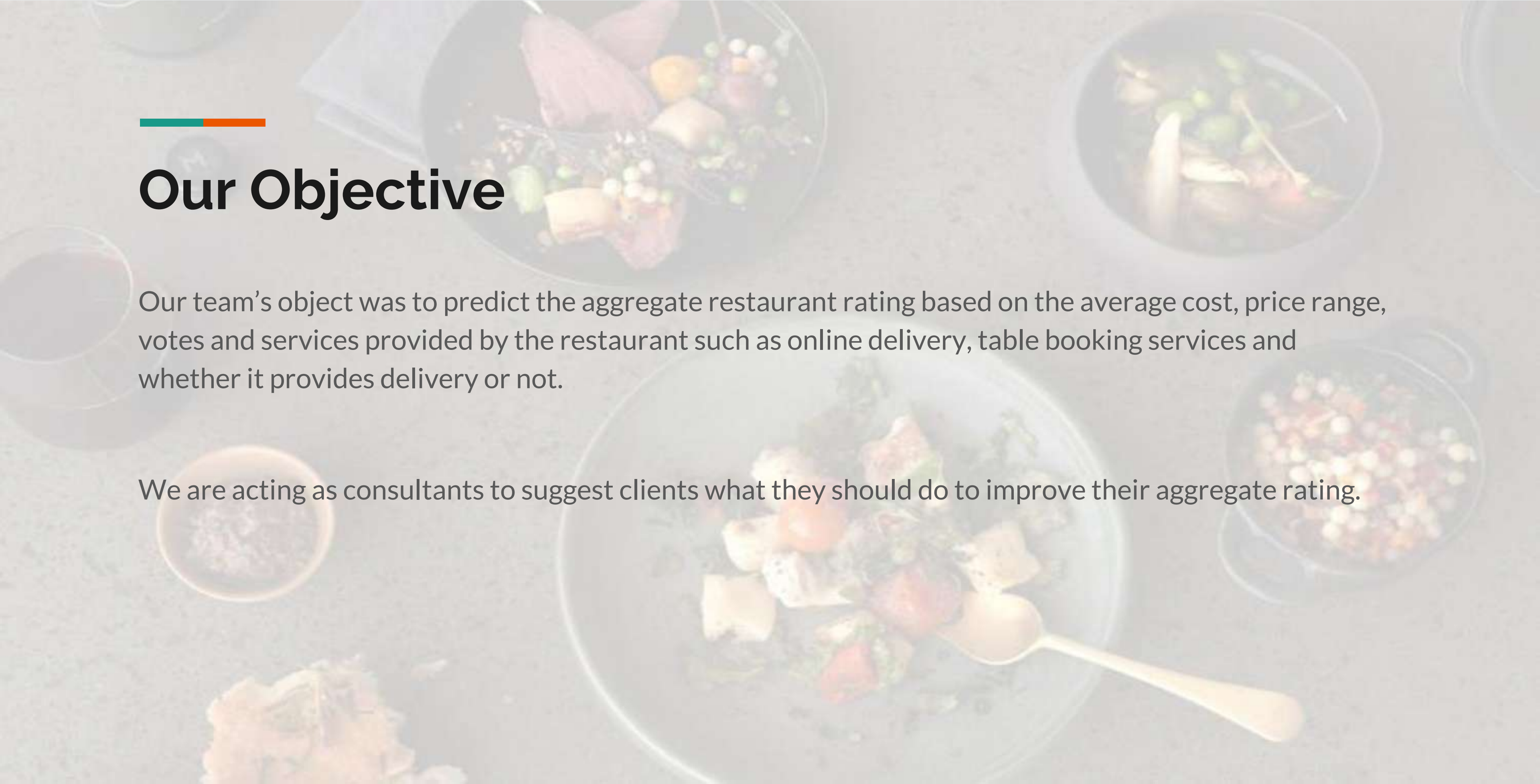
Predicting Restaurant Rating on Zomato.

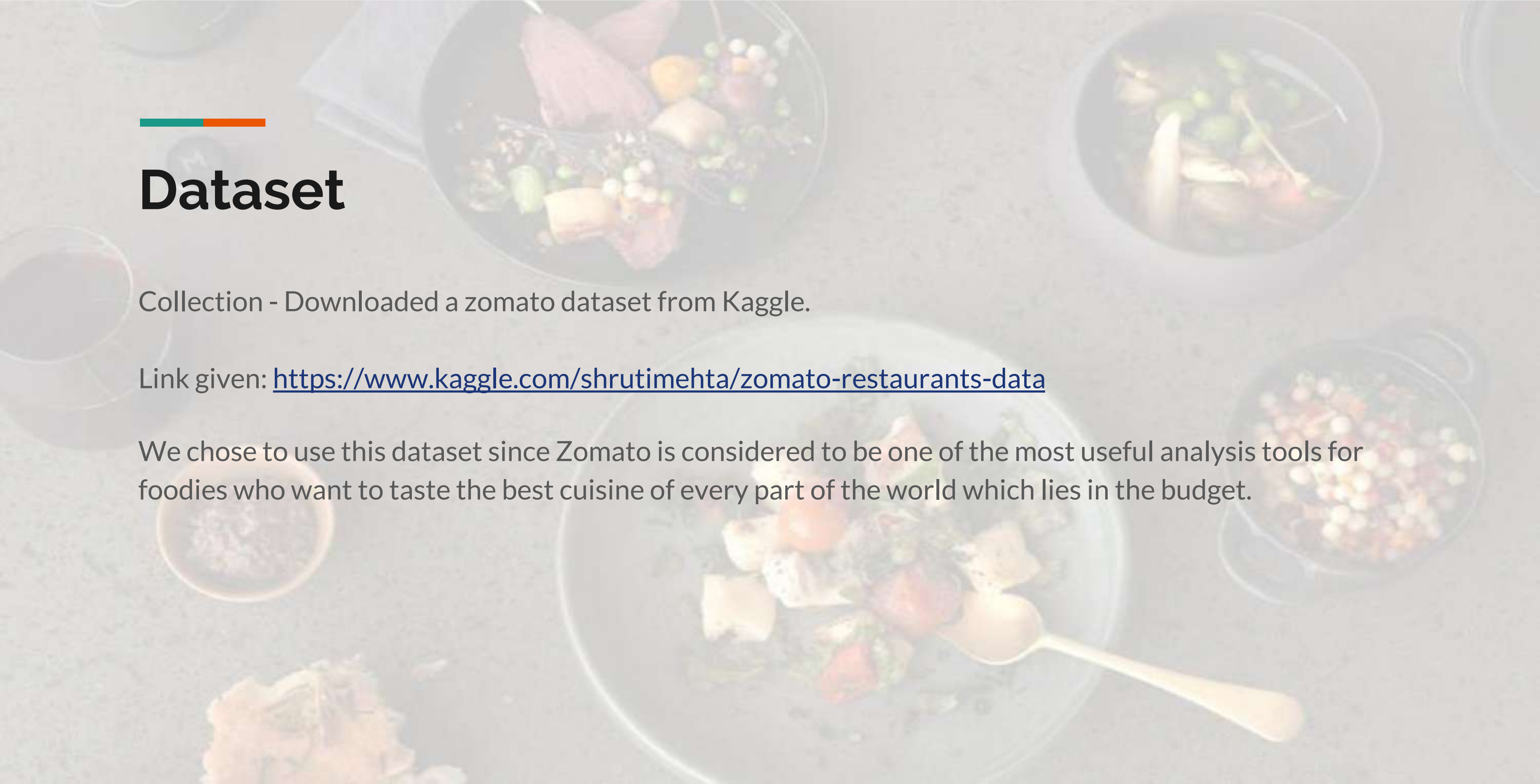
Group 21
William Lombardi
Pranjali Nag
Rohan Mahajan
Shanshan Ma



Our Objective

Our team’s object was to predict the aggregate restaurant rating based on the average cost, price range, votes and services provided by the restaurant such as online delivery, table booking services and whether it provides delivery or not.

