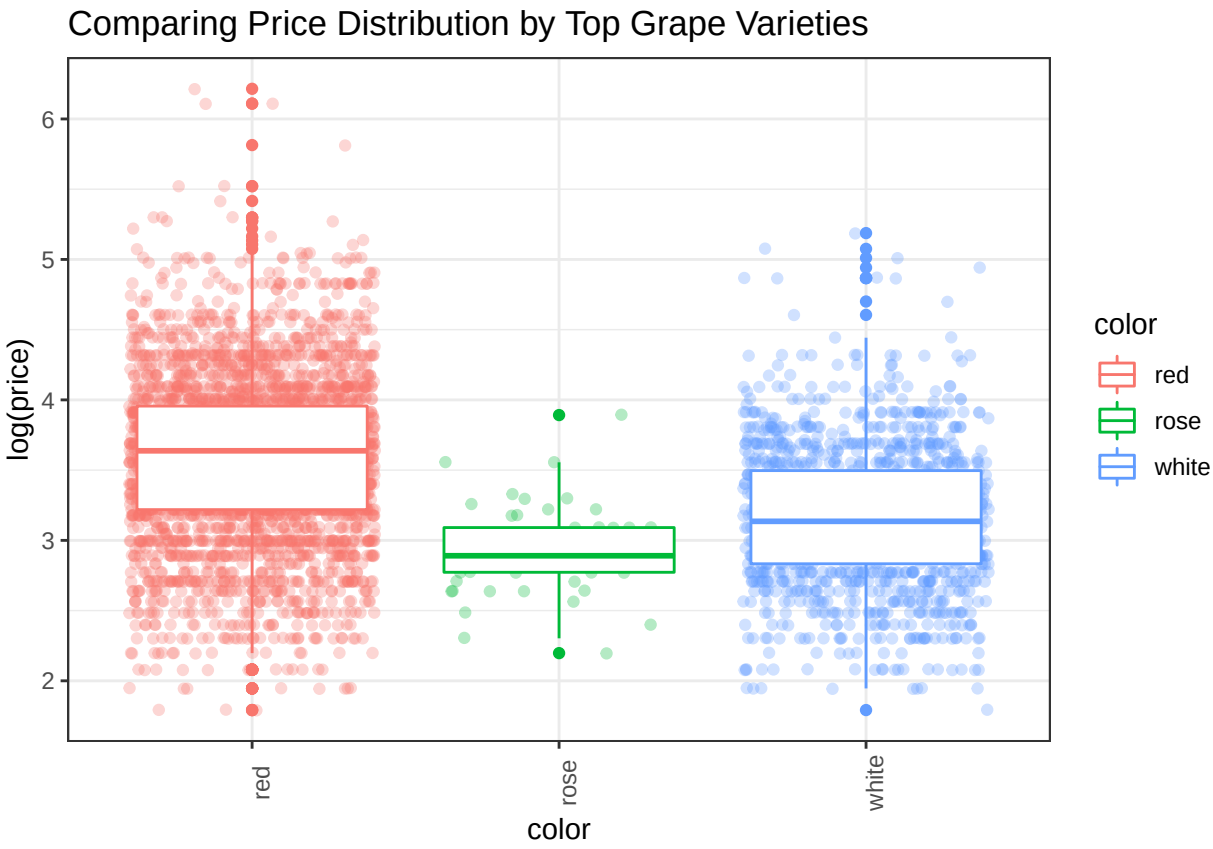
3. consider having a prior hypothesis about the effect of X on Y?

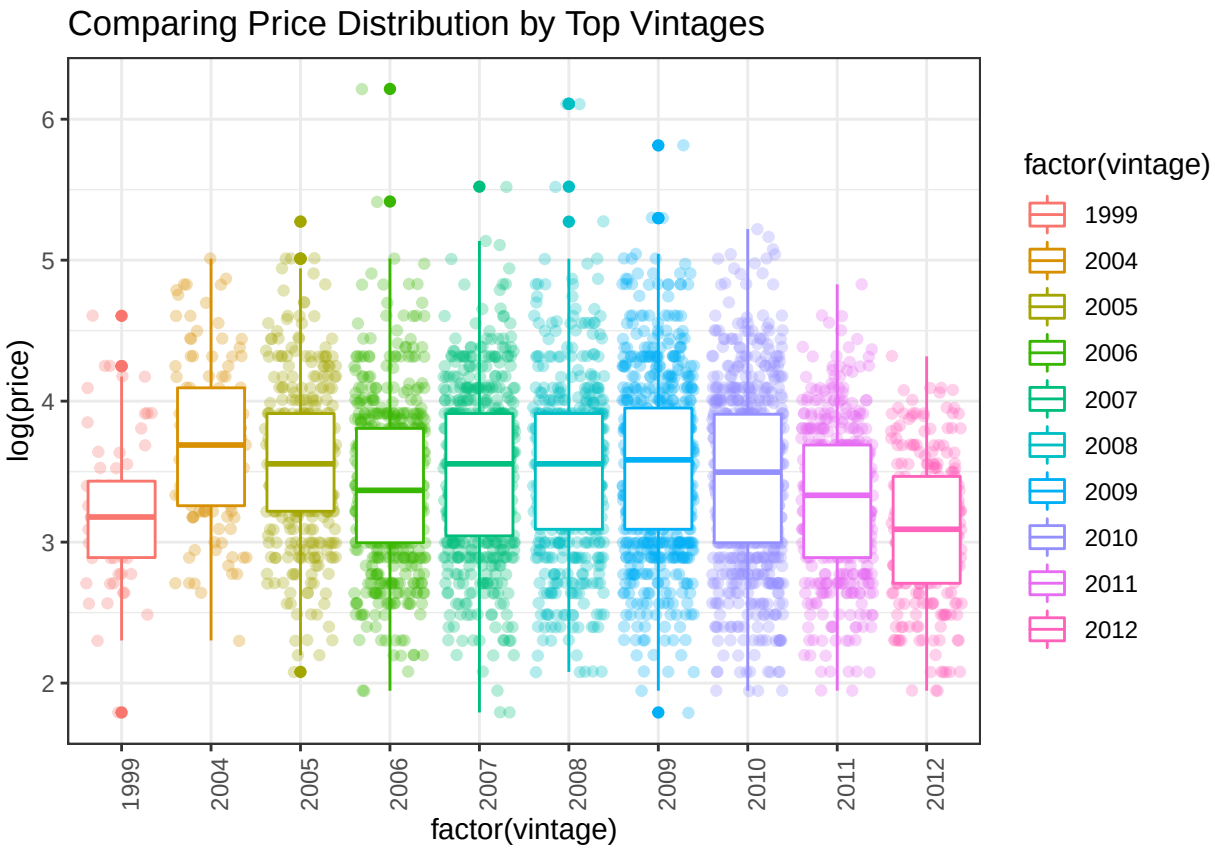
- Write down our theory of the relationship
- Develop first hypothesis test
  - Null Hypothesis: No relationship between price (outcome) and points (predictor)
    - Assume the coefficient of points is equal to zero
  - Alt. Hypothesis: Assume that wine with higher rating will have a higher price controlling for vintage and grape varieties

Additional Covariates: * Vintage * Grape Variety * Region

Scope Decisions: * Global Model: country + vintage + grape variety <- eliminate due to issues w/ IID **Country-specific: region + vintage + grape variety**

