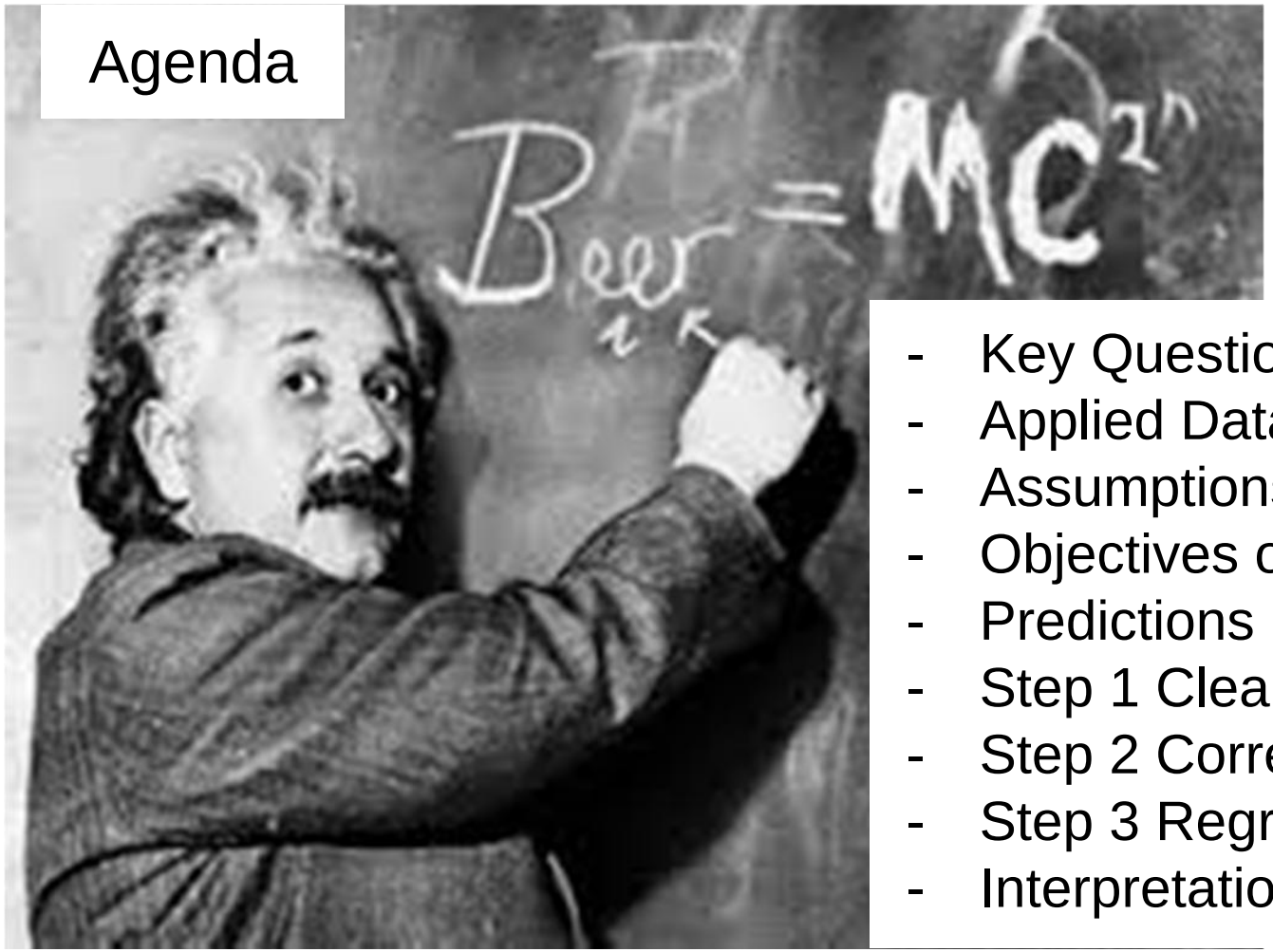
BEER

Gillian Almeida,Steven Benedetti, Ryan Chung, Lee Cleghorn, Ryo Naiki, Aanchal Saxena, Dana Silberstein, Naomi Ueda

Agenda



- Key Questions
- Applied Data
- Assumptions
- Objectives of the project
- Predictions
- Step 1 Cleaning the data
- Step 2 Correlation Matrix
- Step 3 Regression Analysis
- Interpretation

Key Questions

What variables have the most statistical impact on a beer's score?

