Assignment 2 – Which recipes receive higher user ratings?

Table of Contents

Introduction

Food.com is an online social networking service which has a massive collection of recipes that are submitted, rated and reviewed by people who are passionate about food. In this report, our aim is to investigate which recipes receive higher user ratings from the food.com data. By looking into the distribution and summary statistics of the associated variables of the given information, we are capable to filter the dataset and develop a step-by-step approach to find the solutions. We have also used two sample t-tests to investigate if there is any difference in overall average of rating scores by categorising all these related variables. Beside applying statistical analysis, thorough understanding on the relations of data, also help us to present and summarize findings on the solutions.

Methods

The data source of this report is originated from Food.com (Majumder et al 2019) (Reference <R1>) and is modified for the purpose of this assignment. There are 3 csv files:

1) "Recipes.csv**"** contains 231,575 rows of observations, each observation represents a recipe with unique recipe id, recipe name, minutes to prepare, submission date to food.com, contributor id, number of steps and ingredients for preparing the recipes.
2) "nutrition.csv" contains 231,637 rows of observations, each observation describes the nutrition values of a particular recipe.
3) "ratings.csv" contains 1,048,575 rows of observations, each observation represents an individual rating to a particular recipe. User id, recipe id, rating scores and date submitted from user are all being recorded in this file. Ratings are scored from 0 (least) to 5 (highest).

There are 2 additional SAS formatted data file.

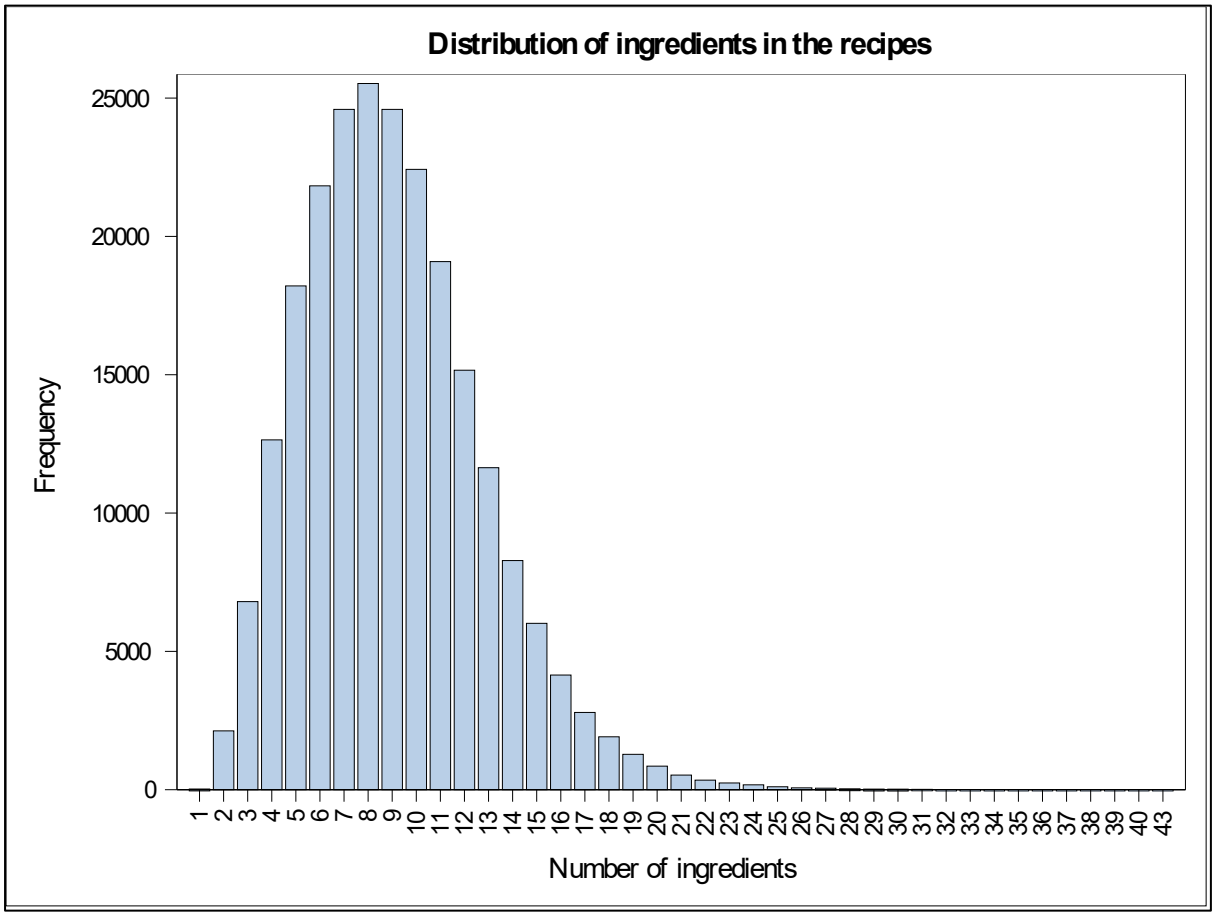
1) "ingredients" – list all the recipes and its associated ingredients id. For example, recipe id 139 have 7 ingredients, there would be 7 distinctive rows for recipe id 139 each with different ingredients id.
2) "ing_ref" – list all the ingredients name for each ingredients id.

SAS studio is used to load in all the csv and SAS formatted data files (SAS Code /* Task 1 */), SAS programming language is accordingly applied to the datasets to conduct this report.

Our ultimate goal is to find out which recipes receive high user ratings equitably. We need to look at factors influencing high user ratings and present an unbiased and informative list in the investigation.

Results

First, we look into the distribution of number of ingredients in the recipes of the data set. (SAS Code /* Task 2 */)



The average number of ingredients is 9.05 with a median of 9. We can categorise the ingredients' needs into 2 groups (above and below average) by rounding up the average to 9 for lateral use. (SAS Output <1>)

Consequentially, we also look into the variables "number of steps" in the recipes file. We worked out that the average number of steps is 9.766 with a similar median of 9. It's safe to divide all the recipes into 2 bins with the average of 10 to classify complexity. (SAS Code /*2b*/, SAS Output <2>)

Next, we look into the average rating given by each user in the dataset. (SAS Code /*2c*/ , SAS Output <3>) We have the following findings:

Number of Users	Average recipes reviewed per user	Minimum Value (i.e. 157,430 users have only reviewed on recipe)		Maximum Value (i.e. One user has reviewed 7,084 recipes)		Distribution
		Value	Occurrence	Value	Occurrence	
214,484	4.88	1	157,430	7,084	1	Heavily Positive skewed

Looking at the percentile, for user who has reviewed more than 5 recipes, would already fall beyond the top 10 percentile. There are 18,556 users who have reviewed more than 5 recipes with an overall average rating score of 4.455. (SAS Code /* Task 3 */, SAS Output <4>)

After knowing the distribution of reviews per user, conversely, we would also like to know the distribution of rating count per recipe. High rating count tends to give unbiased judgement on the recipe. (SAS Code /*3b*/, SAS Output <5>) Below is our findings:

Number of Recipes	Average rating count received per recipe	Minimum Value (i.e. 84,949 recipes have been rated once)		Maximum Value (i.e. One recipe has been rated for 1,613 times)		Distribution
		Value	Occurrence	Value	Occurrence	
214,217	4.895	1	84,949	1,613	1	Heavily Positive skewed

Recipe being rated for more than 15 times would fall beyond the top 5 percentile.

If we subset the data with average rating score higher than 4.455 and being rated for more than 15 times, we have a filtered dataset with high rating scores and being reviewed frequently. (SAS Code /*6a i*/) One step closer to finding our solutions.

Beside finding the recipe with higher user rating, we are also interested to know whether number of steps in the recipe has any inference in calories per serve. As we have already divided the complexity into 2 bins, where low is less than 10 and high complexity is 10 or more steps, we can perform 2 sample t-test to check our hypothesis. (SAS Code and SAS Code /* Task 4 */ and /* Task 5 */, SAS Output <6>)

2-sample Ttest of complexity on Calories Per Serve

The TTEST Procedure

Variable: CaloriesPS

Complexity	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
high		4070	525.9	929.7	14.5727	0	38680.1

Complexity	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
low		6155	419.5	845.0	10.7713	0	38662.3
Diff (1-2)	Pooled		106.4	879.7	17.7730		
Diff (1-2)	Satterthwaite		106.4		18.1213		

Complexity	Method	Mean	97% CL Mean		Std Dev	97% CL Std Dev	
High		525.9	494.3	557.5	929.7	907.8	952.6
Low		419.5	396.1	442.9	845.0	828.8	861.9
Diff (1-2)	Pooled	106.4	67.8093	145.0	879.7	866.6	893.3
Diff (1-2)	Satterthwaite	106.4	67.0519	145.7			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	10223	5.99	<.0001
Satterthwaite	Unequal	8125.9	5.87	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	4069	6154	1.21	<.0001

To perform 2-sample t-test, SAS would check the Homogeneity of Variance, as $p < 0.03$, it's not safe to assume equal variance.

$$H_0: \sigma^2_{high\_complexity} = \sigma^2_{low\_complexity}$$

$$H_A: \sigma^2_{high\_complexity} \neq \sigma^2_{low\_complexity}$$

We need to perform 2-sample t-test assuming unequal variance, result: $p < 0.03$, which means there is statistical significance in the mean of calories per serve between recipes with low and high complexity.

$$H_0: \mu_{high\_complexity} - \mu_{low\_complexity} = 0$$

$$H_A: \mu_{high\_complexity} - \mu_{low\_complexity} \neq 0$$

Similarly, we could divide all the associated variables from the mean or median into 2 groups and perform 2-sample t-test to check if there is any inference on the average rating score. (SAS Code /*6b i*/, SAS Output <7>) Results are summarized in the following table.

