

Abidemi Olaoye

Ben Soumahoro

Oluwapelumi Osunrayi

Omotola Adeoye

July 28, 2020

Agenda

- 1 Introduction
- 2 Literature Review
- 3 Data Understanding and Data Cleaning
- 4 Exploratory Analysis
- 5 Modeling and Evaluation
- 6 Conclusion and Recommendations
- 7 Questions and Answers



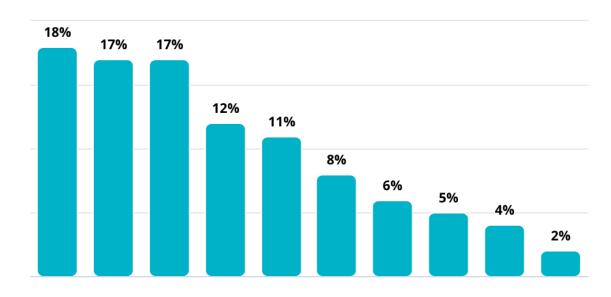


Company Background

- Online crowd-sourced review forum, headquartered in San Francisco, California.
- Founded in 2004 by former PayPal employees Russel Simmons and Jeremy Stoppelman.
- Find local businesses like hair stylists, mechanics, dentist, etc.



Motivation



Reviewed Business by Category

178 million Unique visitors monthly

35 Million mobile app unique users (in Q1'20)

Problem

Increase in new restaurants by 6%

Increase in competition

Increase in closure

	2018	2019
Number Of Restaurants	33,110	35,305
Open	23,118	23,867
Closed	9,992	11,438

Note: This calculation only applies to restaurants on Yelp. Not all restaurants sign up on Yelp

Problem

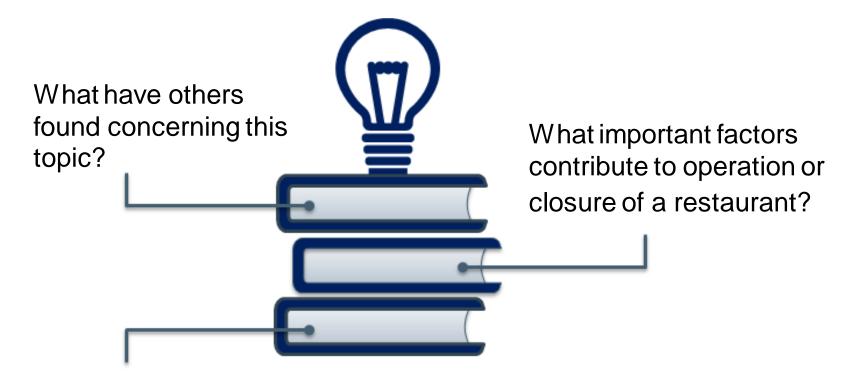
The restaurant industry sees a 1% - 1.2% yearly increase in restaurants.

Restaurants that cannot compete also close each year.

In the spring of 2018, the U.S. restaurant count decreased by 1% from the year before.



Research Questions



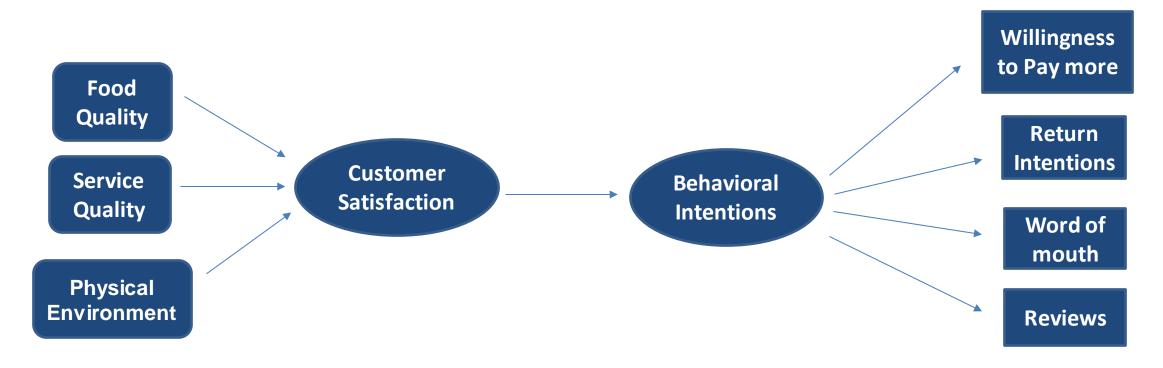
How do these attributes impact the closure of a restaurant?



Literature Review

Predictors of Restaurant Closure

Authors: Dutta & Venkatesh, 2007; Gagić et al., 2013; Namkung & Jang, 2010; Ozdemir & Hewett, 2010; Parsa et al., 2005; Tripathi & Dave, 2016.

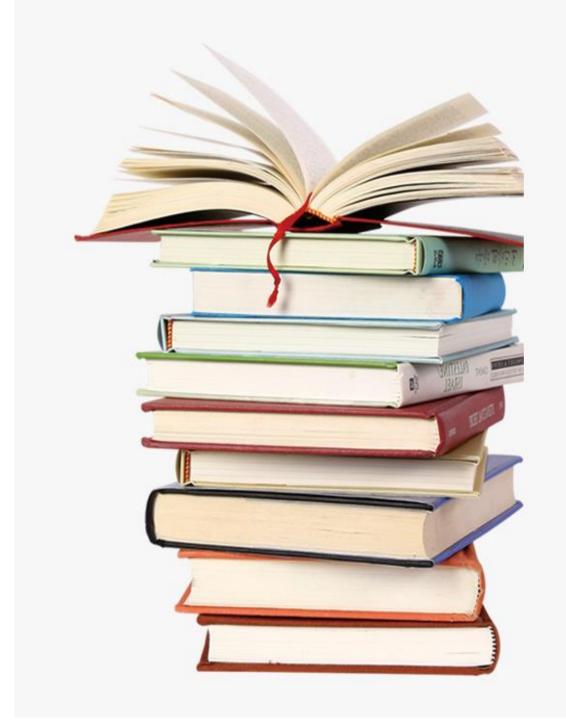


Literature Review

Predictors of Restaurant Closures Closely Related to our Project

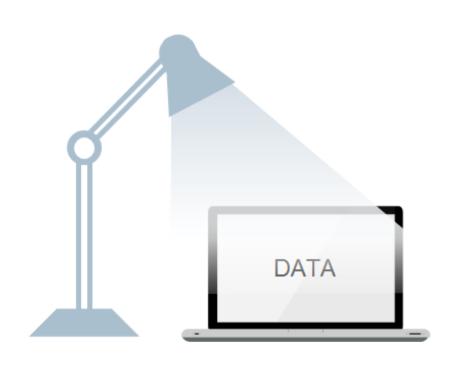
Authors: Kong et al., n.d.; Shellenberger, 2017; Luca, 2016; Snow, 2018.

