SAS® GLOBAL FORUM 2018

USERS PROGRAM

April 8 – 11 | Denver, CO
Colorado Convention Center

#SASGF

Understanding the Factors that Affect Customers' Choice of Sunscreen Products

Zihan Fan & Xiaojing Zhao

Zihan Fan, Graduate Student, Clark University

Xiaojing Zhao, Graduate Student, Clark University

Zihan Fan is a master of science in business analytics in Clark University, who worked in Marketing Department in Amazon Kindle for one year, and started to learn SAS® last summer.

Xiaojing Zhao is now a business analytics student in Clark University, who has been worked in Bank of China, Beijing branch for 3 years. She started to learn SAS ® since last summer on big data course.

Insert your Twitter handle if applicable.

Understanding the Factors that Affect Customers' Choice of Sunscreen Products

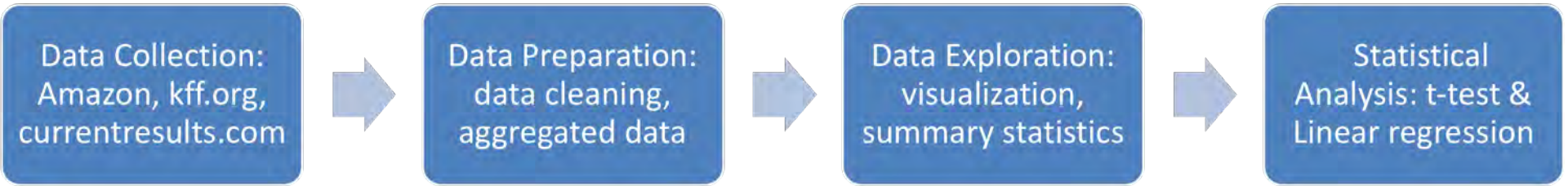
Understanding the Factors that Affect Customers' Choice of Sunscreen Products

INTRODUCTION

- Motivation: We found that the body sunscreen products of the US market usually merely have UV protection and broad spectrum. So, we wonder what are the main concerns for US consumers to purchase specific products.
- Purpose: This study is to analyze the crucial factors that affect consumers' preference of sun care products in different states.
- Crucial factors(variables): Average hour on sunshine in four seasons in 2016, median annual household income in each state in US, population distribution by race in US and reviews written in 2016 from Amazon which categorized by each state in US.
- Research questions:
  - What are consumers' main concerns when choosing sunscreen products?
  - What makes consumers keep loyalty to specific types of sunscreen products?
  - Is there a difference in the demand for sunscreens' features among different race lived in the same region in US?

Understanding the Factors that Affect Customers' Choice of Sunscreen Products

METHODS



Understanding the Factors that Affect Customers' Choice of Sunscreen Products

RESULTS

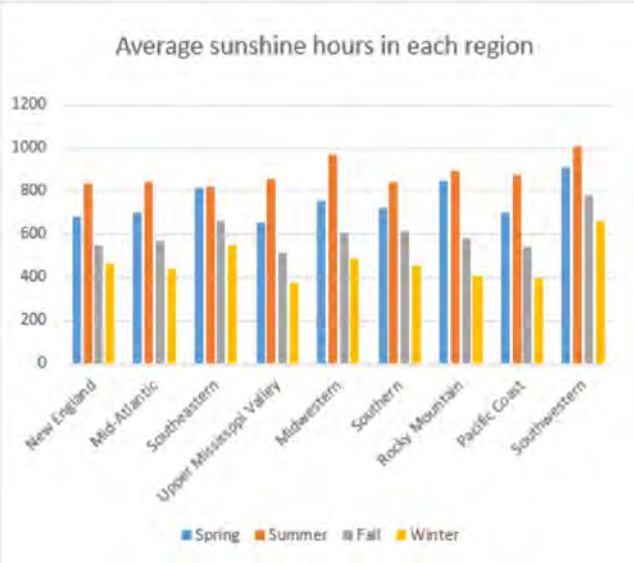


Figure 1 Average sunshine hours in nine regions in US

Variable	Mean	Std Dev	Minimum	Maximum	Range	N
spring	746.3258333	91.5824326	607.0000000	1079.00	472.0000000	50
Summer	898.3869444	100.6665815	729.0000000	1209.00	480.0000000	50
Fall	603.1241667	101.0335956	358.0000000	888.0000000	530.0000000	50
Winter	465.9113889	106.0886990	232.0000000	764.0000000	532.0000000	50

Figure 2 Statistics of average sunshine hours in four seasons in US

- Figure 1 shows that the southwestern region has the highest average hours of sunshine within four seasons of a year.
- Figure 2 illustrates that the range of each data is around 450, which indicates that there is obvious variation from state to state.

Understanding the Factors that Affect Customers' Choice of Sunscreen Products

RESULTS

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Sum-Quantity

Number of Observations Read	49
Number of Observations Used	49

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	11065	1229.40580	10.17	<.0001
Error	39	4715.34778	120.90635		
Corrected Total	48	15780			

Root MSE	10.99574	R-Square	0.7012
Dependent Mean	11.00000	Adj R-Sq	0.6322
Coeff Var	99.96130		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	Intercept	1	60.47743	66.09668	0.91	0.3658	0	0
Spring		1	0.01925	0.04707	0.41	0.6848	0.09816	7.51932
Summer		1	-0.03226	0.03721	-0.87	0.3912	-0.17555	5.35059
Fall		1	-0.08849	0.06253	-1.42	0.1650	-0.46665	14.19440
Winter		1	0.08456	0.04181	2.02	0.0500	0.47890	7.31792
Median Annual household income 2	Median Annual household income 2016	1	-0.00056729	0.00027008	-2.10	0.0422	-0.27274	2.20061
White		1	-0.02230	0.63325	-0.04	0.9721	-0.01831	35.31718
Black		1	-0.23873	0.65063	-0.37	0.7157	-0.14050	19.13739
Hispanic		1	0.14608	0.74013	0.20	0.8446	0.08208	22.57041
Asian		1	4.56386	0.92009	4.96	<.0001	0.81239	3.50090

Generated by the SAS System ('Local', X64_8PRO) on December 06, 2017 at 5:48:40 PM

Figure 3 Linear regression result of sum quantity and all variables in our model in full model