We are acting as consultants to suggest clients what they should do to improve their aggregate rating.



Dataset

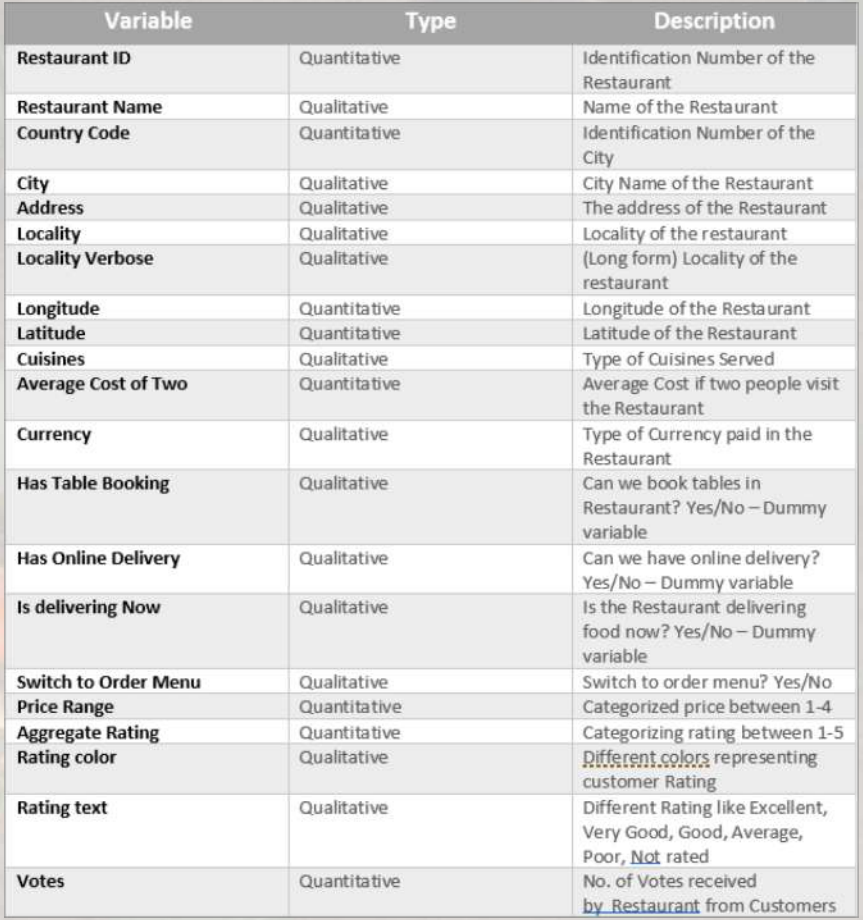
Collection - Downloaded a zomato dataset from Kaggle.

Link given: https://www.kaggle.com/shrutimehta/zomato-restaurants-data

We chose to use this dataset since Zomato is considered to be one of the most useful analysis tools for foodies who want to taste the best cuisine of every part of the world which lies in the budget.

Dataset

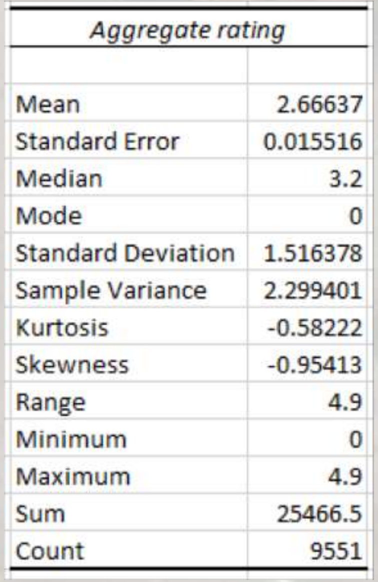
9552 rows and 21 variable columns of data.



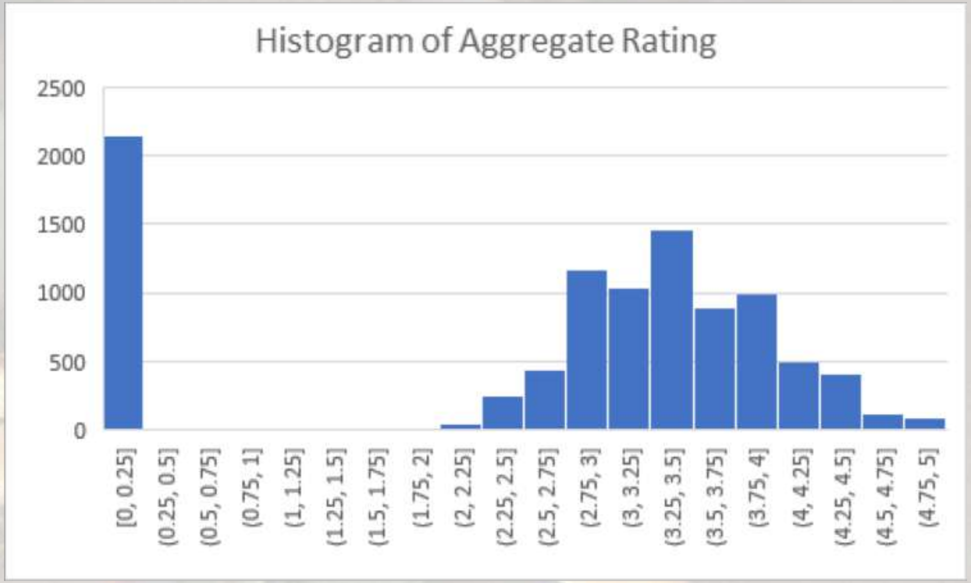
Variable	Type	Description
Restaurant ID	Quantitative	Identification Number of the Restaurant
Restaurant Name	Qualitative	Name of the Restaurant
Country Code	Quantitative	Identification Number of the City
City	Qualitative	City Name of the Restaurant
Address	Qualitative	The address of the Restaurant
Locality	Qualitative	Locality of the restaurant
Locality Verbose	Qualitative	(Long form) Locality of the restaurant
Longitude	Quantitative	Longitude of the Restaurant
Latitude	Quantitative	Latitude of the Restaurant
Cuisines	Qualitative	Type of Cuisines Served
Average Cost of Two	Quantitative	Average Cost if two people visit the Restaurant
Currency	Qualitative	Type of Currency paid in the Restaurant
Has Table Booking	Qualitative	Can we book tables in Restaurant? Yes/No – Dummy variable
Has Online Delivery	Qualitative	Can we have online delivery? Yes/No – Dummy variable
Is delivering Now	Qualitative	Is the Restaurant delivering food now? Yes/No – Dummy variable
Switch to Order Menu	Qualitative	Switch to order menu? Yes/No
Price Range	Quantitative	Categorized price between 1-4
Aggregate Rating	Quantitative	Categorizing rating between 1-5
Rating color	Qualitative	Different colors representing customer Rating
Rating text	Qualitative	Different Rating like Excellent, Very Good, Good, Average, Poor, Not rated
Votes	Quantitative	No. of Votes received by Restaurant from Customers