Class	Average or median used to subclass the variable	Hypothesis of Equal variance	P-Value of 2-sample T-test	Null Hypothesis for no difference in the average rating score between 2 groups
Complexity (number of steps) (SAS Code /* Task 4 */ and /*6b i*/, SAS Output <7a>)	10	Fail to Reject	0.0015	Reject
IngredientsNeed (number of ingredients) (SAS Code /*4a*/ and /*6b ii*/, SAS Output <7b>)	9	Fail to Reject	0.0004	Reject
submittedRange (submission date) (SAS Code /*6a ii*/ and /*6b iii */, SAS Output <7c>)	31/5/2004 (~Median)	Reject	<.0001	Reject
ratingFreq (number of rating counts) (SAS Code Task /*6a i*/ and /*6b iv*/, SAS Output <7d>)	44	Reject	<.0001	Reject

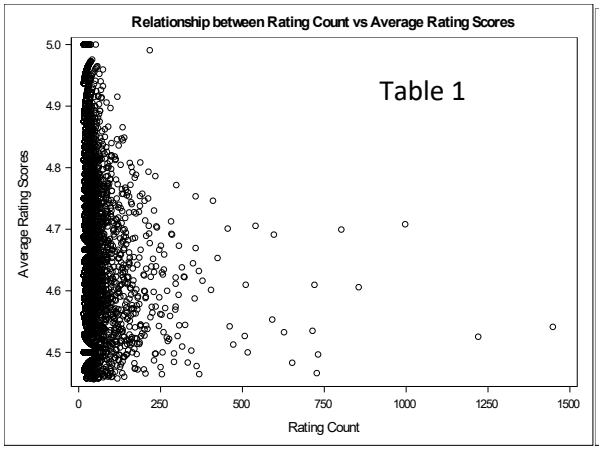
The results show there are statistical significant difference for all the variables on the average of the rating scores.

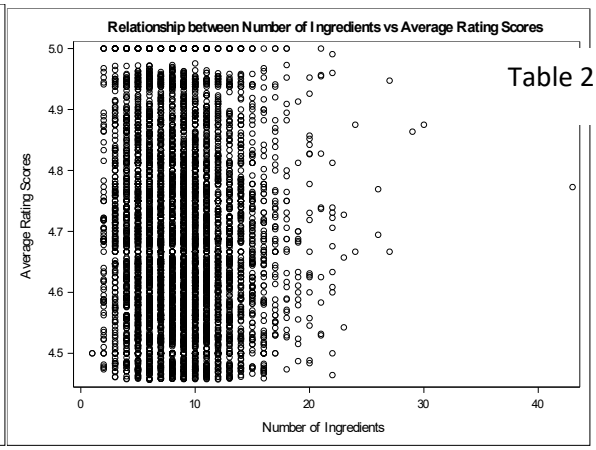
However, to find out recipe with high rating value, we also need to examine the distributions of extreme average rating scores in the datasets. (SAS Code /*6c*/)

The UNIVARIATE Procedure
Variable: AvgRatingScores

Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	4.45652	2	951	4.96667	2
2	4.45714	1	952	4.96774	2
3	4.45763	4	953	4.96875	1
4	4.45833	21	954	4.96970	1
5	4.45902	2	955	4.97143	4
6	4.45946	5	956	4.97222	2
7	4.46000	4	957	4.97297	2
8	4.46032	1	958	4.97561	1
9	4.46043	1	959	4.99078	1
10	4.46067	1	960	5.00000	156

There are 156 recipes with perfect rating (rating score = 5), it's best to select only a few distinctive recipes. Recipes other than perfect rating are also worthwhile to look at, getting perfect score is almost impossible when rating count increase, as the rating received is more likely to be diverged. Let's see whether we could do so by comparing how these 156 recipes distribute against different variables. The following scatter plots (SAS Code /*6h*/) show the relationship and we have summarised the findings in a table:





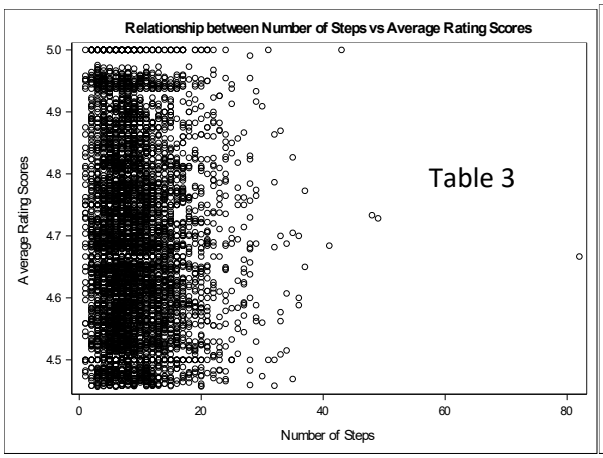


Table 4

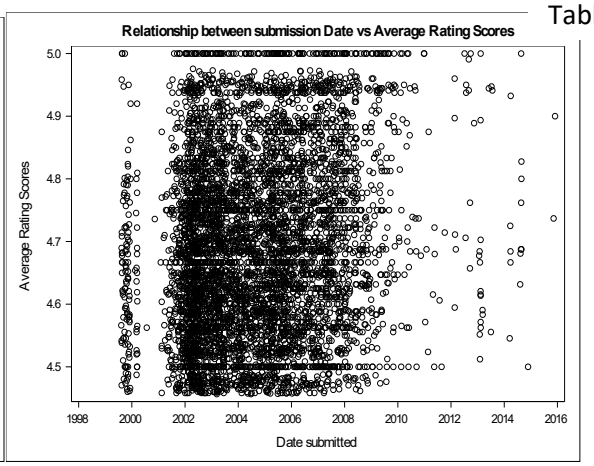
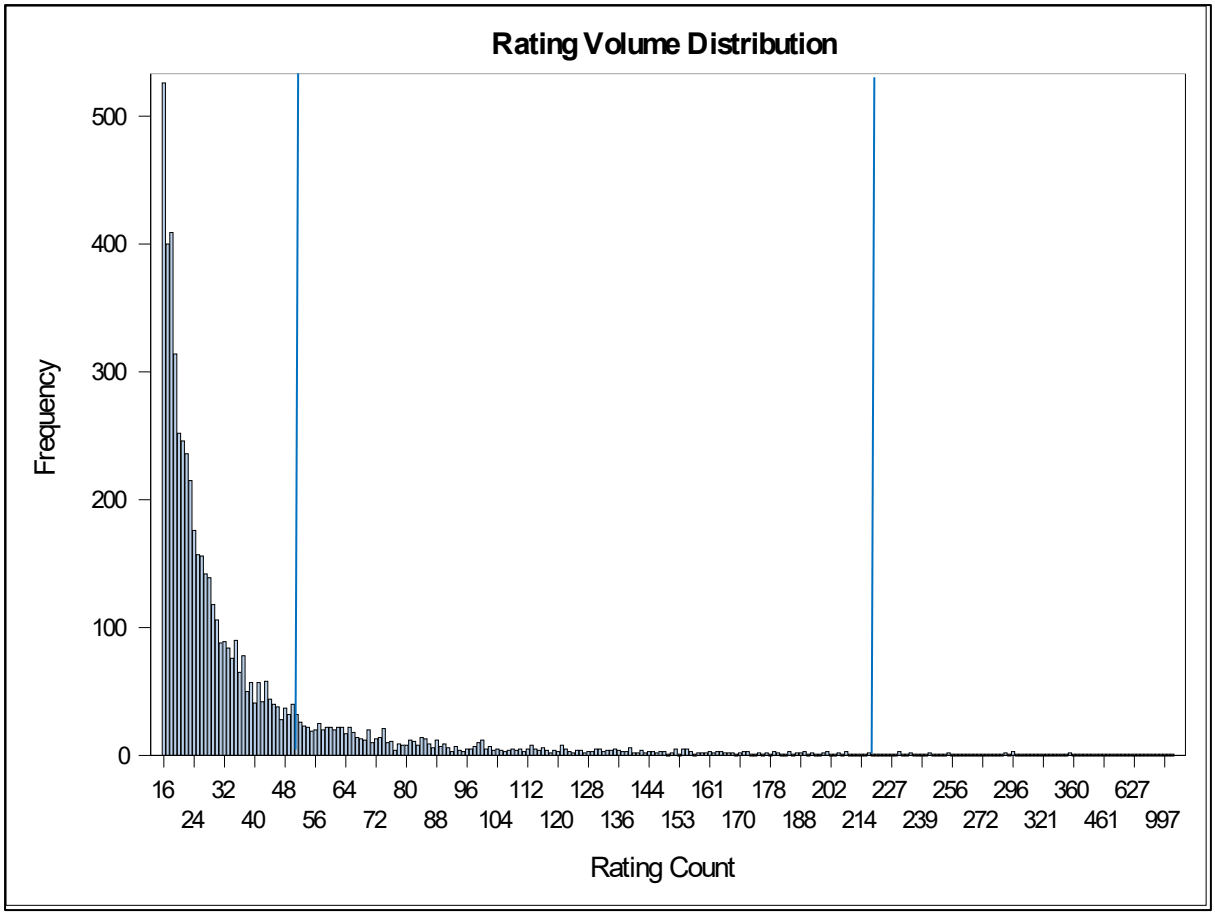