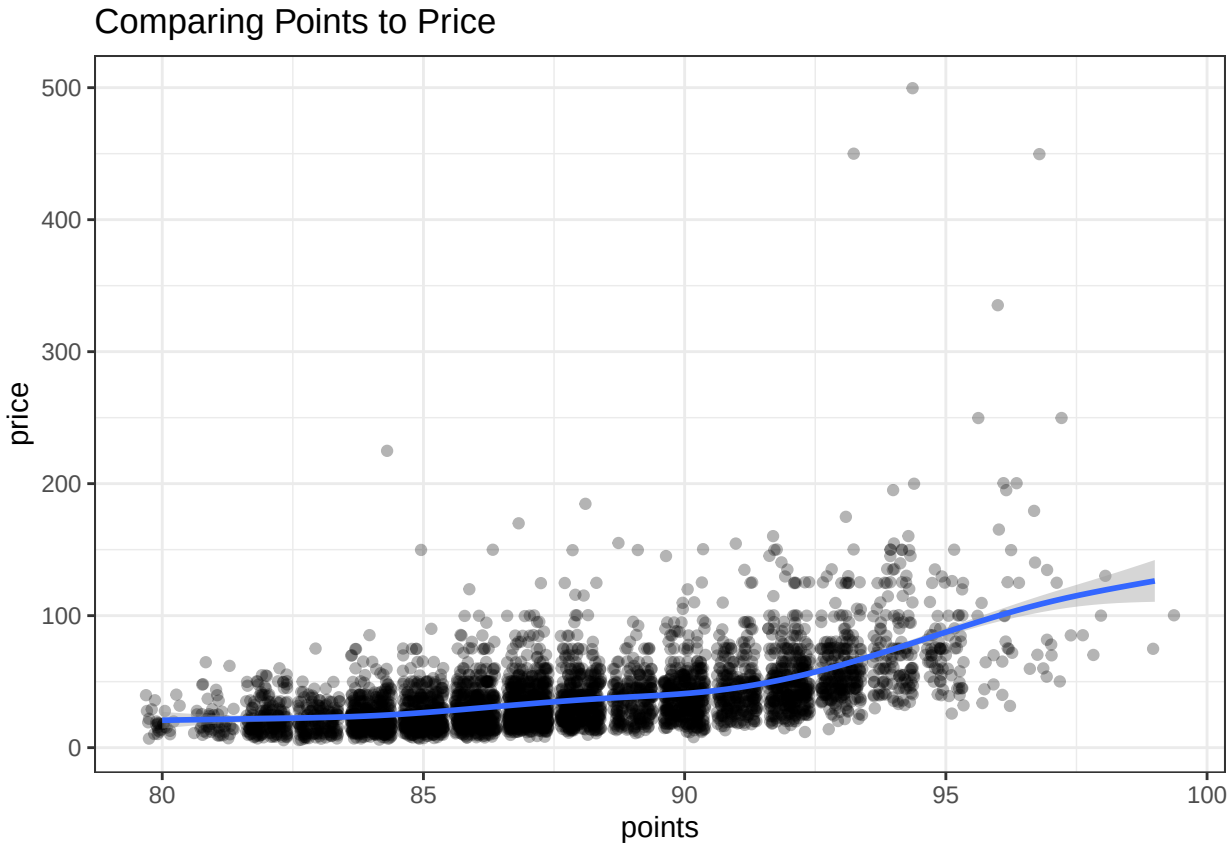

Comparing Points Distribution by Top Regions With Most Wineries
log(price)
6
5
4
3
2
California Other
Central Coast
Central Valley
Napa−Sonoma
North Coast
Sierra Foothills
South Coast
region_3
region_3
California Other
Central Coast
Central Valley
Napa−Sonoma
North Coast
Sierra Foothills
South Coast

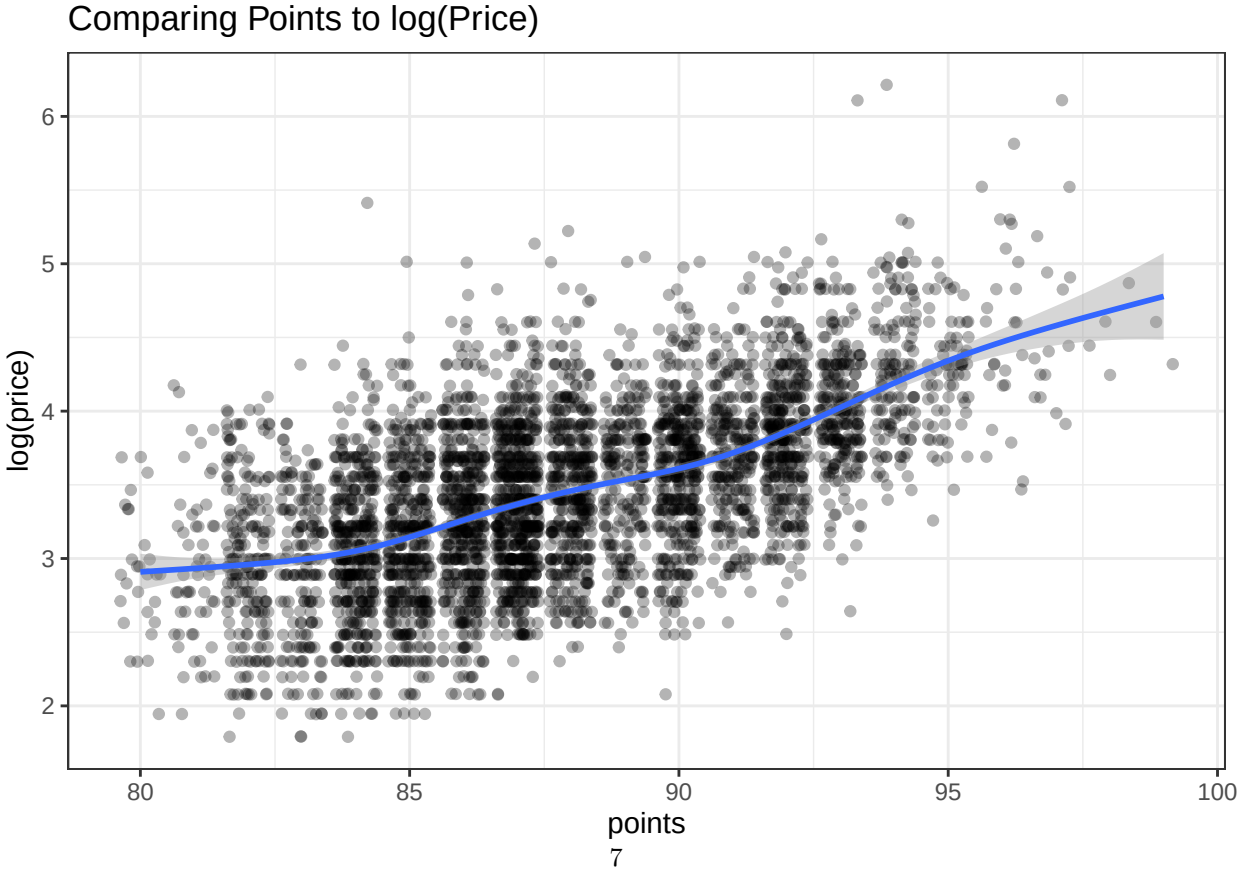

Comparing Price Distribution by Top Grape Varieties
log(price)
6
5
4
3
2
red
rose
white
color
color
red
rose
white