Descriptive Statistics - Aggregate Rating

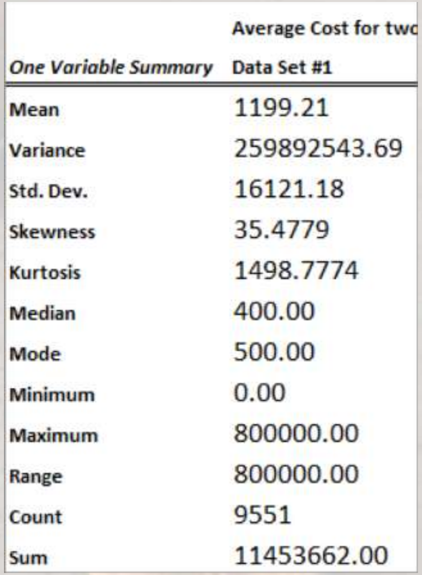
Our data is left skewed.



Aggregate rating	
Mean	2.66637
Standard Error	0.015516
Median	3.2
Mode	0
Standard Deviation	1.516378
Sample Variance	2.299401
Kurtosis	-0.58222
Skewness	-0.95413
Range	4.9
Minimum	0
Maximum	4.9
Sum	25466.5
Count	9551



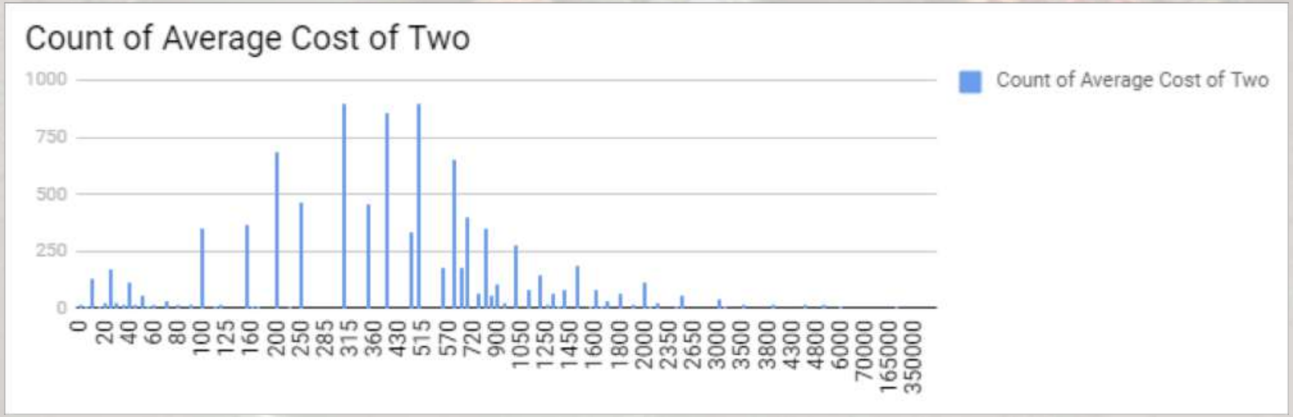
Descriptive Statistics - Average Cost of Two



One Variable Summary	Average Cost for two Data Set #1
Mean	1199.21
Variance	259892543.69
Std. Dev.	16121.18
Skewness	35.4779
Kurtosis	1498.7774
Median	400.00
Mode	500.00
Minimum	0.00
Maximum	800000.00
Range	800000.00
Count	9551
Sum	11453662.00

The mode of the **“Average Cost of Two”** was 500, which was about 9.42% of the data.

Please note that is an interesting variable since there is all different currencies that have been put in play.



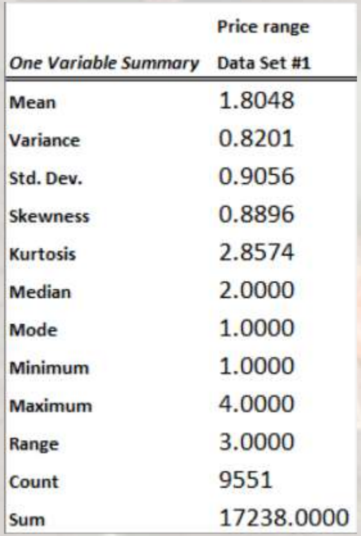
Descriptive Statistics - Price Range

About **46.43%** of our dataset has a price range of 1.

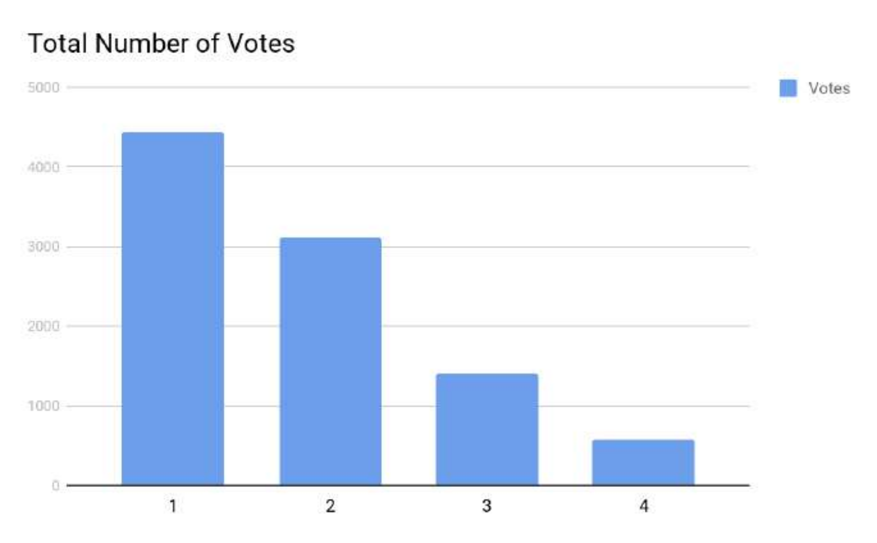
About **32.59%** of our dataset has a price range of 2.

About **14.74%** of our dataset has a price range of 3.

About **6.14%** of our dataset has a price range of 4.