Table	Variables in X Axis	Variables in Y Axis	Maximum X-axis value for Rating Score =5
1	Rating Count	Average Rating Score	52
2	Number of Ingredients	Average Rating Score	21
3	Number of Steps	Average Rating Score	43
4	Submission Date	Average Rating Score	21/08/2014

From the scatterplots, they clearly show that perfect scores are widely spread in table 2, 3 and 4, but they only come up with low rating count in table 1. Also, we can observe that in table 1, rating scores tend to decrease when rating count increases, but for other variables, there seems to be no associations at all.

Using SAS proc SQL procedures, if we filter the dataset with rating count > 52, we find that the next extreme value 4.99078, will occur with a recipe having a rating count of 217.

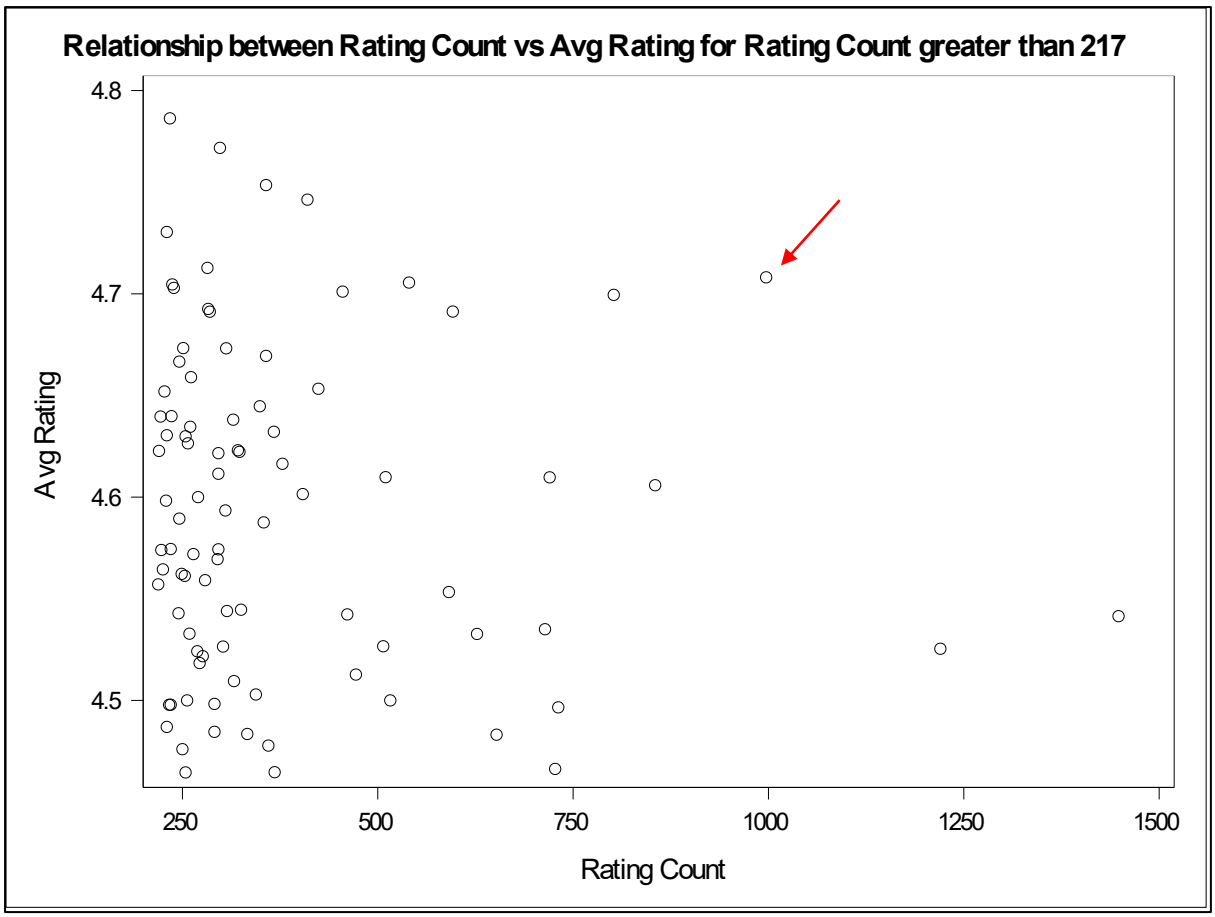
We could divide the dataset into 3 groups as below (SAS Code /*6d*/ and /*6e*/):



We then list out the most distinctive recipe with high user rating in each group as follows:

Group	Recipe Name	Rating Count	Average Rating Scores
[1] Low Rating Count (<= 52)	caprese salad tomatoes italian marinated tomatoes	52	5
[2] Medium Rating Count 52 < Rating Count < =217	mexican stack up rsc	217	4.99078
[3] High Rating Count (> 217)	kittencal s italian melt in your mouth meatballs	997	4.70812

For group 3, we are weighting between rating count and average scores and pick the most significant recipe, this could be better illustrated in the below graph. [SAS Code Task 6 part f]



The top 10 high rating recipes for each group are listed in [SAS Output <8>] (SAS Code/*6g*/)

.

Conclusion

From the above investigation, we could conclude the below findings in this report:

1) Average number of ingredients is 9.05 with a median of 9
2) Average number of steps is 9.766 with a median of 9.
3) There are 18,556 users who have reviewed more than 5 recipes with an overall average rating score of 4.455.
4) Recipe being rated for more than 15 times would fall beyond the top 5 percentile.
5) 156 recipes with perfect rating (rating score = 5)
6) By dividing up rating count into 3 groups, we can list the recipes receive higher user ratings in each group as below:

Group	Recipe Name	Rating Count	Average Rating Scores
[1] Low Rating Count	caprese salad tomatoes italian marinated tomatoes	52	5
[2] Medium Rating Count	mexican stack up rsc	217	4.99078
[3] High Rating Count	kittencal s italian melt in your mouth meatballs	997	4.70812

Base on the above findings, it would be very interesting to extend our investigation to :

1) What kind of recipes would receive high ratings?
2) Which contributor submit high rating recipes?

References

<R1>

Majumder, Bodhisattwa Prasad and Li, Shuyang and Ni, Jianmo and McAuley, Julian. Generating Personalized Recipes from Historical User Preferences. Proceedings EMNLP 2019.

