PREDICTIVE ANALYSIS USING SAS - S19 (BUAN 6337.001)

HOMEWORK 1

GROUP-3

JAINIK PATEL
KANAK KANTI ROY
PRATHAMESH RANJANKUMAR SHINDE
SHRUTI RAJANI
AISHWARYA BADLANI
NITHIN NAIR

DATE : 01/29/2018

Reading Insurance Claims Dataset

```
proc import
datafile = 'h:\car_insurance_19.csv'
out =carins
dbms = CSV ;
run;
```

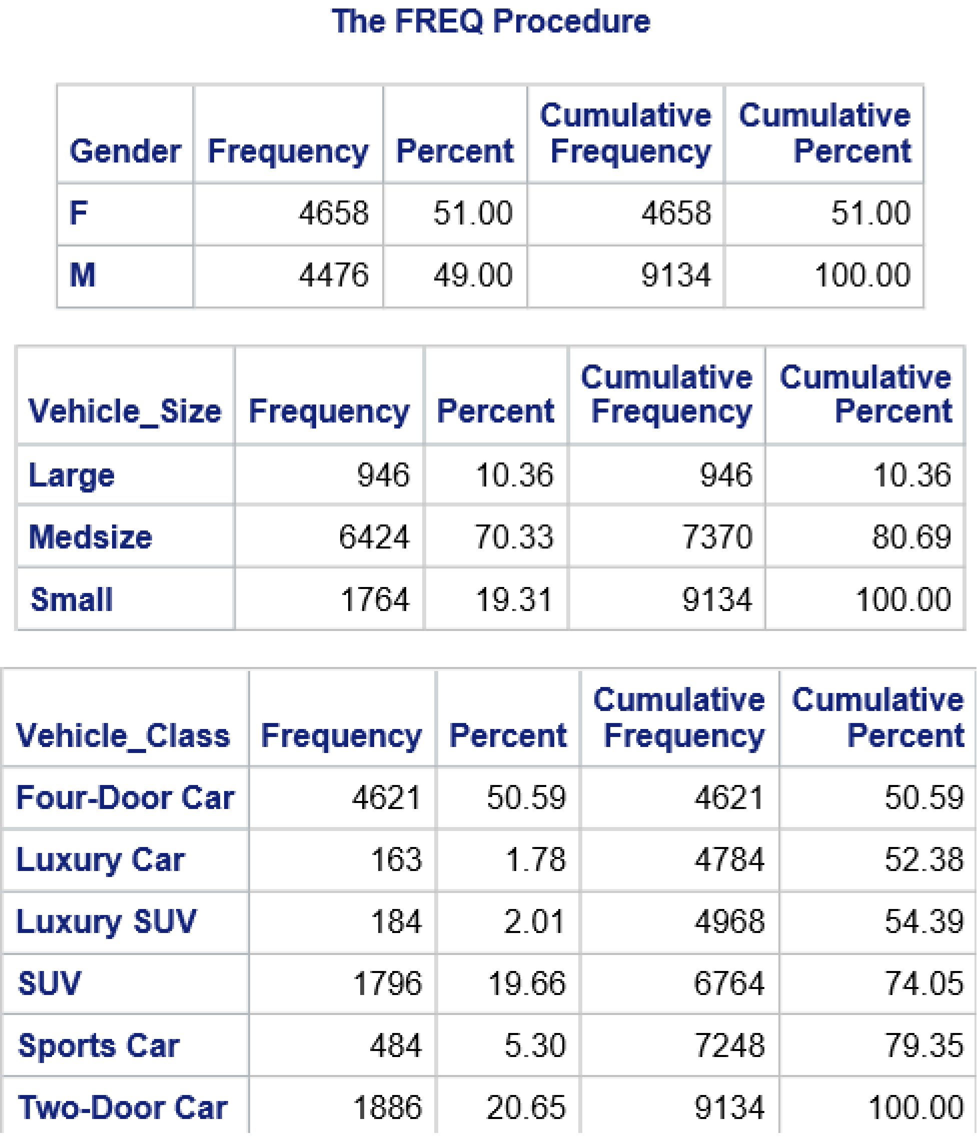
Log Output:

```
9134 rows created in WORK.CARINS from h:\car_insurance_19.csv.

NOTE: WORK.CARINS data set was successfully created.
NOTE: The data set WORK.CARINS has 9134 observations and 24 variables.
NOTE: PROCEDURE IMPORT used (Total process time):
      real time           0.68 seconds
      cpu time            0.37 seconds
```

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

```
proc freq data=carins;
Table Gender Vehicle_Size Vehicle_Class;
run;
```



The FREQ Procedure

Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
F	4658	51.00	4658	51.00
M	4476	49.00	9134	100.00

Vehicle_Size	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Large	946	10.36	946	10.36
Medsize	6424	70.33	7370	80.69
Small	1764	19.31	9134	100.00

Vehicle_Class	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Four-Door Car	4621	50.59	4621	50.59
Luxury Car	163	1.78	4784	52.38
Luxury SUV	184	2.01	4968	54.39
SUV	1796	19.66	6764	74.05
Sports Car	484	5.30	7248	79.35
Two-Door Car	1886	20.65	9134	100.00

**** Gender is almost evenly distributed, however there are more females (51 %) than males (49 %) in the dataset.**

**** Most of the vehicles are Medium-sized (70.33 %). Small sized vehicles account for 19.31% of the data. The number of large vehicles is quite less**

(946) ~ almost 1/7th of the number of medium-sized vehicles and account for only 10.36 % of the data.

