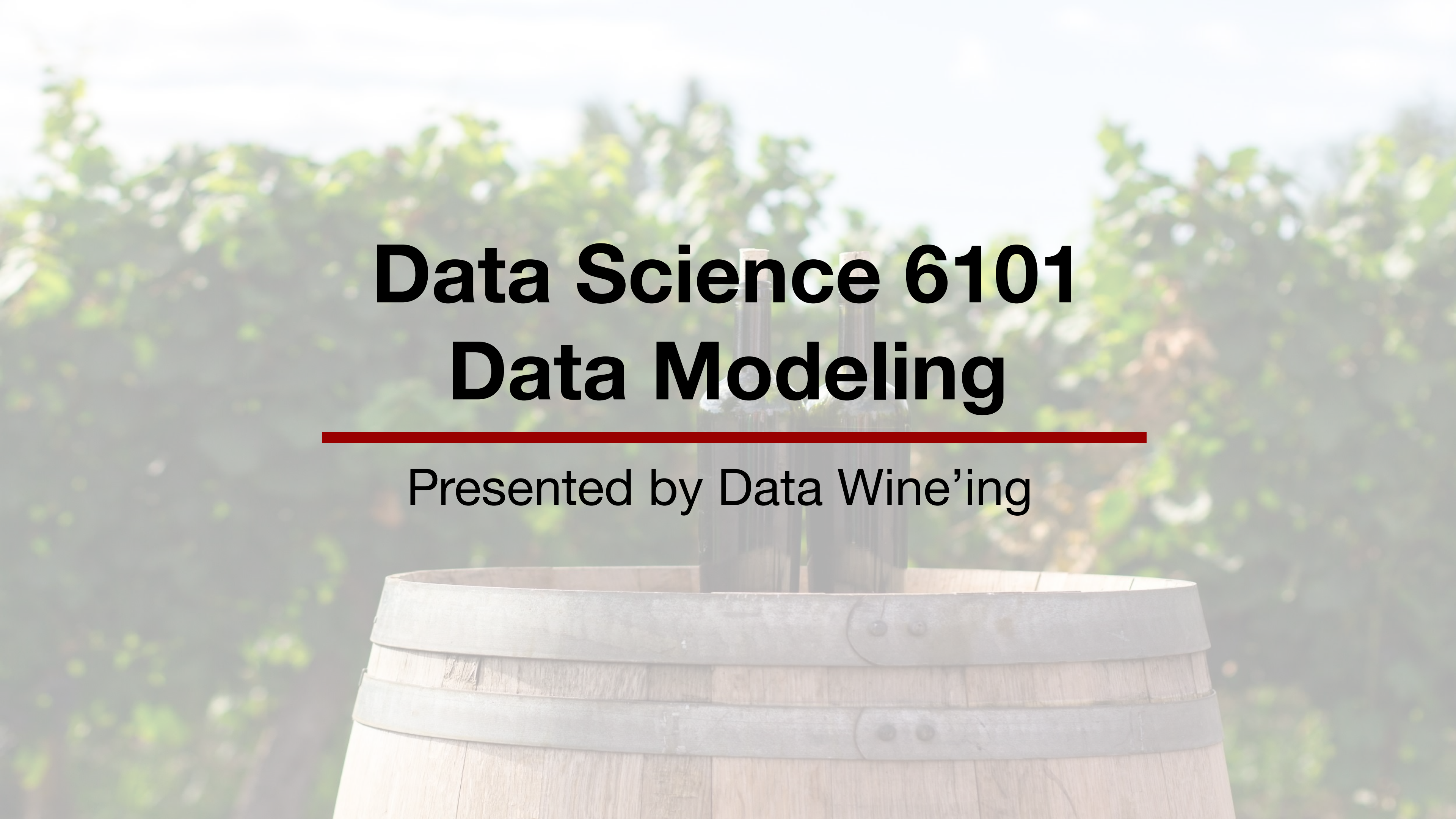
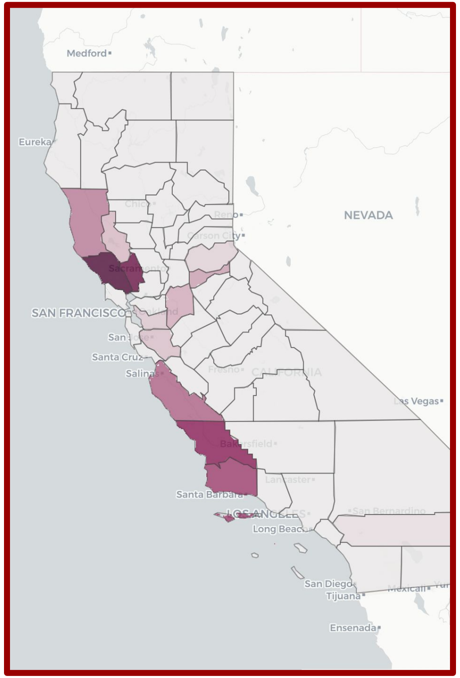


