Homework 1

Group 4

BUAN 6337 Predictive Analytics using SAS

Professor B.P.S. Murthi

Sept 4, 2020

Lin, Shih-Min

Devineni, Rajasekhar

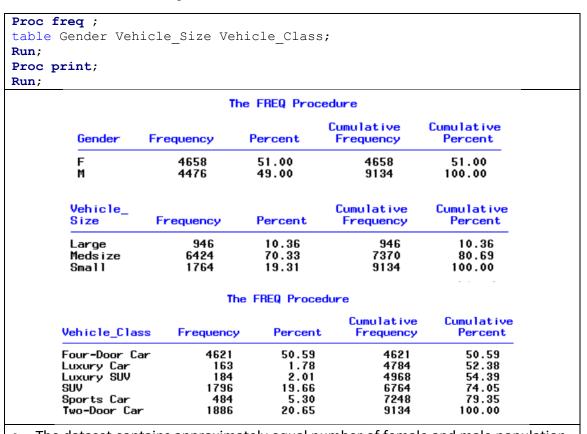
Talreja, Vinita Govind

Zhou, Wei

Read the data

```
Data a1;
Infile 'D:\UTDallas\OneDrive - The University of Texas at
Dallas\OCoursework\O BUAN 6337 - PREDICTIVE ANALYTICS USING
SAS\HW\car_insurance_19.csv' DLM=',' Missover firstobs=2;
Length Customer $ 12 State $ 12 Education $ 20 Employment_Status $ 20 Gender
$ 2 Policy_Type $ 15 Policy $ 15 Sales_Channel $ 12 Vehicle_Class $ 15;
Input Customer$ State$ Customer_Lifetime_Value Response$ Coverage$
Education$ Effective_To_Date$ Employment_Status$ Gender$ Income
Location_Code$ Marital_Status$ Monthly_Premium_Auto Months_Since_Last_Claim
Months_Since_Policy_Inception Number_of_Open_Complaints Number_of_Policies
Policy_Type$ Policy$ Renew_Offer_Type$ Sales_Channel$ Total_Claim_Amount
Vehicle_Class$ Vehicle_Size$;
Informat_Effective_To_Date_mmddyy10.;
Run;
```

1. What is the distribution of gender, vehicle size, and vehicle class?



- The dataset contains approximately equal number of female and male population.
- Medium sized cars are the most bought vehicles with over 70% share.
- Four door cars are popular among the vehicle class with 50% distribution.
- 2. What is the average customer lifetime value of each level of gender, vehicle size, and vehicle class?

```
Proc means;
var Customer_Lifetime_Value; class Gender;
Run;
Proc means;
var Customer_Lifetime_Value; class Vehicle_Size;
Run;
Proc means;
var Customer_Lifetime_Value; class Vehicle_Class;
Run;
```

The MEANS Procedure

Analysis Variable : Customer_Lifetime_Value

Gender	N Obs	N	Mean	Std Dev	Minimum	Maximum
F	4658	4658	8096.60	6956.06	1898.68	73225.96
М	4476	4476	7909.55	6780.74	1898.01	83325.38

The MEANS Procedure

Analysis Variable : Customer_Lifetime_Value

Vehicle_ Size	N Obs	N	Mean	Std Dev	Minimum	Maximum
Large	946	946	7545.00	6625.40	1940.98	60556.19
Medsize	6424	6424	8050.66	6833.10	1898.01	74228.52
Small	1764	1764	8085.10	7127.66	1898.68	83325.38

The MEANS Procedure

Analysis Variable : Customer_Lifetime_Value

Vehicle_Class	N Obs	N	Mean	Std Dev	Minimum	Maximum
Four - Door Car	4621	4621	6631.73	5164.94	1904.00	41787.90
Luxury Car	163	163	17053.35	12542.36	5886.22	83325.38
Luxury SUV	184	184	17123.00	12671.87	6383.61	73225.96
SUV	1796	1796	10443.51	7939.86	2864.82	58753.88
Sports Car	484	484	10750.99	8462.33	3074.11	67907.27
Two-Door Car	1886	1886	6671.03	5163.89	1898.01	38887.90

Classes	Level	Average Customer Lifetime Value		
Gender	Female	8096.60		
	Male	7909.55		
Vehicle size	Large	7545.00		
	Medsize	8050.66		
	Small	8085.10		
Vehicle class	Four-Door Car	6631.73		
	Luxury Car	17053.35		
	Luxury SUV	17123.00		
	SUV	10443.51		
	Sports Car	750.99		
	Two-Door Car	6671.03		

3. Do Large cars have a higher lifetime value than medsize cars. Do a ttest and report on your findings.



H₀: Larger cars have lifetime value less than or equal to medium size cars.