Appendix – SAS Output

<1>

Summary statistics of number of ingredients for each recipe

The MEANS Procedure

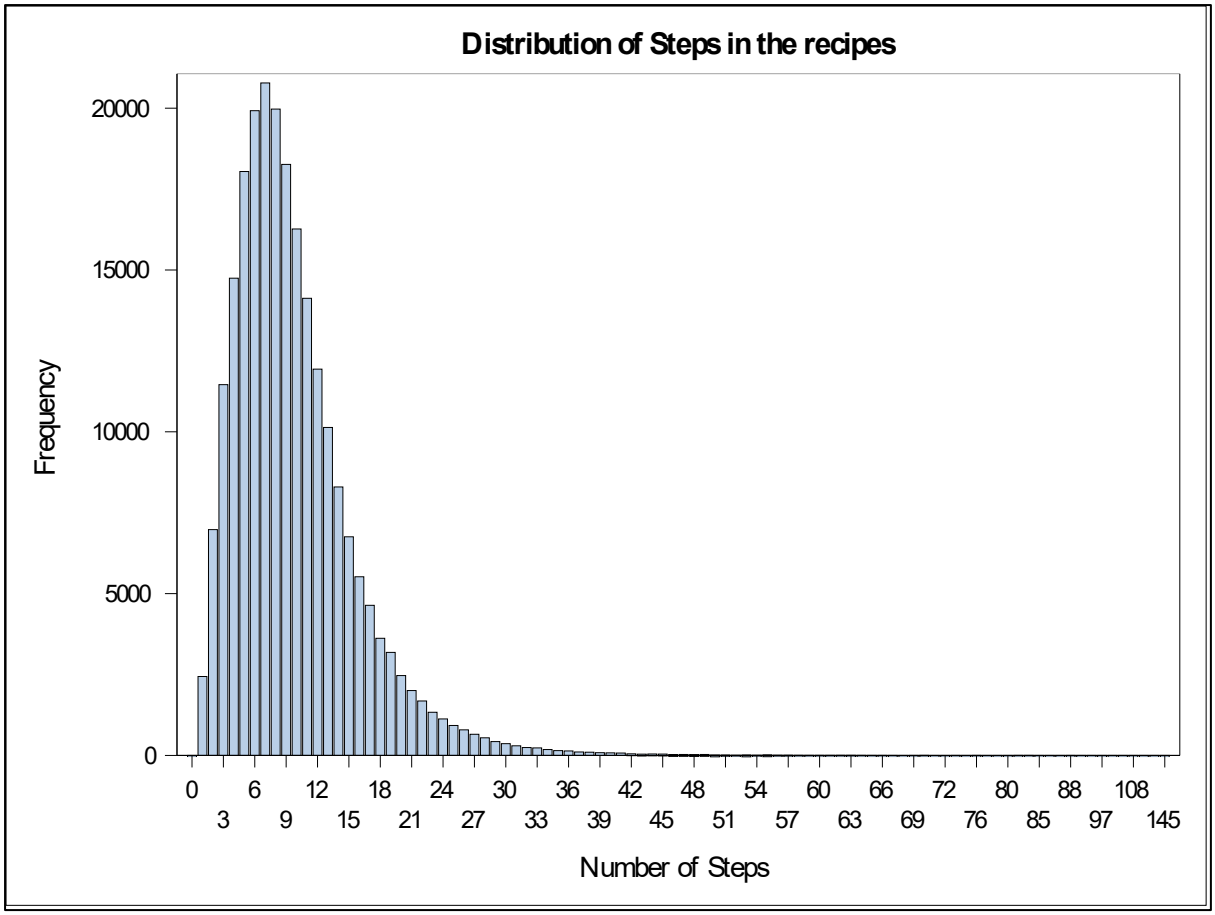
Analysis Variable : nIngredients			
Minimum	**Maximum**	**Mean**	**Median**
1.0000000	43.0000000	9.0514477	9.0000000

<2>

Summary statistics of number of steps for each recipe

The MEANS Procedure
Analysis Variable :
nSteps

Minimum	Maximum	Mean	Median
0	145.0000000	9.7660196	9.0000000



<3>

Summary statistics of reviews number for each individual user

The MEANS Procedure

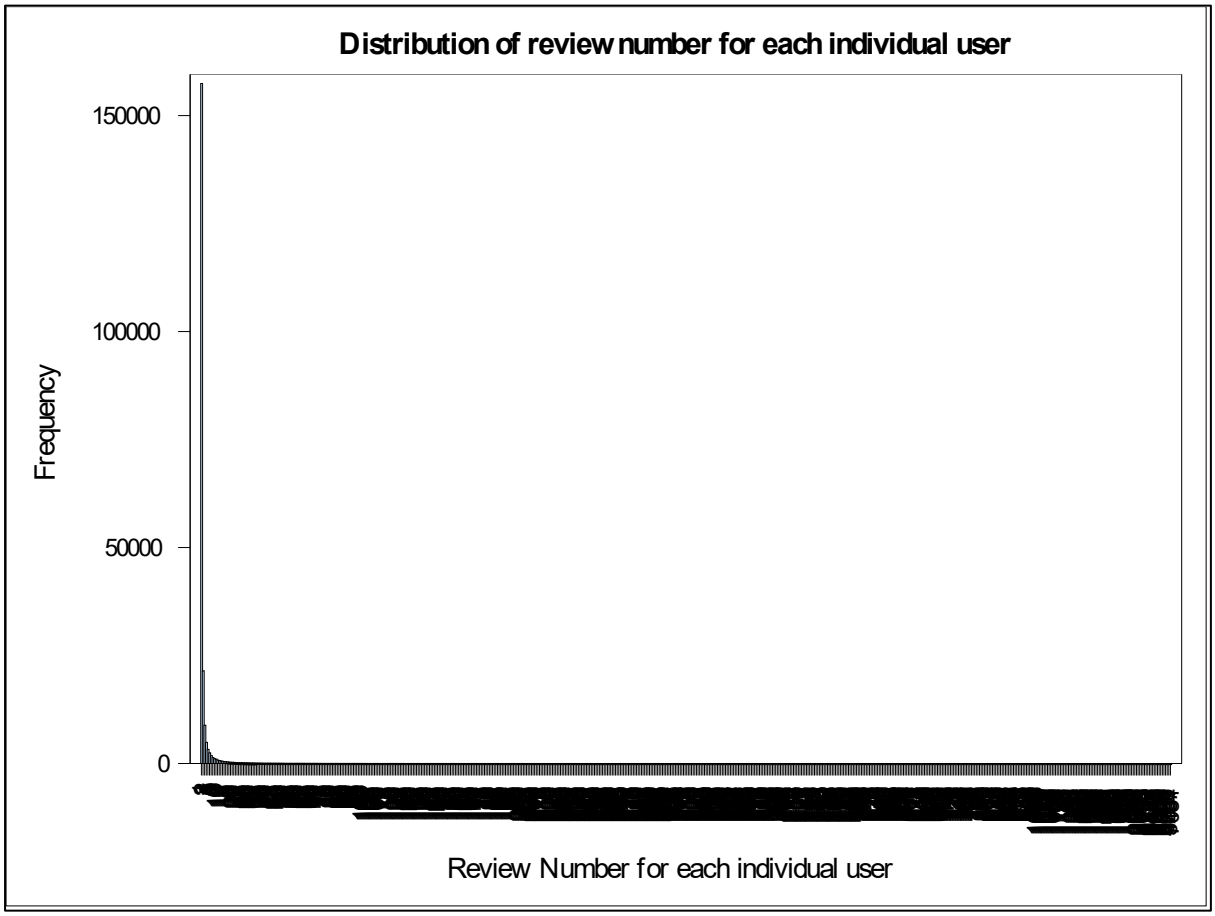
Analysis Variable : reviewsPerUser										
N	Mean	Median	Minimum	Maximum	5th Pctl	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl
214484	4.8888262	1.0000000	1.0000000	7084.00	1.0000000	1.0000000	1.0000000	1.0000000	2.0000000	5.0000000

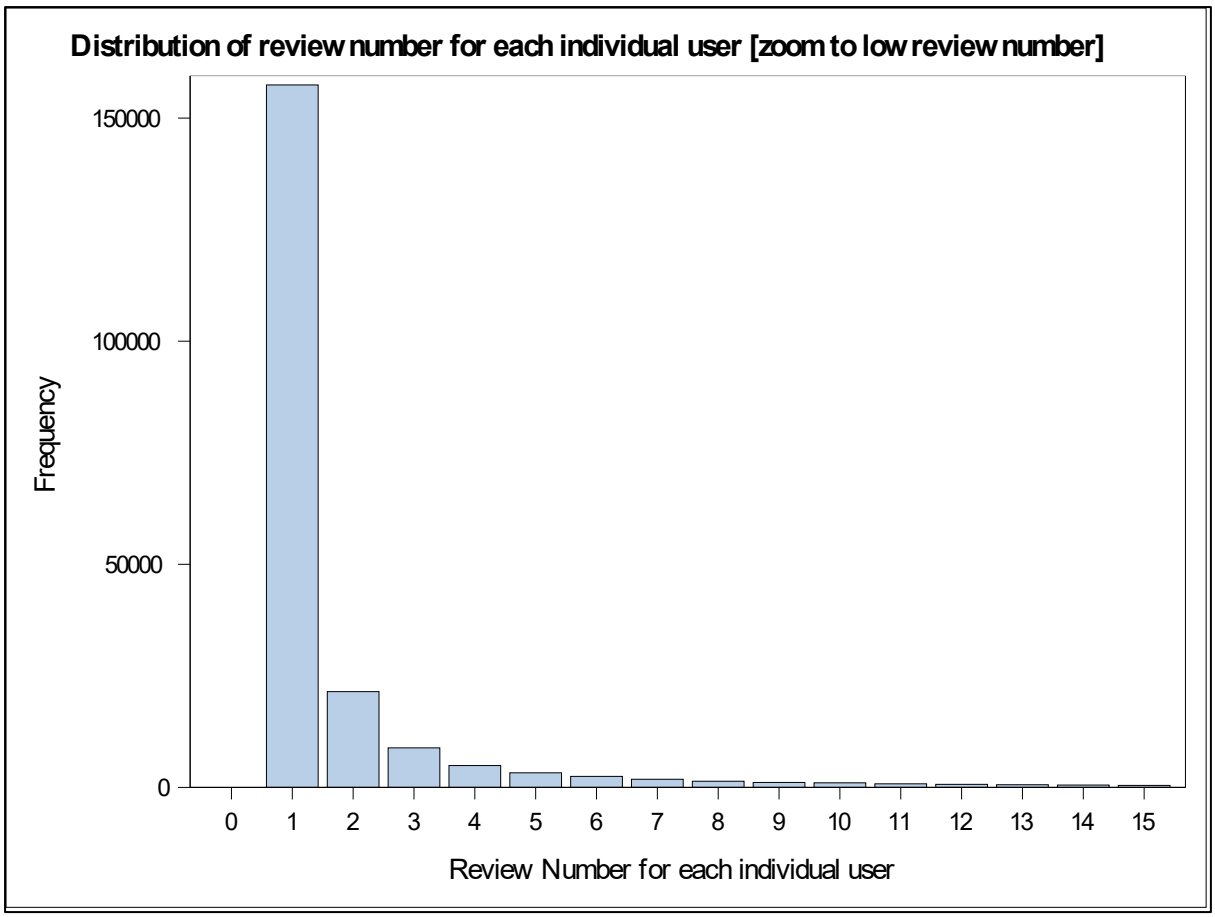
Analysis Variable : reviewsPerUser
95th Pctl
11.0000000