1. Does the location of your brewery influence the quality* of your beer?

2. If we assume that craft brews are largely purchased locally, do the demographics of the local population influence how people rate the local beers (and how much they like them)?

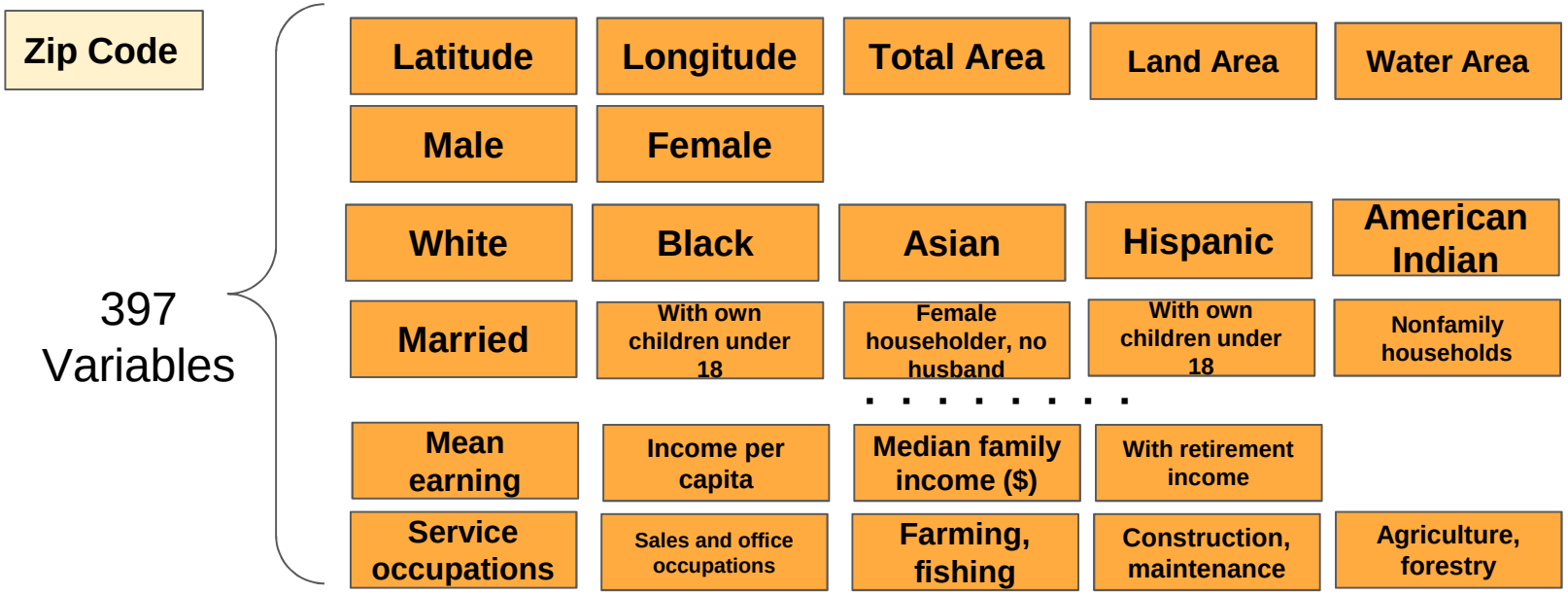
* We used user review scores (out of 5) from BeerAdvocate website database as a proxy for perceived beer quality

Data

Data 1: Beer Score Data per Zip Code (Source: www.beeradvocate.com)

Zip Code | **Score*** | **Style** | **ABV** | **Population**

Data 2: Zip Code Demographics Data (Source: National Census Bureau)



Assumption

1) These craft breweries have mostly local distribution and sales

2) Therefore, the majority of people consuming these beers and writing the reviews are local to the area where the beer is produced