H₁: Larger cars have lifetime value greater than medium size cars.

- Assumption: Null hypothesis is true.
- Since p > 0.05 in Equality of Variances, it means that the variances are not significantly different. Hence, the pooled method applies.
- p=0.0329 for pooled method. Since p < 0.05, we reject the H₀.
- Hence larger cars have a higher lifetime value than medium-sized cars.
- 4. Is there a significant difference between men and women in customer lifetime value?

```
Proc ttest;
var Customer_Lifetime_Value;
class Gender;
Run;
```

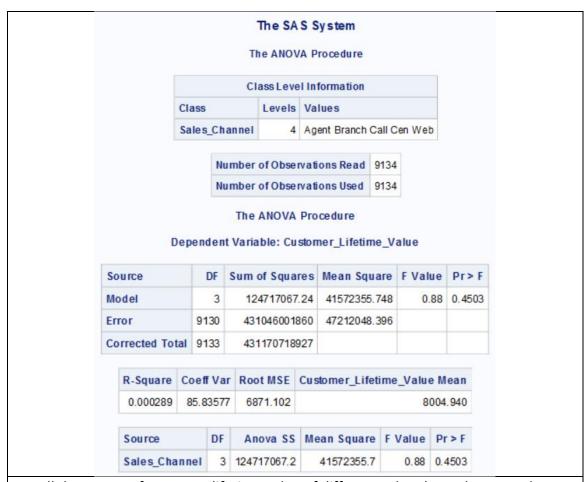
			The S	AS	Sys	tem	1					
		TI	ne TTE	ST	Proc	edur	re					
	Var	iable	: Custo	ome	r_Lif	etim	e_Va	lue				
Gender	Method	N	Mea	n S	td D	ev	Std E	rr	Min	imum	Ma	ximum
F		4658	8096	6	6956	5.1	101	9	4	898.7	7 7	73226.0
M		4476	7909	6	6780).7	101	4	4	898.0) (3325.4
Diff (1-2)	Pooled		187.	1	6870).7	143	8				
Diff (1-2)	Satterthwaite		187.	.1			143	7				
Gender	Method	Me	an	95% (CL M	lean	St	d D	ev	95% (CL St	d Dev
F		809	6.6	7896	.8 8	3296.	.4 (395	6.1	6817	7.6 7	100.3
M		790	9.6	7710	.9 8	3108.	.3 (378	0.7	6643	3.1 6	924.2
Diff (1-2)	Pooled	18	7.1 -9	4.84	77	468.	.9 (687	0.7	6772	2.5	971.8
Diff (1-2)	Satterthwaite	18	7.1 -9	4.70	43	468	.8					
	Method		/arian	ces		-	t Valı					
	Pooled		Equal	_		32			0.1			
	Satterthwa	iite (Jnequa	al	913	0.1	1.	30	0.1	932		
		Е	quality	y of \	Varia	nce	s					
	Method	Nu	m DF	Der	DF	F۷	/alue	P	r > F			
	Folded	F	4657	4	4475		1.05	0.	084	7		

H₀: No significant difference between men and women in customer lifetime value.

H₁: Has a significant difference between men and women in customer lifetime value.

- Equality of variance shows that p >0.05, which implies that we use pooled method.
- Pooled method shows a p-value > 0.05. We cannot reject the null hypothesis.
- There is no significant difference between men and women in customer lifetime value.
- 5. Use ANOVA to test whether there is difference in customer lifetime value across different sales channels. Which sales channel generates the highest lifetime value?

```
Proc ANOVA;
class Sales_Channel;
model Customer_Lifetime_Value=Sales_Channel;
Run;
```

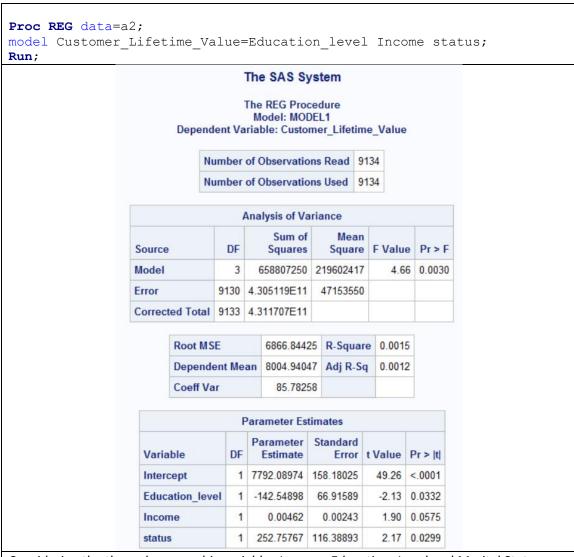


H₀: All the means of customer lifetime value of different sales channel are equal.