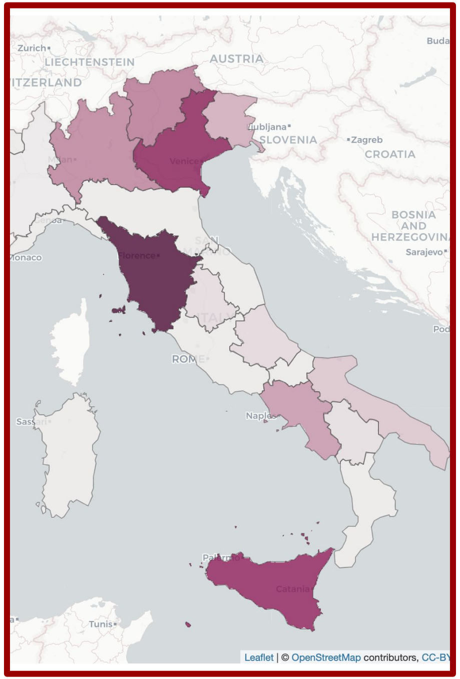
Data Science 6101
Data Modeling

Presented by Data Wine’ing

Previously from Project 1

- Leveraged wine reviews collected from the magazine WineEnthusiast to conduct inferential statistics and uncover insights about wine quality.
- Aimed to answer the following: "Are wines grown in prominent wine producing countries (e.g., Italy and France) rated higher than those grown in California?"





Outline

- SMART Question
- Data Wrangling
- Regression Analysis
  - Correlational Analysis
  - Feature Selection
- Logistic Regression
- K Nearest Neighbors
- K Means
- Discussion

SMART Question

What are the factors that influence wine quality? Does there exist a limited number of factors that consumers can use to reliably choose wines of high quality?

Average Consumer

Source: https://streamable.com/i7ga#

Previous Findings

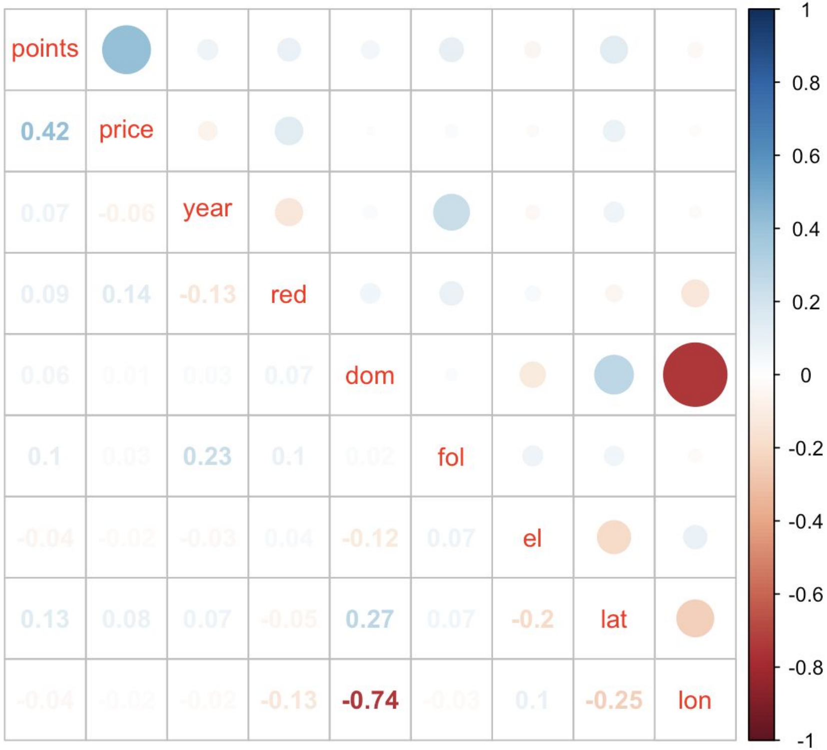
According to 2012 study...