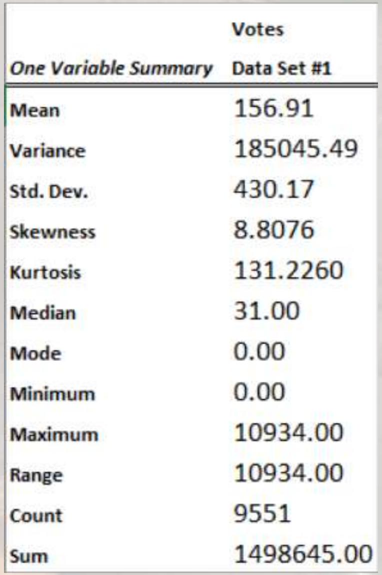
	Price range
One Variable Summary	Data Set #1
Mean	1.8048
Variance	0.8201
Std. Dev.	0.9056
Skewness	0.8896
Kurtosis	2.8574
Median	2.0000
Mode	1.0000
Minimum	1.0000
Maximum	4.0000
Range	3.0000
Count	9551
Sum	17238.0000



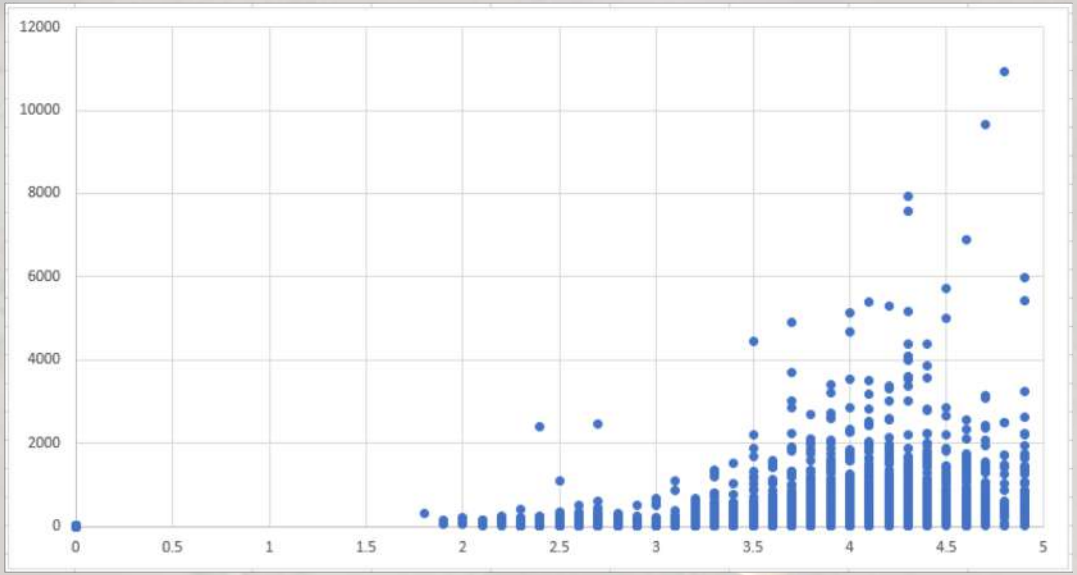
Descriptive Statistics - Votes

There were 1,498,645 votes in total.

Most of the data has 0 votes.

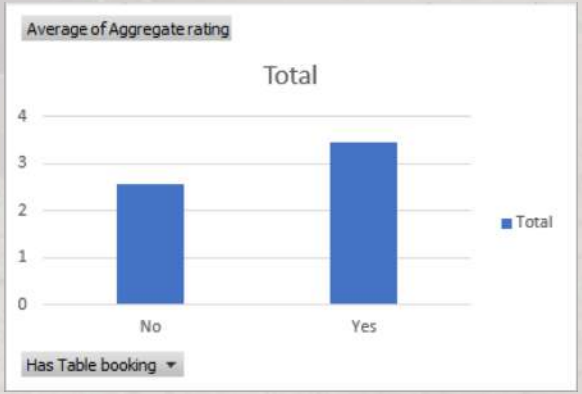


	Votes
One Variable Summary	Data Set #1
Mean	156.91
Variance	185045.49
Std. Dev.	430.17
Skewness	8.8076
Kurtosis	131.2260
Median	31.00
Mode	0.00
Minimum	0.00
Maximum	10934.00
Range	10934.00
Count	9551
Sum	1498645.00

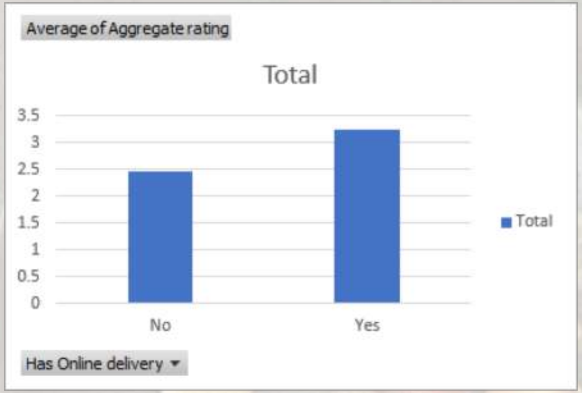


Average of Aggregate Rating

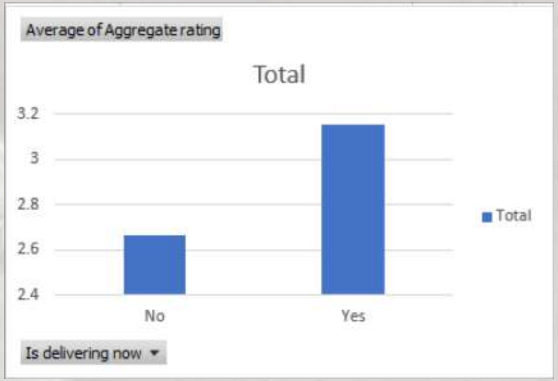
Compared with “Has table booking”, “Has online delivery”, & “Is delivering now”.



Restaurants with **Table booking** has a higher ***Aggregate Rating.***



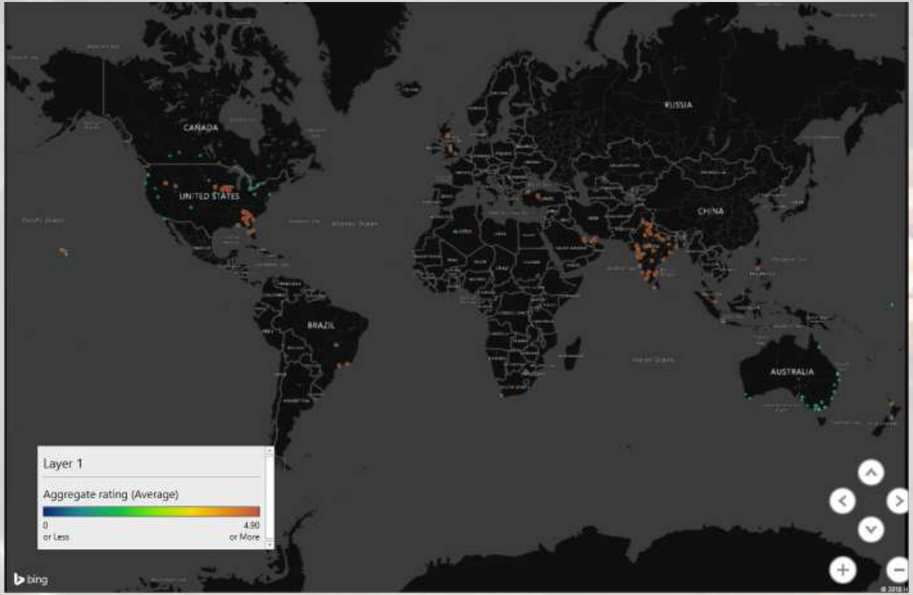
Restaurants with **Online Delivery** has a higher ***Aggregate Rating.***



Restaurants that are **delivering now** has a higher ***Aggregate Rating.***

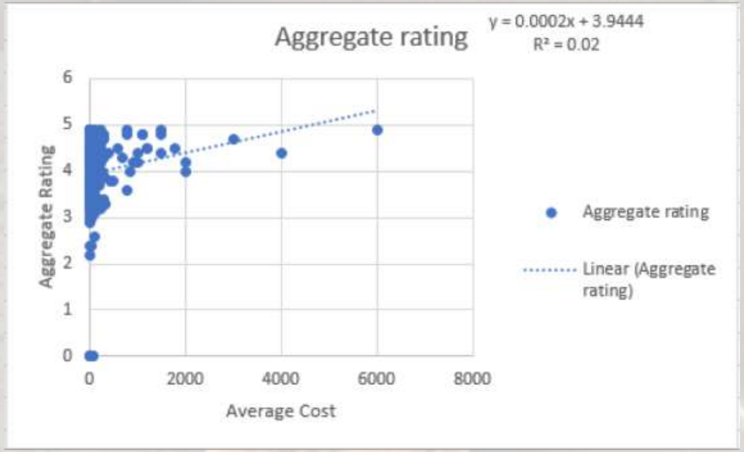
Locality of the Data - Heat Map

Our data is located over the world. But mostly located in the United States and India.