H₁: At least two means of customer lifetime value of different sales channel differ.

- Since p > 0.05, we do not reject the null hypothesis.
- Hence, we conclude that the means of customer lifetime value across different sales channel are equal. In other words, there isn't a specific sales channel that generates more lifetime value than the others.
- 6. What demographic factors (education, income, marital_status) affect customer lifetime value?

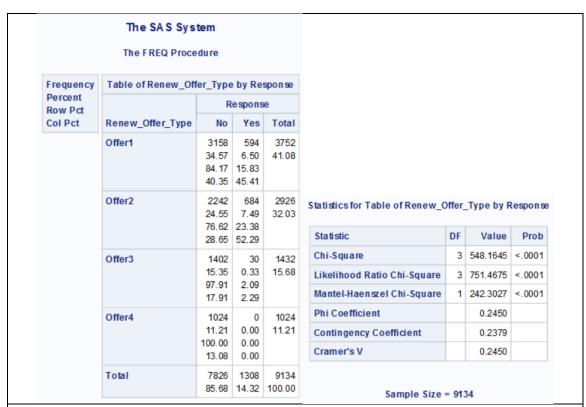
```
DATA A2;
SET a1;
if Marital_Status="Single" then status=0;
if Marital_Status="Married" then status=1;
if Marital_Status="Divorced" then status=2;
if Education="High School or Below" then Education_level=0;
if Education="College" then Education_level=1;
if Education="Bachelor" then Education_level=2;
if Education="Master" then Education_level=3;
if Education="Doctor" then Education_level=4;
run;
```



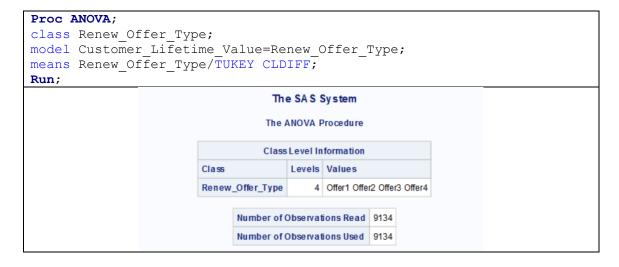
Considering the three demographic variables Income, Education_Level and Marital Status, we find that Education level and Marital Status affect Customer Lifetime Value in 95% of the cases, while Income variable affects Customer Lifetime Value in 93% of the cases.

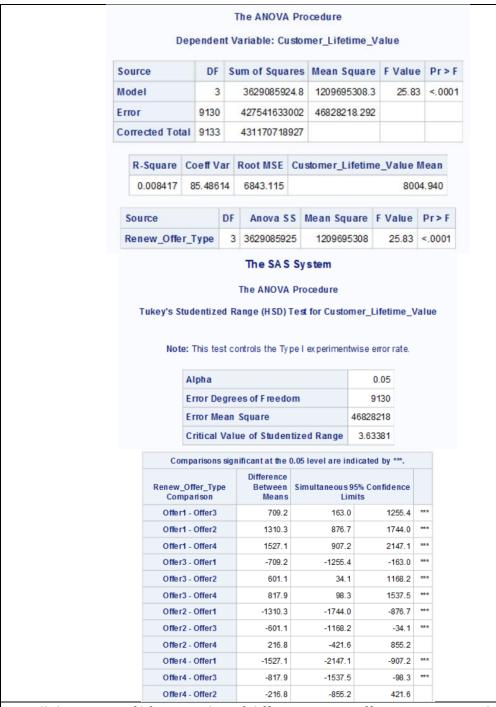
7. Is there a relationship between renew_offer_type and response (use Chi-sq test)? Which offer type generates the highest response rate?

```
Proc freq ;
tables Renew_Offer_Type*Response/chisq;
Run;
```



- There is a relationship between renew offer type and response since the calculated chi-sq value (548.1645) is greater than the table chi-sq value at 95% confidence.
- Offer 2 generates the most "yes" responses compared to other offers with a value of 23.38%.
- 8. Do different renew_offer_types have different lifetime values? Which offer type is the best?



