Modeling Coffee Reviews

Will Rodman

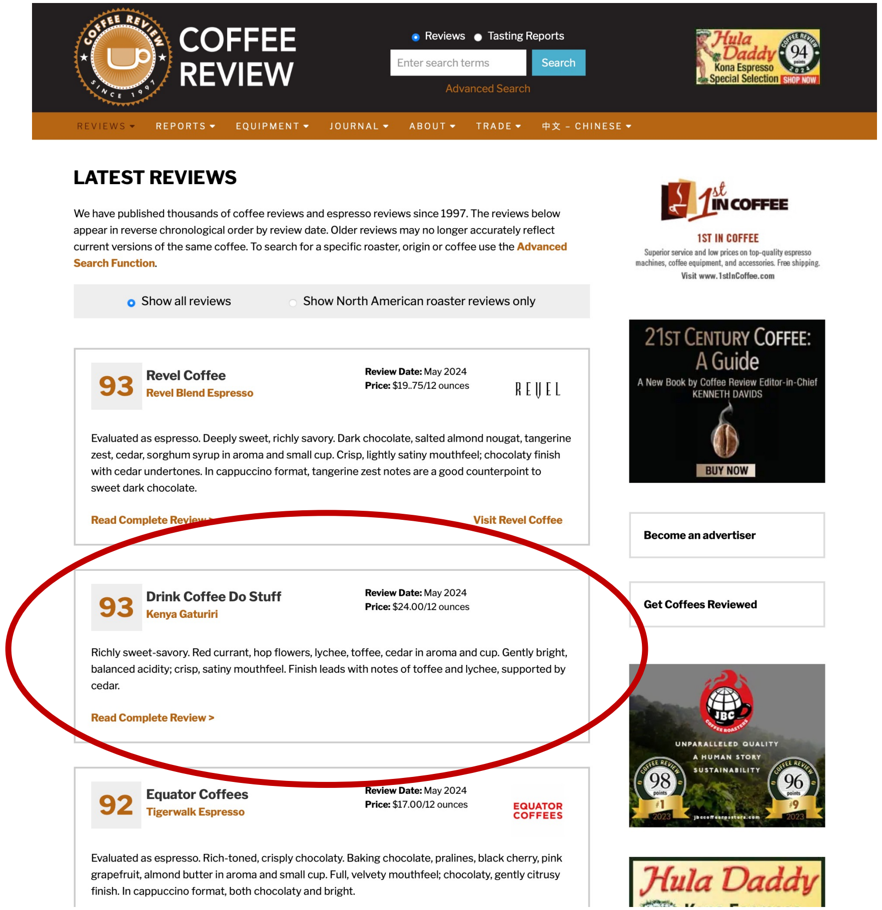
Final Project

MATH-6040 Linear Models

Project Summary

- Modeled a dataset sourced from Kaggle.com.
- The dataset consists of web-scraped coffee bean reviews from CoffeeReview.com.
- The goal was to develop a linear model that predicts users' coffee bean ratings.

Individual coffee bean review.



LATEST REVIEWS

We have published thousands of coffee reviews and espresso reviews since 1997. The reviews below appear in reverse chronological order by review date. Older reviews may no longer accurately reflect current versions of the same coffee. To search for a specific roaster, origin or coffee use the **Advanced Search Function**.

● Show all reviews ○ Show North American roaster reviews only

93 **Revel Coffee**
Revel Blend Espresso
Review Date: May 2024
Price: $19..75/12 ounces
REVEL

Evaluated as espresso. Deeply sweet, richly savory. Dark chocolate, salted almond nougat, tangerine zest, cedar, sorghum syrup in aroma and small cup. Crisp, lightly satiny mouthfeel; chocolaty finish with cedar undertones. In cappuccino format, tangerine zest notes are a good counterpoint to sweet dark chocolate.

Read Complete Review > **Visit Revel Coffee**

93 **Drink Coffee Do Stuff**
Kenya Gaturiri
Review Date: May 2024
Price: $24.00/12 ounces

Richly sweet-savory. Red currant, hop flowers, lychee, toffee, cedar in aroma and cup. Gently bright, balanced acidity; crisp, satiny mouthfeel. Finish leads with notes of toffee and lychee, supported by cedar.

Read Complete Review >

92 **Equator Coffees**
Tigerwalk Espresso
Review Date: May 2024
Price: $17.00/12 ounces
EQUATOR COFFEES

Evaluated as espresso. Rich-toned, crisply chocolaty. Baking chocolate, pralines, black cherry, pink grapefruit, almond butter in aroma and small cup. Full, velvety mouthfeel; chocolaty, gently citrusy finish. In cappuccino format, both chocolaty and bright.