3) Therefore, there is some relevance to demographic data o where the beer is produced to the score that the beer receive

Objectives

- Determine beer and brewery attributes that meaningfully impact the score

- Speculate the score of new beer depending on the location of brewery and other attributes using the regression equation

- Interpret regression results using our existing knowledge of beer and beer lovers

Predictions

- High ABV beers get rated higher due to alcohol making everything taste good
- Definite pattern between geographical data of breweries and scores of the beer produced
- Unclear link between demographics data vs. perceived beer quality

vs

Steps

Step 1
Cleaning the data

Step 2
Correlation Matrix

Step 3
Regression Analysis

1. Removed attributes with incomplete data

2. Removed the attributes that were represented in other attributes

   ex. Male + Female = Total population(Removed)

1. Combined some attributes into one variable

   ex. Consolidated the age data into 1

a. Threshold : 0.9

b. Removed the attribute with values larger than 0.9

Ran regression analyses based on the correlation numbers

Interpretation

Step 1: Cleaning the Data

3,800 Rated Beers

400 Census Attributes

United States™ Census Bureau

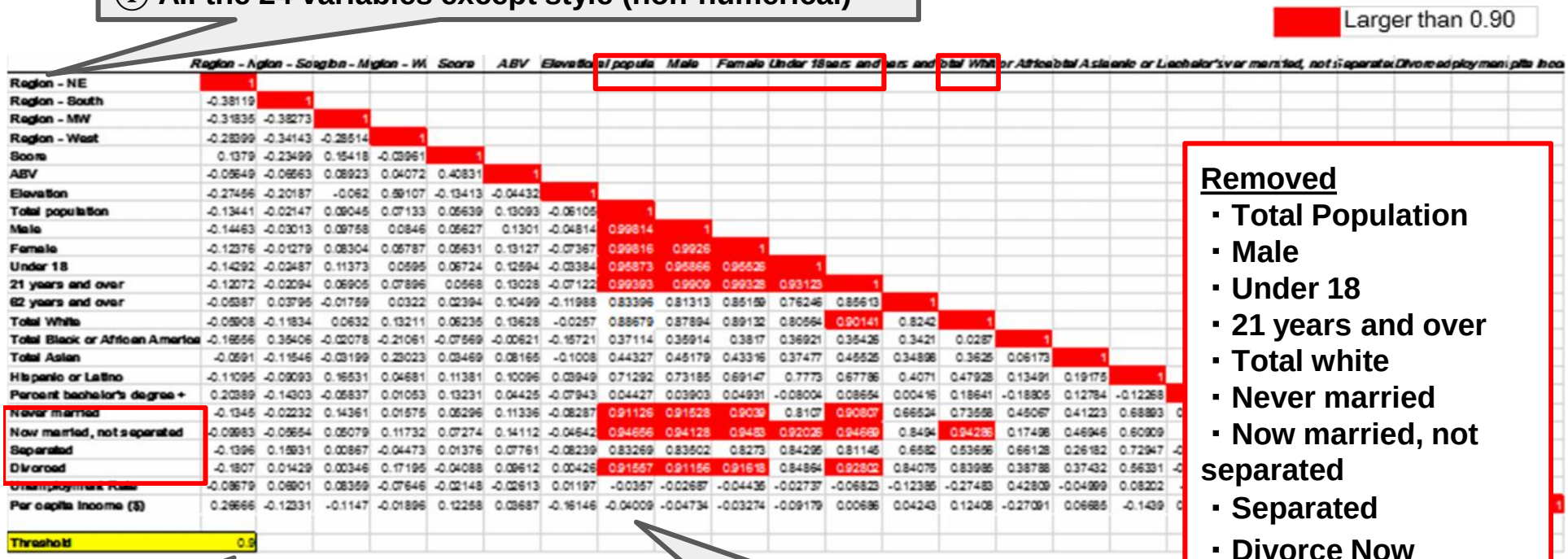
= **OVERLOAD**

1. Python scrape of Beer Advocate - ”States top 100 “ --50 Lists
2. Produced 96 Beer styles---all eliminated as data point from this study (another day)
3. ABV kept as major data point
4. Zip Codes from brewery matched