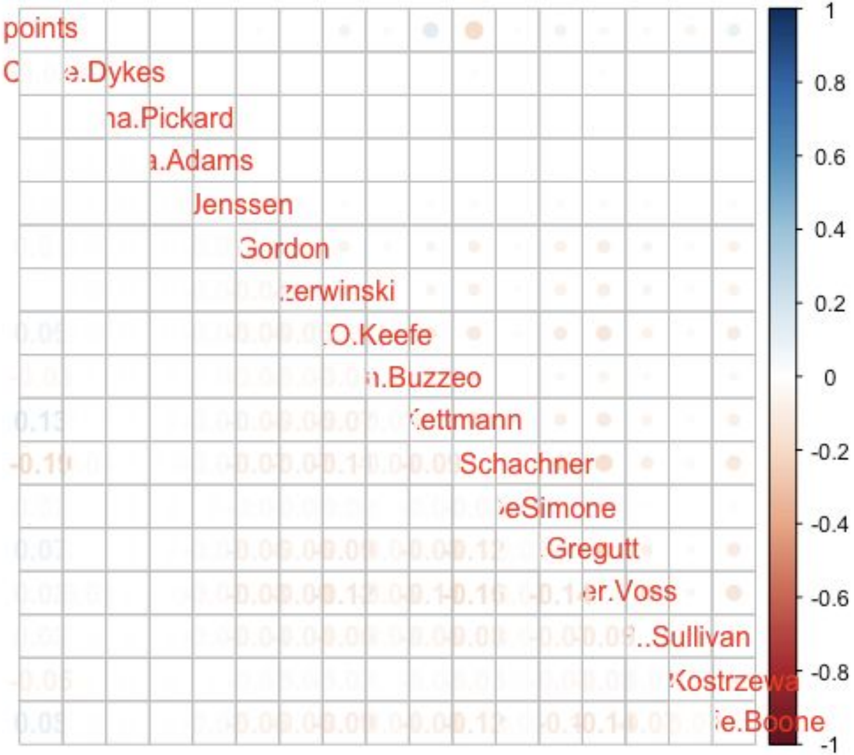
- Only 10% of wine 'experts' can consistently rate the wine from the same bottle in the same way, and they aren't consistent the next year. After analyzing results across wine competitions in California, medals were found to be distributed at random.

Data Wrangling

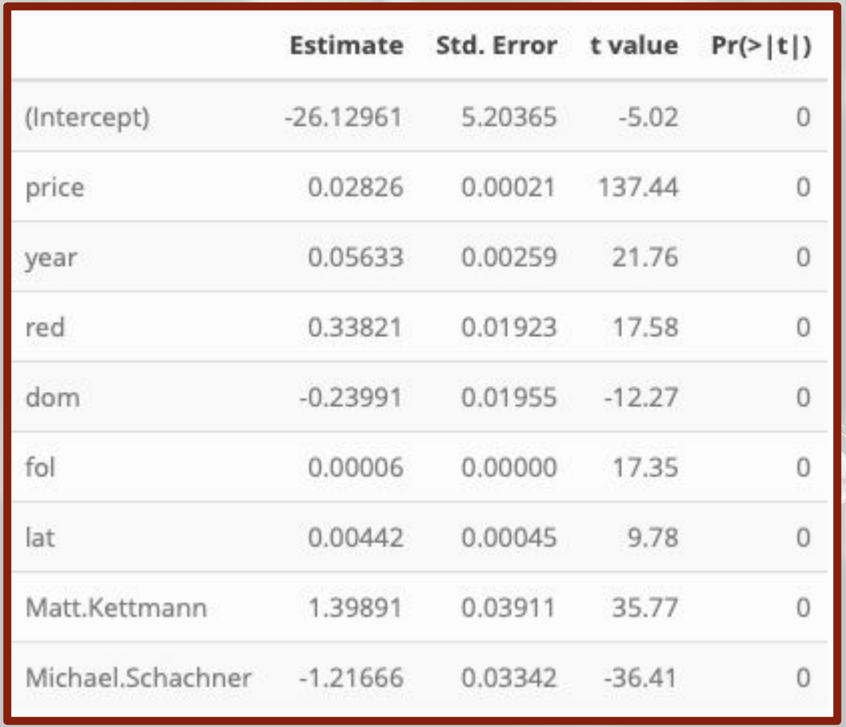
- Leveraged the Google Maps API to associate named locations with longitude, latitude and elevation
- Gathered data from Twitter to measure wine critic’s social media following