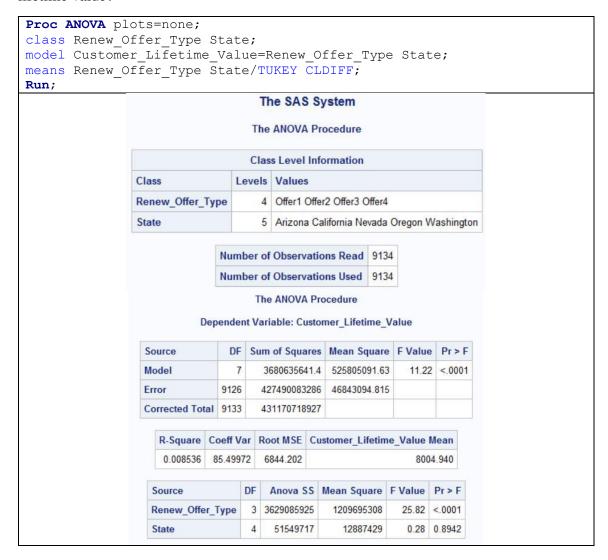


H₀: All the means of lifetime value of different renew offer types are equal.

H₁: At least two means of lifetime value of different renew offer types are <u>different</u>.

- Since p < 0.05, we reject the null hypothesis. Hence, at least of two means of lifetime value of different renew offer types differ.
- Tukey HSD test is used to find the best among the offers. The difference between all possible combinations of different offers shows that Offer1 is the best.

9. Is the effectiveness of renew_offer_type different across different states with respect to lifetime value?



Alpha	0.05
Error Degrees of Freedom	9126
Error Mean Square	46843095
Critical Value of Studentized Range	3.63381

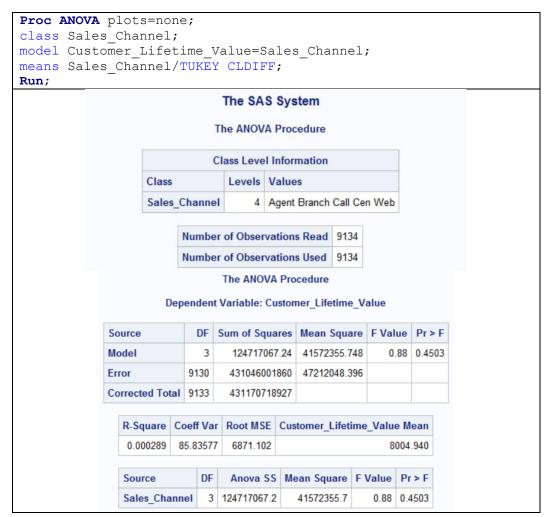
	Confidence	Simultaneous 95%	Difference Between Means	Renew_Offer_Type Comparison
**	1255.5	162.9	709.2	Offer1 - Offer3
**	1744.1	876.6	1310.3	Offer1 - Offer2
**	2147.2	907.1	1527.1	Offer1 - Offer4
**	-162.9	-1255.5	-709.2	Offer3 - Offer1
**	1168.3	34.0	601.1	Offer3 - Offer2
**	1537.7	98.2	817.9	Offer3 - Offer4
**	-876.6	-1744.1	-1310.3	Offer2 - Offer1
**	-34.0	-1168.3	-601.1	Offer2 - Offer3
	855.3	-421.7	216.8	Offer2 - Offer4
**	-907.1	-2147.2	-1527.1	Offer4 - Offer1
**	-98.2	-1537.7	-817.9	Offer4 - Offer3
	421.7	-855.3	-216.8	Offer4 - Offer2

Alpha	0.05
Error Degrees of Freedom	9126
Error Mean Square	46843095
Critical Value of Studentized Range	3.85843