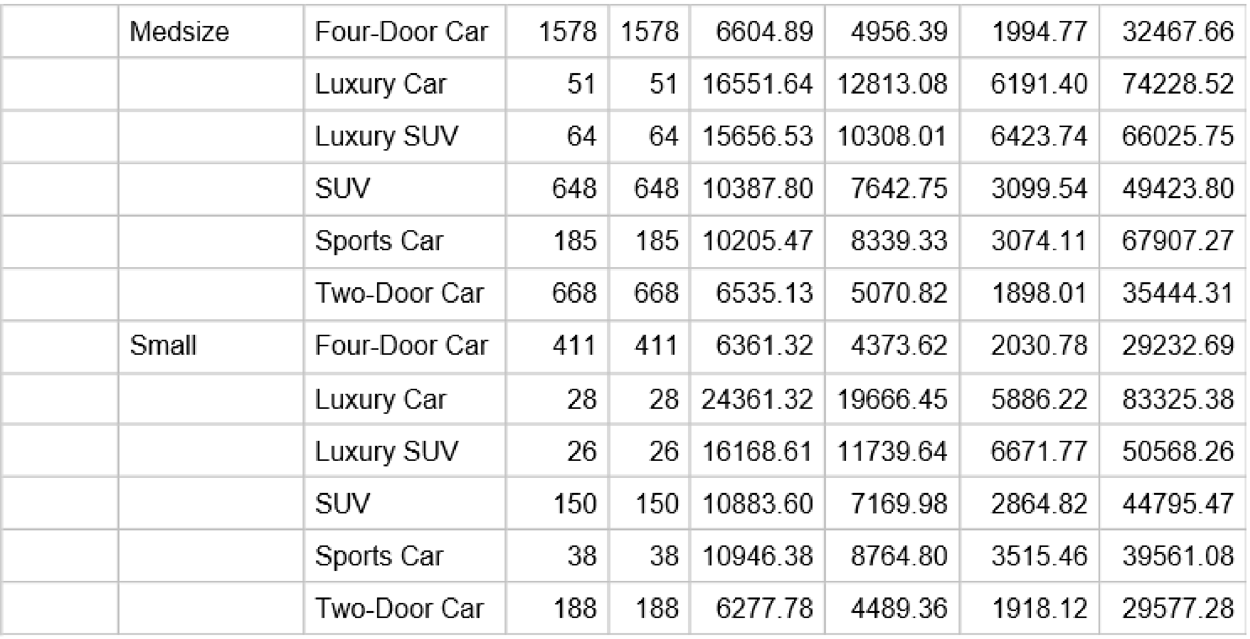
**** Almost half of the vehicles are Four-Door Cars (50.59 %). The majority among the remaining vehicles are- Two-Door Cars (20.65 %) and SUV (19.66 %). The Sports Car, Luxury SUV and Luxury Car are quite rare and account for 5.3 %, 2.01% and 1.78 % of the data respectively.**

2) What is the average customer life time value of each level of gender, vehicle size, and vehicle class?

```
proc means data=carins; var Customer_Lifetime_Value; Class Gender Vehicle_Size Vehicle_Class;Run;
```

The MEANS Procedure

Analysis Variable : Customer_Lifetime_Value								
Gender	Vehicle_Size	Vehicle_Class	N Obs	N	Mean	Std Dev	Minimum	Maximum
F	Large	Four-Door Car	249	249	6596.15	4753.13	2111.99	27564.74
		Luxury Car	7	7	13152.99	5183.70	7373.23	21435.88
		Luxury SUV	7	7	28847.15	21236.57	7449.86	60556.19
		SUV	91	91	9441.19	7539.97	3853.47	51337.91
		Sports Car	30	30	11161.95	6318.59	4062.00	35537.85
		Two-Door Car	121	121	6637.54	5118.68	2336.29	27528.31
	Medsize	Four-Door Car	1659	1659	6748.67	5503.89	1904.00	41787.90
		Luxury Car	55	55	14437.68	7992.32	6698.97	51426.25
		Luxury SUV	61	61	17888.00	13980.05	6991.25	73225.96
		SUV	660	660	10572.28	8322.12	3371.53	58753.88
		Sports Car	181	181	11542.64	9010.80	3595.31	40132.01
		Two-Door Car	614	614	7028.99	5454.13	2147.66	38887.90
	Small	Four-Door Car	498	498	6820.34	5637.46	2004.35	36470.30
		Luxury Car	13	13	18922.65	7945.75	7255.14	25807.06
		Luxury SUV	15	15	16917.91	9972.78	6383.61	46770.95
		SUV	171	171	10436.55	7879.10	3451.10	51016.07
		Sports Car	31	31	9801.49	6596.88	3884.86	26900.27
		Two-Door Car	195	195	6828.67	5781.18	1898.68	35186.26
M	Large	Four-Door Car	226	226	6075.99	4665.63	2052.95	35944.71
		Luxury Car	9	9	13478.59	6256.67	7126.60	22837.14
		Luxury SUV	11	11	16487.56	15022.67	6674.18	58207.13
		SUV	76	76	10147.42	9132.98	3123.08	46611.87
		Sports Car	19	19	9030.71	9463.37	3954.34	40636.67
		Two-Door Car	100	100	5853.44	3610.24	1940.98	22563.62