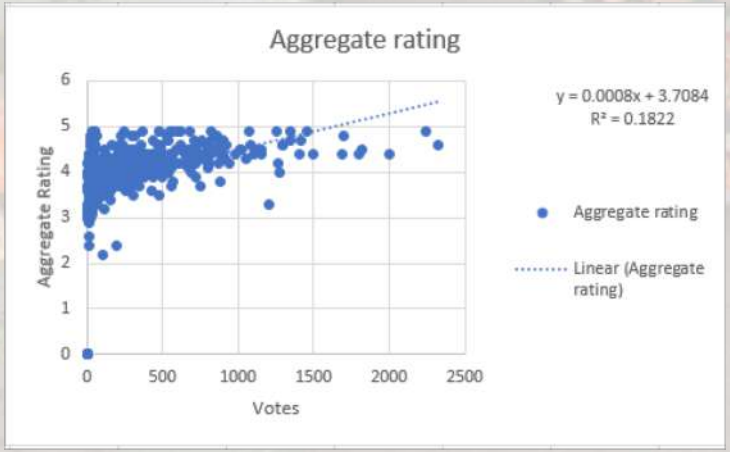


Single Regression Analysis

Positive Relationship between the Average Cost of Two and Aggregate Rating.

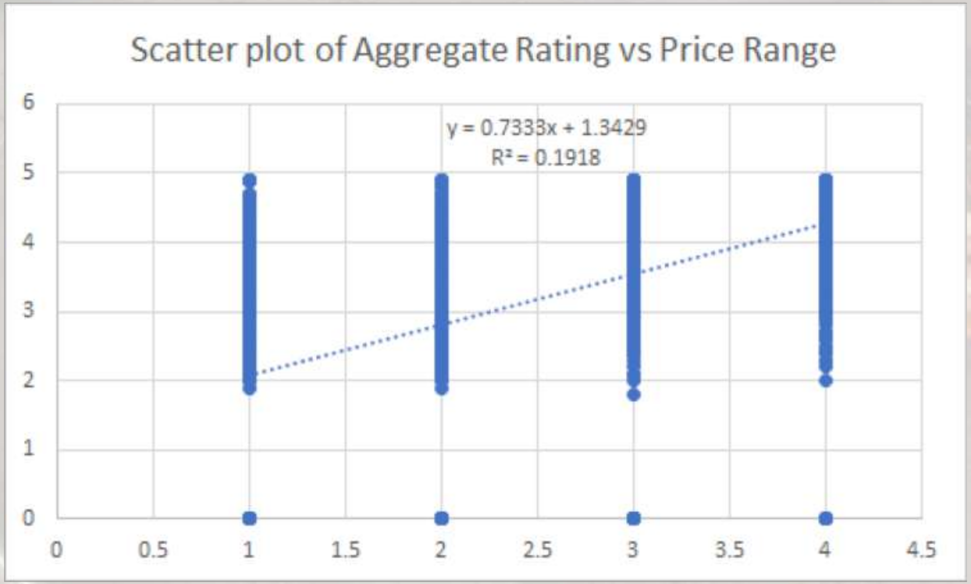


Positive Relationship between Votes and Aggregate Rating.

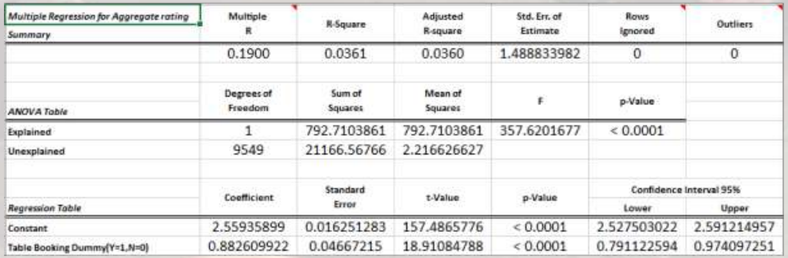


Single Regression Analysis

Positive Relationship between the Price Range and Aggregate Rating.



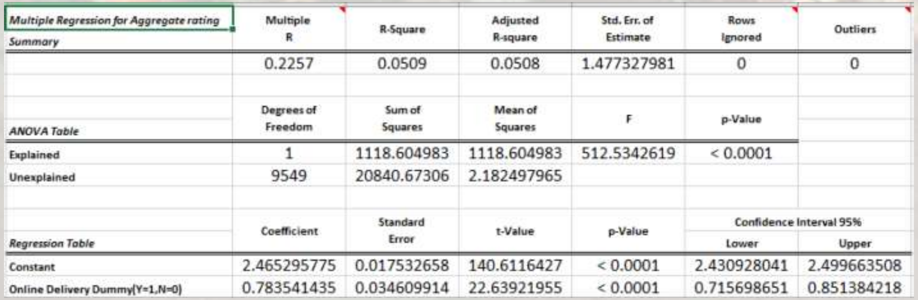
Dummy Variables.



Multiple Regression for Aggregate rating *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.1900	0.0361	0.0360	1.488833982	0	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	1	792.7103861	792.7103861	357.6201677	< 0.0001	
Unexplained	9549	21166.56766	2.216626627			
		Standard			Confidence Interval 95%	
Regression Table	Coefficient	Error	t-Value	p-Value	Lower	Upper
Constant	2.55935899	0.016251283	157.4865776	< 0.0001	2.527503022	2.591214957
Table Booking Dummy(Y=1,N=0)	0.882609922	0.04667215	18.91084788	< 0.0001	0.791122594	0.974097251

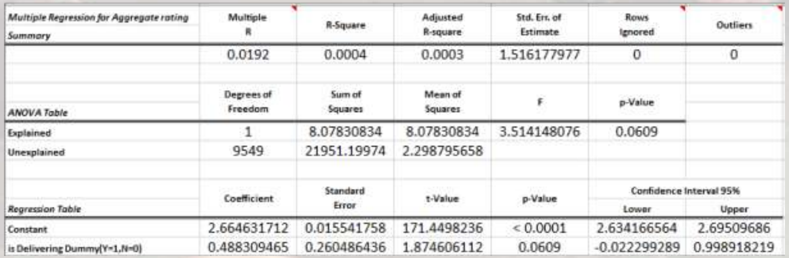
The table booking dummy variable gave us a Adjusted R^2 Value of 0.031, which is low. This make senses because if a place has table booking, there won't be much of an effect in terms of rating the entire restaurant.