Correlational Analysis

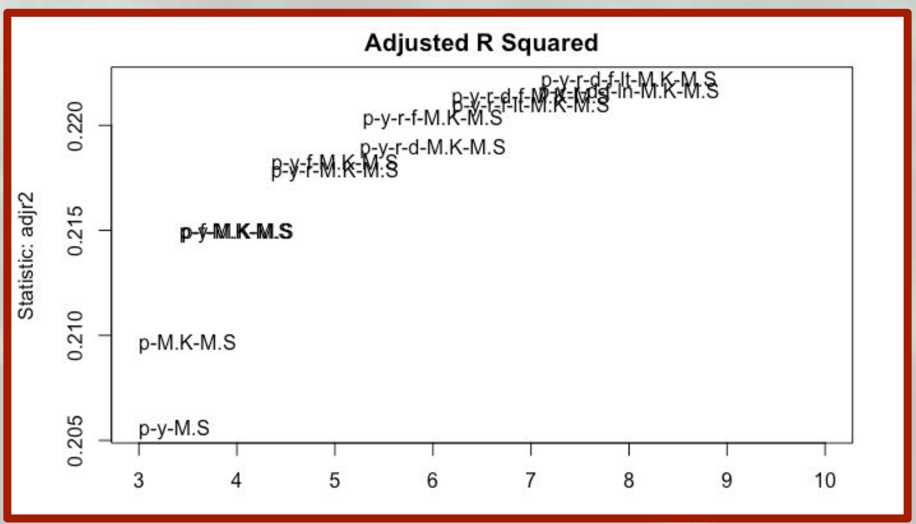




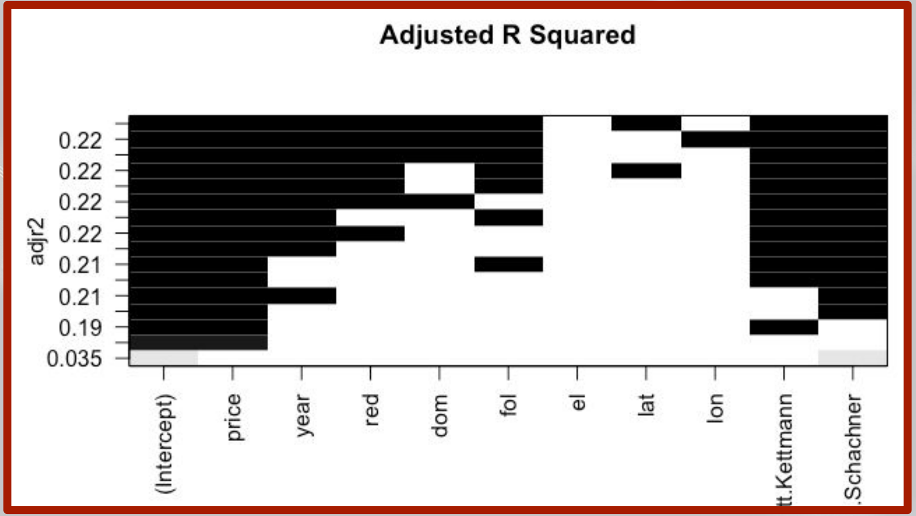
Feature Selection



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-26.12961	5.20365	-5.02	0
price	0.02826	0.00021	137.44	0
year	0.05633	0.00259	21.76	0
red	0.33821	0.01923	17.58	0
dom	-0.23991	0.01955	-12.27	0
fol	0.00006	0.00000	17.35	0
lat	0.00442	0.00045	9.78	0
Matt.Kettmann	1.39891	0.03911	35.77	0
Michael.Schachner	-1.21666	0.03342	-36.41	0