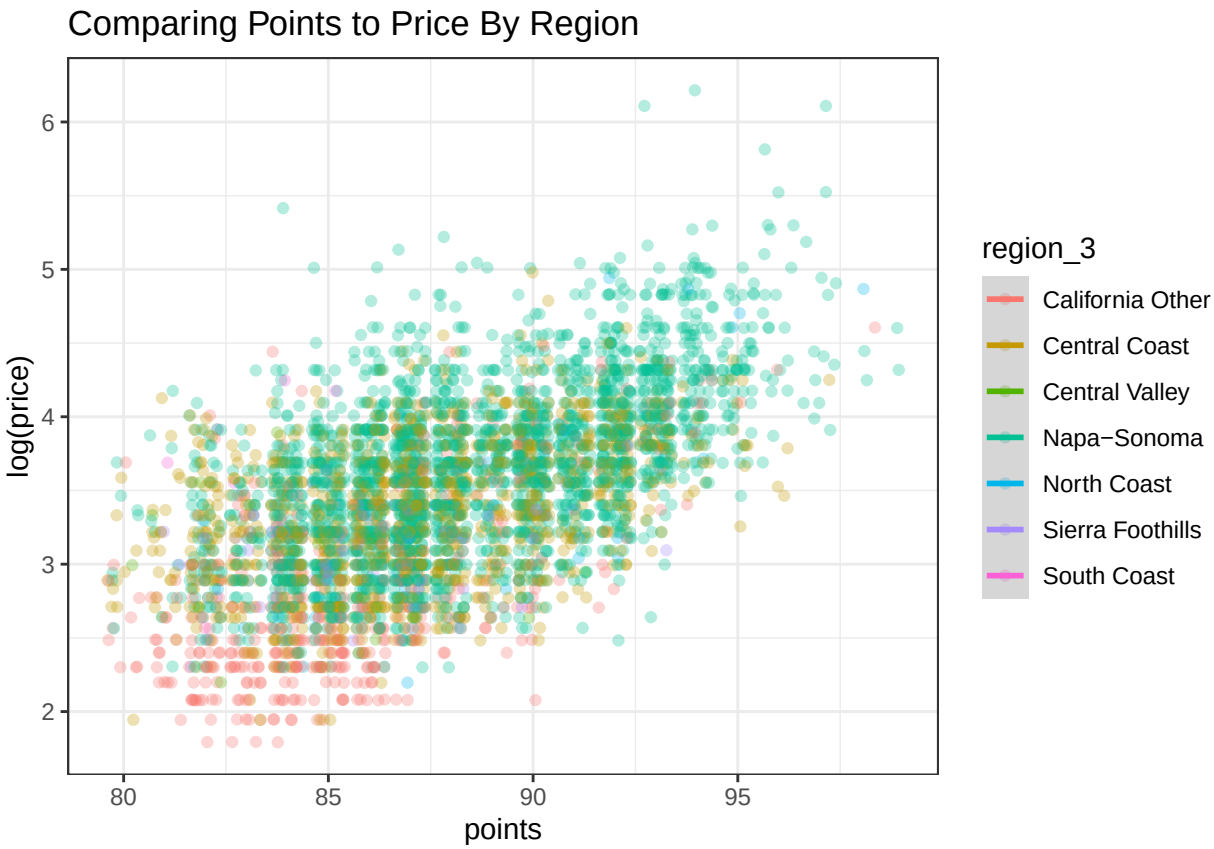


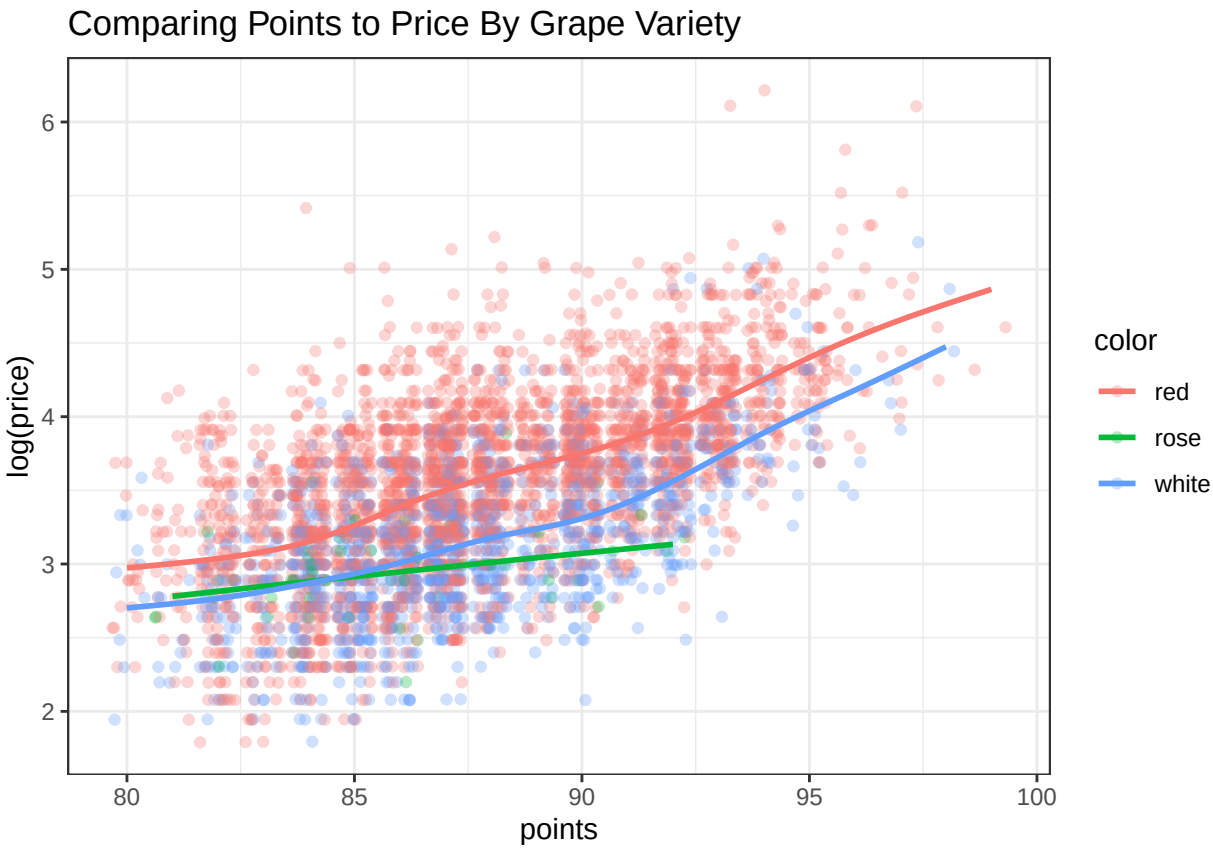
Our Theory: * type of grape likely covaries with region due to temperature and soil requirements * some variation could be explained by variety and by region * vintage may explain some variation, but it could be limited

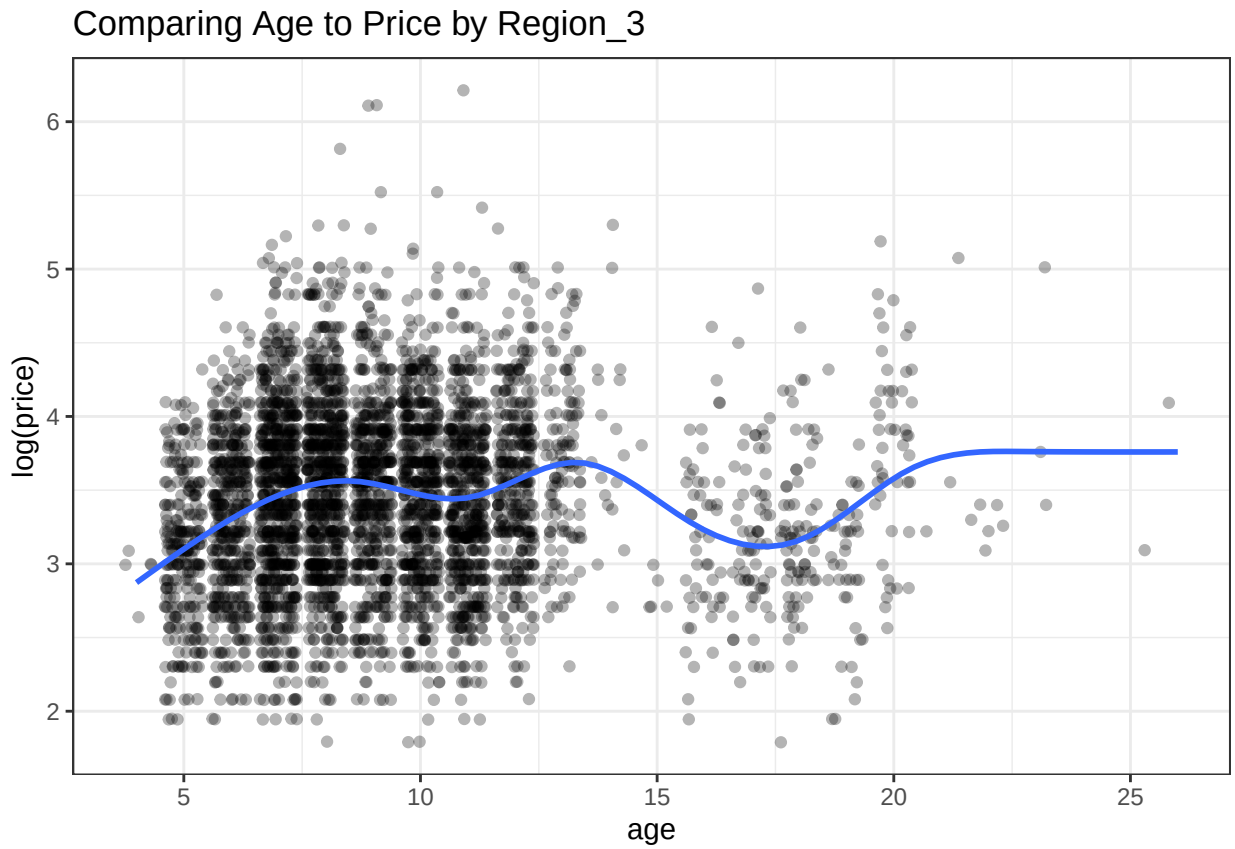
Main Variable Comparison






Comparing Points to Price By Region
log(price)
points
6
5
4
3
2
80
85
90
95
region_3
California Other
Central Coast
Central Valley
Napa−Sonoma
North Coast
Sierra Foothills
South Coast


Comparing Points to Price By Grape Variety
log(price)
points
color
red
rose
white
6
5
4
3
2
80
85
90
95
100