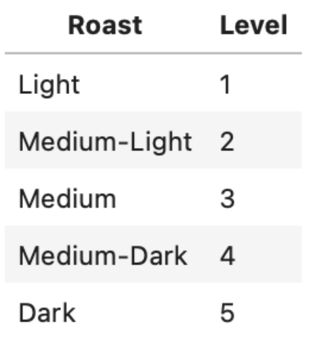
Become an advertiser

Get Coffees Reviewed

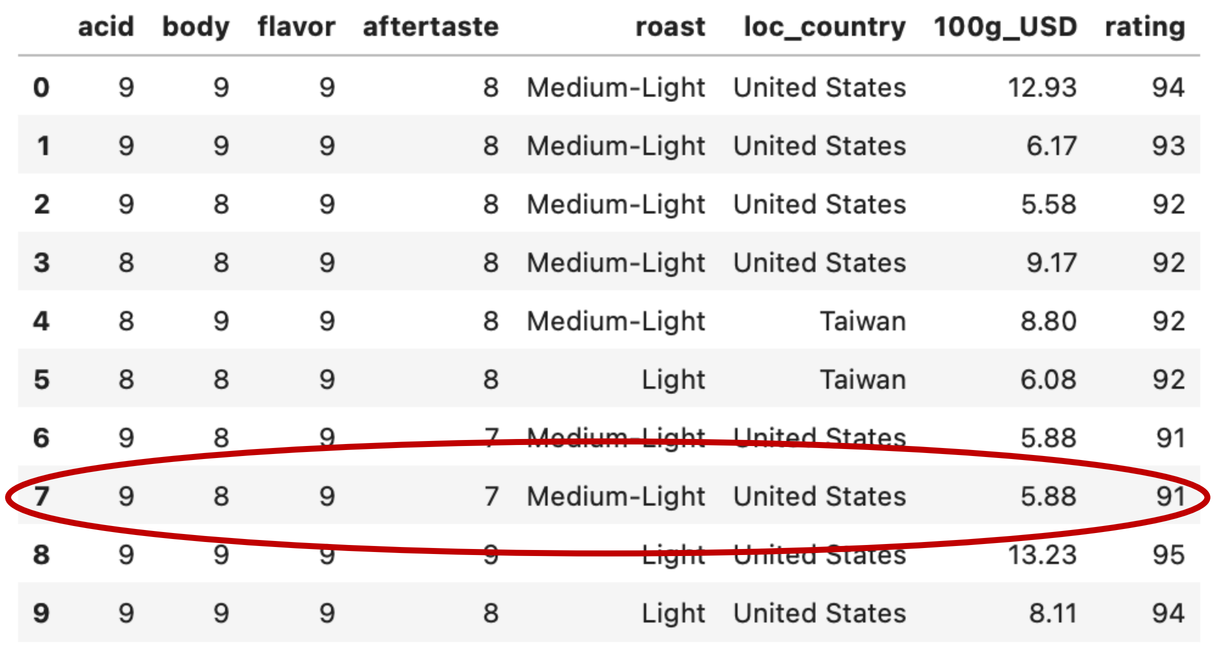
Figure 1: coffeereview.com reviews webpage.

Metadata

- Number of Reviews: 1779
- Feature Descriptions:
  - **acid**: Acidity level from 1 - 10.
  - **body**: Body characteristic from 1 - 10.
  - **flavor**: Strength of flavor from 1 - 10.
  - **aftertaste**: Aftertaste persistence from 1 - 10.
  - **roast**: Level of roast.
  - **loc_country**: Location of a users rating.
  - **100g_USD**: Price per 100 grams in USD.
  - **rating**: Overall rating from 1 - 100.



Roast	Level
Light	1
Medium-Light	2
Medium	3
Medium-Dark	4
Dark	5



	acid	body	flavor	aftertaste	roast	loc_country	100g_USD	rating
0	9	9	9	8	Medium-Light	United States	12.93	94
1	9	9	9	8	Medium-Light	United States	6.17	93
2	9	8	9	8	Medium-Light	United States	5.58	92
3	8	8	9	8	Medium-Light	United States	9.17	92
4	8	9	9	8	Medium-Light	Taiwan	8.80	92
5	8	8	9	8	Light	Taiwan	6.08	92
6	9	8	9	7	Medium-Light	United States	5.88	91
7	9	8	9	7	Medium-Light	United States	5.88	91
8	9	9	9	9	Light	United States	13.23	95
9	9	9	9	8	Light	United States	8.11	94

Figure 2: Dataset heading after data cleaning.

Individual coffee bean review.