The online delivery booking dummy variable gave us a Adjusted R^2 Value of 0.050, which is low. This make senses because a place that offers online delivery encourages people to order food.



Multiple Regression for Aggregate rating *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.2257	0.0509	0.0508	1.477327981	0	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	1	1118.604983	1118.604983	512.5342619	< 0.0001	
Unexplained	9549	20840.67306	2.182497965			
		Standard			Confidence Interval 95%	
Regression Table	Coefficient	Error	t-Value	p-Value	Lower	Upper
Constant	2.465295775	0.017532658	140.6116427	< 0.0001	2.430928041	2.499663508
Online Delivery Dummy(Y=1,N=0)	0.783541435	0.034609914	22.63921955	< 0.0001	0.715698651	0.851384218

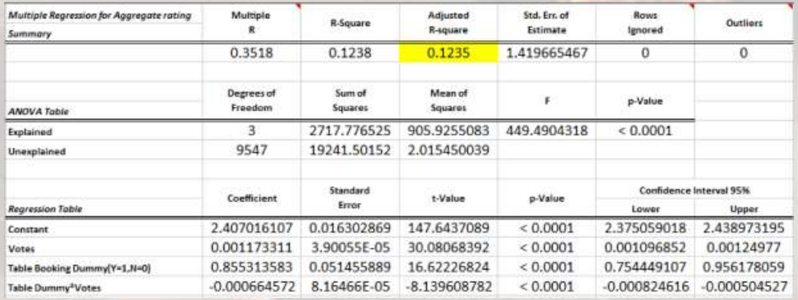
Dummy Variables.



Multiple Regression for Aggregate rating *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.0192	0.0004	0.0003	1.516177977	0	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	1	8.07830834	8.07830834	3.514148076	0.0609	
Unexplained	9549	21951.19974	2.298795658			
					Confidence Interval 95%	
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Lower	Upper
Constant	2.664631712	0.015541758	171.4498236	< 0.0001	2.634166564	2.69509686
is Delivering Dummy(Y=1,N=0)	0.488309465	0.260486436	1.874606112	0.0609	-0.022299289	0.998918219

The online delivery booking dummy variable gave us a Adjusted R$^{\wedge 2}$ Value of 0.003, which is the lowest. This makes sense because this variable is not clear of what it means.

Interaction Variables



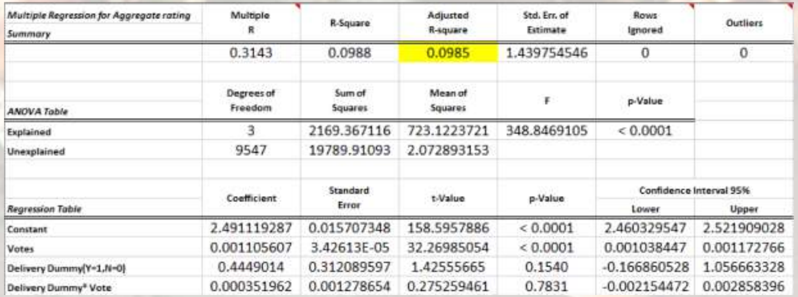
Multiple Regression for Aggregate rating *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.3518	0.1238	0.1235	1.419665467	0	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	3	2717.776525	905.9255083	449.4904318	< 0.0001	
Unexplained	9547	19241.50152	2.015450039			
					Confidence Interval 95%	
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Lower	Upper
Constant	2.407016107	0.016302869	147.6437089	< 0.0001	2.375059018	2.438973195
Votes	0.001173311	3.90055E-05	30.08068392	< 0.0001	0.001096852	0.00124977
Table Booking Dummy(Y=1,N=0)	0.855313583	0.051455889	16.62226824	< 0.0001	0.754449107	0.956178059
Table Dummy*Votes	-0.000664572	8.16466E-05	-8.139608782	< 0.0001	-0.000824616	-0.000504527

The Adjusted R-square is 0.1235, which is much higher than just the dummy variable. So it's better explained.

Regression equation: Predicted Aggregate Rating = 2.407+0.001*Votes + 0.855* Dummy - 0.0007Dummy*Vote.

For online delivery, the Adjusted R-square is 0.0985, which is better explained.

Regression equation: Predicted Aggregate Rating = 2.491+0.001*Votes + 0.445* Dummy +0.0004Dummy*Vote.

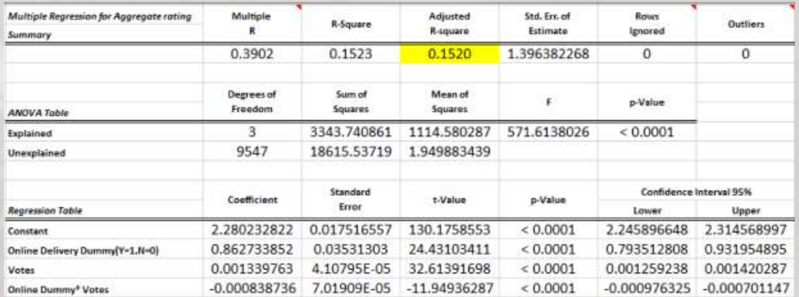


Multiple Regression for Aggregate rating *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.3143	0.0988	0.0985	1.439754546	0	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	3	2169.367116	723.1223721	348.8469105	< 0.0001	
Unexplained	9547	19789.91093	2.072893153			
					Confidence Interval 95%	
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Lower	Upper
Constant	2.491119287	0.015707348	158.5957886	< 0.0001	2.460329547	2.521909028
Votes	0.001105607	3.42613E-05	32.26985054	< 0.0001	0.001038447	0.001172766
Delivery Dummy(Y=1,N=0)	0.4449014	0.312089597	1.42555665	0.1540	-0.166860528	1.056663328
Delivery Dummy* Vote	0.000351962	0.001278654	0.275259461	0.7831	-0.002154472	0.002858396

Interaction Variables

The Adjusted R-square is 0.1520, which is much higher than the dummy variable of Table Booking.

Regression equation: Predicted Aggregate Rating = 2.280+0.001*Votes + 0.863* Dummy - 0.0008Dummy*Vote.



Multiple Regression for Aggregate rating Summary	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.3902	0.1523	0.1520	1.396382268	0	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	3	3343.740861	1114.580287	571.6138026	< 0.0001	
Unexplained	9547	18615.53719	1.949883439			
Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95%	
					Lower	Upper
Constant	2.280232822	0.017516557	130.1758553	< 0.0001	2.245896648	2.314568997
Online Delivery Dummy(Y=1,N=0)	0.862733852	0.03531303	24.43103411	< 0.0001	0.793512808	0.931954895
Votes	0.001339763	4.10795E-05	32.61391698	< 0.0001	0.001259238	0.001420287
Online Dummy* Votes	-0.000838736	7.01909E-05	-11.94936287	< 0.0001	-0.000976325	-0.000701147

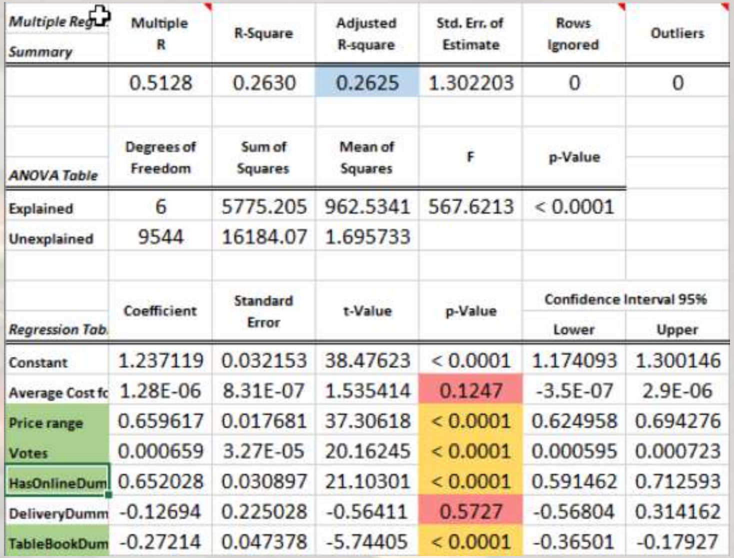
Multiple Regression Model

The regression model predicts that Price Range, Votes, Online Delivery and Table Booking are statistically significant which make them good predictors as their p-values are low.