Comparing Age to Price by Region_3
log(price)
6
5
4
3
2
5
10
15
20
25
age

```
## [1] 0.5888724
```

What Models Do We Want To Build

Model #1: Our Primary Relationship

- log(Price) ~ Points
- Selecting the log-linear model to use points + other covariates to explain **changes in price** caused by a change in points (rating)

Model #2:

- log(Price) ~ Points + variety

Model #3:

- log(Price) ~ Points + variety + region_1

Model #4:

- log(Price) ~ Points + variety + region_1 + vintage
  - Vintage as metric: we're defining relationship between price x vintage
  - Vintage as ordinal: distinct intercepts between points x price relationship
    - ordinal can be subset of metric
  - Expect the relationship: older vintage (smaller #), higher the price

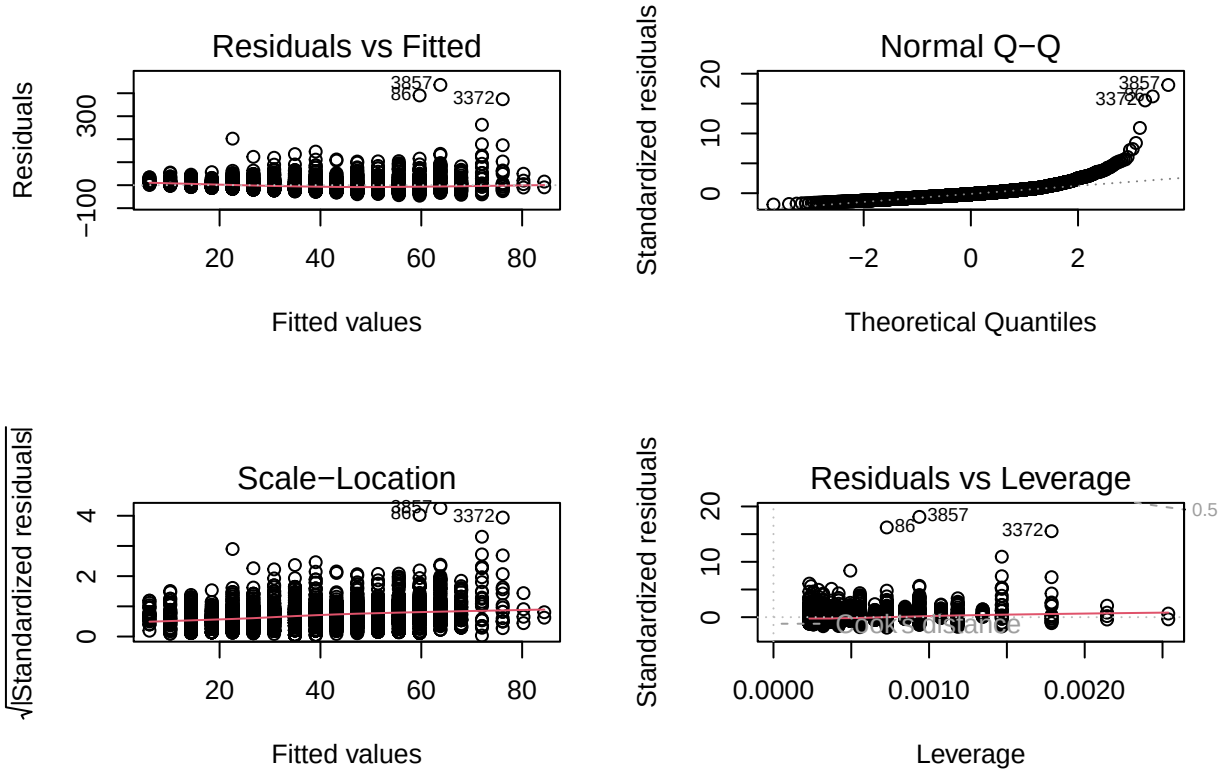
What Is Needed Before Building Models

- Build hypothesis test + theory of relationship between points and log(price) ~ X and Y
- Identify specific layers of covariates to include in regression model

Begin Model Building

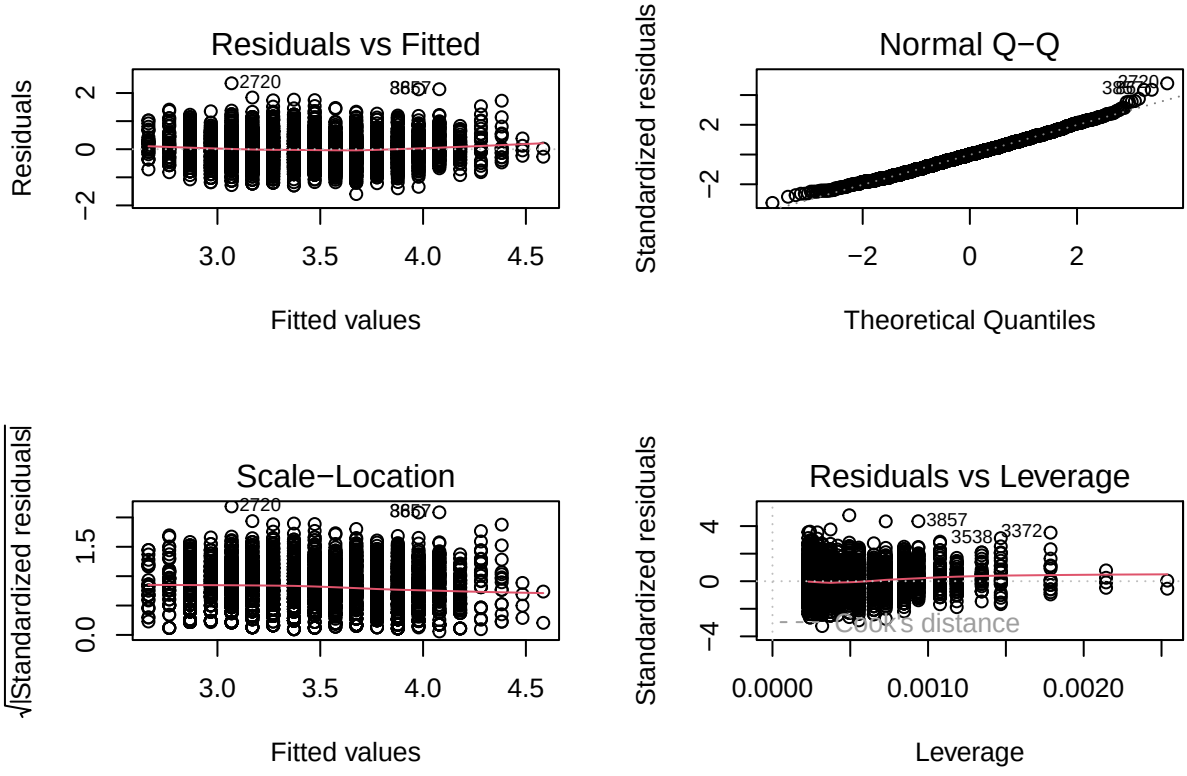
Comparing Level-Level and Log-Level Model

```
##
## Call:
## lm(formula = price ~ points, data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -45.66 -13.68  -4.06   8.30 436.22
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -323.5353     9.0723  -35.66   <2e-16 ***
## points         4.1204     0.1033   39.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.09 on 4349 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.268,  Adjusted R-squared:  0.2679
## F-statistic:  1592 on 1 and 4349 DF,  p-value: < 2.2e-16
```