Figure 3 shows that Sum Quantity of reviews = - 0.00056729 * Median annual household income + 4.56386 * percentage of Asian resident + 60.47743

Understanding the Factors that Affect Customers' Choice of Sunscreen Products

RESULTS

t Test

The TTEST Procedure

Difference: Body - Face

N	Mean	Std Dev	Std Err	Minimum	Maximum
49	7.8163	12.7045	1.8149	-1.0000	76.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
7.8163	4.1672	11.4655	12.7045	10.5945	15.8717

DF	t Value	Pr > |t|
48	4.31	<.0001

H_0: $\mu_{body} - \mu_{face} = 0$
H_a: $\mu_{body} - \mu_{face} \neq 0$

The results show that the p-value is 0.0001, which is less than α(0.05). We can reject the null hypothesis because we have strong statistical evidence from our sample that there is difference between sales of body sunscreen products and face sunscreen products.

t Test

The TTEST Procedure

Difference: SPF50- - SPF55+

N	Mean	Std Dev	Std Err	Minimum	Maximum
49	4.1020	8.7945	1.2564	-6.0000	50.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
4.1020	1.5760	6.6281	8.7945	7.3339	10.9870

DF	t Value	Pr > |t|
48	3.27	0.0020

H_0: $\mu_{SPF50-} - \mu_{SPF55+} = 0$
H_a: $\mu_{SPF50-} - \mu_{SPF55+} \neq 0$

The results show that the p-value is 0.0020, which is less than α(0.05). We can reject the null hypothesis because we have strong statistical evidence from our sample that there is difference between sales of SPF50- sunscreen products and SPF55+ sunscreen products.

t Test

The TTEST Procedure

Difference: Lotion - Spray

N	Mean	Std Dev	Std Err	Minimum	Maximum
49	5.1224	10.1891	1.4556	-4.0000	66.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
5.1224	2.1958	8.0491	10.1891	8.4969	12.7292

DF	t Value	Pr > |t|
48	3.52	0.0010

H_0: $\mu_{lotion} - \mu_{spray} = 0$
H_a: $\mu_{lotion} - \mu_{spray} \neq 0$

The results show that the p-value is 0.0010, which is less than α(0.05). We can reject the null hypothesis because we have strong statistical evidence from our sample that there is difference between sales of lotion sunscreen products and spray sunscreen products.

t Test

The TTEST Procedure

Difference: sensitive - Normal

N	Mean	Std Dev	Std Err	Minimum	Maximum
49	-7.2449	12.3635	1.7662	-76.0000	0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-7.2449	-10.7961	-3.6937	12.3635	10.3102	15.4457

DF	t Value	Pr > |t|
48	-4.10	0.0002

H_0: $\mu_{sensitive} - \mu_{normal} = 0$
H_a: $\mu_{sensitive} - \mu_{normal} \neq 0$

The results show that the p-value is 0.0002, which is less than α(0.05). We can reject the null hypothesis because we have strong statistical evidence from our sample that there is difference between sales of sensitive sunscreen products and normal sunscreen products.

Understanding the Factors that Affect Customers' Choice of Sunscreen Products

CONCLUSIONS

- The reviews on Amazon of sunscreen products are negatively correlated with median annual household income, which might be caused by that consumer with high income would require higher quality of service.
- Reviews on Amazon of sunscreen products are positively correlated with percentage of Asian residents within a state.
- In each state, reviews on Amazon of sunscreen products with different features are different, so consumers have preferences toward different features.
- Based on our descriptive analysis , we recommend focusing on Southwestern region and trying to explore more types of demand about sunscreen products in this region.
- According to our t-test analysis and linear regression model, we recommend that the manufacturers could pay more attention to targeting Asian consumers' needs in each state to acquire more sales potential.
- We believe that this project will help the sunscreen product manufacturers come up with more accurate manufacturing strategies and marketing plans.

Paper #2862-2018

Understanding the Factors that Affect Customers' Choice of Sunscreen Products

Zihan Fan and Xiaojing Zhao
Professor: Dr. Pankush Kalgotra
Clark University, Worcester, MA

ABSTRACT:

Recently, we found that the body sunscreen products of the US market usually merely have UV protection and broad spectrum. So, we wonder what are the main concerns for US consumers to purchase specific products. The purpose of this study is to analyze the crucial factors that affect consumers' preference of sun care products in different states. We come up with several factors and collect the relative data from historical data on sunscreens and study the relationship between the factors and sun care products sales in each state in US. We use SAS Enterprises Guide to perform analysis. We believe that this project will help the sunscreen product manufacturers come up with more accurate manufacturing strategies and marketing plans. In our research, we use data from websites and integrate the information in the form we need. According to the data analysis, we can mainly give the manufacturer three suggestions, which are, firstly, focusing on Southwestern region and trying to explore more types of demand about sunscreen products in this region. Secondly, focusing more on daily sunscreen products, rather than other featured type of sun care products. Thirdly, the manufacturers could pay more attention to targeting Asian consumers' need in each state to acquire more sales potential.

INTRODUCTION:

Northeastern Asians pay more attention to sunscreen, since over sunlight may cause tanning, accelerating skin aging and sunburn, even skin diseases. Recently, we found that the body sunscreen products of the US market usually merely have UV protection and broad spectrum. So, we wonder what is the main concern for US consumers to purchase specific sunscreen products. We believe that this conclusion will make the products more in line with market expectations of the sunscreen product and provide a more effective way to prevent skin diseases. Besides, we also believe that this project will help the sunscreen product manufacturers come up with more accurate manufacturing strategies and better marketing plans. Moreover, people now pay increased attention to health-related problems. Skin diseases as a prevalent health problem, may lead to skin cancer. It is important to promote more effective sunscreen products to prevent skin diseases caused by sunburn.

The management question we focus on is what marketing and production strategy should manufactures take to improve their sales in each state in the United States and help consumers prevent skin diseases caused by sunburn. Consequently, the research questions we have taken into consideration are listed below:

1. What are the factors that cause the different preferences of consumers in different regions, i.e. what are their main concerns when choosing sunscreen products?
2. How do the consumers weigh the performance of the sunscreen products?
3. What is the main reason for consumers to choose the specific products? In other words, what makes them keep loyalty to specific types of sunscreen products?
4. How do people in targeting region perceive sunscreen products?
5. Are existing sun care products meeting all of the most important customer needs in targeting regions?
6. Is there a difference in the demand for sunscreens' features among different race lived in the same region in US?