The UNIVARIATE Procedure

Variable: reviewsPerUser

Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	1	157430	596	3621	1
2	2	21465	597	3766	1
3	3	8864	598	4281	1
4	4	4897	599	5167	1
5	5	3272	600	7084	1





<4>

No. of users reviewed > 5 recipes	Frequent Reviewers' rating mean
18556	4.455069

<5>

Summary statistics of rating count for each recipe

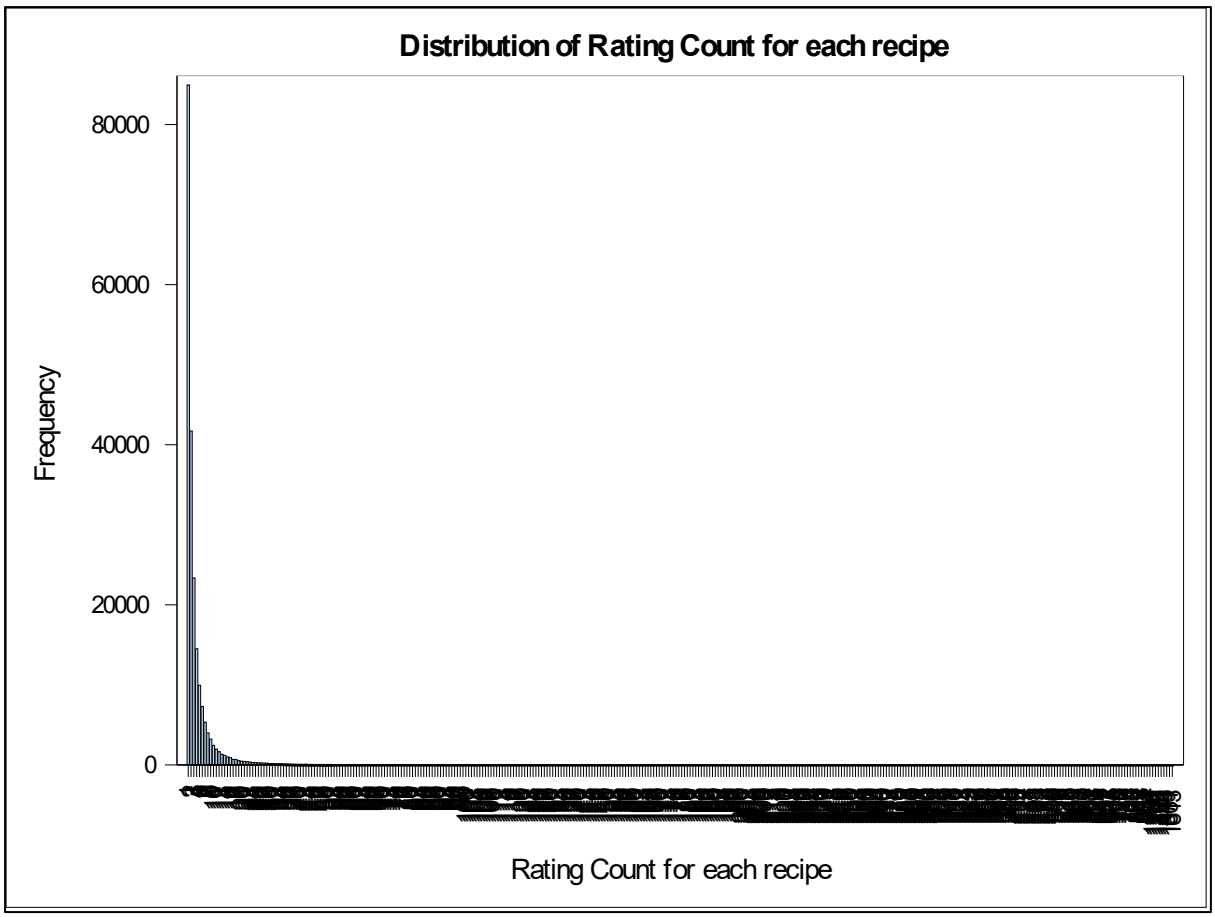
The MEANS Procedure

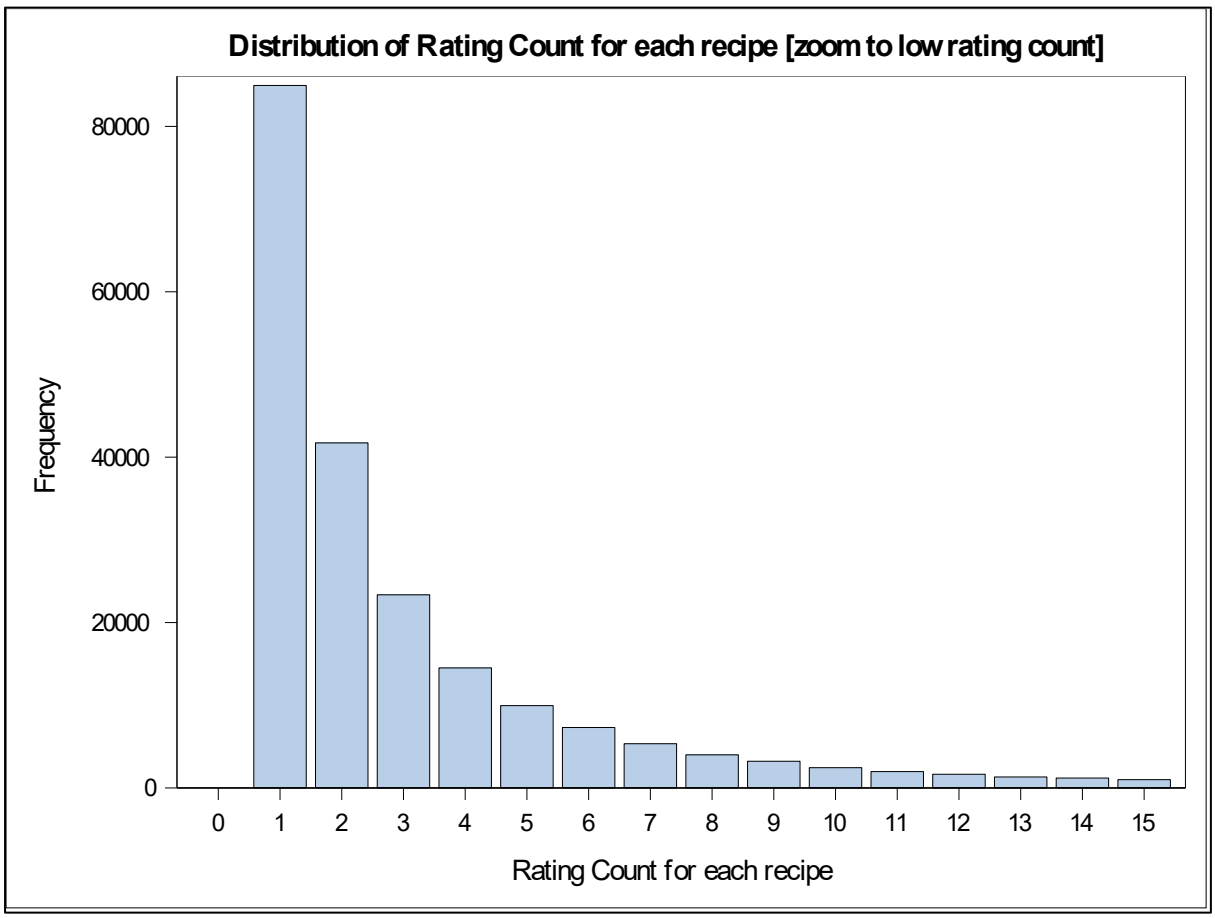
Analysis Variable : RatingCount										
N	Mean	Median	Minimum	Maximum	5th Pctl	10th Pctl	25th Pctl	50th Pctl	75th Pctl	90th Pctl
214217	4.8949196	2.0000000	1.0000000	1613.00	1.0000000	1.0000000	1.0000000	2.0000000	4.0000000	9.0000000