Location
 Number of Reviews
 Ethnicity
 Noise Level
 Entertainment





The Dataset



- Yelp Dataset Challenge 2019
- 209,393 businesses, 15 variables
- 35,305 restaurants
- Arizona, Nevada, North Carolina, Ohio, Pennsylvania

Data Cleaning

Irrelevant columns such as "Longitude," "Latitude," and "Hours" were dropped

Foreign symbols such as commas and brackets were removed from "Categories" and "Attributes" columns

Binary and interval variables created from "Names", "Attributes" and "Categories" columns

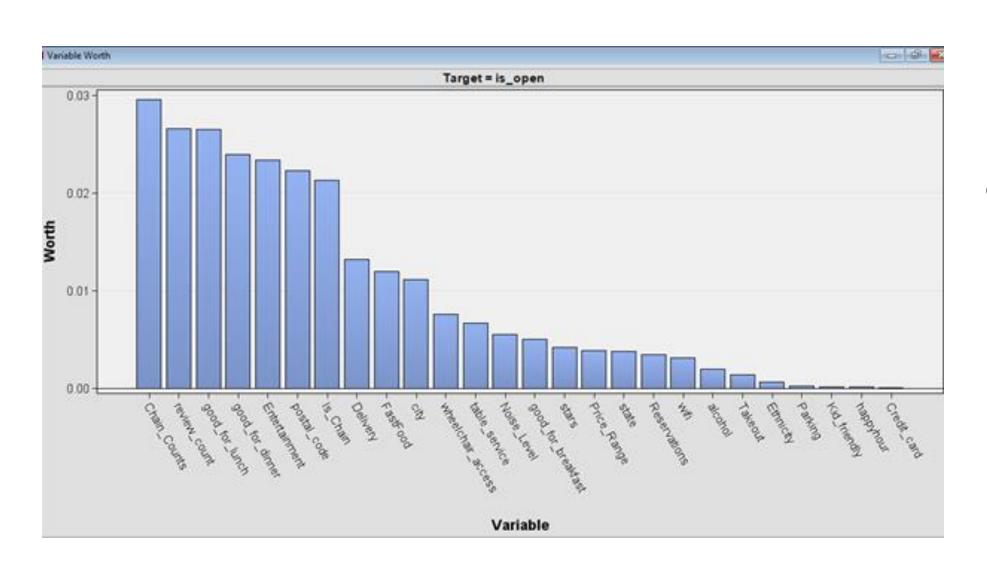
Column Name	Description	Role	Measurement Level	Data Source
Delivery	Dichotomous indicator of a restaurant that includes "delivery" as an attribute (1=Yes, 0=No)	Input	Binary	Extracted
Alcohol	Dichotomous indicator of a restaurant that offers alcohol as an attribute (1=Yes, 0=No)	Input	Binary	Extracted
Attributes	Describes certain amenities and services available at each business. For example; parking availability, table reservations, price range, kid friendly meals, ambience (outdoor seating), availability of alcohol, acceptance of credit cards etc.	Rejected	Nominal	Yelp Dataset Challenge
Business ID	Displays a unique id for each business.	ID	Nominal	Yelp Dataset Challenge
FastFood	Dichotomous indicator of a restaurant that includes "fast food" as an attribute (1=Yes, 0=No)	Input	Binary	Extracted
Categories	Mentions the specific services that each business provides, such as Restaurants, Fast Food, Pizza, etc.	Rejected	Nominal	Yelp Dataset Challenge
Credit_ card	Dichotomous indicator of a restaurant that has credit card services as an attribute (1=Yes, 0=No)	Input	Binary	Extracted
Entertainment	Dichotomous indicator of a restaurant that has entertainment services (such as background music, television, games, photo booth, etc) as an attribute (1=Yes, O=No)	Input	Binary	Extracted

Column Name	Description	Role	Measurement Level	Data Source
Good_for_ breakfast	Dichotomous indicator of a restaurant that is "good for breakfast" as an attribute (1=Yes, 0=No)	Input	Binary	Extracted
Good_for_ dinner	Dichotomous indicator of a restaurant that is "good for dinner as an attribute (1=Yes, 0=N:)	Input	Binary	Extracted
Good_for_ lunch	Dichotomous indicator of a restaurant that is "good for lunch' as an attribute (1=Yes, 0=No)	Input	Binary	Extracted
Happyhour	Dichotomous indicator of a restaurant that includes "happy hour" as an attribute (1=Yes; 0=No)	Input	Binary	Extracted
Address	Displays street addresses of businesses	Rejected	Nominal	Yelp Dataset Challenge
Chain_ Counts	Count of all restaurants that are a part of the same franchise and have the same name	Input	Interval	Extracted
City	Lists the city where the business is located	Rejected	Nominal	Yelp Dataset Challenge
Is_Chain	Dichotomous indicator of a restaurant that has a value of 4 or above in the "Chain_Counts" variable. Indicating restaurant is a chain in the terms of this project (1=Yes, 0=No)	Input	Binary	Extracted

Column Name	Description	Role	Measurement Level	Data Source
Price_Range	Nominal indicator of the price range per person at a restaurant (1= under S10 , 2= 511-530 cheap: 3 = \$31-\$60 expensive, 4 = S60 and above)	Input	Nominal	Extracted
State	Lists state abbreviations e.g. NJ, OK, NV _: etc.	Input	Nominal	Yelp Dataset Challenge
Is_ open (Target Variable)	Consists of Binary numbers specifying whether that business is functioning (1) or out-of-business (0).	Target	Binary	Yelp Dataset Challenge
Kid_Friendly	Dichotomous indicator of a restaurant that has kid friendly services as an attribute (1=Yes, 0=No)	Input	Binary	Extracted
Name	Displays name of each business	Rejected	Nominal	Yelp Dataset Challenge
Noise_Level	Nominal indicator of noise level at the restaurant (1 = quiet, 2= average, 3= loud, 4=very loud)	Input	Ordinal	Extracted
Parking	Dichotomous indicator of a restaurant that has parking services as an attribute (1=Yes, 0=No)	Input	Binary	Extracted
Postal Code	Displays zip code of each business	Rejected	Nominal	Yelp Dataset Challenge