In the article *"Sunscreen Product Performance and Other Determinants of Consumer Preferences"*, the writers determined the characteristics and the most commonly cited positive and negative features of highly rated sunscreens described by consumers on Amazon. The analysis result could be used as suggestions to dermatologists making recommendations to their consumers. In the *"Assessing the current market of sunscreen: A cross-sectional study of sunscreen availability in metropolitan counties in the United States"*, the writer's analyzed sunscreen availability in three large metropolitan counties to determine the relationship between availability and community demographics. As for *"Trends in sunscreen recommendation among US physicians"*, to evaluate trends in sunscreen recommendation among physicians to determine whether they are following suggested patients - education guidelines regarding sun protection, and to assess data physician sunscreen recommendations to determine the association with patient demographic, physician specialty, and physician diagnosis.

As far as what we have dug into in this field, we have found several studies about the relationship between sunscreen products and skin diseases, and classification of different features on sunscreens that are popular in different regions, as well as analysis on several factors that affect consumers' preferences. However, we are focusing on the main factors that would mostly influence consumers' preference of choosing sun care products in different regions of the United States, where the weather and climate are different, population compositions are various, and the economic development is on different level. Based on the former researches, we are aiming at coming up with a consulting result for manufacturers to

help them change their current marketing strategy to improve their sales and profit, as well as people's acceptance toward sunscreen products.

METHODS AND ANALYSIS

We have mined average hours of sunshine, median annual household income and population distribution by race, as well as quantity of reviews on Amazon of sunscreen products with different features, i.e. used for face or body, whether it is sensitive or not, SPF under 50 or above 55, spray or lotion, in each state for our data to be used. We look into review writers' profiles on Amazon to find their state information. All the data is from the www.currentresults.com, www.amazon.com and www.kff.org websites.

For variables, we chose the average hour on sunshine in four seasons in 2016, median annual household income in each state, population distribution by race in US as independent variables to figure out what influence will the sunshine has on consumption of sunscreen products, and reviews written in 2016 from Amazon which categorized by each state in US. We combined the data of different seasons in one sheet and deleted the invalid data points that cannot be estimated in an appropriate way.

We use paired T-Test to do the hypothesis test and linear regression analysis to figure out whether the factors have relationship with each other. The analysis is processed via SAS Enterprises Guide.

RESULT

The Figure 1 shows the average sunshine hours in nine different regions in the United States. It is clear that the southwestern region has the highest average hours of sunshine within four seasons of a year. So, it is wise for producers to focus on this region and try to explore more types of demand about sunscreen products in this region.

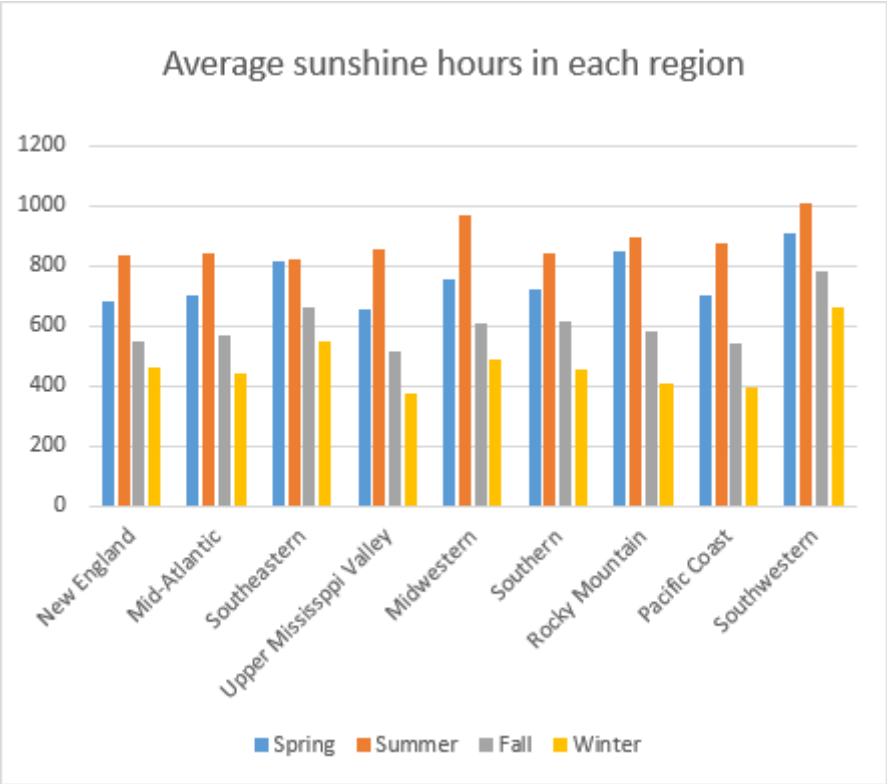


Figure 1 Average sunshine hours in nine regions in US

The Figure 2 illustrates that the range of each data is around 450, which indicates that there is obvious variation from state to state. As a consequence, we suggest that the producers' marketing strategy should different according to the different environmental situation of different states.

Variable	Mean	Std Dev	Minimum	Maximum	Range	N
spring	746.3258333	91.5824326	607.0000000	1079.00	472.0000000	50
Summer	898.3869444	100.6665815	729.0000000	1209.00	480.0000000	50
Fall	603.1241667	101.0335956	358.0000000	888.0000000	530.0000000	50
Winter	465.9113889	106.0886990	232.0000000	764.0000000	532.0000000	50

Figure 2 Statistics of average sunshine hours in four seasons in US

The descriptive analysis in Figure 3 shows that the population distribution composition is white, black, Hispanic and Asian. Among these four races, the white accounts for 69.32%. It can be concluded from the chart above that the main consumers of sunscreen products are white people. As a result, the manufacturer should focus on white people's preferences and their special need toward sunscreen products. However, to improve the market share, we suggest the manufacturer to try to reach out to people from different races.