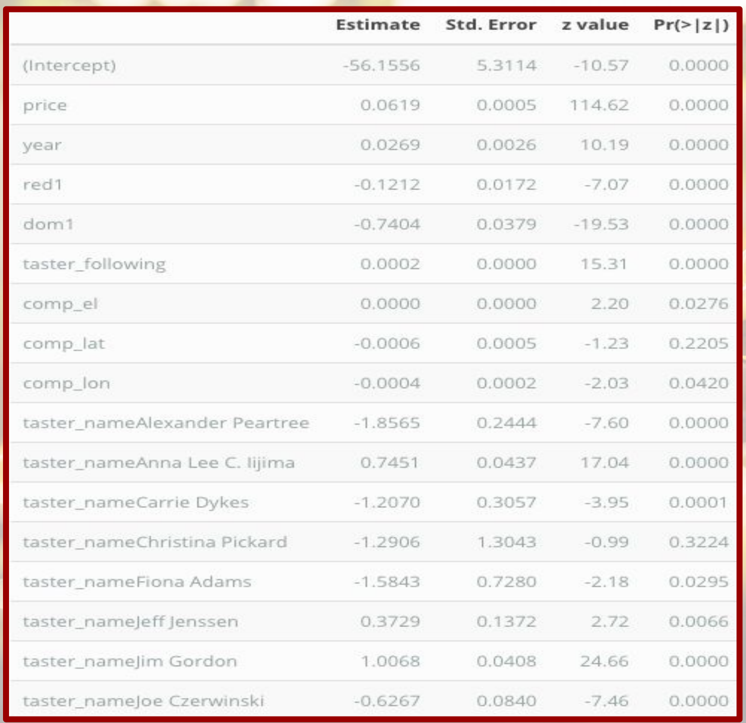
Adjusted R Squared



Adjusted R Squared

```
glm(formula = p88 ~ . -points, data = wine_reviews.df, family = binomial(link="logit"))
```

Logistic Regression

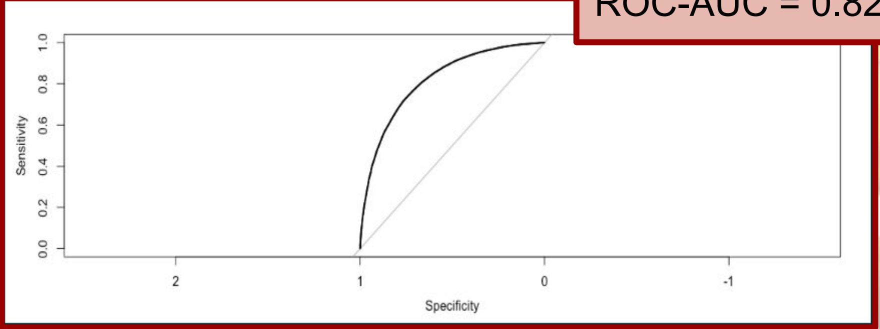


	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-56.1556	5.3114	-10.57	0.0000
price	0.0619	0.0005	114.62	0.0000
year	0.0269	0.0026	10.19	0.0000
red1	-0.1212	0.0172	-7.07	0.0000
dom1	-0.7404	0.0379	-19.53	0.0000
taster_following	0.0002	0.0000	15.31	0.0000
comp_el	0.0000	0.0000	2.20	0.0276
comp_lat	-0.0006	0.0005	-1.23	0.2205
comp_lon	-0.0004	0.0002	-2.03	0.0420
taster_nameAlexander Peartree	-1.8565	0.2444	-7.60	0.0000
taster_nameAnna Lee C. Iijima	0.7451	0.0437	17.04	0.0000
taster_nameCarrie Dykes	-1.2070	0.3057	-3.95	0.0001
taster_nameChristina Pickard	-1.2906	1.3043	-0.99	0.3224
taster_nameFiona Adams	-1.5843	0.7280	-2.18	0.0295
taster_nameJeff Jenssen	0.3729	0.1372	2.72	0.0066
taster_nameJim Gordon	1.0068	0.0408	24.66	0.0000
taster_nameJoe Czerwinski	-0.6267	0.0840	-7.46	0.0000

**See report for full list*

- Model was run with 73.5% accuracy
- Significant coefficient for all predictor variables, with the exception of \`comp_lat\`, \`taster_nameChristina Pickard\` and \`taster_nameVirginie Boone\`
- \`taster_following\`, \`comp_el\`, \`comp_lat\`, \`comp_lon\` provide no or a negligible effect on the odds-ratio