Numerical Distributions

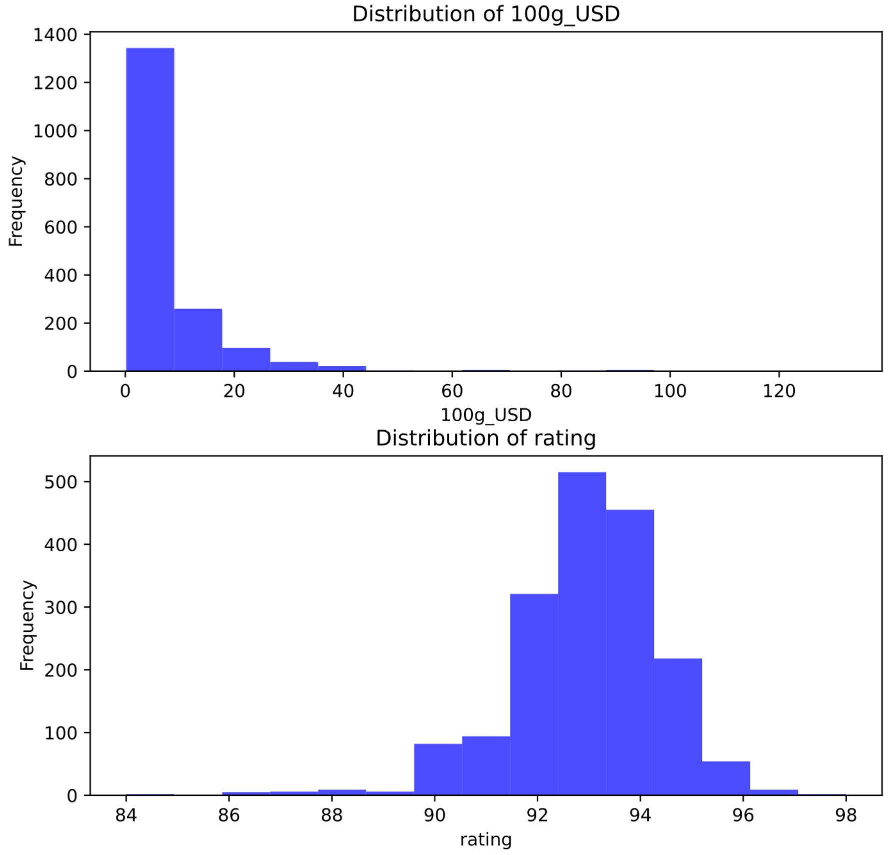


Figure 3: Distributions of price and rating values.

Box-Cox Transformation
$\lambda \approx -0.3$

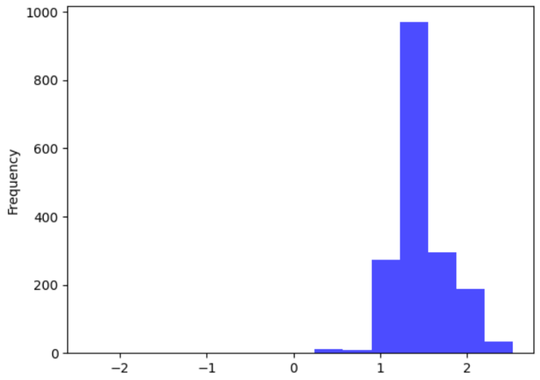
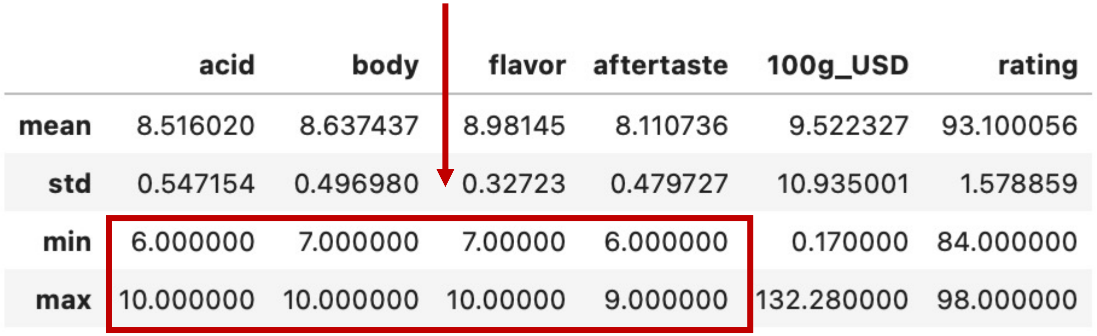


Figure 4: Distribution price after transformation.

The distributions do not cover the entire range of scores.



	acid	body	flavor	aftertaste	100g_USD	rating
mean	8.516020	8.637437	8.98145	8.110736	9.522327	93.100056
std	0.547154	0.496980	0.32723	0.479727	10.935001	1.578859
min	6.000000	7.000000	7.00000	6.000000	0.170000	84.000000
max	10.000000	10.000000	10.00000	9.000000	132.280000	98.000000

Figure 5: Distribution statistics.

Numerical Distributions

Caused by uniform distribution.