1. 40,000 rows of census data matched to breweries
2. 400 Census Attributes, reduced to ~50
3. Eliminated much of the economic data because it was very repetitive

Step 2: Full Correlation Matrix

① All the 24 variables except style (non-numerical)



② 0.90 as the threshold

③ Removed the attributes with values larger than 0.90

Removed

- Total Population
- Male
- Under 18
- 21 years and over
- Total white
- Never married
- Now married, not separated
- Separated
- Divorce Now married, not separated
- Separated
- Divorced

Step 2: Reduced Correlation Matrix

	Region - N	gion - So	egion - M	gion - W	Score	ABV	Elevation	Female	or Africa	otal Asia	anic or L	achelor's	ployment	pita incor
Region - NE	1													
Region - South	-0.3812	1												
Region - MW	-0.3183	-0.3827	1											
Region - West	-0.284	-0.3414	-0.2851	1										
Score	0.1379	-0.235	0.1542	-0.0396	1									
ABV	-0.0565	-0.0656	0.0892	0.0407	0.4083	1								
Elevation	-0.2746	-0.2019	-0.062	0.5911	-0.1341	-0.0443	1							
Female	-0.1238	-0.0128	0.083	0.0579	0.0563	0.1313	-0.0737	1						
Total Black or African	-0.1656	0.3541	-0.0208	-0.2106	-0.0757	-0.0062	-0.1572	0.3817	1					
Total Asian	-0.0591	-0.1155	-0.032	0.2302	0.0347	0.0817	-0.1008	0.4332	0.0617	1				
Hispanic or Latino	-0.1109	-0.0909	0.1653	0.0468	0.1138	0.101	0.0395	0.6915	0.1349	0.1918	1			
Percent bachelor's de	0.2039	-0.143	-0.0584	0.0105	0.1323	0.0443	-0.0794	0.0493	-0.1881	0.1278	-0.1227	1		
Unemployment Rate	-0.0868	0.069	0.0836	-0.0765	-0.0215	-0.0261	0.012	-0.0443	0.4281	-0.05	0.082	-0.3974	1	
Per capita income ($)	0.2667	-0.1233	-0.1147	-0.019	0.1226	0.0369	-0.1615	-0.0327	-0.2709	0.0669	-0.1439	0.788	-0.5692	1

Biggest correlation is 0.768

Step 3: Regression Analysis - Results in Excel

Regression Statistics	
Multiple R	0.50796456
R Square	0.258027994
Adjusted R Square	0.254847344
Standard Error	0.221916531
Observations	3361

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	14	57.32107546	4.094362533	89.53475847	8.4111E-219
Residual	3347	164.8295313	0.049246947		
Total	3361	222.1506068			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3.659950272	0.026219456	139.5891028	0	3.608542493	3.711358051	3.608542493	3.711358051
Region - NE	0	0	65535	#NUM!	0	0	0	0
Region - South	-0.131783313	0.011343046	-11.61798261	#NUM!	-0.154023317	-0.109543308	-0.154023317	-0.109543308
Region - MW	-0.001737167	0.011661644	-0.14896419	0.881590864	-0.024601839	0.021127504	-0.024601839	0.021127504
Region - West	-0.029765757	0.015107103	-1.970315345	0.048884546	-0.059385847	-0.000145668	-0.059385847	-0.000145668
ABV	4.143015623	0.163459525	25.34581957	8.4338E-130	3.822524945	4.463506302	3.822524945	4.463506302
Elevation	-2.68929E-05	3.77278E-06	-7.128134377	1.24062E-12	-3.42901E-05	-1.94957E-05	-3.42901E-05	-1.94957E-05
Female	-1.26719E-05	6.29883E-06	-2.011793179	0.04432174	-2.50219E-05	-3.21997E-07	-2.50219E-05	-3.21997E-07
Total White	2.51644E-06	1.6432E-06	1.531426682	0.12575848	-7.05336E-07	5.73821E-06	-7.05336E-07	5.73821E-06
Total Black or African American	2.75743E-06	1.71181E-06	1.610822954	0.10731261	-5.98878E-07	6.11373E-06	-5.98878E-07	6.11373E-06
Total Asian	6.32572E-07	1.91415E-06	0.330471745	0.741064223	-3.12045E-06	4.38559E-06	-3.12045E-06	4.38559E-06
Hispanic or Latino	6.86768E-06	1.84183E-06	3.728732143	0.000195633	3.25646E-06	1.04789E-05	3.25646E-06	1.04789E-05
Percent bachelor's degree +	0.001793107	0.000458978	3.906742549	9.53979E-05	0.000893202	0.002693012	0.000893202	0.002693012
Unemployment Rate	0.163722341	0.123129529	1.329675686	0.183715761	-0.077694403	0.405139084	-0.077694403	0.405139084
Per capita income ($)	4.89471E-07	9.31805E-07	0.525293024	0.599414301	-1.33749E-06	2.31643E-06	-1.33749E-06	2.31643E-06