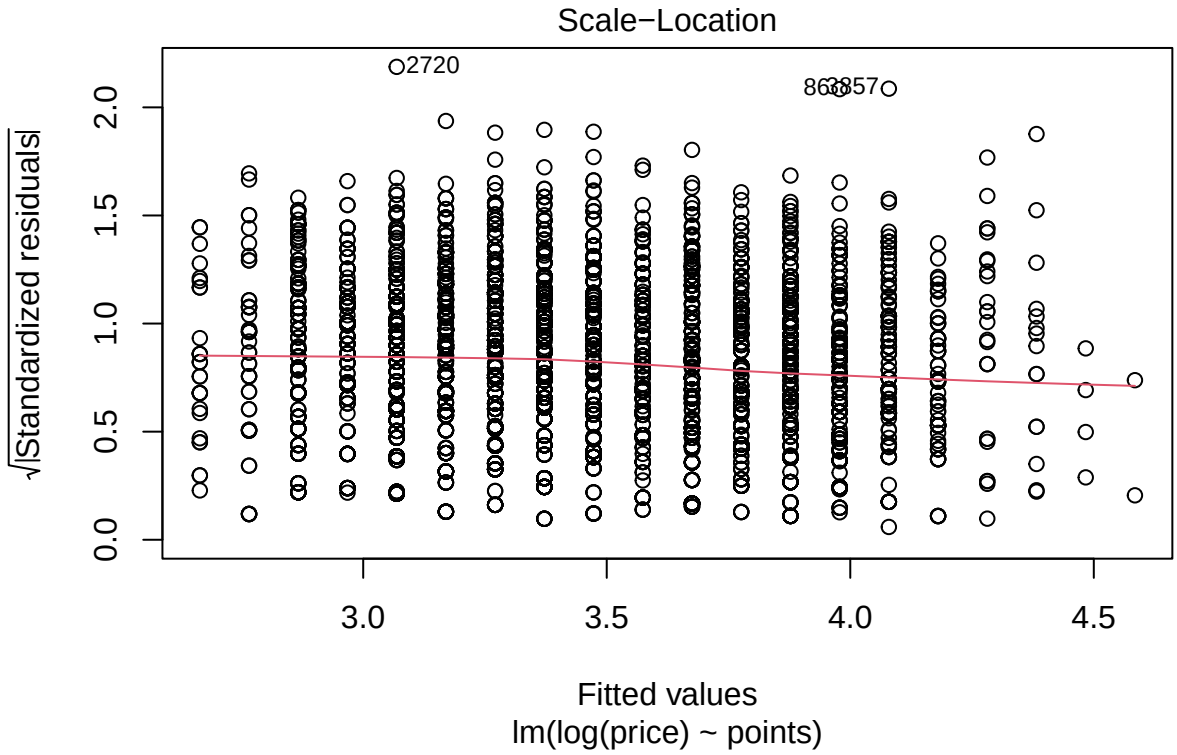

```
##
```

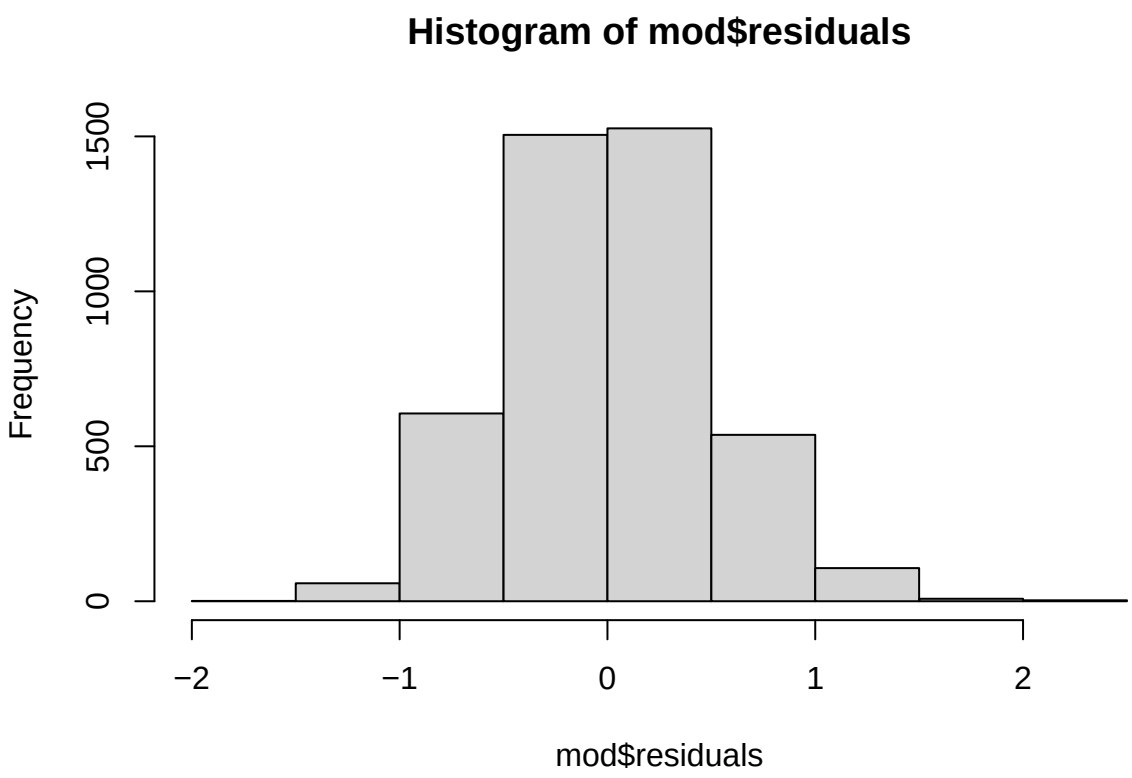
```
## Call:
## lm(formula = log(price) ~ points, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.59562 -0.34285  0.00696  0.32167  2.34733
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.419263   0.184777  -29.33   <2e-16 ***
## points       0.101048   0.002103   48.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4907 on 4349 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.3468, Adjusted R-squared:  0.3466
## F-statistic:  2309 on 1 and 4349 DF,  p-value: < 2.2e-16
```