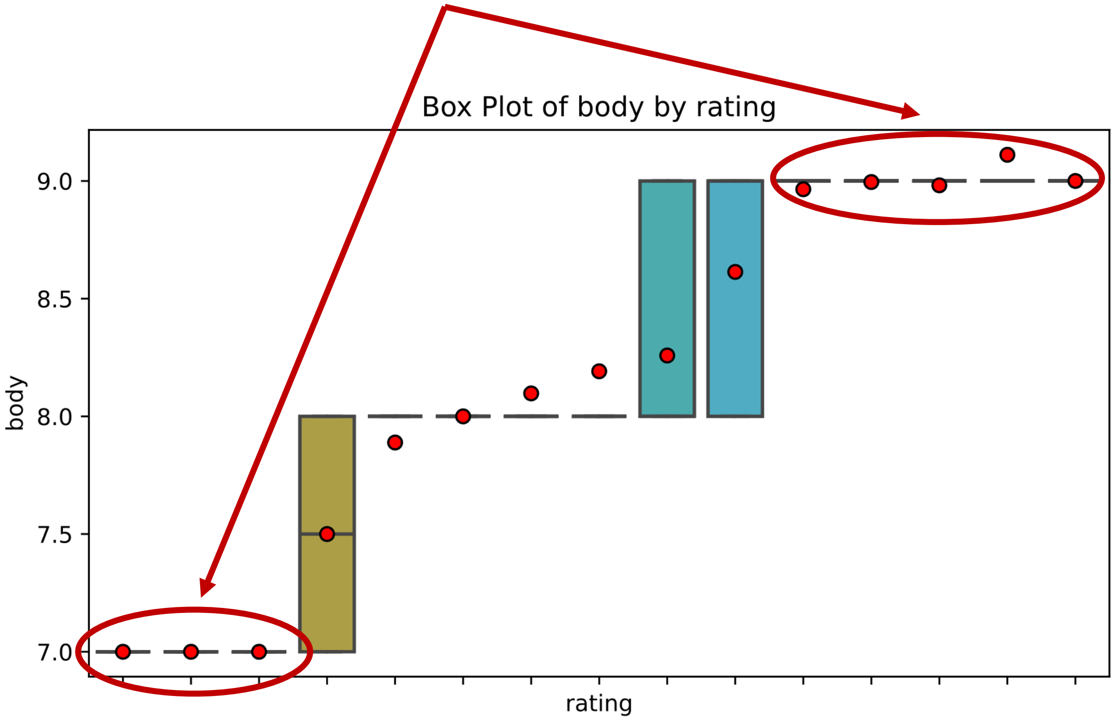


Figure 6: Box plot distribution of aftertaste by rating.

Exponential relationship relationship between price and rating.

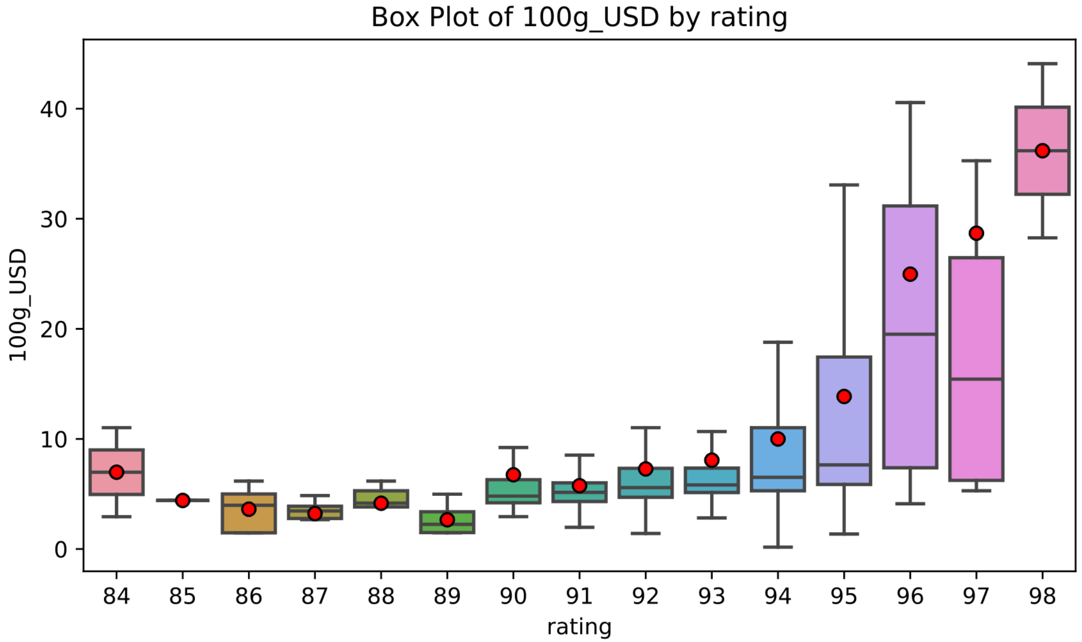


Figure 7: Box plot distribution of price by rating.

Categorical Features

Number of roast Observations by Category:

Category	Observations
Medium-Light	1304 73% of data.
Light	280
Medium	175
Medium-Dark	16
Dark	4

Top 5 location Observations by Category:

Category	Observations
United States	1210 68% of data.
Taiwan	402
Hawaii	82
Guatemala	27
Canada	21

* 17 total location categories.

Categorical Features

No significant variance in the averages by location.