Variable	Mean	Std Dev	Minimum	Maximum	N
White	69.3200000	14.8027025	37.0000000	94.0000000	50
Black	11.1200000	10.6265743	1.0000000	46.0000000	50
Hispanic	11.9800000	10.1085739	1.0000000	46.0000000	50
Asian	3.3200000	3.2035694	0	15.0000000	50
Two or more races	1.9800000	1.3626205	0	7.0000000	50

Figure 3 Statistics of population distribution of races in US

In 2016 annual median household income dataset, according to our statistical analysis result, which presented as Figure 4 and Figure 5, we can see the distribution is approaching normal, which means that most common annual household show in the interval between 50,000 and 60,000. Consequently, we suggest that manufacturer can notice this point when they make their pricing strategy.

Analysis Variable : Median Annual household income 2 Median Annual household income 2016					
Mean	Std Dev	Minimum	Maximum	Range	N
58919.92	8962.08	41099.00	76260.00	35161.00	50

Figure 4 Statistic of annual median household income in 2016

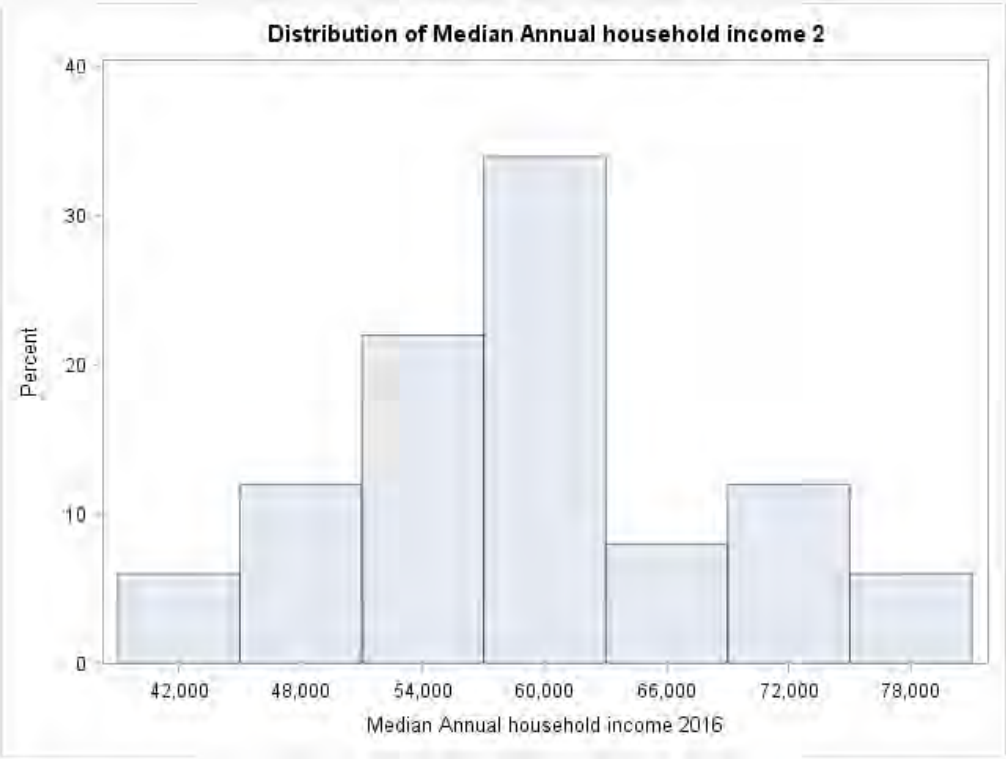


Figure 5 Histogram of median annual household income

In figure 6, it is obvious that the sales of sunscreen products aimed to body are more than to face. The sunscreen lotion is more popular than sunscreen spray. The sales of sunscreen products spf50- are more than those of spf55+. In a conclusion, we suggest manufacturers to produce normal spf50- sunscreen lotion.

Variable	Mean	Std Dev	Sum	N
BODY	9.4285714	15.3595790	462.0000000	49
FACE	1.5918367	2.9645009	78.0000000	49
SPF50-	7.5714286	13.2177280	371.0000000	49
SPF55+	3.4489796	5.2956792	169.0000000	49
LOTION	8.0816327	13.9997570	396.0000000	49
SPRAY	2.9387755	4.4693780	144.0000000	49
SENSITIVE	1.8775510	3.4134089	92.0000000	49
NORMAL	9.2040816	15.0982609	451.0000000	49

Figure 6 Statistics of features of sunscreen products

From figure 7, we make the following analysis.

H_0: $\beta_1=\beta_2=\beta_3=\beta_4= 0$

H_a: At least one of the regression coefficients is $\neq 0$

Conclusion: The overall model is significant as p-value is less than α (0.05). So, we reject the null and say at least one of the regression coefficients is not equal to zero. Unstandardized estimate is the change in the amount of dependent variable if we change the independent variable by one unit whereas standardized estimate is the change in the amount of the standard deviation of dependent variable due to a change of one standard deviation of independent variable. In this model unstandardized estimate of Asian is 2.81, that means one-unit change in sales of lotion sunscreen product will change quality of life by 2.81 unit. The estimate is different from the previous model since this one is a multiple regression model keeping other three predictors constant.

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Lotion

Number of Observations Read	49
Number of Observations Used	49

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	5389.06303	1347.26576	14.70	<.0001
Error	44	4033.75329	91.67621		
Corrected Total	48	9422.81633			

Root MSE	9.57477	R-Square	0.5719
Dependent Mean	8.06122	Adj R-Sq	0.5330
Coeff Var	118.77562		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.94780	49.76029	0.08	0.9371
White	1	-0.08563	0.52280	-0.16	0.8706
Black	1	-0.07802	0.49964	-0.16	0.8766
Hispanic	1	0.14233	0.60204	0.24	0.8142
Asian	1	2.80801	0.56538	4.97	<.0001

Figure 7 Linear regression result of lotion and different races

From figure 8, we make the following analysis.

H_0: $\beta_1=\beta_2=\beta_3=\beta_4= 0$

H_a: At least one of the regression coefficients is ≠ 0

Conclusion: The overall model is significant as p-value is less than α (0.05). So, we reject the null and say at least one of the regression coefficients is not equal to zero. Unstandardized estimate is the change in the amount of dependent variable if we change the independent variable by one unit whereas standardized estimate is the change in the amount of the standard deviation of dependent variable due to a change of one standard deviation of independent variable. In this model unstandardized estimate of Asian is 0.876, that means one-unit change in sales of spray sunscreen product will change quality of life by 0.876 unit. The estimate is different from the previous model since this one is a multiple regression model keeping other three predictors constant.