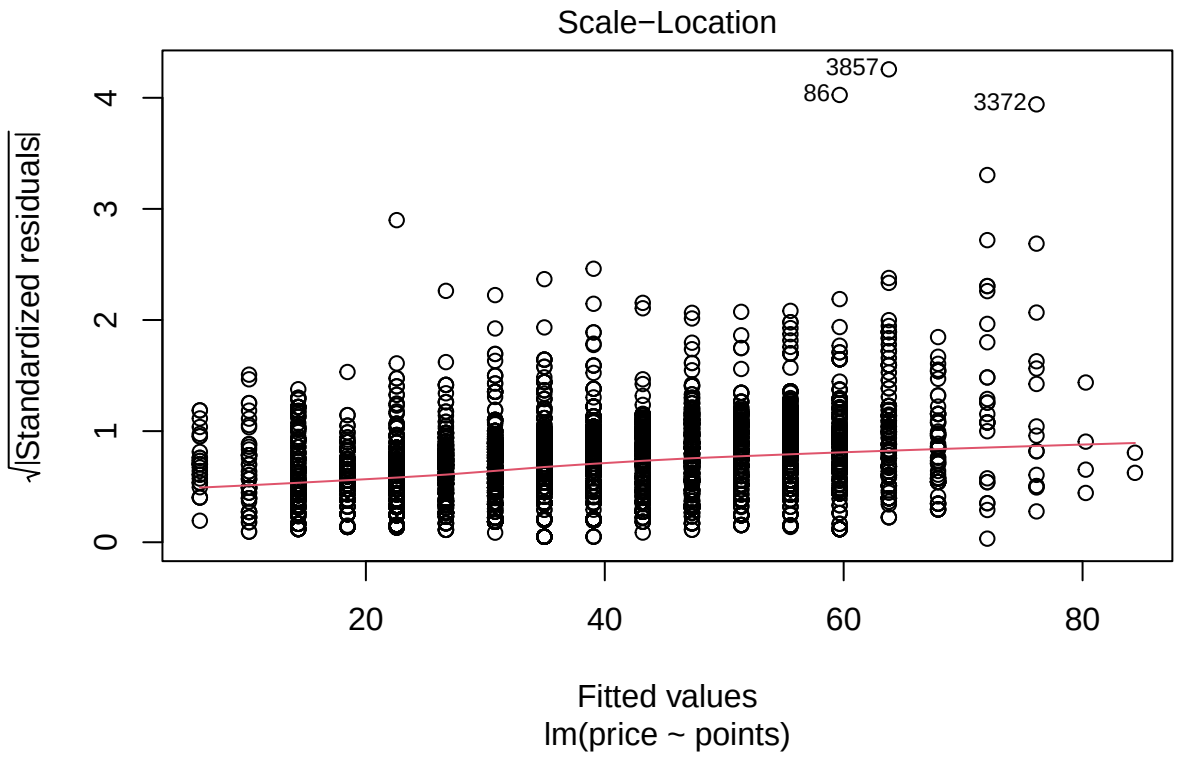


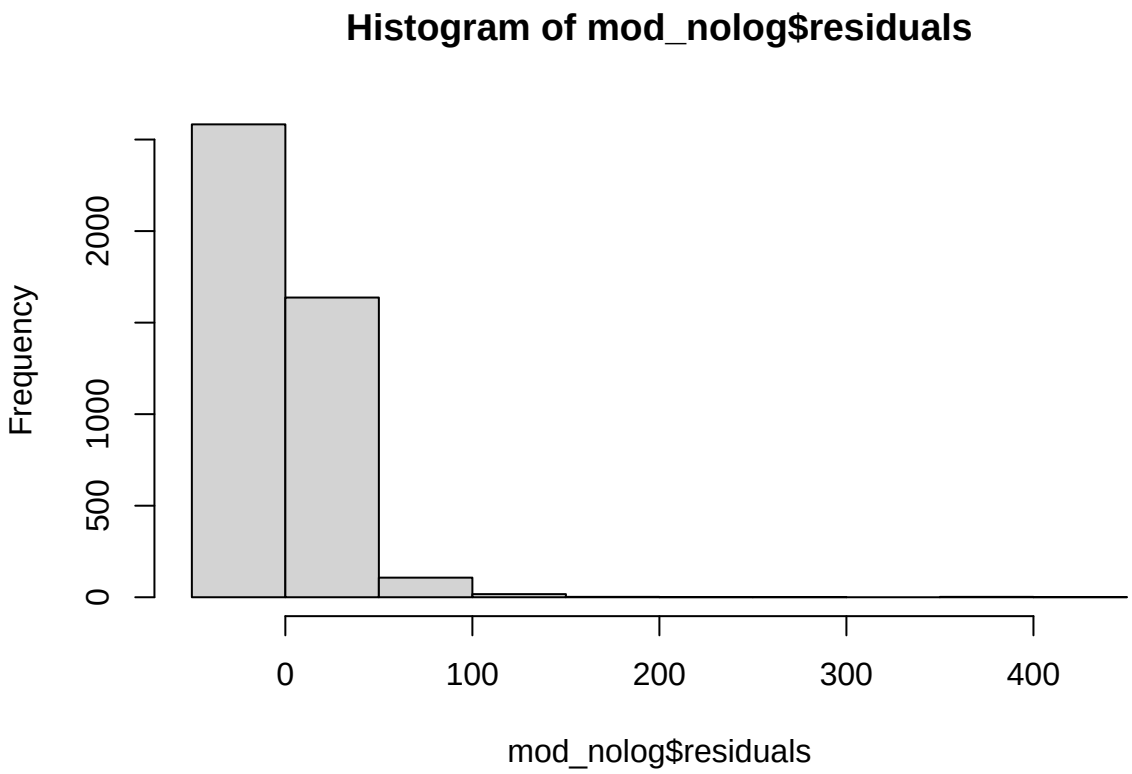

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -5.4192630  0.1859504 -29.144 < 2.2e-16 ***
## points       0.1010480  0.0021095  47.902 < 2.2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




Histogram of mod$residuals
Frequency
0
500
1000
1500
−2
−1
0
1
2
mod$residuals

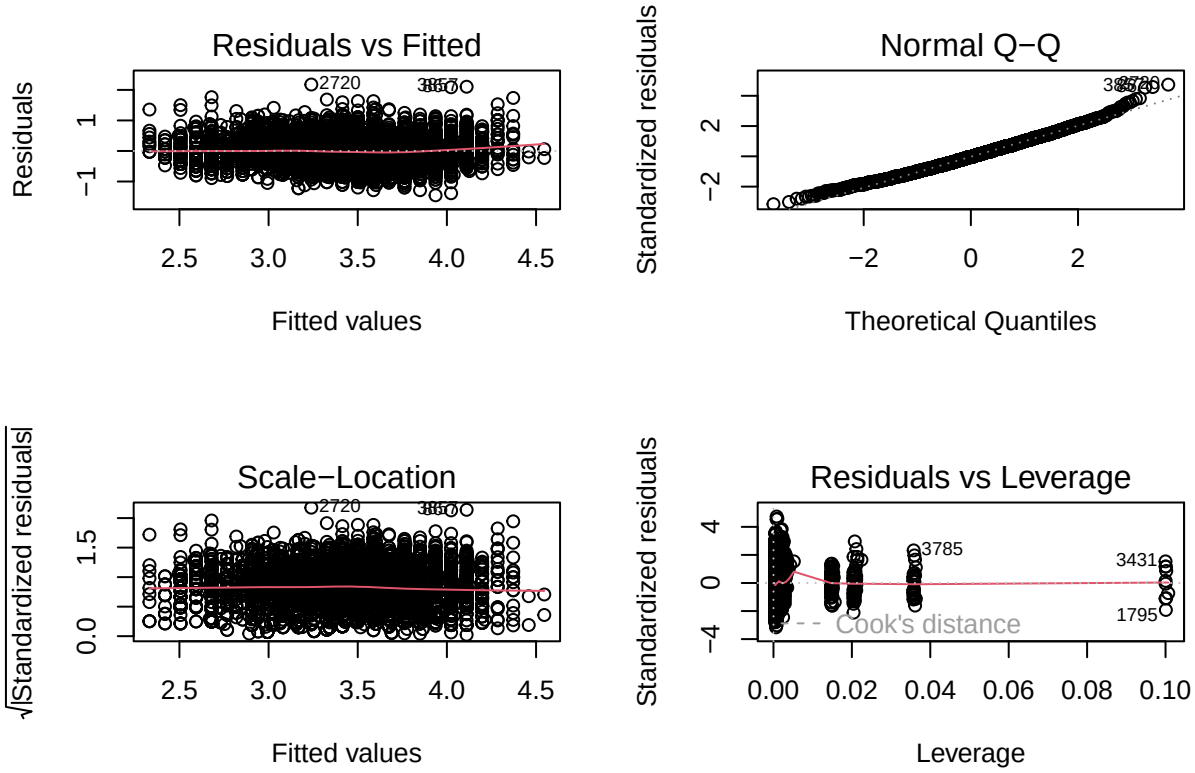

Scale−Location
3857
86
3372
√|Standardized residuals|
0
1
2
3
4
20
40
60
80
Fitted values
lm(price ~ points)



Does some regional detail add insight? Yes, it explains +7% of variation in the model

```
##
## Call:
## lm(formula = log(price) ~ points + region_3, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45151 -0.32591 -0.02077  0.30573  2.17645
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -4.635697   0.177464 -26.122  < 2e-16 ***
## points                     0.087096   0.002061  42.253  < 2e-16 ***
## region_3Central Coast      0.400277   0.024909  16.070  < 2e-16 ***
## region_3Central Valley     0.162928   0.089327   1.824   0.0682 .
## region_3Napa-Sonoma        0.559307   0.023656  23.644  < 2e-16 ***
## region_3North Coast        0.216872   0.069074   3.140   0.0017 **
## region_3Sierra Foothills   0.247087   0.059526   4.151 3.37e-05 ***
## region_3South Coast        0.896183   0.146860   6.102 1.14e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4596 on 4343 degrees of freedom
##   (3 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.4278, Adjusted R-squared:  0.4269
## F-statistic: 463.8 on 7 and 4343 DF,  p-value: < 2.2e-16
```