Step 3: Regression Analysis - Results in Minitab

Ran best subsets with selected 14 vars
Highlighted row shows the best model

- Highest R-Sq (adj) of 25.2
- Mallows CP of 9.3 closest to number of predictors + constant (10)

Therefore, best model includes:

- ABV
- Elevation
- Female
- Hispanic or Latino
- Unemployment rate
- Per capita income
- Region - NE
- Region - South
- Region - MW

Stepwise regression (alpha = 0.15), forward selection (alpha to enter = 0.25) and backward elimination (alpha to remove =0.1) all yield the same set of predictors

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Cp	S	ABV	Elev	Female	Black/African American	Asian	Hispanic/Latino	Unemployment Rate	Per Capita Income	Region - NE	Region - South	Region - MW
1	16.7	16.6	16.6	384.3	0.23476	X										
1	5.6	5.5	5.5	883	0.24991										X	
2	21.1	21	20.9	189.2	0.22852	X									X	
2	19.3	19.2	19.1	270.1	0.23112	X								X		
3	23.7	23.6	23.5	72.9	0.22471	X	X								X	
3	22.5	22.4	22.3	126.7	0.22647	X								X		X
4	24.1	24	23.8	58.6	0.22421	X	X				X				X	
4	24	23.9	23.7	62.4	0.22433	X	X						X		X	
5	24.6	24.5	24.3	37.3	0.22347	X	X	X			X				X	
5	24.5	24.3	24.2	42.6	0.22365	X	X				X		X		X	
6	25.1	24.9	24.7	17.9	0.2228	X	X	X			X		X		X	
6	24.8	24.7	24.4	29.7	0.22319	X	X			X	X		X		X	
7	25.2	25	24.8	13.6	0.22262	X	X	X			X	X	X		X	
7	25.1	25	24.7	16.4	0.22272	X	X	X		X	X		X		X	
8	25.3	25.1	24.8	11.8	0.22253	X	X	X		X	X	X	X		X	
8	25.3	25.1	24.8	12.9	0.22257	X	X	X			X		X	X	X	X
9	25.4	25.2	24.9	9.3	0.22241	X	X	X			X	X	X	X	X	X
9	25.3	25.1	24.8	12.4	0.22252	X	X	X		X	X	X	X		X	X
10	25.4	25.2	24.9	10	0.22241	X	X	X		X	X	X	X	X	X	X
10	25.4	25.2	24.9	11.2	0.22244	X	X	X	X		X	X	X	X	X	X
11	25.4	25.2	24.8	12	0.22244	X	X	X	X	X	X	X	X	X	X	X

Step 3: Regression Analysis - Results in Minitab

ANOVA

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	9	56.383	6.2647	126.64	0.000
ABV	1	31.816	31.8164	643.17	0.000
Elevation	1	2.123	2.1227	42.91	0.000
Female	1	0.929	0.9287	18.77	0.000
Hispanic or Latino	1	2.007	2.0074	40.58	0.000
Unemployment Rate	1	0.279	0.2787	5.63	0.018
Per capita income ($)	1	1.325	1.3253	26.79	0.000
Region - NE	1	0.308	0.3081	6.23	0.013
Region - South	1	2.474	2.4739	50.01	0.000
Region - MW	1	0.367	0.3675	7.43	0.006
Error	3351	165.768	0.0495		
Lack-of-Fit	2636	154.396	0.0586	3.68	0.000
Pure Error	715	11.373	0.0159		
Total	3360	222.151			