Analysis Variable : RatingCount
95th Pctl
15.0000000

The UNIVARIATE Procedure
Variable: RatingCount

Extreme Values					
Lowest			Highest		
Order	Value	Freq	Order	Value	Freq
1	1	84949	348	1322	1
2	2	41723	349	1448	1
3	3	23355	350	1579	1
4	4	14513	351	1601	1
5	5	9955	352	1613	1


Distribution of Rating Count for each recipe
Frequency
80000
60000
40000
20000
0
Rating Count for each recipe


Distribution of Rating Count for each recipe [zoom to low rating count]
Frequency
80000
60000
40000
20000
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Rating Count for each recipe

<6>

2-sample Ttest of complexity on Calories Per Serve

The TTEST Procedure

Variable: CaloriesPS

Complexity	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
high		4070	525.9	929.7	14.5727	0	38680.1
low		6155	419.5	845.0	10.7713	0	38662.3
Diff (1-2)	**Pooled**		106.4	879.7	17.7730		
Diff (1-2)	**Satterthwaite**		106.4		18.1213		

Complexity	Method	Mean	97% CL Mean		Std Dev	97% CL Std Dev	
high		525.9	494.3	557.5	929.7	907.8	952.6
low		419.5	396.1	442.9	845.0	828.8	861.9
Diff (1-2)	**Pooled**	106.4	67.8093	145.0	879.7	866.6	893.3
Diff (1-2)	**Satterthwaite**	106.4	67.0519	145.7			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	10223	5.99	<.0001
Satterthwaite	Unequal	8125.9	5.87	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	4069	6154	1.21	<.0001

<7>

<7a>

Frequency Distribution of Complexity

The FREQ Procedure

Complexity	Frequency	Percent	Cumulative Frequency	Cumulative Percent
high	2233	38.55	2233	38.55
low	3560	61.45	5793	100.00

2-sample Ttest of complexity on AvgRatingScores

The TTEST Procedure

Variable: AvgRatingScores

Complexity	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
high		2233	4.6843	0.1397	0.00296	4.4571	5.0000
low		3560	4.6966	0.1456	0.00244	4.4565	5.0000
Diff (1-2)	Pooled		-0.0123	0.1433	0.00387		
Diff (1-2)	Satterthwaite		-0.0123		0.00383		

Complexity	Method	Mean	97% CL Mean		Std Dev	97% CL Std Dev	
high		4.6843	4.6779	4.6907	0.1397	0.1353	0.1444
low		4.6966	4.6913	4.7019	0.1456	0.1419	0.1494
Diff (1-2)	Pooled	-0.0123	-0.0207	-0.00391	0.1433	0.1405	0.1463
Diff (1-2)	Satterthwaite	-0.0123	-0.0206	-0.00399			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	5791	-3.18	0.0015
Satterthwaite	Unequal	4885.4	-3.21	0.0013

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	3559	2232	1.09	0.0329

<7b>

Frequency Distribution of IngredientsNeed

The FREQ Procedure

IngredientsNeed	Frequency	Percent	Cumulative Frequency	Cumulative Percent
few	2933	50.63	2933	50.63
lots	2860	49.37	5793	100.00

2-sample Ttest of IngredientsNeed on AvgRatingScores

The TTEST Procedure

Variable: AvgRatingScores

IngredientsNeed	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
few		2933	4.6985	0.1448	0.00267	4.4565	5.0000
lots		2860	4.6851	0.1417	0.00265	4.4571	5.0000
Diff (1-2)	**Pooled**		0.0133	0.1433	0.00377		
Diff (1-2)	**Satterthwaite**		0.0133		0.00377		

IngredientsNeed	Method	Mean	97% CL Mean		Std Dev	97% CL Std Dev	
few		4.6985	4.6927	4.7043	0.1448	0.1408	0.1491
lots		4.6851	4.6794	4.6909	0.1417	0.1378	0.1459
Diff (1-2)	**Pooled**	0.0133	0.00517	0.0215	0.1433	0.1405	0.1463
Diff (1-2)	**Satterthwaite**	0.0133	0.00517	0.0215			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	5791	3.54	0.0004
Satterthwaite	Unequal	5790.9	3.54	0.0004

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	2932	2859	1.04	0.2473

<7c>

Frequency Distribution of submittedRage

The FREQ Procedure

submittedRage	Frequency	Percent	Cumulative Frequency	Cumulative Percent
After or in May 2014	2873	49.59	2873	49.59
Before May2014	2920	50.41	5793	100.00

2-sample Ttest of Date Submitted on AvgRatingScores

The TTEST Procedure

Variable: AvgRatingScores

submittedRage	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
After or in May 2014		2873	4.7078	0.1490	0.00278	4.4565	5.0000
Before May2014		2920	4.6762	0.1360	0.00252	4.4565	5.0000
Diff (1-2)	**Pooled**		0.0316	0.1426	0.00375		
Diff (1-2)	**Satterthwaite**		0.0316		0.00375		

submittedRage	Method	Mean	97% CL Mean		Std Dev	97% CL Std Dev	
After or in May 2014		4.7078	4.7018	4.7139	0.1490	0.1448	0.1534
Before May2014		4.6762	4.6708	4.6817	0.1360	0.1323	0.1400
Diff (1-2)	Pooled	0.0316	0.0235	0.0397	0.1426	0.1398	0.1455
Diff (1-2)	Satterthwaite	0.0316	0.0235	0.0397			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	5791	8.43	<.0001
Satterthwaite	Unequal	5725.5	8.43	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	2872	2919	1.20	<.0001

<7d>

Frequency Distribution of ratingFreq

The FREQ Procedure

ratingFreq	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Average	4467	77.11	4467	77.11
High	1326	22.89	5793	100.00

2-sample Ttest of ratingFreq on AvgRatingScores

The TTEST Procedure

Variable: AvgRatingScores

ratingFreq	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
Average		4467	4.7084	0.1469	0.00220	4.4571	5.0000
High		1326	4.6361	0.1148	0.00315	4.4565	5.0000
Diff (1-2)	**Pooled**		0.0723	0.1402	0.00439		
Diff (1-2)	**Satterthwaite**		0.0723		0.00384		

ratingFreq	Method	Mean	97% CL Mean		Std Dev	97% CL Std Dev	
Average		4.7084	4.7037	4.7132	0.1469	0.1436	0.1504
High		4.6361	4.6293	4.6430	0.1148	0.1102	0.1199
Diff (1-2)	**Pooled**	0.0723	0.0628	0.0818	0.1402	0.1374	0.1431
Diff (1-2)	**Satterthwaite**	0.0723	0.0639	0.0806			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	5791	16.49	<.0001
Satterthwaite	Unequal	2734.5	18.81	<.0001

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	4466	1325	1.64	<.0001

<8>

Top 10 recipes with AvgRatingScores = 5 <perfect score> order by number of Rating count

Obs	recipeId	name	RatingCount	AvgRatingScores	stdRating	submitted
1	55309	caprese salad tomatoes italian marinated tomatoes	52	5	0	01/03/2003
2	24768	berry cream cheese coffee cake	37	5	0	08/04/2002
3	121941	kittencal s soft white baguette style bread	36	5	0	13/05/2005
4	62754	linda s special potato salad	32	5	0	21/05/2003
5	199171	broiled cinnamon toast	31	5	0	04/12/2006
6	269136	earth bread	31	5	0	30/11/2007
7	45107	strawberry cheese ring	28	5	0	31/10/2002
8	94087	cranberry nut swirls	28	5	0	23/06/2004
9	495202	crunchy valley chicken rsc	28	5	0	11/02/2013
10	58687	chocolate covered strawberries	27	5	0	07/04/2003

Obs	Minutes	nSteps	nIngredients	Complexity	IngredientsNeed	ratingFreq
1	10	13	6	high	few	High
2	85	19	12	high	lots	Average
3	145	24	9	high	lots	Average
4	70	6	14	low	lots	Average
5	8	9	4	low	few	Average
6	75	7	15	low	lots	Average
7	190	6	8	low	few	Average
8	35	15	12	high	lots	Average
9	55	21	15	high	lots	Average
10	30	4	3	low	few	Average