Our regression is being explained by 6 variables. The adjusted R-square is 0.2625. For a large dataset, this is a good prediction model.

Regression equation:

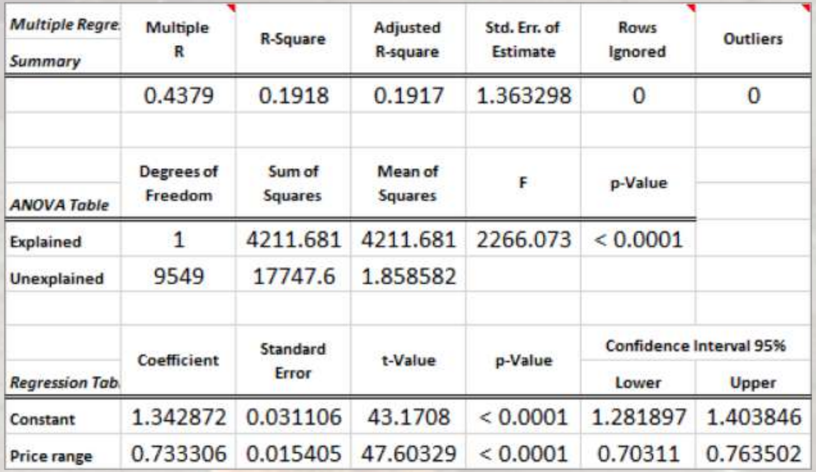
$$Y = 1.237119 + (1.28E{-}06)\mathbf{X1} + (0.659617)\mathbf{X2} + (0.00659)\mathbf{X3} + (0.652028)\mathbf{X4} + (-0.12694)\mathbf{X5} + (-0.27214)\mathbf{X6}$$



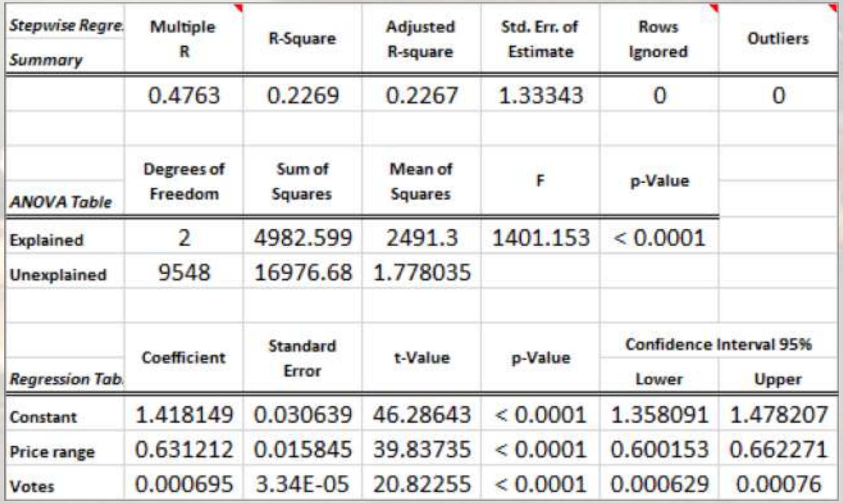
Multiple Regr. *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.5128	0.2630	0.2625	1.302203	0	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	6	5775.205	962.5341	567.6213	< 0.0001	
Unexplained	9544	16184.07	1.695733			
					Confidence Interval 95%	
Regression Tab.	Coefficient	Standard Error	t-Value	p-Value	Lower	Upper
Constant	1.237119	0.032153	38.47623	< 0.0001	1.174093	1.300146
Average Cost fc	1.28E-06	8.31E-07	1.535414	0.1247	-3.5E-07	2.9E-06
Price range	0.659617	0.017681	37.30618	< 0.0001	0.624958	0.694276
Votes	0.000659	3.27E-05	20.16245	< 0.0001	0.000595	0.000723
HasOnlineDum	0.652028	0.030897	21.10301	< 0.0001	0.591462	0.712593
DeliveryDumm	-0.12694	0.225028	-0.56411	0.5727	-0.56804	0.314162
TableBookDum	-0.27214	0.047378	-5.74405	< 0.0001	-0.36501	-0.17927

Stepwise Regression

To make our prediction more accurate, we decided to use Stepwise Regression.



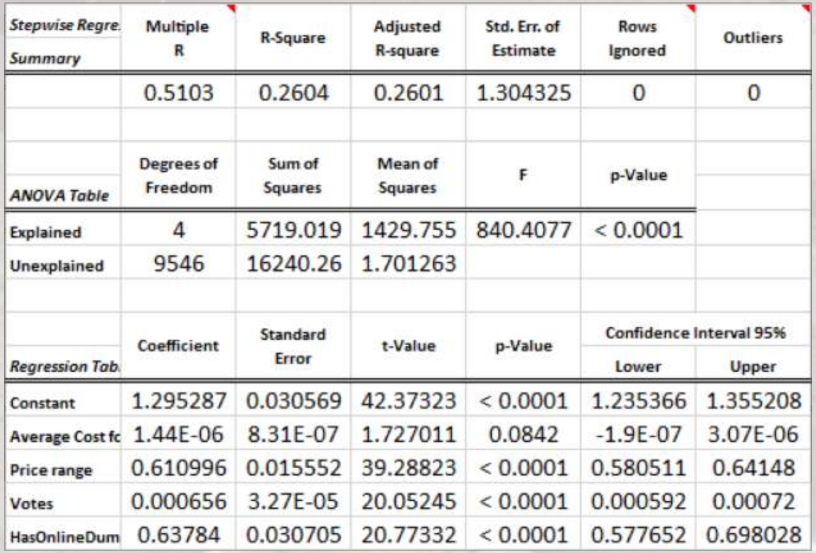
Multiple Regre. *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.4379	0.1918	0.1917	1.363298	0	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	1	4211.681	4211.681	2266.073	< 0.0001	
Unexplained	9549	17747.6	1.858582			
					Confidence Interval 95%	
Regression Tab.	Coefficient	Standard Error	t-Value	p-Value	Lower	Upper
Constant	1.342872	0.031106	43.1708	< 0.0001	1.281897	1.403846
Price range	0.733306	0.015405	47.60329	< 0.0001	0.70311	0.763502