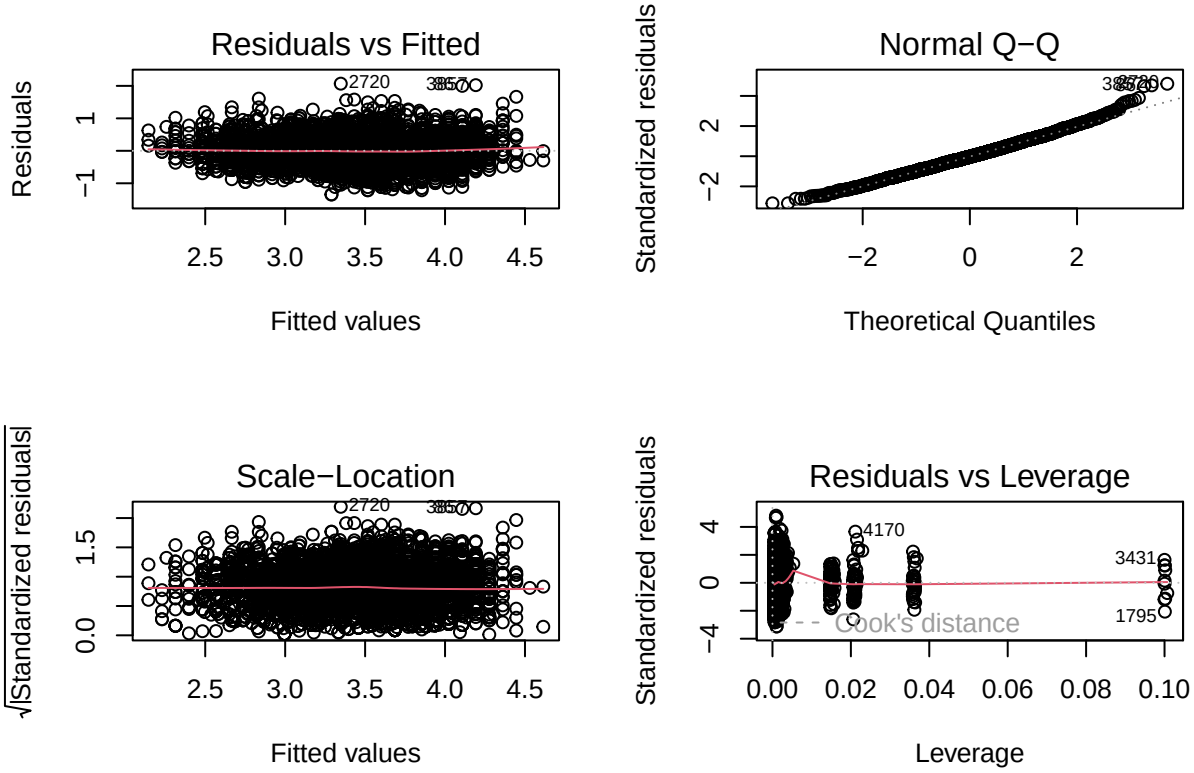

```
##              GVIF Df GVIF^(1/(2*Df))
## points   1.095214  1        1.046525
## region_3 1.095214  6        1.007608
```

```
##
## t test of coefficients:
##
##                             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)              -4.6356969  0.1781912 -26.0153 < 2.2e-16 ***
## points                    0.0870957  0.0020843  41.7865 < 2.2e-16 ***
## region_3Central Coast     0.4002767  0.0257563  15.5409 < 2.2e-16 ***
## region_3Central Valley    0.1629285  0.0888429   1.8339  0.066738 .
## region_3Napa-Sonoma       0.5593066  0.0250995  22.2836 < 2.2e-16 ***
## region_3North Coast       0.2168715  0.0748281   2.8983  0.003771 **
## region_3Sierra Foothills  0.2470870  0.0445693   5.5439 3.134e-08 ***
## region_3South Coast       0.8961833  0.1726085   5.1920 2.176e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Wald test
##
## Model 1: log(price) ~ points
```

```
## Model 2: log(price) ~ points + region_3
##   Res.Df Df      F    Pr(>F)
## 1   4349
## 2   4343  6 94.499 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = log(price) ~ points + region_3 + red, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34391 -0.29533 -0.01597  0.27804  2.06824
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -4.609296   0.166267 -27.722  < 2e-16 ***
## points                    0.084430   0.001934  43.651  < 2e-16 ***
## region_3Central Coast     0.370272   0.023368  15.845  < 2e-16 ***
## region_3Central Valley    0.112716   0.083714   1.346 0.178230
## region_3Napa-Sonoma       0.512745   0.022243  23.052  < 2e-16 ***
## region_3North Coast       0.223586   0.064715   3.455 0.000556 ***
## region_3Sierra Foothills  0.144781   0.055923   2.589 0.009660 **
## region_3South Coast       0.741673   0.137734   5.385 7.63e-08 ***
## red                       0.352310   0.014313  24.614  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4306 on 4342 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.4979, Adjusted R-squared:  0.4969
## F-statistic: 538.1 on 8 and 4342 DF,  p-value: < 2.2e-16
```




```
##              GVIF Df GVIF^(1/(2*Df))
## points   1.098659  1        1.048169
## region_3 1.109168  6        1.008672
## red      1.019015  1        1.009463
```

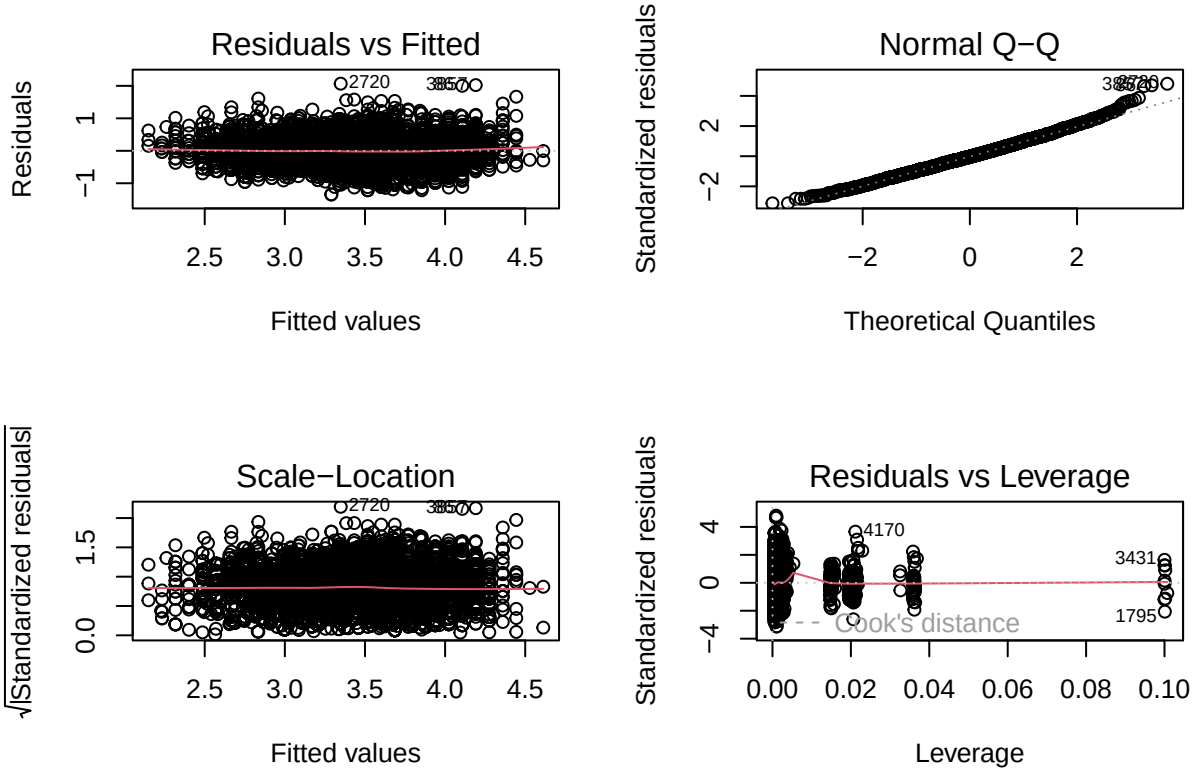
```
##
## t test of coefficients:
##
##                             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)              -4.6092960  0.1682426 -27.3967 < 2.2e-16 ***
## points                    0.0844298  0.0019682  42.8976 < 2.2e-16 ***
## region_3Central Coast     0.3702725  0.0246953  14.9936 < 2.2e-16 ***
## region_3Central Valley    0.1127158  0.0849374   1.3270  0.184563
## region_3Napa-Sonoma       0.5127453  0.0242228  21.1679 < 2.2e-16 ***
## region_3North Coast       0.2235857  0.0803212   2.7836  0.005398 **
## region_3Sierra Foothills  0.1447814  0.0444864   3.2545  0.001145 **
## region_3South Coast       0.7416735  0.1721126   4.3092 1.674e-05 ***
## red                       0.3523104  0.0134790  26.1377 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Wald test
##
## Model 1: log(price) ~ points
## Model 2: log(price) ~ points + region_3
```

```
##   Res.Df Df      F    Pr(>F)
## 1   4349
## 2   4343  6 94.499 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What about wine color?