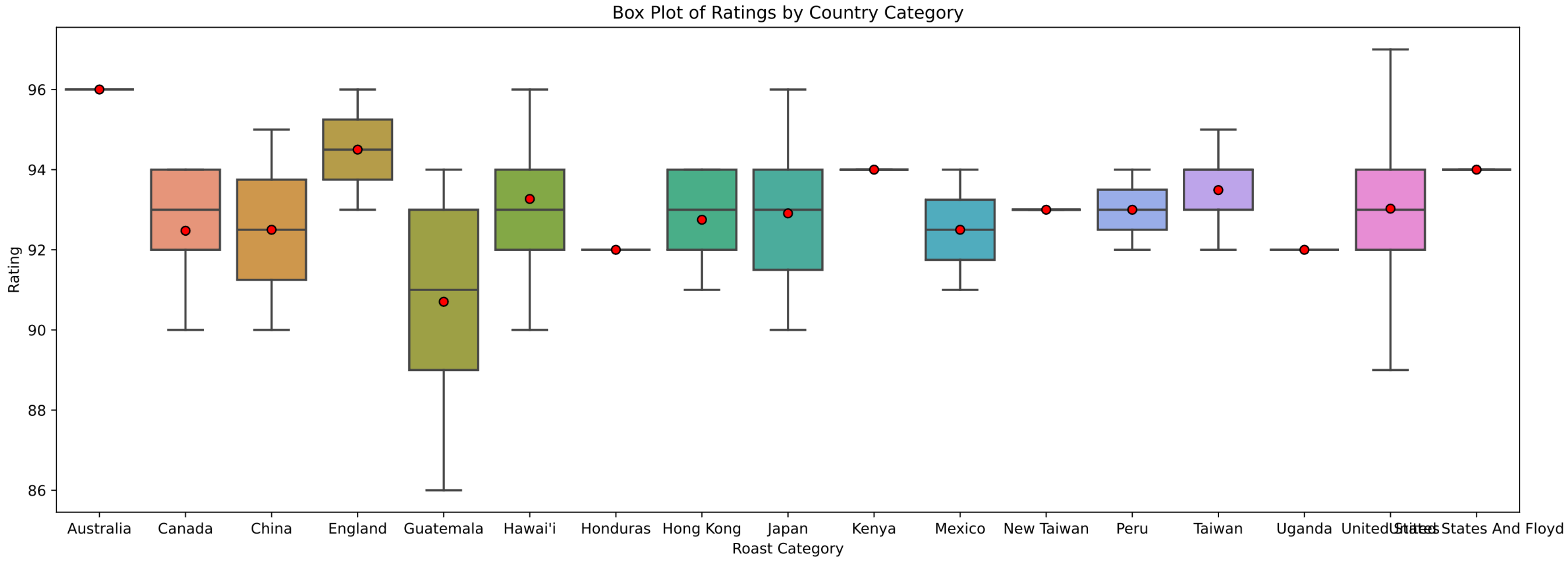


Figure 8: Box plot distribution of rating by location.

Average rating decreases as the level of roast increases.

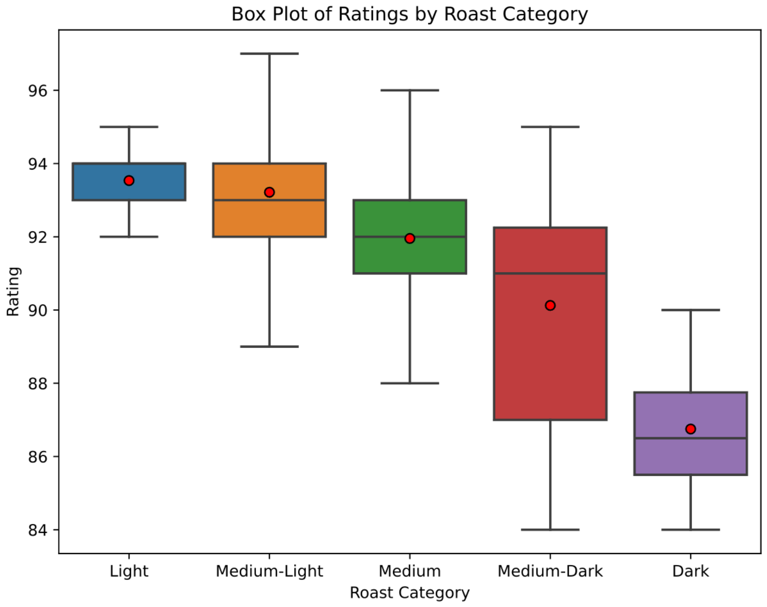
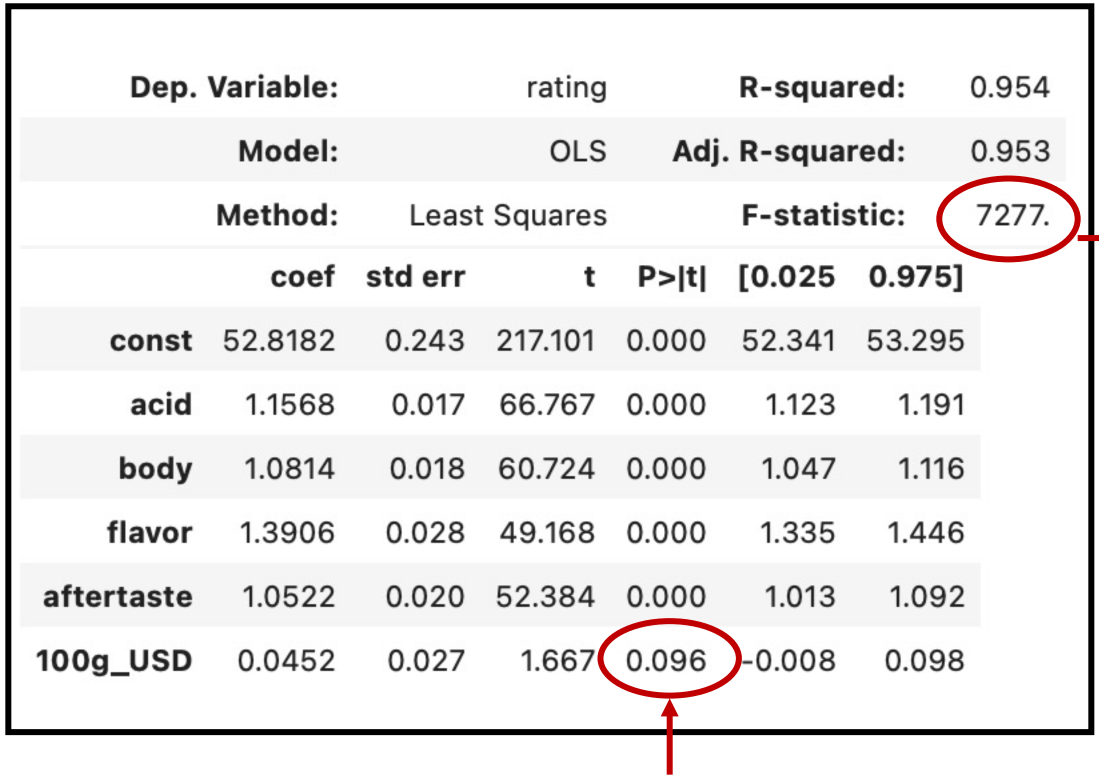


Figure 9: Box plot distribution of rating by roast.