Top 10 recipes with highest rating scores with rating count > 52 and <= 217, order by rating scores

Obs	recipeId	name	RatingCount	AvgRatingScores	stdRating	submitted
1	486261	mexican stack up rsc	217	4.99078	0.09578	01/09/2012
2	42038	syrup for blueberry pancakes	57	4.96491	0.18564	02/10/2002
3	21033	toffee dip with apples	55	4.96364	0.18892	01/03/2002
4	248495	make your own boursin cheese paula deen	53	4.96226	0.19238	24/08/2007
5	63621	mango salsa 1	74	4.95946	0.25851	02/06/2003
6	59776	scott hibb s amazing whiskey grilled baby back ribs	63	4.95238	0.21467	15/04/2003
7	37455	boboli pizza crust clone	67	4.94030	0.23872	16/08/2002
8	515167	simply irresistible tropical potato salad sp5	74	4.93243	0.58124	31/03/2014
9	186029	the best creole cajun seasoning mix	68	4.92647	0.60634	13/09/2006
10	43332	paige s buttercream frosting	53	4.92453	0.26668	16/10/2002

Obs	Minutes	nSteps	nIngredients	Complexity	IngredientsNeed	ratingFreq
1	75	28	22	high	lots	High
2	15	3	3	low	few	High
3	5	8	8	low	few	High
4	10	6	9	low	lots	High
5	20	2	12	low	lots	High
6	180	17	18	high	lots	High
7	70	6	9	low	lots	High
8	13	5	5	low	few	High
9	5	3	10	low	lots	High
10	20	5	4	low	few	High

Top 10 recipes with highest rating scores with rating count > 217, order by rating scores

Obs	recipeId	name	RatingCount	AvgRatingScores	stdRating	submitted
1	87925	kittencal s extreme chocolate brownies	234	4.78632	0.79515	31/03/2004
2	8739	wholly guacamole	298	4.77181	0.87325	09/04/2001
3	106251	roasted cauliflower 16 roasted cloves of garlic	357	4.75350	0.83852	16/12/2004
4	31128	yummy crunchy apple pie	410	4.74634	0.94548	13/06/2002
5	104086	the best chicken tortilla soup	230	4.73043	0.92818	15/11/2004
6	20616	irish rosie s irish soda bread	282	4.71277	1.02935	25/02/2002
7	69173	kittencal s italian melt in your mouth meatballs	997	4.70812	1.01399	20/08/2003
8	43509	crumb topped banana muffins	540	4.70556	0.98239	18/10/2002
9	80413	homemade hamburger buns bread machine	237	4.70464	1.04825	07/01/2004
10	57130	awesome slow cooker pot roast	239	4.70293	0.94359	25/03/2003

Obs	Minutes	nSteps	nIngredients	Complexity	IngredientsNeed	ratingFreq
1	50	13	9	high	lots	High
2	20	13	8	high	few	High
3	40	4	6	low	few	High
4	70	15	12	high	lots	High
5	375	5	20	low	lots	High
6	70	9	9	low	lots	High
7	50	5	10	low	lots	High
8	35	9	11	low	lots	High
9	19	9	7	low	few	High
10	490	3	4	low	few	High

Top 10 recipes with highest rating scores with rating count > 217, order by rating count

Obs	recipeId	name	RatingCount	AvgRatingScores	stdRating	submitted	Minutes
1	39087	creamy cajun chicken pasta	1448	4.54144	1.25842	02/09/2002	25
2	32204	whatever floats your boat brownies	1220	4.52541	1.25770	25/06/2002	35
3	69173	kittencal s italian melt in your mouth meatballs	997	4.70812	1.01399	20/08/2003	50
4	82102	kittencal s moist cheddar garlic oven fried chicken breast	855	4.60585	1.03915	26/01/2004	60
5	28148	oven fried chicken chimichangas	802	4.69950	0.92448	13/05/2002	45
6	10744	delicious chicken pot pie	731	4.49658	1.35926	07/08/2001	105
7	33671	crock pot whole chicken	727	4.46630	1.34580	09/07/2002	1380
8	63689	my family s favorite sloppy joes pizza joes	720	4.60972	1.05174	05/06/2003	45
9	150863	panera s cream cheese potato soup	714	4.53501	1.26093	08/01/2006	60
10	32614	fudge crinkles a great 4 ingredient cake mix cookie	652	4.48313	1.26928	30/06/2002	15

Obs	nSteps	nIngredients	Complexity	IngredientsNeed	ratingFreq
1	4	12	low	lots	High
2	14	14	high	lots	High
3	5	10	low	lots	High
4	8	10	low	lots	High
5	7	9	low	lots	High
6	11	12	high	lots	High
7	8	10	low	lots	High
8	5	11	low	lots	High
9	6	6	low	few	High
10	20	4	high	few	High

Appendix - SAS Code

/* Task 1 */

```
libname a2 "/home/u41107333/Assignment2";


data nutrition;
infile "/home/u41107333/Assignment2/nutrition.csv" delimiter="," firstobs =4  dsd missover;
input id CaloriesPS Fat   Sugars   Sodium Protein SatFat   Carb;
run;


data ratings;
infile "/home/u41107333/Assignment2/ratings.csv" delimiter="," firstobs =2  dsd missover;
informat userId $15. recipeId 15.          date ddmmyy10.          rating;
```

```
input userId      recipeId          date      rating;
format date ddmmyy10.;
run;

data recipes;
infile "/home/u41107333/Assignment2/Recipes.csv" delimiter="," firstobs =2  dsd missover;
informat name $100.     id 15.    Minutes 10.      contributorId $15.        submitted ddmmyy10.
          nSteps  nIngredients;
input name       id         Minutes          contributorId    submitted        nSteps  nIngredients;
format submitted ddmmyy10.;
run;

/* Task 2 */
title "Summary statistics of number of ingredients for each recipe";
proc means data=work.recipes min max mean median;
var nIngredients;
run;
title;

ods graphics on;
proc sgplot data=work.recipes;
vbar nIngredients;
xaxis label="Number of ingredients" valuesrotate=vertical;
yaxis label="Frequency";
title "Distribution of ingredients in the recipes";
run;
title;

/*2b*/
title "Summary statistics of number of steps for each recipe";
proc means data=work.recipes min max mean median;
var nSteps;
```

```
run;
title;

ods graphics on;
proc sgplot data=work.recipes;
vbar nSteps;

xaxis fitpolicy=staggerthin label="Number of Steps" valuesrotate=vertical;
yaxis label="Frequency";
title "Distribution of Steps in the recipes";



run;
title;

/*2c*/
proc sql;
create table NumReview as
select userId, count(userId) as reviewsPerUser
from ratings
group by userId;
quit;



title "Summary statistics of reviews number for each individual user";
proc means data=NumReview n mean median min max p5 p10 p25 p50 p75 p90 p95;
var reviewsPerUser;
run;
title;

ods select ExtremeValues;
```

```
proc univariate data=NumReview  NEXTRVAL=5;
var reviewsPerUser;
run;

ods graphics on;
proc sgplot data=NumReview;
vbar reviewsPerUser;
xaxis label="Review Number for each individual user" valuesrotate=vertical;
yaxis label="Frequency";
title "Distribution of review number for each individual user";
run;
title;

ods graphics on;
proc sgplot data=NumReview;
vbar reviewsPerUser;
xaxis label="Review Number for each individual user" valuesrotate=vertical values=(0 to 15 by 1)
valueshint;
yaxis label="Frequency";
title "Distribution of review number for each individual user [zoom to low review number]";
run;
title;
```

/* Task 3 */

```
proc sql;
create table FreqReviewer as
select userId, count(userId) as noOfReviews
from ratings
group by userId
having noOfReviews > 5;
quit;
```

```
proc sql;
create table FreqReviewerRating as
select r.userId, mean(r.rating) as AvgRatingScores
from ratings as r, FreqReviewer as f
where r.userId=f.userId
group by r.userId;
quit;

proc sql;
select count(userId) label="No. of users reviewed > 5 recipes", mean(fr.AvgRatingScores)
label="Frequent Reviewers' rating mean"
from FreqReviewerRating as fr;
quit;
```