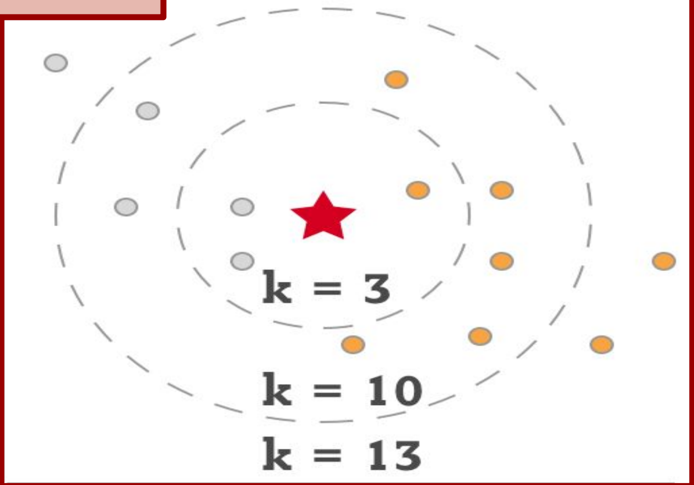
ROC-AUC = 0.82

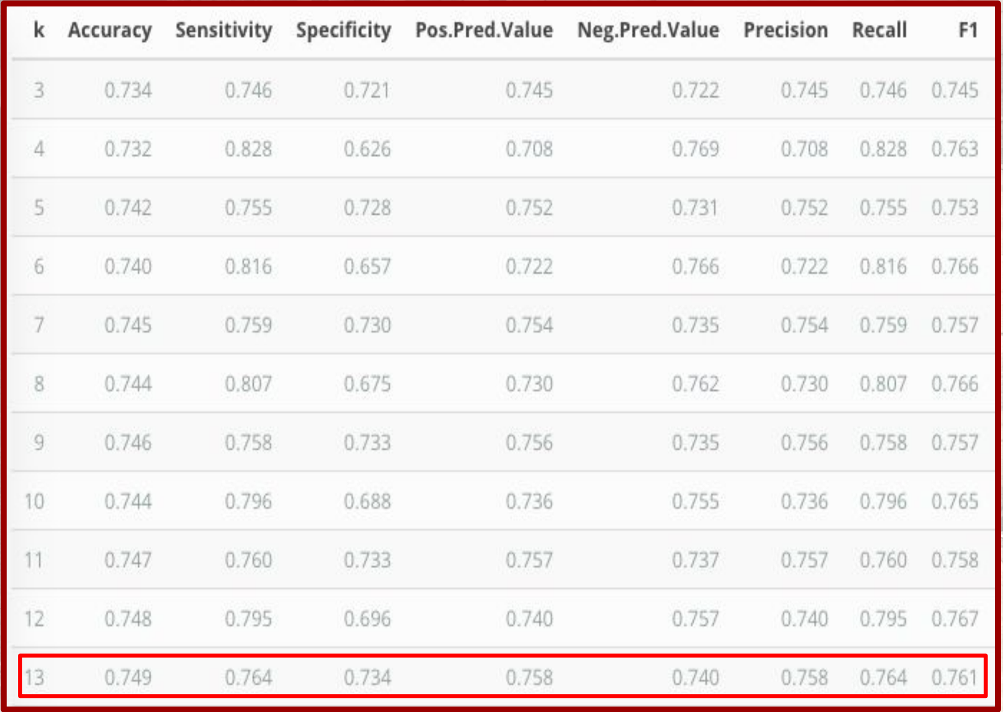


K-Nearest Neighbor (KNN)