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Spray

Number of Observations Read	49
Number of Observations Used	49

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	451.38350	112.84587	9.78	<.0001
Error	44	507.43283	11.53256		
Corrected Total	48	958.81633			

Root MSE	3.39596	R-Square	0.4708
Dependent Mean	2.93878	Adj R-Sq	0.4227
Coeff Var	115.55707		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-8.85861	17.64890	-0.50	0.6182
White	1	0.08446	0.18542	0.46	0.6510
Black	1	0.10420	0.17721	0.59	0.5596
Hispanic	1	0.15494	0.21353	0.73	0.4719
Asian	1	0.87603	0.20053	4.37	<.0001

Figure 8 Linear regression result of spray and different races

From figure 9, we make the following analysis.

H_0: $\beta_1=\beta_2=\beta_3=\beta_4= 0$

H_a: At least one of the regression coefficients is ≠ 0

Conclusion: The overall model is significant as p-value is less than α (0.05). So, we reject the null and say at least one of the regression coefficients is not equal to zero. Unstandardized estimate is the change in the amount of dependent variable if we change the independent variable by one unit whereas standardized estimate is the change in the amount of the standard deviation of dependent variable due to a change of one standard deviation of

independent variable. In this model unstandardized estimate of Asian is 2.69, that means one-unit change in sales of SPF50- sunscreen product will change quality of life by 2.69 unit. The estimate is different from the previous model since this one is a multiple regression model keeping other three predictors constant.

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: SPF50-

Number of Observations Read	49
Number of Observations Used	49

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4762.01726	1190.50432	14.40	<.0001
Error	44	3638.10519	82.68421		
Corrected Total	48	8400.12245			

Root MSE	9.09309	R-Square	0.5669
Dependent Mean	7.55102	Adj R-Sq	0.5275
Coeff Var	120.42195		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3.03490	47.25697	-0.06	0.9491
White	1	-0.01079	0.49650	-0.02	0.9828
Black	1	0.00387	0.47451	0.01	0.9935
Hispanic	1	0.20198	0.57175	0.35	0.7256
Asian	1	2.69416	0.53694	5.02	<.0001

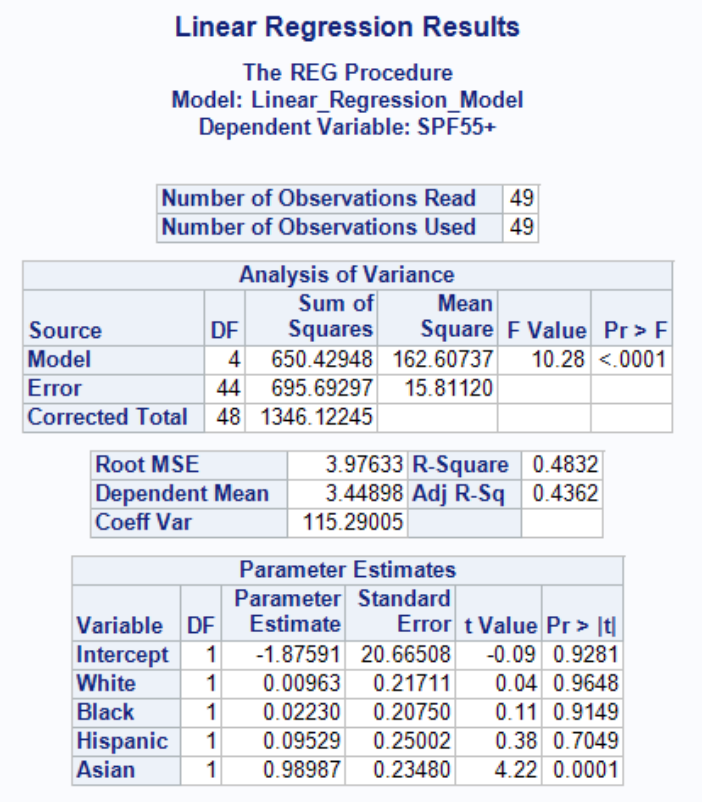
Figure 9 Linear regression result of SPF50- and different races

From figure 10, we make the following analysis.

H_0: $\beta_1=\beta_2=\beta_3=\beta_4= 0$

H_a: At least one of the regression coefficients is ≠ 0

Conclusion: The overall model is significant as p-value is less than α (0.05). So, we reject the null and say at least one of the regression coefficients is not equal to zero. Unstandardized estimate is the change in the amount of dependent variable if we change the independent variable by one unit whereas standardized estimate is the change in the amount of the standard deviation of dependent variable due to a change of one standard deviation of independent variable. In this model unstandardized estimate of Asian is 0.99, that means one-unit change in sales of SPF55+ sunscreen product will change quality of life by 0.99 unit. The estimate is different from the previous model since this one is a multiple regression model keeping other three predictors constant.



Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: SPF55+

Number of Observations Read	49
Number of Observations Used	49

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	650.42948	162.60737	10.28	<.0001
Error	44	695.69297	15.81120		
Corrected Total	48	1346.12245			

Root MSE	3.97633	R-Square	0.4832
Dependent Mean	3.44898	Adj R-Sq	0.4362
Coeff Var	115.29005		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.87591	20.66508	-0.09	0.9281
White	1	0.00963	0.21711	0.04	0.9648
Black	1	0.02230	0.20750	0.11	0.9149
Hispanic	1	0.09529	0.25002	0.38	0.7049
Asian	1	0.98987	0.23480	4.22	0.0001

Figure 10 Linear regression result of SPF55+ and different races

The null hypothesis in this case would be that there is no difference between sales of body sunscreen products and face sunscreen products. The alternate hypothesis would be there is difference between sales of body sunscreen products and face sunscreen products.