State Comparison	Difference Between Means	Simultaneous 95% Confidence Limits			
Oregon - Nevada	21.2	-706.4	748.8		
Oregon - Washington	56.4	-699.2	812.1		
Oregon - California	74.3	-420.5	569.0		
Oregon - Arizona	216.6	-365.5	798.6		
Nevada - Oregon	-21.2	-748.8	706.4		
Nevada - Washington	35.2	-877.1	947.5		
Nevada - California	53.1	-658.3	764.4		
Nevada - Arizona	195.4	-579.3	970.0		
Washington - Oregon	-56.4	-812.1	699.2		
Washington - Nevada	-35.2	-947.5	877.1		
Washington - California	17.8	-722.2	757.9		
Washington - Arizona	160.1	-640.9	961.2		
California - Oregon	-74.3	-569.0	420.5		
California - Nevada	-53.1	-764.4	658.3		
California - Washington	-17.8	-757.9	722.2		
California - Arizona	142.3	-419.3	703.9		
Arizona - Oregon	-216.6	-798.6	365.5		
Arizona - Nevada	-195.4	-970.0	579.3		
Arizona - Washington	-160.1	-961.2	640.9		
Arizona - California	-142.3	-703.9	419.3		

Higher the customer lifetime value, more effective is the renew offer. From the ANOVA test results, we conclude that the effectiveness of renew offer types is not significantly different across states.

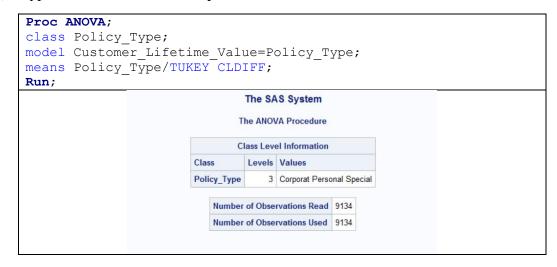
- 10. What other interesting insights that are useful to the company in terms of action can be obtained from the data?
 - a. Write any three (3) hypotheses. The hypotheses should be useful to the insurance firm.
 - b. Do appropriate statistical tests or analysis.
 - c. Report what you found in each case and also write how management can use this information to improve their operations.
 - 1) Hypotheses 1: Different sales channel have different lifetime values.

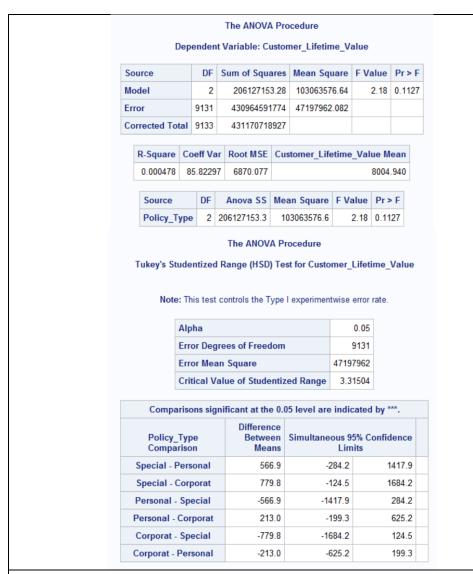




H₀: All the means of lifetime value of different sales channel are equal. H1: At least two means of lifetime value of different sales channel are different.

- Since p > 0.05, we do not reject the null hypothesis.
- There are no significant differences between different sales channel. The insurance company can invest equally among different sales channel.
- 2) Hypotheses 2: Different auto policies have different lifetime values.



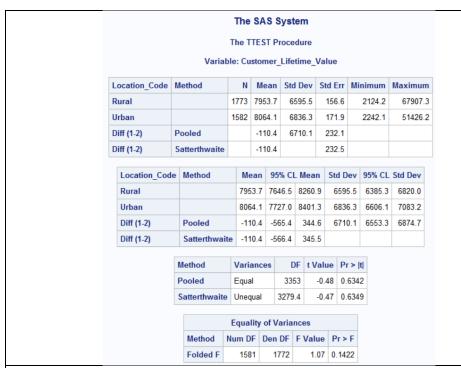


H₀: All the means of customer lifetime value of different auto policies are equal.

H1: At least two means of customer lifetime value of different auto policies are different.

- Since p > 0.05 we do not reject the null hypothesis.
- There are no significant differences between different auto policies. The insurance company can develop equally among different auto policies.
- 3) Hypotheses 3: Rural customers are less valuable than Urban customers

```
Proc ttest;
var Customer_Lifetime_Value;
class Location_Code; where Location_Code in ("Rural", "Urban");
Run;
```



H₀: No significant difference between rural and urban customers in terms of customer lifetime value.

H₁: There is a significant difference between rural and urban customers in terms of customer lifetime value.

- Equality of variance shows that p >0.05 which means that it is a pooled variance.
- Pooled method shows a p-value > 0.05. We cannot reject the null hypothesis.
- We conclude that there is no significant difference in rural and urban customers in terms of customer lifetime value. Hence, the insurance company should focus on both rural and urban customers equally.