Models

1) Regression including all numerical features:

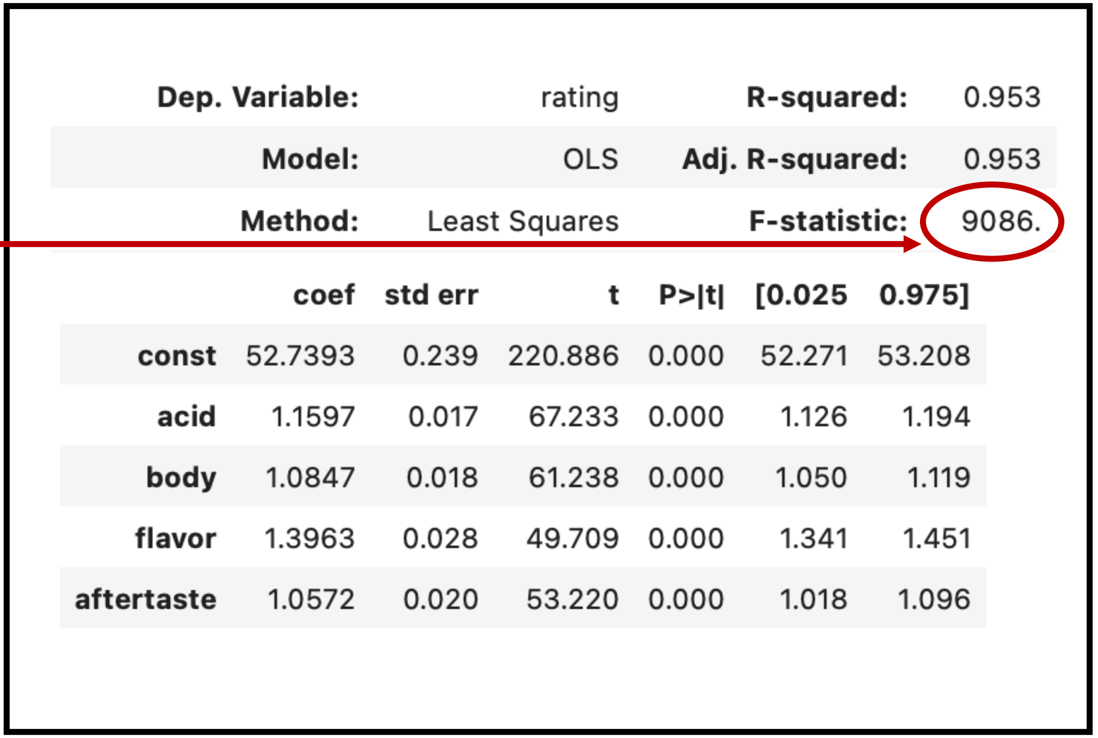


Dep. Variable:	rating	**R-squared:**	0.954
Model:	OLS	**Adj. R-squared:**	0.953
Method:	Least Squares	**F-statistic:**	7277.

	coef	std err	t	P>|t|	[0.025	0.975]
const	52.8182	0.243	217.101	0.000	52.341	53.295
acid	1.1568	0.017	66.767	0.000	1.123	1.191
body	1.0814	0.018	60.724	0.000	1.047	1.116
flavor	1.3906	0.028	49.168	0.000	1.335	1.446
aftertaste	1.0522	0.020	52.384	0.000	1.013	1.092
100g_USD	0.0452	0.027	1.667	0.096	-0.008	0.098

$$0.096 > \alpha = 0.05$$

2) Regression excluding the price feature:

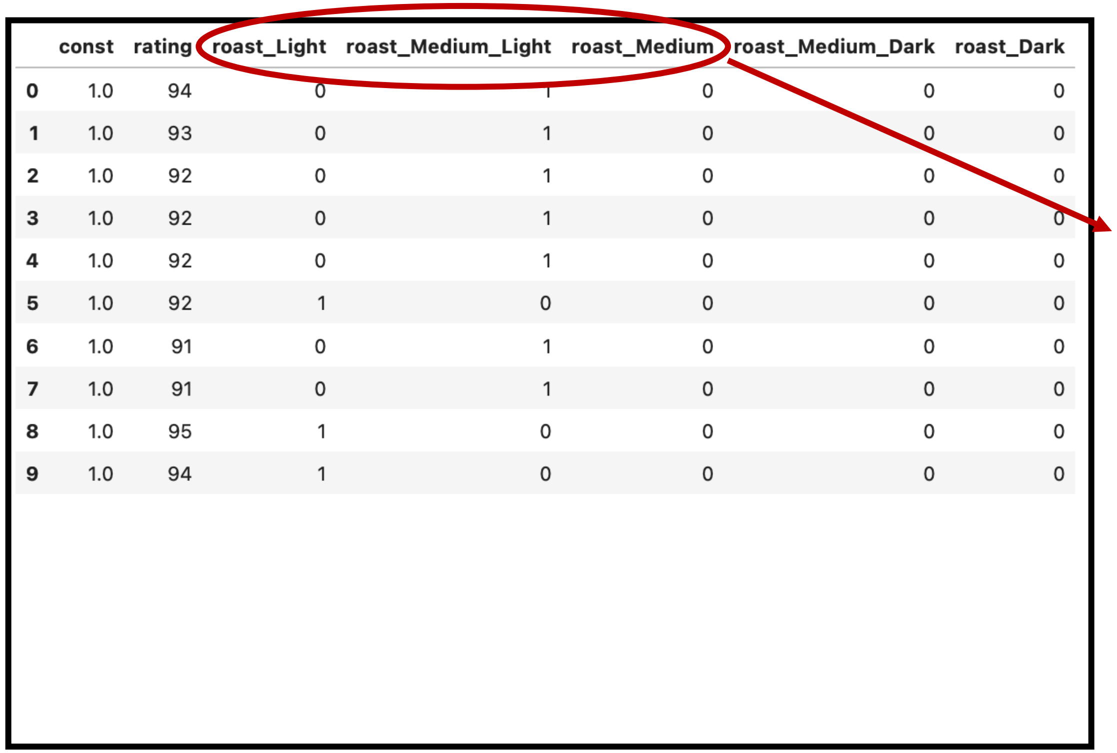


Dep. Variable:	rating	**R-squared:**	0.953
Model:	OLS	**Adj. R-squared:**	0.953
Method:	Least Squares	**F-statistic:**	9086.

	coef	std err	t	P>|t|	[0.025	0.975]
const	52.7393	0.239	220.886	0.000	52.271	53.208
acid	1.1597	0.017	67.233	0.000	1.126	1.194
body	1.0847	0.018	61.238	0.000	1.050	1.119
flavor	1.3963	0.028	49.709	0.000	1.341	1.451
aftertaste	1.0572	0.020	53.220	0.000	1.018	1.096

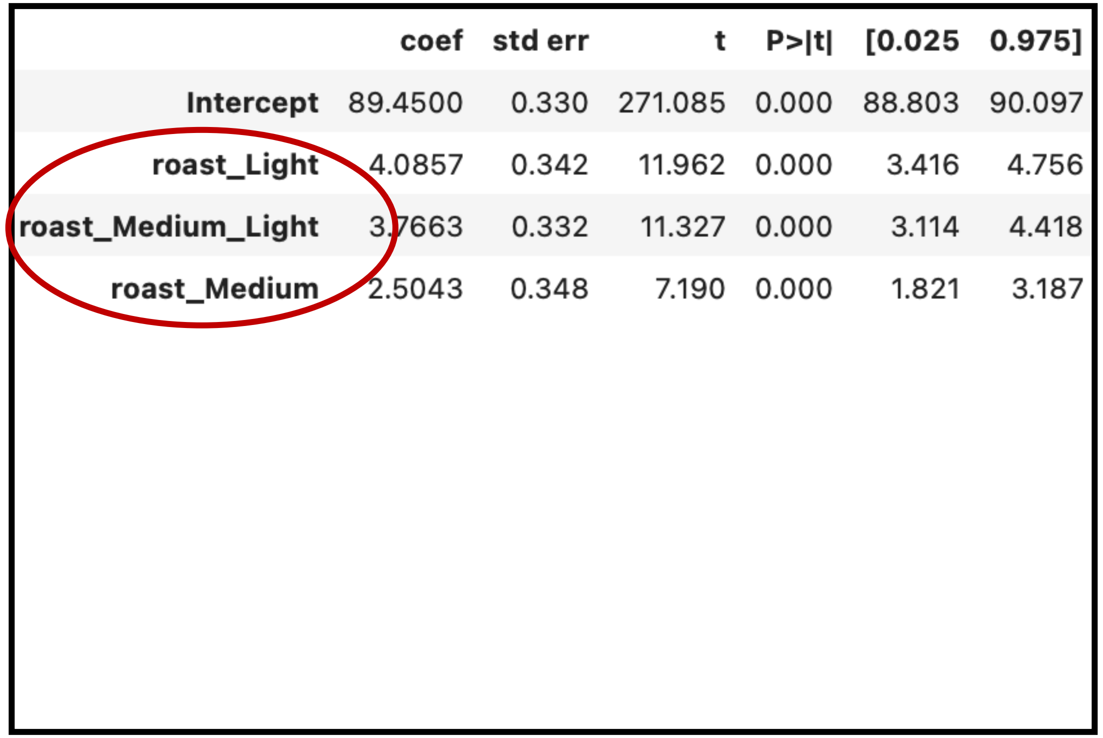
Models

Design Matrix for Roast:



	const	rating	roast_Light	roast_Medium_Light	roast_Medium	roast_Medium_Dark	roast_Dark
0	1.0	94	0	1	0	0	0
1	1.0	93	0	1	0	0	0
2	1.0	92	0	1	0	0	0
3	1.0	92	0	1	0	0	0
4	1.0	92	0	1	0	0	0
5	1.0	92	1	0	0	0	0
6	1.0	91	0	1	0	0	0
7	1.0	91	0	1	0	0	0
8	1.0	95	1	0	0	0	0
9	1.0	94	1	0	0	0	0

3) Analysis of Light Light/Medium and Medium Roasts:

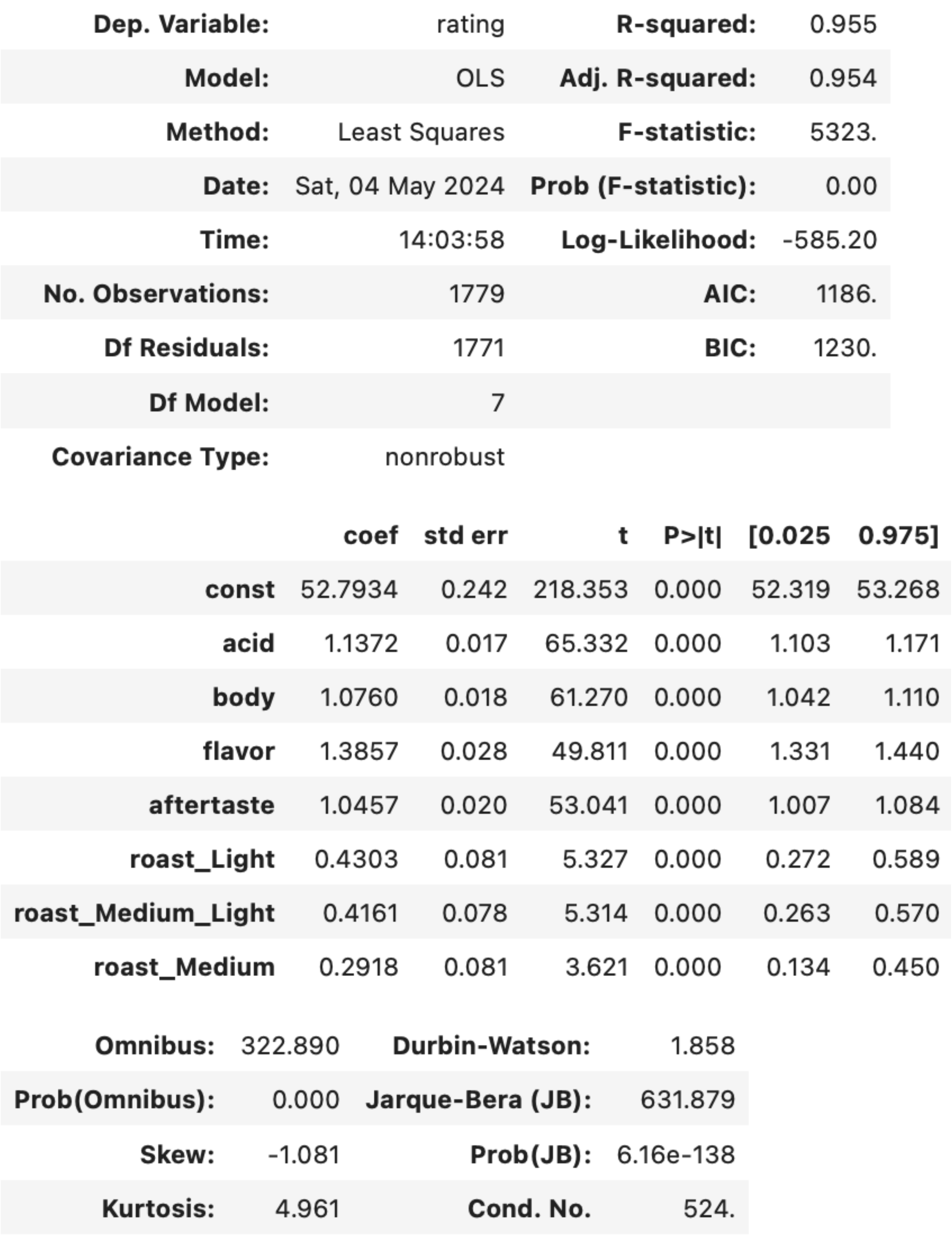


	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	89.4500	0.330	271.085	0.000	88.803	90.097
roast_Light	4.0857	0.342	11.962	0.000	3.416	4.756
roast_Medium_Light	3.7663	0.332	11.327	0.000	3.114	4.418
roast_Medium	2.5043	0.348	7.190	0.000	1.821	3.187

Models

4) Combining Categorical and Numerical Features:

- R-squared of 95%.
- All coefficients are positive.
- Improvement in Log Likelihoods:
    1. Model 1: -606.38
    2. Model 2: -607.78
    3. Model 4: -585.20



Dep. Variable:	rating	**R-squared:**	0.955
Model:	OLS	**Adj. R-squared:**	0.954
Method:	Least Squares	**F-statistic:**	5323.
Date:	Sat, 04 May 2024	**Prob (F-statistic):**	0.00
Time:	14:03:58	**Log-Likelihood:**	-585.20
No. Observations:	1779	**AIC:**	1186.
Df Residuals:	1771	**BIC:**	1230.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	52.7934	0.242	218.353	0.000	52.319	53.268
acid	1.1372	0.017	65.332	0.000	1.103	1.171
body	1.0760	0.018	61.270	0.000	1.042	1.110
flavor	1.3857	0.028	49.811	0.000	1.331	1.440
aftertaste	1.0457	0.020	53.041	0.000	1.007	1.084
roast_Light	0.4303	0.081	5.327	0.000	0.272	0.589
roast_Medium_Light	0.4161	0.078	5.314	0.000	0.263	0.570
roast_Medium	0.2918	0.081	3.621	0.000	0.134	0.450

Omnibus:	322.890	**Durbin-Watson:**	1.858
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	631.879
Skew:	-1.081	**Prob(JB):**	6.16e-138
Kurtosis:	4.961	**Cond. No.**	524.