H_0: $\mu_{body} - \mu_{face} = 0$

H_a: $\mu_{body} - \mu_{face} \neq 0$

This is a two-tail hypothesis. Consider $\alpha = 0.05$. This hypothesis can be tested using the paired t-test. The results from the TTEST procedure are given below. The results show that the p-value is 0.0001, which is less than α. Therefore, we can reject the null hypothesis because we have strong statistical evidence from our sample that there is difference between sales of body sunscreen products and face sunscreen products.

t Test

The TTEST Procedure

Difference: Body - Face

N	Mean	Std Dev	Std Err	Minimum	Maximum
49	7.8163	12.7045	1.8149	-1.0000	76.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
7.8163	4.1672	11.4655	12.7045	10.5945	15.8717

DF	t Value	Pr > |t|
48	4.31	<.0001

Figure 11 Paired t Test of body sunscreen products and face sunscreen products

The null hypothesis in this case would be that there is no difference between sales of SPF50- sunscreen products and SPF55+ sunscreen products. The alternate hypothesis would be there is difference between sales of SPF50- sunscreen products and SPF55+ sunscreen products.

H_0: $\mu_{SPF50-} - \mu_{SPF55+} = 0$

H_a: $\mu_{SPF50-} - \mu_{SPF55+} \neq 0$

This is a two-tail hypothesis. Consider $\alpha = 0.05$. This hypothesis can be tested using the paired t-test. The results from the TTEST procedure are given below. The results show that the p-value is 0.0001, which is less than α. Therefore, we can reject the null hypothesis because we have strong statistical evidence from our sample that there is difference between sales of SPF50- sunscreen products and SPF55+ sunscreen products.

t Test

The TTEST Procedure

Difference: SPF50- - SPF55+

N	Mean	Std Dev	Std Err	Minimum	Maximum
49	4.1020	8.7945	1.2564	-6.0000	50.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
4.1020	1.5760	6.6281	8.7945	7.3339	10.9870

DF	t Value	Pr > |t|
48	3.27	0.0020

Figure 12 Paired t Test of SPF50- sunscreen products and SPF55+ sunscreen products

The null hypothesis in this case would be that there is no difference between sales of lotion sunscreen products and spray sunscreen products. The alternate hypothesis would be there is difference between sales of lotion sunscreen products and spray sunscreen products.

H_0: $\mu_{lotion} - \mu_{spray} = 0$

H_a: $\mu_{lotion} - \mu_{spray} \neq 0$

This is a two-tail hypothesis. Consider $\alpha = 0.05$. This hypothesis can be tested using the paired t-test. The results from the TTEST procedure are given below. The results show that the p-value is 0.0001, which is less than α. Therefore, we can reject the null hypothesis because we have strong statistical evidence from our sample that there is difference between sales of lotion sunscreen products and spray sunscreen products.

t Test

The TTEST Procedure

Difference: Lotion - Spray

N	Mean	Std Dev	Std Err	Minimum	Maximum
49	5.1224	10.1891	1.4556	-4.0000	66.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
5.1224	2.1958	8.0491	10.1891	8.4969	12.7292

DF	t Value	Pr > |t|
48	3.52	0.0010

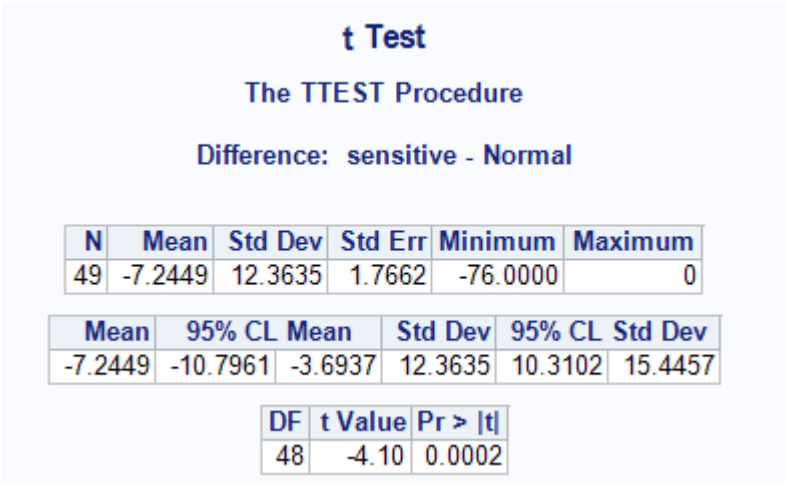
Figure 13 Paired t Test of lotion sunscreen products and spray sunscreen products

The null hypothesis in this case would be that there is no difference between sales of sensitive sunscreen products and normal sunscreen products. The alternate hypothesis would be there is difference between sales of sensitive sunscreen products and normal sunscreen products.

H_0: $\mu_{sensitive} - \mu_{normal} = 0$

H_a: $\mu_{sensitive} - \mu_{normal} \neq 0$

This is a two-tail hypothesis. Consider $\alpha = 0.05$. This hypothesis can be tested using the paired t-test. The results from the TTEST procedure are given below. The results show that the p-value is 0.0001, which is less than α. Therefore, we can reject the null hypothesis because we have strong statistical evidence from our sample that there is difference between sales of sensitive sunscreen products and normal sunscreen products.



t Test

The TTEST Procedure

Difference: sensitive - Normal

N	Mean	Std Dev	Std Err	Minimum	Maximum
49	-7.2449	12.3635	1.7662	-76.0000	0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
-7.2449	-10.7961	-3.6937	12.3635	10.3102	15.4457

DF	t Value	Pr > |t|
48	-4.10	0.0002

Figure 14 Paired t Test of sensitive sunscreen products and normal sunscreen products

From figure 14 we can see that based on the p-value in analysis of variance, which is less than 0.0001, we can reject the null hypothesis that all beta in the model are equal. As a result, the R-square shows that there are 70.12% of the data can be explained by the model. In the parameter estimates chart, we can see that only two variables' p-value are less than 0.05. Consequently, our linear model would be:

Sum Quantity of reviews = - 0.00056729 * Median annual household income + 4.56386 * percentage of Asian resident + 60.47743

The explanation of the model is that with keeping the median annual household income as a constant, increasing one percentage of Asian resident would increase one unit of sum quantity of reviews on Amazon about sunscreen products. If we keep the percentage of Asian residents as constant, one unit increment in median annual household income will decrease 0.00056729 units' reviews on Amazon about sunscreen product.