- Optimal K=13, 74.9% accuracy
- Supervised Learning Algorithm

Example

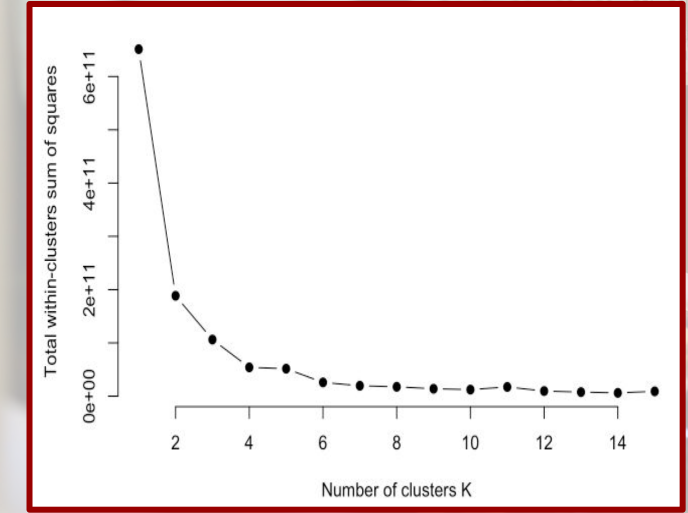


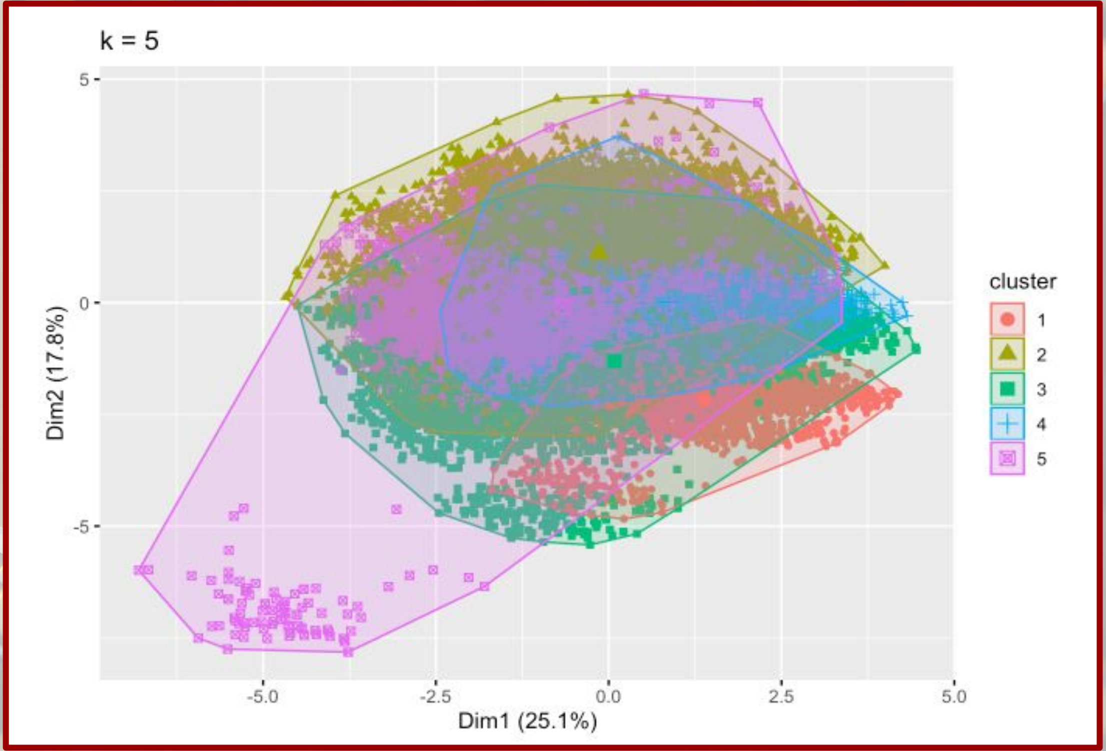


k	Accuracy	Sensitivity	Specificity	Pos.Pred.Value	Neg.Pred.Value	Precision	Recall	F1
3	0.734	0.746	0.721	0.745	0.722	0.745	0.746	0.745
4	0.732	0.828	0.626	0.708	0.769	0.708	0.828	0.763
5	0.742	0.755	0.728	0.752	0.731	0.752	0.755	0.753
6	0.740	0.816	0.657	0.722	0.766	0.722	0.816	0.766
7	0.745	0.759	0.730	0.754	0.735	0.754	0.759	0.757
8	0.744	0.807	0.675	0.730	0.762	0.730	0.807	0.766
9	0.746	0.758	0.733	0.756	0.735	0.756	0.758	0.757
10	0.744	0.796	0.688	0.736	0.755	0.736	0.796	0.765
11	0.747	0.760	0.733	0.757	0.737	0.757	0.760	0.758
12	0.748	0.795	0.696	0.740	0.757	0.740	0.795	0.767
13	0.749	0.764	0.734	0.758	0.740	0.758	0.764	0.761