P values are low, therefore we can reject the null hypothesis

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.222415	25.38%	25.18%	24.92%

Step 3: Regression Analysis - Results in Minitab

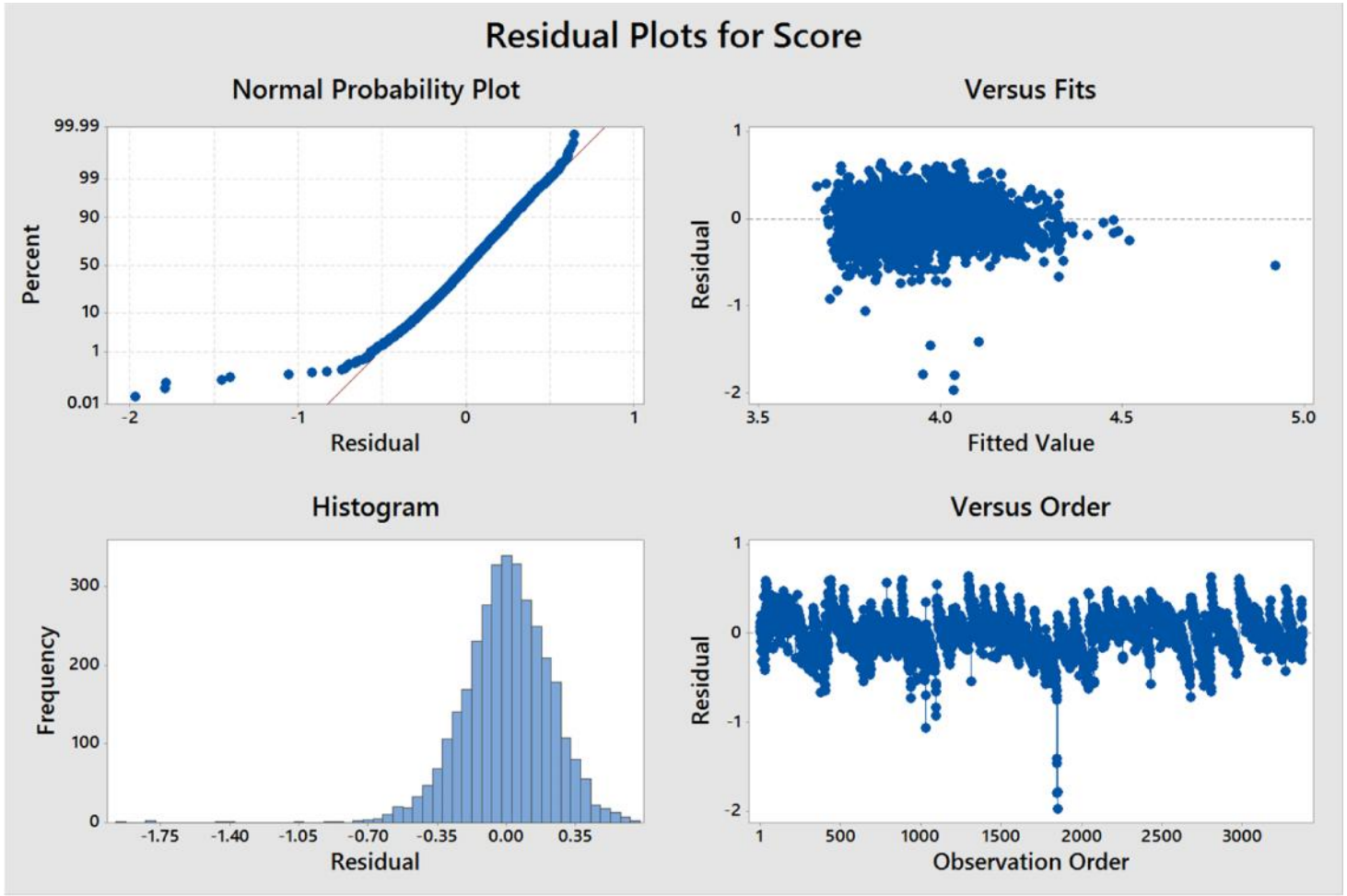
Coefficients

Term	Coef		SE Coef	T-Value	P-Value	VIF
Constant	3.6033		0.0273		131.91	0.000
ABV	4.152		0.164	25.36	0.000	1.04
Elevation		-0.000024	0.000004	-6.55	0.000	1.70
Female		-0.000003	0.000001	-4.33	0.000	2.07
Hispanic or Latino	0.000004	0.000001	6.37		0.000	2.07
Unemployment Rate	0.255		0.107	2.37	0.018	1.52
Per capita income ($)	0.000003	0.000001	5.18		0.000	1.65
Region - NE	0.0360		0.0144	2.50	0.013	2.59
Region - South	-0.0941	0.0133	-7.07		0.000	2.59
Region - MW	0.0363		0.0133	2.73	0.006	2.22

VIF values close to 1 tell us there is very little multicollinearity present in this analysis

Regression Equation

Score = 3.6033 + 4.152 ABV - 0.000024 Elevation - 0.000003 Female + 0.000004 Hispanic or Latino + 0.255 Unemployment Rate + 0.000003 Per capita income ($) + 0.0360 Region - NE - 0.0941 Region - South + 0.0363 Region - MW


Residual Plots for Score
Normal Probability Plot
Percent
99.99
99
90
50
10
1
0.01
-2
-1
0
1
Residual
Versus Fits
Residual
1
0
-1
-2
3.5
4.0
4.5
5.0
Fitted Value
Histogram
Frequency
300
200
100
0
-1.75
-1.40
-1.05
-0.70
-0.35
0.00
0.35
Residual
Versus Order
Residual
1
0
-1
-2
1
500
1000
1500
2000
2500
3000
Observation Order

Interpretation

Observations	Effect	Potential Explanation