Column Name	Description	Role	Measurement Level	Data Source
Reservations	Dichotomous indicator of a restaurant that has reservation services as an attribute (1=Yes. 0=No)	Input	Binary	Extracted
Restaurant	Dichotomous indicator of a business categorized as a restaurant (1=Yes 0=No)	Rejected	Unary	Extracted
Review_ Count	Shows the number of online reviews that each business has received.	Input	Interval	Yelp Dataset Challenge
Stars	Shows the number of star-ratings each business has achieved.	Input	Interval	Yelp Dataset Challenge
Ethnicity	Indicates if a restaurant is labeled as American, Chinese, Italian, Japanese, Mexican, or Other	Rejected	Nominal	Extracted
Table_ sercice	Dichotomous indicator of a restaurant that includes table services as an attribute (1=Yes; 0=No)	Input	Binary	Extracted
Takeout	Dichotomous indicator of a restaurant that has takeout as an attribute (1=Yes, 0=No)	Input	Binary	Extracted
Wheelchairacce pted	Dichotomous indicator of a restaurant that is wheelchair accessible as an attribute (1=Yes: 0=No)	Input	Binary	Extracted
Wifi	Dichotomous indicator of a restaurant that includes Wifi as an attribute (1=Yes, 0 =No)	Input	Binary	Extracted

Variable Worth Plot



Chain_Counts

Review_Count

Good_for_lunch

Good_for_dinner

Entertainment

Postal_code

Is_chain

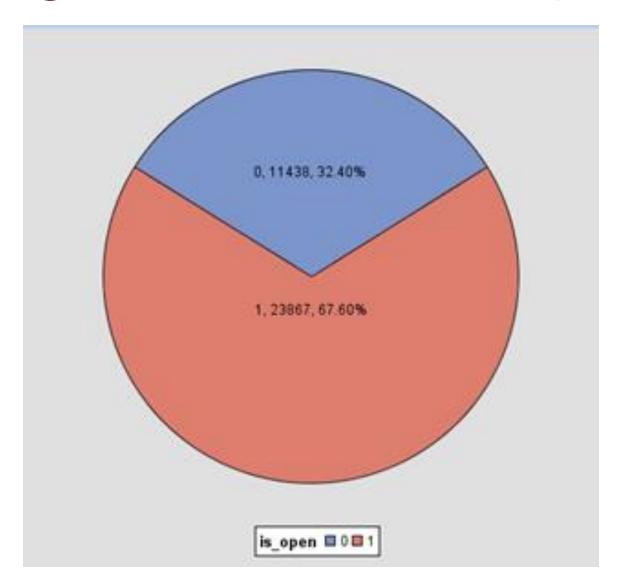
Delivery

FastFood

City



Target Variable "Is_Open"

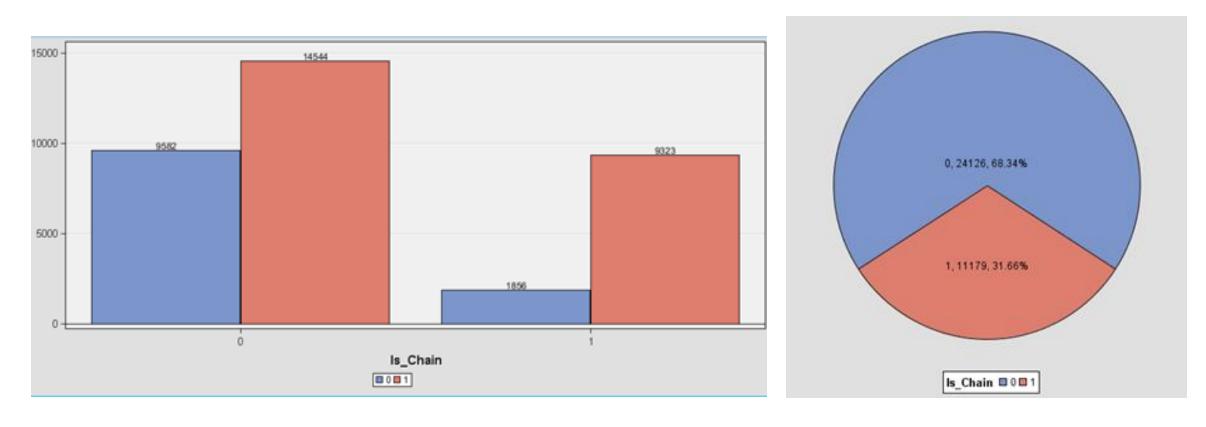


1=Is Open

0=Is Not Open (or Closed)

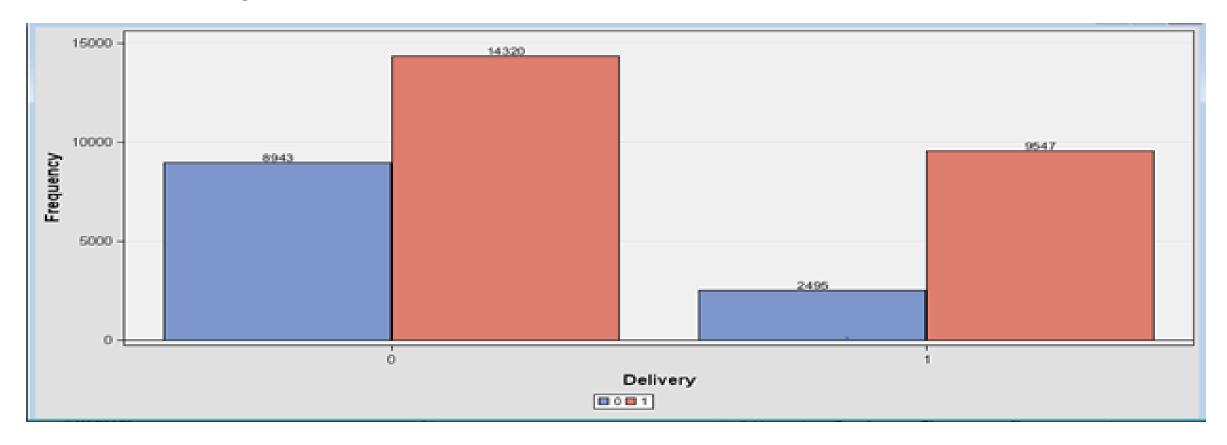
67.60% of the target variable represent open restaurants, while 32.40% represent closed restaurants