Also, we noticed that some of the variance inflation values that are larger than 10, such as average sunshine hour in fall, percentage of white people, black people and Hispanic people. So, we conclude that there is multicollinearity in our model.

Linear Regression Results

The REG Procedure
Model: Linear_Regression_Model
Dependent Variable: Sum-Quantity

Number of Observations Read	49
Number of Observations Used	49

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	11065	1229.40580	10.17	<.0001
Error	39	4715.34778	120.90635		
Corrected Total	48	15780			

Root MSE	10.99574	R-Square	0.7012
Dependent Mean	11.00000	Adj R-Sq	0.6322
Coeff Var	99.96130		

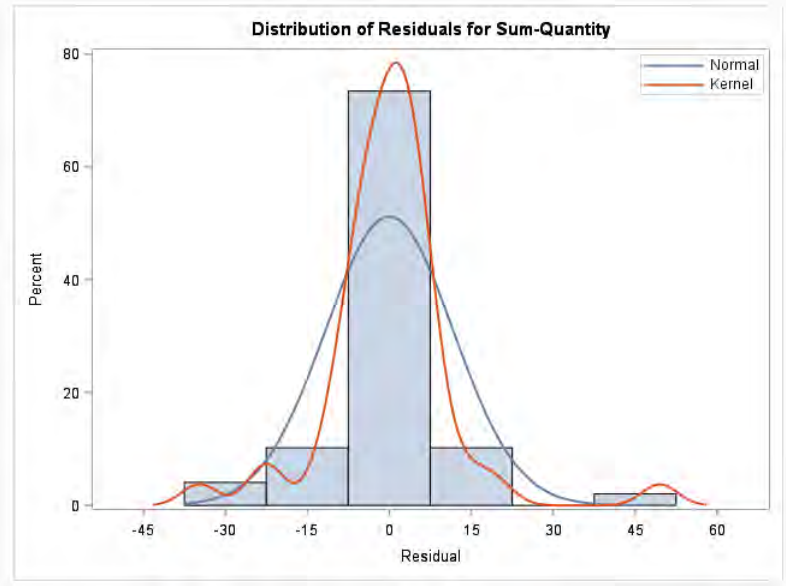
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	Intercept	1	60.47743	66.09668	0.91	0.3658	0	0
Spring		1	0.01925	0.04707	0.41	0.6848	0.09816	7.51932
Summer		1	-0.03226	0.03721	-0.87	0.3912	-0.17555	5.35059
Fall		1	-0.08849	0.06253	-1.42	0.1650	-0.46665	14.19440
Winter		1	0.08456	0.04181	2.02	0.0500	0.47890	7.31792
Median Annual household income 2	Median Annual household income 2016	1	-0.00056729	0.00027008	-2.10	0.0422	-0.27274	2.20061
White		1	-0.02230	0.63325	-0.04	0.9721	-0.01831	35.31718
Black		1	-0.23873	0.65063	-0.37	0.7157	-0.14050	19.13739
Hispanic		1	0.14608	0.74013	0.20	0.8446	0.08208	22.57041
Asian		1	4.56386	0.92009	4.96	<.0001	0.81239	3.50090

Generated by the SAS System ('Local', X64_8PRO) on December 06, 2017 at 5:48:40 PM

Figure 15 Linear regression result of sum quantity and all variables in our model in full model

From the histogram and Q-Q plot in figure 15, we can see that the data of sum quantity is normally distributed. The RStudent chart shows that there are several outliers.



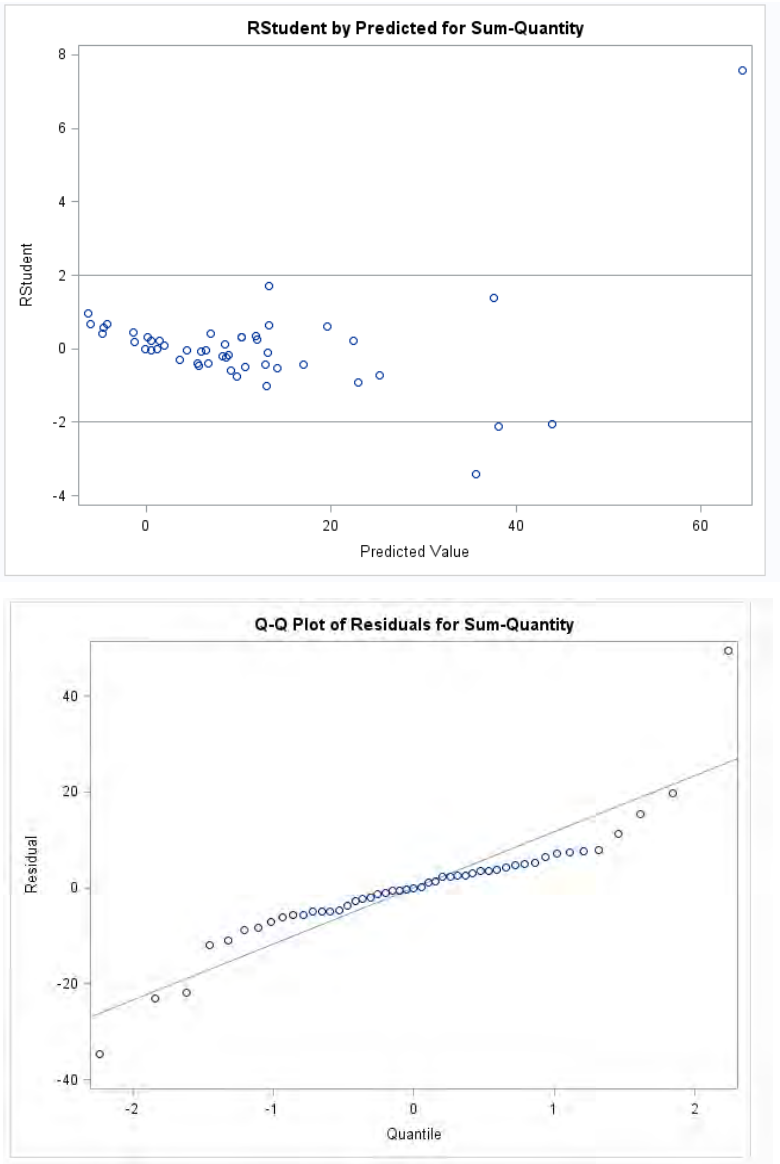


Figure 16 Histogram, RStudent and Q-Q plot for sum quantity

The results of our correlation analysis are shown in figure 17, from which we found out that average sunshine hour in spring 2016 is highly associated with average sunshine hour in fall and winter 2016. It has moderate association with average sunshine hour in summer 2016, percent of white people and Hispanic people. Median annual household income in 2016 and percentage of black people and Asian are low associated with it.