/*3b*/

```
proc sql;
create table NumRating as
select recipeId, count(rating) as RatingCount
from ratings
group by recipeId;
quit;

title "Summary statistics of rating count for each recipe";
proc means data=NumRating n mean median min max p5 p10 p25 p50 p75 p90 p95;
var RatingCount;
run;
title;

ods select ExtremeValues;
```

```
proc univariate data=NumRating  NEXTRVAL=5;
var RatingCount;
run;



proc sql;
create table NumRatingGT15 as
select recipeId, count(rating) as RatingCount
from ratings
group by recipeId
having ratingCount > 15;
quit;

title "Summary statistics of rating count for each recipe with count no. greater than 15 ";
proc means data=NumRatingGT15 n mean median min max p5 p10 p25 p50 p75 p90 p95;
var RatingCount;
run;
title;

ods graphics on;
proc sgplot data=NumRating;
vbar RatingCount;
xaxis label="Rating Count for each recipe" valuesrotate=vertical;
yaxis label="Frequency";
title "Distribution of Rating Count for each recipe";
run;
title;

ods graphics on;
proc sgplot data=NumRating;
vbar RatingCount;
```

```
xaxis label="Rating Count for each recipe" valuesrotate=vertical values=(0 to 15 by 1) valueshint;;
yaxis label="Frequency";
title "Distribution of Rating Count for each recipe [zoom to low rating count]";
run;
title;
```

```
/* Task 4 */
data work.recipes;
set work.recipes;
length Complexity $5. ;
if nSteps < 10 then Complexity='low';
else if nSteps >= 10 then Complexity='high';
else Complexity='Unknown';
run;

title 'Frequency Distribution of Complexity';
proc freq data=work.recipes;
table Complexity;
run;
title;
```

```
/*4a*/
data work.recipes;
set work.recipes;
length IngredientsNeed $5. ;
if nIngredients < 9 then IngredientsNeed='few';
else if nIngredients >= 9 then IngredientsNeed='lots';
else IngredientsNeed='Unknown';
run;
```

/* Task 5 */

```
proc sql;
create table highReview as
select recipeId, count(rating) as RatingCount
from ratings
group by recipeId
having RatingCount > 15;
quit;

proc sql;
create table CalComplexity as
select r.Id, r.nSteps, r.Complexity, n.CaloriesPS
from highReview as h inner join recipes as r
on h.recipeId = r.Id inner join nutrition as n
on n.id = r.Id;
quit;


title '2-sample Ttest of complexity on Calories Per Serve';
proc ttest data=CalComplexity alpha=0.03;
var CaloriesPS;
class Complexity;
run;
title;
```

/* Task 6 */

```
/*part a*/

proc sql;

create table recipeRating as

select recipeId, count(rating) as RatingCount, mean(rating) as AvgRatingScores, std(rating) as
stdRating

from ratings

group by recipeId

having AvgRatingScores > 4.455 and RatingCount > 15

order by AvgRatingScores desc, ratingcount desc;

quit;
```

/*6a i*/

```
data work.recipeRating;

set work.recipeRating;

length ratingFreq $8. ;

if RatingCount < 44 then ratingFreq='Average';

else if RatingCount >= 44 then ratingFreq='High';

else RatingCount='Unknown';

run;


/* Join with the recipe table to get recipe info such as submission date, contributor etc */

proc sql;

create table recipeRatingWithInfo as

select rr.*, r.submitted, r.contributorID, r.name, r.Minutes, r.nSteps, r.nIngredients, r.complexity,
r.IngredientsNeed

from recipeRating as rr inner join recipes r

on rr.recipeId = r.ID

order by r.submitted;

quit;
```

```
/*6a ii*/

proc tabulate data=recipeRatingWithInfo;
var submitted;
table submitted,
 n nmiss (min max median)*f=mmddyy10. range;
run;

data work.recipeRatingWithInfo;
set work.recipeRatingWithInfo;
length   submittedRage $20. ;
if submitted < '31MAY2004'd then submittedRage='Before May2014';
else if submitted >= '31MAY2004'd then submittedRage='After or in May 2014';
else submittedRage='Unknown';
run;

/*part b*/
/*6b i*/
title 'Frequency Distribution of Complexity';
proc freq data=work.recipeRatingWithInfo;
table Complexity;
run;
title;

title '2-sample Ttest of complexity on AvgRatingScores';
proc ttest data=recipeRatingWithInfo alpha=0.03;
var AvgRatingScores;
class Complexity;
run;
title;
```

```
/*6b ii*/
title 'Frequency Distribution of IngredientsNeed';
proc freq data=work.recipeRatingWithInfo;
table IngredientsNeed;
run;
title;

title '2-sample Ttest of IngredientsNeed on AvgRatingScores';
proc ttest data=recipeRatingWithInfo alpha=0.03;
var AvgRatingScores;
class IngredientsNeed;
run;
title;

/*6b iii */

title 'Frequency Distribution of submittedRage';
proc freq data=work.recipeRatingWithInfo;
table submittedRage;
run;
title;

title '2-sample Ttest of Date Submitted on AvgRatingScores';
proc ttest data=recipeRatingWithInfo alpha=0.03;
var AvgRatingScores;
class submittedRage;
run;
title;
```

```
/*6b iv*/
title 'Frequency Distribution of ratingFreq';
proc freq data=work.recipeRatingWithInfo;
table ratingFreq;
run;
title;


title '2-sample Ttest of ratingFreq on AvgRatingScores';
proc ttest data=recipeRatingWithInfo alpha=0.03;
var AvgRatingScores;
class ratingFreq;
run;
title;



/*6c*/
ods select ExtremeValues;
proc univariate data=recipeRatingWithInfo  NEXTRVAL=10;
var AvgRatingScores;
run;

proc means data=recipeRatingWithInfo n nmiss min max median range maxdec=1;
var RatingCount;
run;

/*6d*/
/* Use graph to review Rating Volume distribution */
title "Rating Volume Distribution";

ods graphics on;
```

```
proc sgplot data=recipeRatingWithInfo;

vbar RatingCount ;
xaxis fitpolicy=staggerthin label="Rating Count";
yaxis label="Frequency";

run;
title;
```

/*6e*/

```
/* Grouping rating count into different brand */
proc sql;
create table rating5 as
select *
from recipeRatingWithInfo
where AvgRatingScores = 5
order by ratingCount desc;
quit;


proc sql;
create table ratingCntGt52 as
select *
from recipeRatingWithInfo
where ratingCount > 52
order by AvgRatingScores desc;
quit;
```

```
proc sql;
create table ratingCntGt217 as
select *
from recipeRatingWithInfo
where ratingCount > 217
order by AvgRatingScores desc;
quit;

proc sql;
create table ratingCntGt217OrderByCont as
select *
from recipeRatingWithInfo
where ratingCount > 217
order by ratingCount desc;
quit;
```

/*6f*/

```
ods graphics on;
proc sgplot data=ratingCntGt217;
scatter x=ratingCount y=AvgRatingScores;
xaxis label="Rating Count";
yaxis label="Avg Rating";
title "Relationship between Rating Count vs Avg Rating for Rating Count greater than 217";

run;
title;
```

/*6g*/

```
title "Top 10 recipes with AvgRatingScores = 5 <perfect score> order by number of Rating count";

proc print data=rating5 (obs=10);

var recipeId name RatingCount AvgRatingScores stdRating submitted Minutes     nSteps
        nIngredients Complexity IngredientsNeed ratingFreq;

run;

title;


title "Top 10 recipes with highest rating scores with rating count > 52 and <= 217, order by rating
scores";

proc print data=ratingCntGt52 (obs=10);

var recipeId name RatingCount AvgRatingScores stdRating submitted Minutes     nSteps
        nIngredients Complexity IngredientsNeed ratingFreq;

run;

title;


title "Top 10 recipes with highest rating scores with rating count > 217, order by rating scores";

proc print data=ratingCntGt217 (obs=10);

var recipeId name RatingCount AvgRatingScores stdRating submitted Minutes     nSteps
        nIngredients Complexity IngredientsNeed ratingFreq;

run;

title;


title "Top 10 recipes with highest rating scores with rating count > 217, order by rating count";

proc print data=ratingCntGt217OrderByCont (obs=10);

var recipeId name RatingCount AvgRatingScores stdRating submitted Minutes     nSteps
        nIngredients Complexity IngredientsNeed ratingFreq;

run;

title;
```

```
/*6h*/

ods graphics on;
proc sgplot data=recipeRatingWithInfo;
scatter x=RatingCount y=AvgRatingScores;
xaxis label="Rating Count";
yaxis label="Average Rating Scores";
title "Relationship between Rating Count vs Average Rating Scores";

run;
title;


ods graphics on;
proc sgplot data=recipeRatingWithInfo;
scatter x=nIngredients y=AvgRatingScores;
xaxis label="Number of Ingredients";
yaxis label="Average Rating Scores";
title "Relationship between Number of Ingredients vs Average Rating Scores";

run;
title;

ods graphics on;
proc sgplot data=recipeRatingWithInfo;
scatter x=nSteps y=AvgRatingScores;
xaxis label="Number of Steps";
yaxis label="Average Rating Scores";
title "Relationship between Number of Steps vs Average Rating Scores";

run;
```

```
title;


ods graphics on;
proc sgplot data=recipeRatingWithInfo;
scatter x=submitted y=AvgRatingScores;
xaxis label="Date submitted";
yaxis label="Average Rating Scores";
title "Relationship between submission Date vs Average Rating Scores";

run;
title;
```