Is_Chain



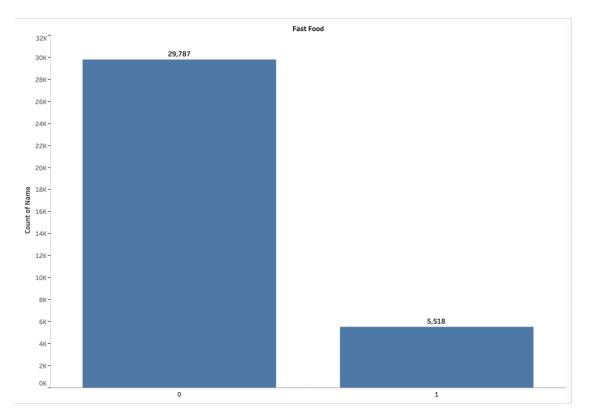
- ❖ There are 11,179 (31.66%) chain restaurants and 24,126 (68.34%) restaurants that are not classified as chain restaurants.
- ❖ "Large chains have the resources to ride out a protracted shutdown, but independent restaurants" find it harder to survive in same climate (Severson & Yaffe-Bellany, 2020).

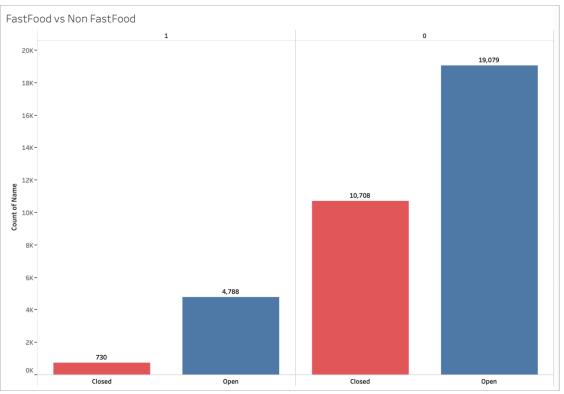
Delivery



- ❖ That 20.72% of restaurants that offer delivery are closed (12,042 total delivery restaurants) while 38.44% of restaurants that do not offer delivery are closed (23,263 total non-delivery restaurants).
- ❖ Delivery services are a popular dining option with U.S. consumers, as a November 2016 survey found that 20 percent of respondents use food delivery at least once a week.

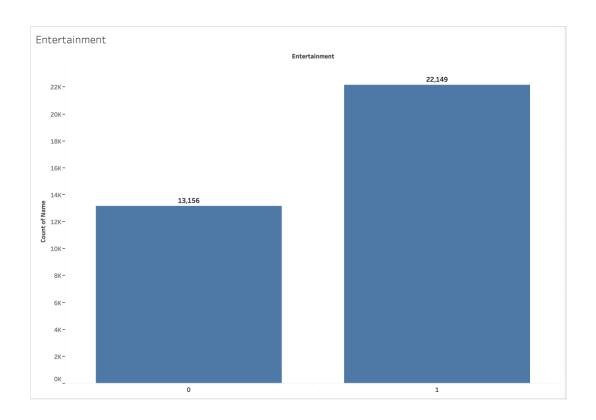
Fast Food

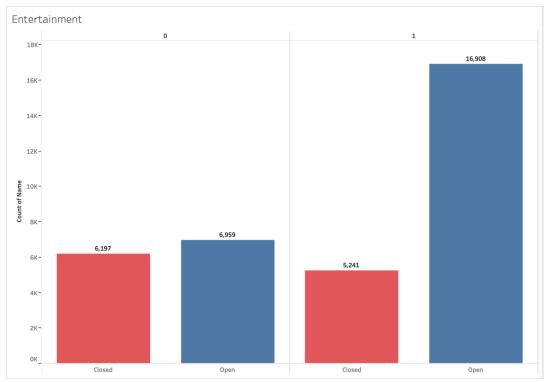




- ❖ There are 5,518 (15.63%) fast food restaurants in the dataset while 29,787 (84.37%) of them are not fast food restaurants i.e., they are dining restaurants, café, buffet and so on.
- ❖ Almost a three times higher rate of closure in non fast food restaurants than in fast food restaurants.

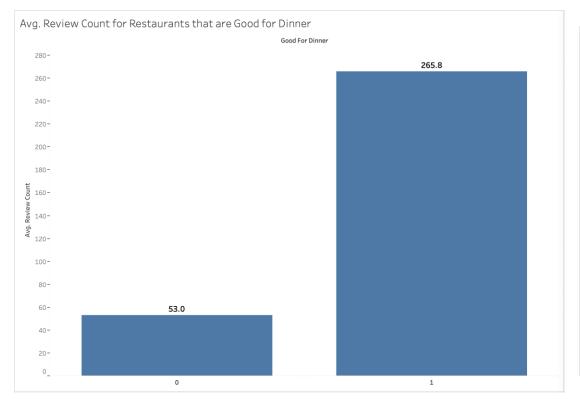
Entertainment

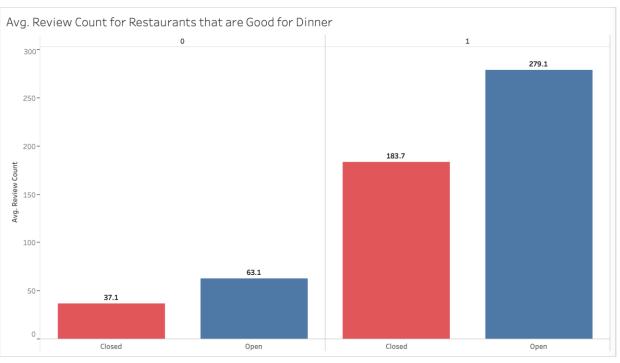




- In the dataset, 22,149 (62.74%) restaurants offer one or more forms of entertainment such as Background Music, Live Music, Jukebox and karaoke while 13,156 (37.26%) restaurants do not offer any entertainment.
- There is a higher rate of closure in restaurants that do not provide entertainment. Experiences such as background music and other forms of entertainment are statistically significant predictors of satisfaction and repeat patronage (DiPietro 2016).