Higher ABV usually leads to higher score	+1% ABV = +0.04 score	Drunk tasters are more generous
Higher unemployment rate leads to higher score	+1% rate = +0.0025 score	People looking for work are more appreciative, or…. Breweries built on cheap property
Beers from the Northeast and Midwest breweries show higher scores	~ 0.04 boost in scores	Breweries in NE and MW just better
Southern beers show lower scores	~ 0.09 penalty in scores	Beers from the South are just not good

National Beer League Half-Time Report

Proud States

1	California	4.33
2	Vermont	4.26
3	Massachusetts	4.23
4	Oregon	4.17
5	Illinois	4.15
6	New York	4.13
7	Michigan	4.12
8	Ohio	4.12
9	Florida	4.10
10	Colorado	4.09
11	Minnesota	4.08
12	Pennsylvania	4.07
13	Indiana	4.05
14	Maine	4.02
15	Texas	4.02
16	North Carolina	4.00

Well-behaved

17	Missouri	3.98
18	Wisconsin	3.98
19	Washington	3.95
20	Wyoming	3.95
21	Connecticut	3.94
22	Virgina	3.93
23	Maryland	3.91
24	Alabama	3.90
25	New Hampshire	3.90
26	Oklahoma	3.88
27	Iowa	3.87
28	Georgia	3.86
29	DC	3.85
30	New Jersey	3.80
31	South Carolina	3.79
32	Kansas	3.78
33	New Mexico	3.76

Critical Condition

34	Delaware	3.74
35	Arkansas	3.73
36	Idaho	3.72
37	Utah	3.72
38	Hawaii	3.70
39	Mississippi	3.70
40	Nebraska	3.70
41	North Dakota	3.69
42	Rhode Island	3.68
43	Kentucky	3.66
44	Montana	3.65
45	Tennessee	3.61
46	Louisiana	3.60
47	West Virginia	3.60
48	South Dakota	3.48
49	Nevada	3.33

****according to our broken data set, if you live in Alaska or Arizona, you are without beer altogether
We recommend moving somewhere else.