```
##
## Call:
## lm(formula = log(price) ~ points + region_3 + color, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.34431 -0.29449 -0.01634  0.27744  2.06741
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -4.242968   0.167226 -25.373  < 2e-16 ***
## points                    0.084259   0.001937  43.489  < 2e-16 ***
## region_3Central Coast     0.371177   0.023374  15.880  < 2e-16 ***
## region_3Central Valley    0.112292   0.083703   1.342 0.179812
## region_3Napa-Sonoma       0.513906   0.022254  23.092  < 2e-16 ***
## region_3North Coast       0.223080   0.064707   3.448 0.000571 ***
## region_3Sierra Foothills  0.148971   0.055989   2.661 0.007826 **
## region_3South Coast       0.742026   0.137716   5.388 7.50e-08 ***
## colorrose                -0.438260   0.060417  -7.254 4.77e-13 ***
## colorwhite               -0.348778   0.014513 -24.031  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4306 on 4341 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.4981, Adjusted R-squared:  0.4971
## F-statistic: 478.7 on 9 and 4341 DF,  p-value: < 2.2e-16
```



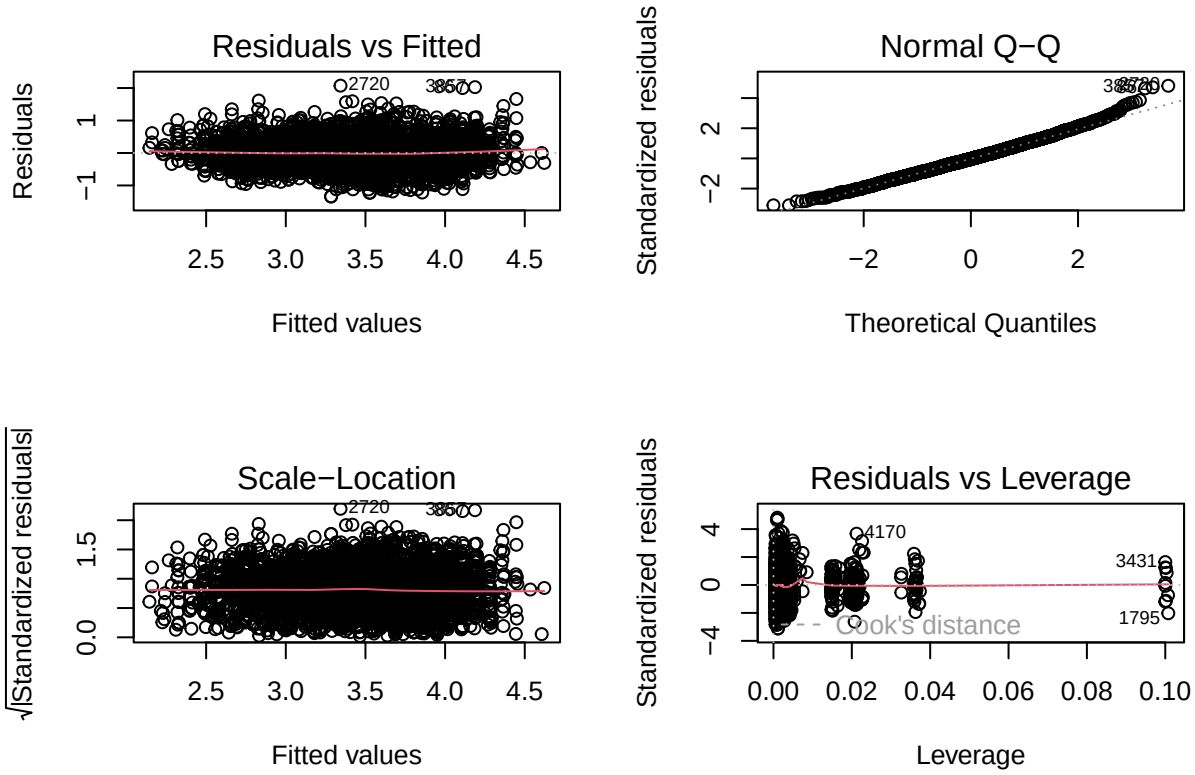

```
##
## t test of coefficients:
##
##                             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)              -4.2429683  0.1705124 -24.8836 < 2.2e-16 ***
## points                    0.0842589  0.0019741  42.6825 < 2.2e-16 ***
## region_3Central Coast     0.3711766  0.0247063  15.0236 < 2.2e-16 ***
## region_3Central Valley    0.1122921  0.0848609   1.3232 0.1858221
## region_3Napa-Sonoma       0.5139063  0.0242298  21.2097 < 2.2e-16 ***
## region_3North Coast       0.2230800  0.0802784   2.7788 0.0054790 **
## region_3Sierra Foothills  0.1489714  0.0446190   3.3387 0.0008487 ***
## region_3South Coast       0.7420260  0.1720736   4.3123 1.652e-05 ***
## colorrose                -0.4382596  0.0429087 -10.2138 < 2.2e-16 ***
## colorwhite               -0.3487776  0.0136918 -25.4734 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Wald test
##
## Model 1: log(price) ~ points
## Model 2: log(price) ~ points + region_3 + color
##   Res.Df Df      F    Pr(>F)
## 1   4349
## 2   4341  8 162.28 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##              GVIF Df GVIF^(1/(2*Df))
## points   1.102659  1        1.050076
## region_3 1.113027  6        1.008964
## color    1.025657  2        1.006353
```

What about vintage?

```
##
## Call:
## lm(formula = log(price) ~ points + region_3 + color + age, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33829 -0.29432 -0.01554  0.27563  2.07265
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -4.225230   0.167606 -25.209  < 2e-16 ***
## points                    0.084414   0.001940  43.516  < 2e-16 ***
## region_3Central Coast     0.369777   0.023388  15.811  < 2e-16 ***
## region_3Central Valley    0.120937   0.083883   1.442 0.149447
## region_3Napa-Sonoma       0.513436   0.022253  23.072  < 2e-16 ***
## region_3North Coast       0.227126   0.064752   3.508 0.000457 ***
## region_3Sierra Foothills  0.157697   0.056273   2.802 0.005096 **
## region_3South Coast       0.752591   0.137870   5.459 5.06e-08 ***
## colorrose                -0.445296   0.060584  -7.350 2.36e-13 ***
## colorwhite               -0.351433   0.014616 -24.045  < 2e-16 ***
## age                      -0.003228   0.002120  -1.523 0.127820
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4305 on 4340 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.4984, Adjusted R-squared:  0.4972
## F-statistic: 431.2 on 10 and 4340 DF,  p-value: < 2.2e-16
```




```
##
## t test of coefficients:
##
##                            Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)              -4.2252297  0.1716078 -24.6214 < 2.2e-16 ***
## points                    0.0844137  0.0019736  42.7712 < 2.2e-16 ***
## region_3Central Coast     0.3697771  0.0247404  14.9463 < 2.2e-16 ***
## region_3Central Valley    0.1209373  0.0849769   1.4232  0.154756
## region_3Napa-Sonoma       0.5134364  0.0242756  21.1503 < 2.2e-16 ***
## region_3North Coast       0.2271262  0.0805786   2.8187  0.004844 **
## region_3Sierra Foothills  0.1576970  0.0451080   3.4960  0.000477 ***
## region_3South Coast       0.7525910  0.1705711   4.4122 1.048e-05 ***
## colorrose                -0.4452965  0.0432264 -10.3015 < 2.2e-16 ***
## colorwhite               -0.3514331  0.0138599 -25.3561 < 2.2e-16 ***
## age                      -0.0032283  0.0022515  -1.4339  0.151682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Wald test
##
## Model 1: log(price) ~ points
## Model 2: log(price) ~ points + region_3 + color + age
##   Res.Df Df      F    Pr(>F)
## 1   4349
## 2   4340  9 143.95 < 2.2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##              GVIF Df GVIF^(1/(2*Df))
## points   1.105694  1        1.051520
## region_3 1.142413  6        1.011157
## color    1.045207  2        1.011115
## age      1.050274  1        1.024829
```

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Wed, Jul 27, 2022 - 8:34:48 PM

Build Regression Models on Test Dataset

Table 1:

	Dependent variable:		
	log(price)		
	(1)	(2)	(3)
points	0.101***	0.087***	0.084***
	(0.002)	(0.002)	(0.002)
region_3Central Coast		0.400***	0.371***
		(0.025)	(0.023)
region_3Central Valley		0.163*	0.112
		(0.089)	(0.084)
region_3Napa-Sonoma		0.559***	0.514***
		(0.024)	(0.022)
region_3North Coast		0.217***	0.223***
		(0.069)	(0.065)
region_3Sierra Foothills		0.247***	0.149***
		(0.060)	(0.056)
region_3South Coast		0.896***	0.742***
		(0.147)	(0.138)
colorrose			−0.438***
			(0.060)
colorwhite			−0.349***
			(0.015)
age			
Constant	−5.419***	−4.636***	−4.243***
	(0.185)	(0.177)	(0.167)
Observations	4,351	4,351	4,351
R^2	0.347	0.428	0.498
Adjusted R^2	0.347	0.427	0.497
Residual Std. Error	0.491 (df = 4349)	0.460 (df = 4343)	0.431 (df = 4341)
F Statistic	2,308.693*** (df = 1; 4349)	463.848*** (df = 7; 4343)	478.697*** (df = 9; 4341)

Note: *p<0.1;