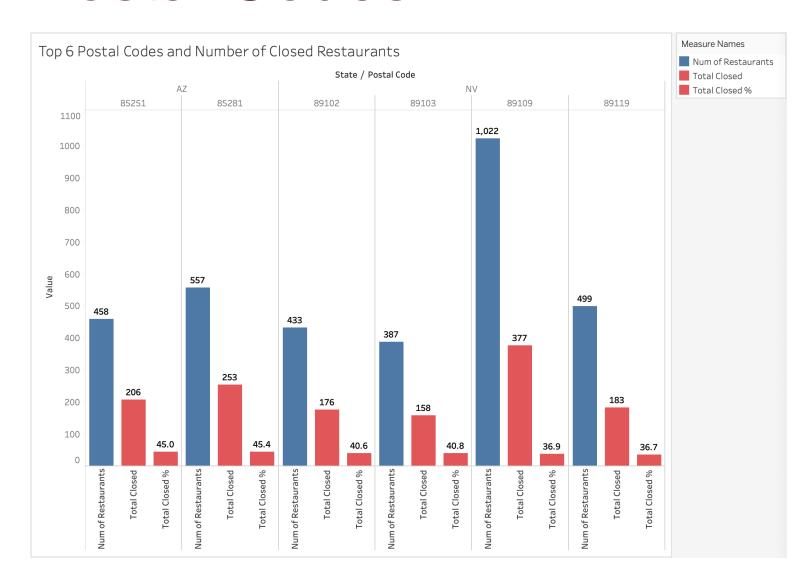
Good for Dinner





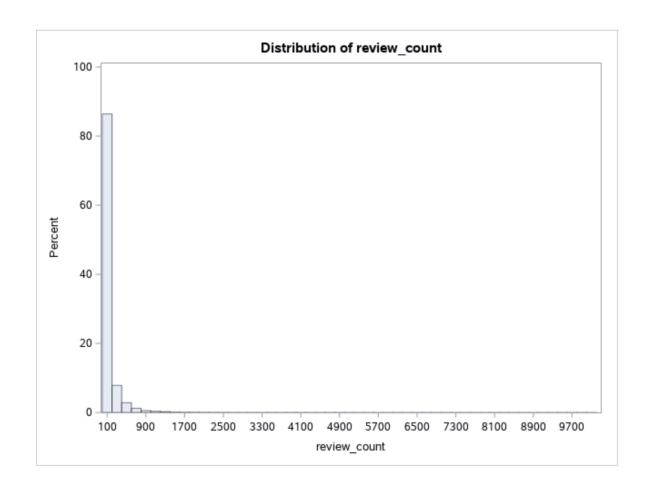
- ❖ There are 212.8 more reviews on average for restaurants that are good for dinner time in comparison to restaurants that do not offer dinner services with an average of 53 reviews.
- High review counts and closed restaurants are mutually exclusive, meaning they cannot occur at the same time. On the other hand, high review counts and open restaurants are mutually inclusive, meaning they occur together.

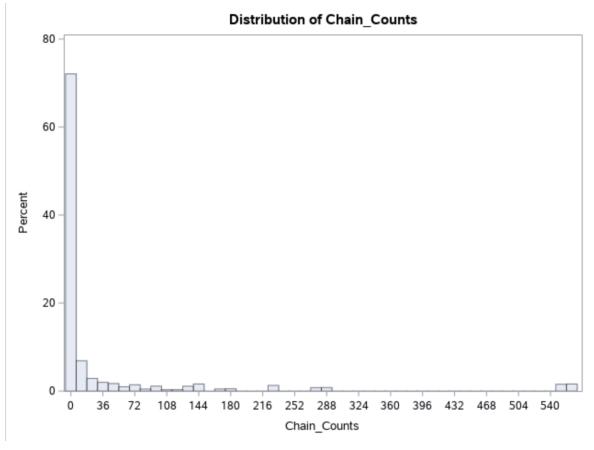
Postal Codes



4 (89102, 89103, 89109, 89139) out of the top 6 postal codes have the highest rate of closure overall. All 4 postal codes are Las Vegas postal codes which are probably home to various restaurants who want to profit from the bustling market.

Review Count and Chain Count





Review Count and Chain Count

Variable Name	Mean	Std. Dev	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Chain_Counts	38.26903	107.4537	0	1	1	568	3.838634	14.83487
Review_count	109.1025	255.7983	0	3	36	10129	10.69444	222.7171

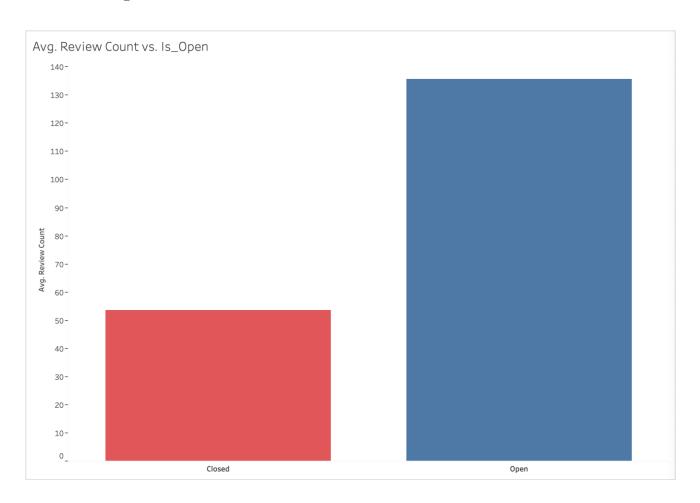
Transformation

Source	Method	Variable Name	Formula 🛦	Non Missing	Missing	Minimum	Maximum	Mean	Standard	Skewness	Kurtosis	Label
									Deviation			
Input	Original	Chain_Counts		17652	0	1	568	39.11432	109.321	3.752594	14.03053	Chain_Cou
Input	Original	review_count		17652	0	3	9264	109.6016	259.9702	10.99932	223.7574	review_cou
Output	Computed	LOG_Chain_Counts	log(Chain	17652	0	0.693147	6.34388	1.731068	1.655816	1.494342	0.884064	Transform
Output	Computed	LOG_review_count	log(review	17652	0	1.386294	9.133999	3.664527	1.410104	0.305811	-0.56394	Transform