For average sunshine hour in summer 2016, average sunshine hour in fall and percentage of black people and Hispanic are moderately associated, while others are low associated.

For average sunshine hour in fall 2016, average sunshine hour in winter is highly associated, while percentage of white people and Hispanic are moderately associated. And median annual household income in 2016, percentage of black people and Asian are low associated.

Correlation Analysis

The CORR Procedure

9 Variables: Spring Summer Fall Winter Median Annual household income 2 White Black Hispanic Asian

Pearson Correlation Coefficients, N = 49
Prob > |r| under H0: Rho=0

	Spring	Summer	Fall	Winter	Median Annual household income 2	White	Black	Hispanic	Asian
Spring	1.00000	0.69108	0.90120	0.74434	-0.13934	-0.36785	-0.05862	0.53145	-0.05764
		<.0001	<.0001	<.0001	0.3396	0.0093	0.6891	<.0001	0.6940
Summer	0.69108	1.00000	0.56758	0.28909	0.04841	0.05041	-0.44719	0.30559	-0.11794
	<.0001		<.0001	0.0439	0.7412	0.7309	0.0013	0.0327	0.4196
Fall	0.90120	0.56758	1.00000	0.85980	-0.21385	-0.46414	0.05098	0.56501	-0.02855
	<.0001	<.0001		<.0001	0.1401	0.0008	0.7279	<.0001	0.8456
Winter	0.74434	0.28909	0.85980	1.00000	-0.09897	-0.48163	0.05335	0.57752	0.02441
	<.0001	0.0439	<.0001		0.4987	0.0005	0.7158	<.0001	0.8678
Median Annual household income 2	-0.13934	0.04841	-0.21385	-0.09897	1.00000	-0.04210	-0.17805	0.14012	0.56501
Median Annual household income 2016	0.3396	0.7412	0.1401	0.4987		0.7740	0.2210	0.3369	<.0001
White	-0.36785	0.05041	-0.46414	-0.48163	-0.04210	1.00000	-0.55408	-0.71562	-0.52533
	0.0093	0.7309	0.0008	0.0005	0.7740		<.0001	<.0001	0.0001
Black	-0.05862	-0.44719	0.05098	0.05335	-0.17805	-0.55408	1.00000	-0.14950	0.06714
	0.6891	0.0013	0.7279	0.7158	0.2210	<.0001		0.3052	0.6467
Hispanic	0.53145	0.30559	0.56501	0.57752	0.14012	-0.71562	-0.14950	1.00000	0.48526
	<.0001	0.0327	<.0001	<.0001	0.3369	<.0001	0.3052		0.0004
Asian	-0.05764	-0.11794	-0.02855	0.02441	0.56501	-0.52533	0.06714	0.48526	1.00000
	0.6940	0.4196	0.8456	0.8678	<.0001	0.0001	0.6467	0.0004	

Generated by the SAS System ('Local', X64_8PRO) on December 06, 2017 at 5:54:29 PM

Figure 17 Result of correlation analysis

CONCLUSION

From current analysis, first of all, the reviews on Amazon of sunscreen products are negatively correlated with annual household income, which might be caused by that consumer with high income would require higher quality of service. Secondly, reviews on Amazon of sunscreen products are positively correlated with percentage of Asian residents within a state. Thirdly, in each state, reviews on Amazon of sunscreen products with different features are different, so consumers have preferences toward different features. Moreover, Amazon reviews of sunscreen products with different features have a positive linear relationship with Asian residents in US. However, we cannot conclude the relationship between sunscreen sales and hours of sunshine, which is indicated by amount of reviews on Amazon in 2016.

Based on our descriptive analysis, we recommend, firstly, focusing on Southwestern region and trying to explore more types of demand about sunscreen products in this region. Secondly, focusing more on daily sunscreen products, rather than other featured types of sun care products.

According to our t-test analysis and linear regression model, we recommend that the manufacturers could pay more attention to targeting Asian consumers' needs in each state to acquire more sales potential. In the meantime, the various sale channels' concentration should be weighed by states that have more races and states that have fewer races. What's more, it is necessary for manufacturers to produce paired features sunscreen products, such as lotion/spray, sensitive/normal, SPF50-/SPF55+ and body/face. In addition, Amazon could try to improve their services in states that have rather high annual income and target consumers that have rather high income.

LIMITATION

From our current analysis, we cannot conclude the relationship between hours of sunshine and sunscreen sales, which is indicated by amount of reviews on Amazon in 2016. Besides, consumers do not tend to write reviews every time they consume products and not every customer fills in their Amazon profile with states. Even if they have states information, there is still a chance that they are not they were in 2016, when they wrote reviews. We have not collected all the data we plan to have, which are either real sales data of sunscreen products from various channel or some data like quantity of reviews from various websites. Whether the amount of reviews could reflect sales information properly should be reconsidered, since the reviews might neither be written by consumers nor write to show their opinions about the products. For future study, we would find sale data on various consuming channel to get the original targeted results.

REFERENCE

1. Xu, S., Kwa, M., Agarwal, A., Rademaker, A., & Kundu, R. V. (2016). Sunscreen product performance and other determinants of consumer preferences. *JAMA dermatology*, *152*(8), 920-927.

2. Amber, K. T., Bloom, R., Staropoli, P., Dhiman, S., & Hu, S. (2014). Assessing the current market of sunscreen: a cross-sectional study of sunscreen availability in three metropolitan counties in the United States. *Journal of skin cancer*, *2014*.

3. Akamine, K. L., Gustafson, C. J., Davis, S. A., Levender, M. M., & Feldman, S. R. (2014). Trends in sunscreen recommendation among US physicians. *JAMA dermatology*, *150*(1), 51-55.

4. https://www.currentresults.com

5. https://www.kff.org