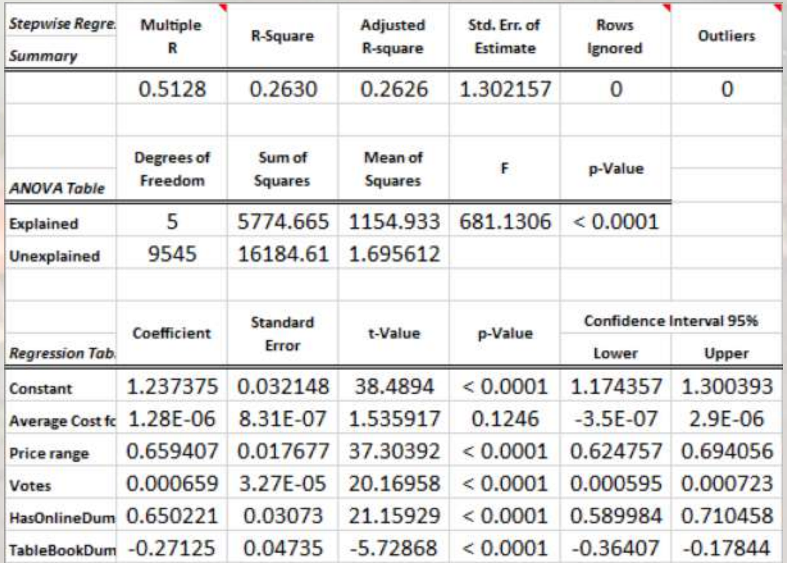
Stepwise Regre. *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.4763	0.2269	0.2267	1.33343	0	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	2	4982.599	2491.3	1401.153	< 0.0001	
Unexplained	9548	16976.68	1.778035			
					Confidence Interval 95%	
Regression Tab.	Coefficient	Standard Error	t-Value	p-Value	Lower	Upper
Constant	1.418149	0.030639	46.28643	< 0.0001	1.358091	1.478207
Price range	0.631212	0.015845	39.83735	< 0.0001	0.600153	0.662271
Votes	0.000695	3.34E-05	20.82255	< 0.0001	0.000629	0.00076

Stepwise Regression

To make our prediction more accurate, we decided to use Stepwise Regression.



Stepwise Regre. *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.5103	0.2604	0.2601	1.304325	0	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	4	5719.019	1429.755	840.4077	< 0.0001	
Unexplained	9546	16240.26	1.701263			
	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95%	
Regression Tab.					Lower	Upper
Constant	1.295287	0.030569	42.37323	< 0.0001	1.235366	1.355208
Average Cost fc	1.44E-06	8.31E-07	1.727011	0.0842	-1.9E-07	3.07E-06
Price range	0.610996	0.015552	39.28823	< 0.0001	0.580511	0.64148
Votes	0.000656	3.27E-05	20.05245	< 0.0001	0.000592	0.00072
HasOnlineDum	0.63784	0.030705	20.77332	< 0.0001	0.577652	0.698028



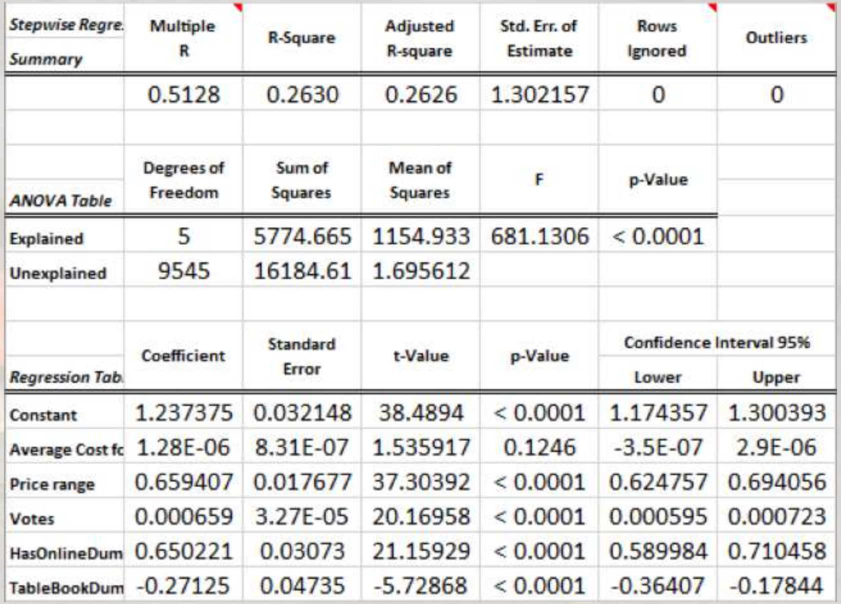
Stepwise Regre. *Summary*	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.5128	0.2630	0.2626	1.302157	0	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	5	5774.665	1154.933	681.1306	< 0.0001	
Unexplained	9545	16184.61	1.695612			
	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95%	
Regression Tab.					Lower	Upper
Constant	1.237375	0.032148	38.4894	< 0.0001	1.174357	1.300393
Average Cost fc	1.28E-06	8.31E-07	1.535917	0.1246	-3.5E-07	2.9E-06
Price range	0.659407	0.017677	37.30392	< 0.0001	0.624757	0.694056
Votes	0.000659	3.27E-05	20.16958	< 0.0001	0.000595	0.000723
HasOnlineDum	0.650221	0.03073	21.15929	< 0.0001	0.589984	0.710458
TableBookDum	-0.27125	0.04735	-5.72868	< 0.0001	-0.36407	-0.17844

Stepwise Regression

What we learn from this is that our prediction is at its best when Average Cost, Price range, votes and Delivery and table booking factors are considered.

Final Regression equation:

Y=1.237375 + (1.28E-06)**X1** + (0.659407)**X2** + (0.000659)**X3** + (0.650221)**X4** + (-0.27125)**X5**



Stepwise Regre. Summary	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate	Rows Ignored	Outliers
	0.5128	0.2630	0.2626	1.302157	0	0
ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value	
Explained	5	5774.665	1154.933	681.1306	< 0.0001	
Unexplained	9545	16184.61	1.695612			
	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95%	
Regression Tab.					Lower	Upper
Constant	1.237375	0.032148	38.4894	< 0.0001	1.174357	1.300393
Average Cost fc	1.28E-06	8.31E-07	1.535917	0.1246	-3.5E-07	2.9E-06
Price range	0.659407	0.017677	37.30392	< 0.0001	0.624757	0.694056
Votes	0.000659	3.27E-05	20.16958	< 0.0001	0.000595	0.000723
HasOnlineDum	0.650221	0.03073	21.15929	< 0.0001	0.589984	0.710458
TableBookDum	-0.27125	0.04735	-5.72868	< 0.0001	-0.36407	-0.17844

Actionable Insights

Insight 1	● Add online booking service.
Insight 2	● Add a delivery option.
Insight 3	● Focus on adding online deliveries and.

Who will use our analysis?

Restaurant Managers

Restaurant managers can use our prediction analysis to work on what could increase their ratings.

Delivery Services

As delivery is an important factor, delivery services can tie in with restaurants to work with them to increase their ratings.

Review Websites(Zomato)

Zomato can focus on these variables to highlight them on restaurant information and could improve their interface for users to access.

Reviewers

Restaurant reviewers and food bloggers can use this data to focus on what points to write in their reviews(Services like delivery.).