Review Count vs. Is_Open T-test

is_open	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		11438	53.6736	126.5	1.1832	3.0000	5494.0
1		23867	135.7	294.9	1.9086	3.0000	10129.0
Diff (1-2)	Pooled		-81.9926	252.9	2.8761		
Diff (1-2)	Satterthwaite		-81.9926		2.2456		

Method	Variano	es D	FtV	alue	Pr > t
Pooled	Equal	3530	3 -2	8.51	<.0001
Satterthwait	e Unequa	al 3496	0 -3	6.51	<.0001
	Equalit	y of Varia	nces		
Method	Equalit	y of Varia	rces F Val	ue	Pr > F

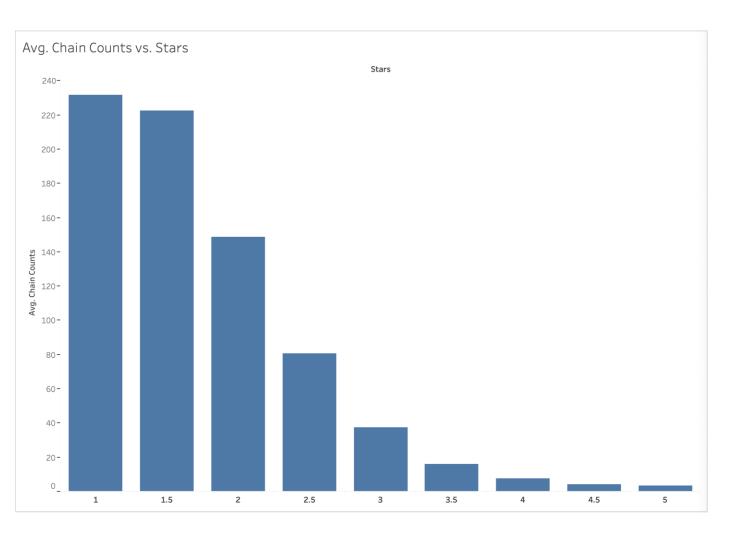


Hypothesis:

 H_0 : μ_0 - μ_c = 0 (with O= open and C= Closed)

 H_a : μ_o - $\mu_c \neq 0$

One Way Anova Test



Le		est for Homogeneit A of Squared Deviat	A STATE OF THE PARTY OF THE PAR		ce
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
stars	8	4.407E12	5.509E11	459.22	<.0001
Error	35296	4.234E13	1.1996E9		

Welch's	ANOVA	for Chain_	Counts
Source	DF	F Value	Pr > F
stars	8.0000	535.38	<.0001
Error	3860.2		

Level of		Chain_Counts			
stars	N	Mean	Std Dev		
1	256	222.472656	222.155164		
2	2297	146.286461	178.325278		
3	6028	36.808394	97.457261		
4	8587	7.531850	41.165797		
5	955	3.364398	27.492122		
1.5	968	218.816116	204.860563		
2.5	3625	79.738207	144.250867		
3.5	7978	15.551893	60.546685		
4.5	4611	4.193017	29.176775		

POST HOC Analysis

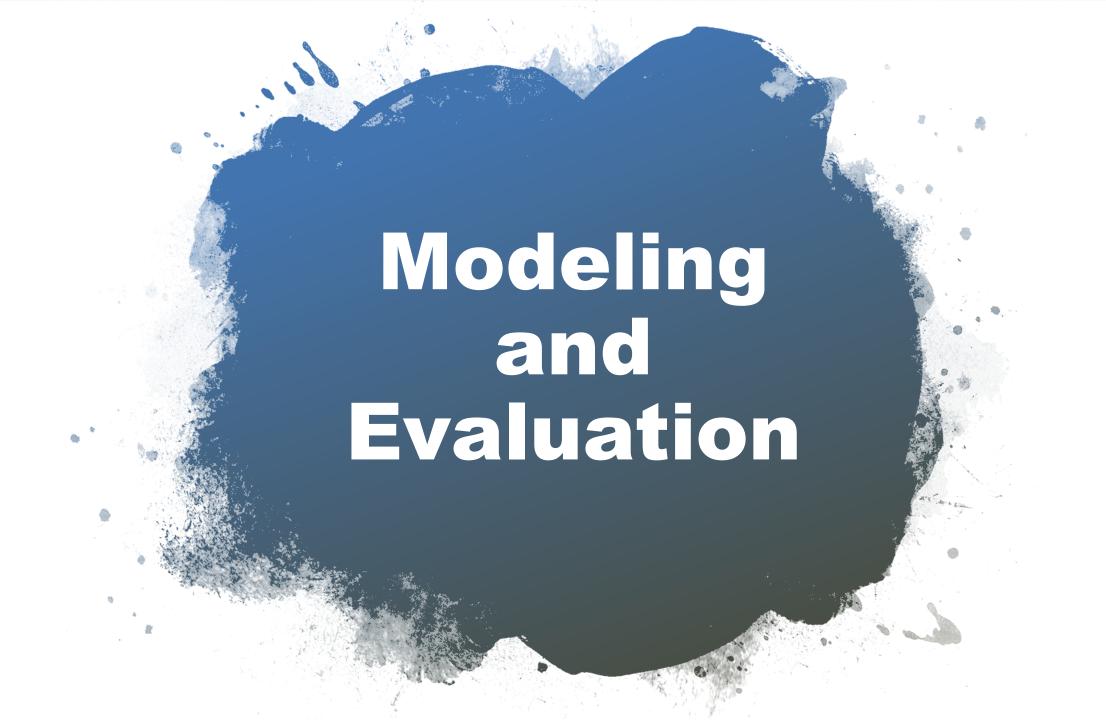
Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer

stars	Chain_Counts LSMEAN	LSMEAN Number
1	222.472656	1
2	146.286461	2
3	36.808394	3
4	7.531850	4
5	3.364398	5
1.5	218.816116	6
2.5	79.738207	7
3.5	15.551893	8
4.5	4.193017	9