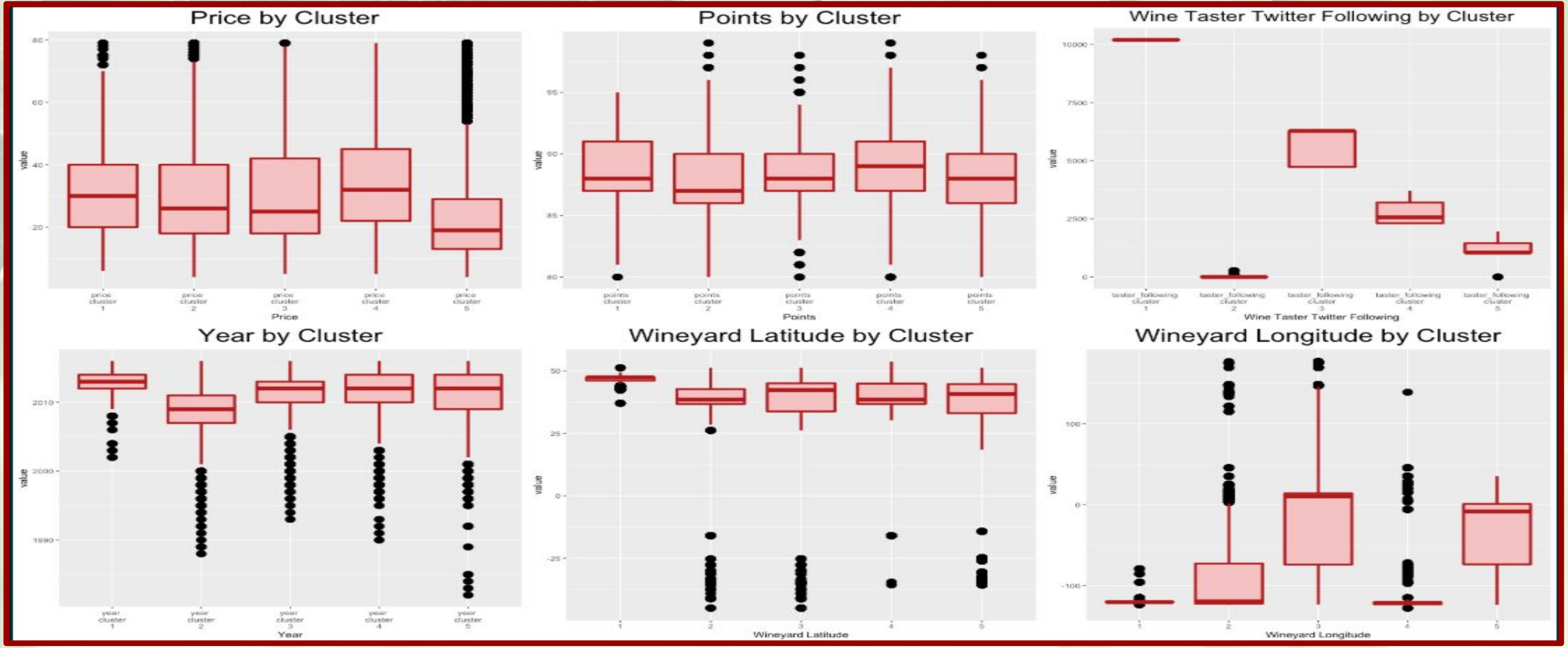
K-Means

- Unsupervised Learning Algorithm
- Optimal K=5





Resulting Cluster Description



Discussion

- Regression analysis showed that who critiqued the wine had an influence on
- the point value assigned to it; there was little influence by where the wine came from.
- KNN recommends using 13 closest data points to classify wine quality.
- K-means analysis shows that the data could be bucketed into 5 groups.meaning a change in any of those sub-characteristics results in a change in points.

Fun Fact

13 (of 16) of the minerals that are essential for life: Calcium, chloride, chromium, copper, iodine, iron, magnesium, phosphorus, potassium, selenium, sodium, and zinc are minerals that essential for life - they can all be found in wine.