	Least Squares Means for effect stars Pr > t for H0: LSMean(i)=LSMean(j)								
	Dependent Variable: Chain_Counts								
i/j	1	2	3	4	5	6	7	8	9
1		<.0001	<.0001	<.0001	<.0001	0.9998	<.0001	<.0001	<.0001
2	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
3	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
4	<.0001	<.0001	<.0001		0.9289	<.0001	<.0001	<.0001	0.5716
5	<.0001	<.0001	<.0001	0.9289		<.0001	<.0001	0.0043	1.0000
6	0.9998	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
7	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
8	<.0001	<.0001	<.0001	<.0001	0.0043	<.0001	<.0001		<.0001
9	<.0001	<.0001	<.0001	0.5716	1.0000	<.0001	<.0001	<.0001	

No significant difference in regard to chain counts between group means of:

- 1 Star and 1.5 stars (p-values 0.9998)
- 4 Stars and 5 stars (p-values 0.9289)
- 4 Star and 4.5 star (p-values 0.5716)
- 4.5 Star and 5 star (p-value 1.000)



Modeling

37 Models were created				
HP SVM Poly	HP SVM (RBF)			
Selection Tree with Neural Network	Stepwise Misclassification Regression			
Stepwise Regression with Neural Network	LARS with Regression			
Neural Network	LASSO with Regression			
LASSO with Neural Network	PLS 0.2 with Auto Neural Network			
LARS with Neural Network	Adaptive LASSO with Regression			
Adaptive LASSO with Neural Network	PCA with Neural Network			
Auto Neural (AOV 16)	PLS 0.2 with Regression			
Variable selection with Neural Network (AOV 16)	Variable Selection with Regression			
HP Forest larger	Variable Clustering (Cluster Component) with Regression			
PLS 0.2 with Neural Network	Variable Clustering (Best Variables) with Regression			
Variable Selection with Regression	Decision Tree			
Variable Selection with Neural Network	PCA with Regression			
HP Forest	PLS with Neural Network			
Adaptive LASSO with Auto Neural Network	PLS with Auto Neural Network			
HP SVM Linear	PLS with Regression			
Stepwise Regression with Auto Neural Network	HP SVM Sigmoid			
LARS with Auto Neural Network	Variable Selection with with Auto Neural Network			
LASSO with Auto Neural Network				

Decision Tree

		Variable Importance Plot					
Variable Name	Number of Spliting Rules	Importance	Validation Importance	Role of Validation to Training Importance			
Chain_Counts	2	1.0000	1.0000	1.0000			
Entertainment	1	0.9962	0.9637	0.9674			
Review_Count	2	0.7247	0.6323	0.8725			
Good_for_Dinner	2	0.6855	0.7689	1.1217			
Price_Range	1	0.6348	0.7088	1.1166			
Wheelchair_Access	3	0.6299	0.6680	1.0604			
Noise Level	2	0.4959	0.5299	1.0686			
Good_for_Lunch	2	0.4794	0.6001	1.2519			
Reservations	1	0.2152	0.2164	1.0056			
The other variables have 0 spliting rules and 0 importance							



False Negative



True **Negative**



False Positive



True Positive

- ❖ The Decision tree has 17 leaves
- The Validation Misclassification rate is 0.225287

Stepwise Regression

Variables Selected for the final model					
Credit_ card	Reservations	Wifi	Stars		
Delivery	Takeout	Wheelchair_access	LOG_review_ count		
Entertainment	Alcohol	Price_Range	Noise_Level		
FastFood	Good_for_ breakfast	Table_service	Happyhour		
ls_chain	Good_for_ dinner	Parking	LOG_chain_ counts		
Kid_ friendly	Good_for_ lunch	State			

- All 23 variables were selected
- The Validation Misclassification rate is 0.199909



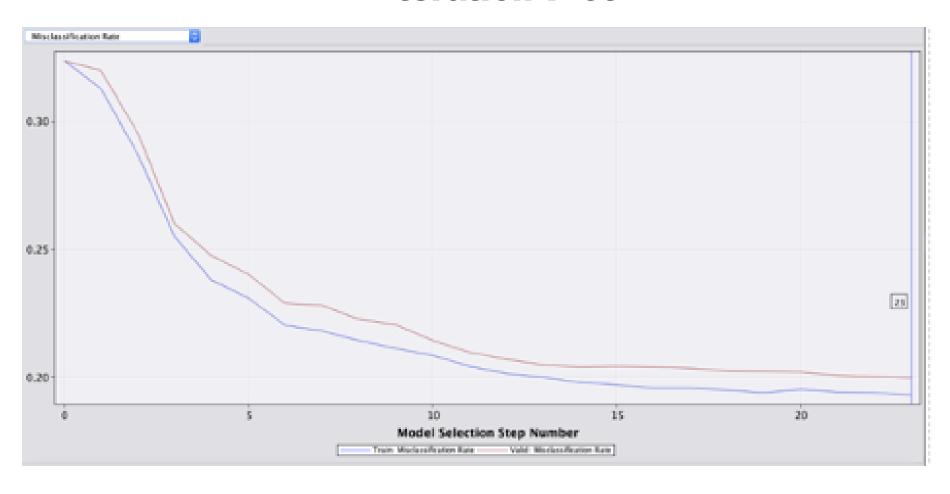






Stepwise Regression

Iteration Plot



Stepwise Regression

Odds Ratio Estimates

LOG_Chain_ Counts 1.847

For each additional restaurant of a chain, the odds of staying open change by a factor of 1.847, or an 84.7% increase.

LOG_Review_Count 2.174

For each additional review, the odds of staying open is 2.174 higher.

Stars 1.337

For each additional unit of star, the odds of staying open changes by a factor of 1.337, or a 33.7% increase.

Entertainment 0.259

For restaurants without entertainment, the odds of staying open are 0.259 times lower than the odds of staying open for restaurants with entertainment.

Good_for_dinner 0.518

For restaurants that are not good for dinner, the odds of staying open are 0.518 times lower than the odds of staying open for restaurants that are good for dinner.

Evaluation

Model	False Negative	True Negative	False Positive	True Positive
HP SVM Poly	1117	3526	2194	10816
Selection Tree with Neural Network	1147	3521	2199	10786
Stepwise Regression with Neural Network	1147	3521	2199	10786
Neural Network	1147	3521	2199	10786
LASSO with Neural Network	1265	3624	2096	10668
LARS with Neural Network	1265	3624	2096	10668
Adaptive LASSO with Neural Network	1238	3591	2129	10695
Variable selection with Auto Neural (AOV 16)	1368	3679	2041	10565
Variable selection with Neural Network (AOV 16)	1294	3603	2117	10639
HP Forest larger	877	3178	2542	11056
Variable Selection with Regression AOV16	1298	3589	2131	10635
Stepwise Misclassification Regression	1332	3523	2197	10601
Decision Tree	1717	3460	2260	10216

Evaluation

Model	Validation Misclassification Rate	ROC Index	Accuracy (%)	Sensitivity (%)	Specificity (%)
Selection Tree with HP SVM Poly	0.18756	0.863	81.24	90.64	61.64
Selection Tree with Neural Network	0.189543	0.863	81.05	90.39	61.56
Stepwise Regression with Neural Network	0.189543	0.863	81.05	90.39	61.56
Neural Network	0.189543	0.863	81.05	90.39	61.56
LASSO with Neural Network	0.190393	0.861	80.96	89.40	63.36
LARS with Neural Network	0.190393	0.861	80.96	89.40	63.36
Adaptive LASSO with Neural Network	0.190732	0.863	80.93	89.63	62.78
Variable selection with Auto Neural (AOV 16)	0.193112	0.859	80.69	88.54	64.32
Variable selection with Neural Network (AOV 16)	0.193225	0.861	80.68	89.16	62.99
HP Forest Larger	0.193678	0.857	80.63	92.65	55.56
Variable Selection with Regression AOV16	0.194245	0.86	80.58	89.12	62.74
Stepwise Misclassification Regression	0.199909	0.849	80.01	88.84	61.59
Decision Tree	0.225287	0.799	77.47	85.61	60.49

The Champion Model

Selection Tree with High Performance SVM Poly

Number of Support Vectors (Train)	7363
Number of Support Vectors on Margin (Train)	6919
Accuracy	0.830897
Sensitivity	0.920647
Specificity	0.643582
Validation Misclassification rate	0.18756



False Negative



True Negative



False Positive



True Positive

The Champion Model

Variable Importance Plot of the Selection Tree

Variable Name	Number of Splitting Rules	Number of Surrogate Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
LOG_review_ count	9.0	5.0	1.0	1.0	1.0
LOG_chain_ Counts	5.0	7.0	0.8642	0.9686	1.1206
Entertainment	1.0	0.0	0.8302	0.9131	1.0998
ls_Chain	1.0	2.0	0.8075	0.9188	1.1378
good_for_dinner	3.0	1.0	0.7793	0.9199	1.1803
Price_Range	3.0	5.0	0.7051	0.8622	1.2229
good_for_lunch	2.0	1.0	0.6421	0.8527	1.3278
wheelchair_access	4.0	3.0	0.5835	0.6892	1.1811
Credit_card	0.0	4.0	0.5387	0.6186	1.1483
table_service	2.0	4.0	0.4832	0.5650	1.1693



Conclusion

The best model for predicting restaurant closure is Selection Tree with High Performance Support Vector Machine Poly.

Based on the Selection Tree Variable Importance Plot (VIP), the **5 most importance variables** for predicting restaurant closure include:

Review Counts (the number of online reviews that each business has received)

Chain Counts (count of all restaurants that are a part of the same franchise and have the same name)

Entertainment (Indicating if the restaurant provides entertainment or not)

Is_Chain (Indicating if the restaurant is a chain or not) **Good_for_dinner** (Indicating if the restaurant is good for dinner or not)



Recommendations

- ❖ Independent restaurants at risk of closing can expand their business in order to grow in different locations or join a franchise to avoid closure. If this is not feasible, restaurants can also adopt the following recommendations.
- ❖ Restaurants should provide entertainment such as background music, live music and TV (DiPietro, 2016) or improve the existing ones.
- ❖ Promotions could be put in place to encourage customers to leave reviews on yelp after visiting the restaurant. ("The restaurant owner who asked for 1-star Yelp reviews", 2020).
- ❖ Restaurants that experience low traffic at dinner time and are considered "not good for dinner" can tailor their menu options to best suit customer needs.



References

- Ames, M., and G. Berle. n.d. What are the major reasons for small business failure? US Small Business Administration. http://www.sba.gov/content/what-are-major-reasons-small business-failure (accessed October 2, 2011).
- DiPietro, R. (2016). Restaurant and foodservice research. International Journal of Contemporary Hospitality

 Management, 29(4), 1203 1234.doi:http://dx.doi.org.vortex3.uco.edu/10.1108/IJCHM-01-2016-0046
- Gagić, S., Tešanović, D., & Jovičić, A. [2013]. The vital components of restaurant quality that affect guest satisfaction.

 Turizam, 17(4), 166-176
- McDonnell, S., 2020. What Percentage of Sales Are from Drive Through Windows at Fast Food Restaurants? [online]

 Small Business Chron.com. Available at: https://smallbusiness.chron.com/percentage-sales-drive-through-windows-fast-food-restaurants-75713.html

References

- McLynn, K. (2018, August 22). Total US Restaurant Count Stands At 660,755 in Spring 2018. Retrieved March 21,2020, from https://www.npd.com/wps/portal/npd/us/news/press_releases/2018/total-us-restaurant-count-at-660755-in-spring-2018-a-one-percent-drop_from-last-year-reports-npd/
- Parsa, H. G., Self, J. T., Njite, D., & King, T. (2005). Why restaurants fail. Cornell Hotel and Restaurant Administration Quarterly, 46(3), 304-322. Retrieved from http://vortex3.uco.edu/ login?url=https://search-proquestcom.vortex3.uco.edu/docview/209705978?accountid=14516
- Severson, K., & Yaffe-Bellany, D. (2020). Independent Restaurants Brace for the Unknown. Nytimes.com. Retrieved 16 July 2020, from https://www.nytimes.com/2020/03/20/dining/local-restaurants-coronavirus.html.
- Wrulich, M., Hirschberg, C., Rajko, A., & Schumacher, T. (2020). McKinsey. McKinsey&Company. Retrieved from https://www.mckinsey.com/industries/technology-media-and telecommunications/our-insights/the-changing-market-for-food-delivery#

Question and Answers





Thank You!