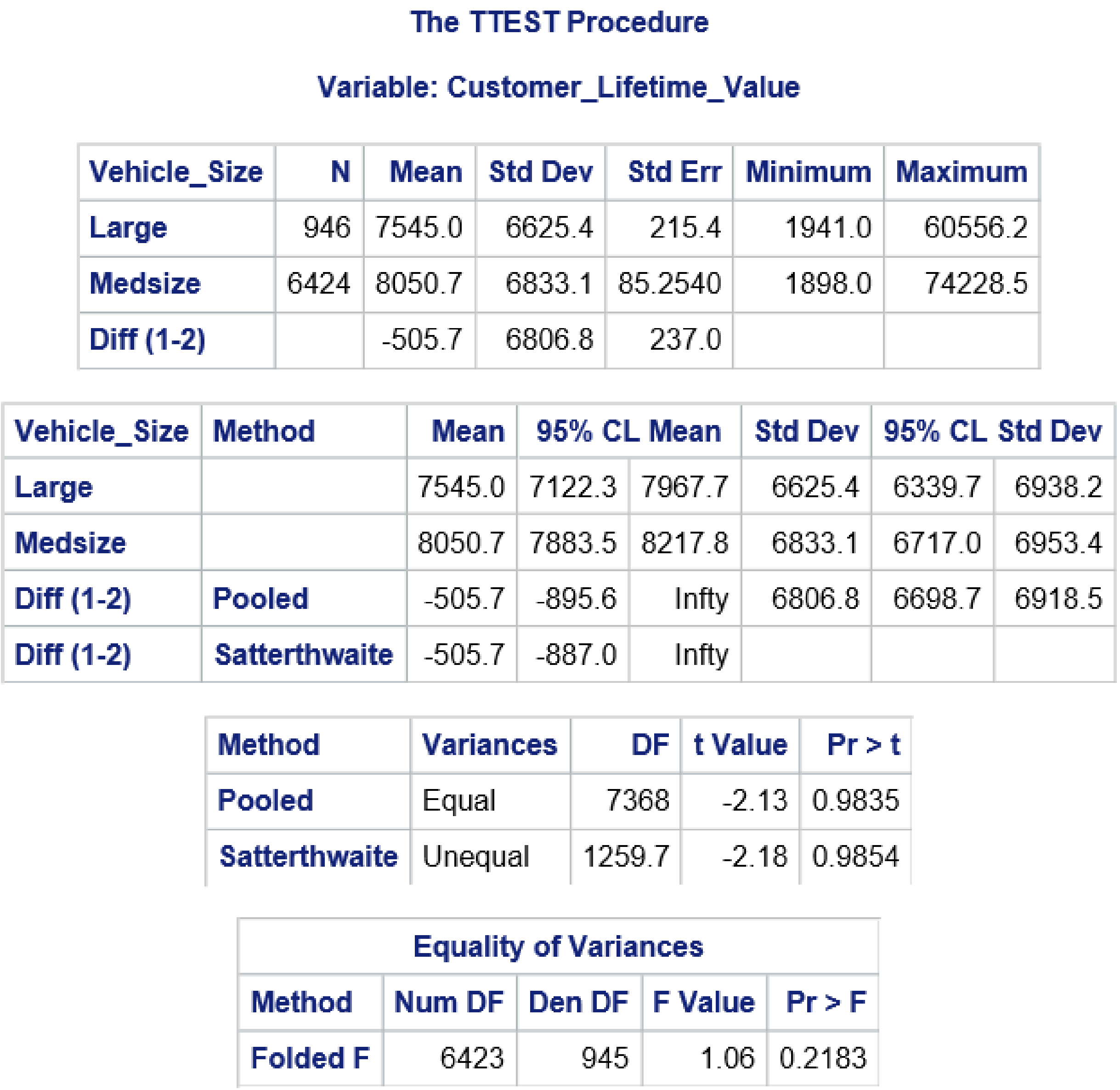


	Medsize	Four-Door Car	1578	1578	6604.89	4956.39	1994.77	32467.66
		Luxury Car	51	51	16551.64	12813.08	6191.40	74228.52
		Luxury SUV	64	64	15656.53	10308.01	6423.74	66025.75
		SUV	648	648	10387.80	7642.75	3099.54	49423.80
		Sports Car	185	185	10205.47	8339.33	3074.11	67907.27
		Two-Door Car	668	668	6535.13	5070.82	1898.01	35444.31
	Small	Four-Door Car	411	411	6361.32	4373.62	2030.78	29232.69
		Luxury Car	28	28	24361.32	19666.45	5886.22	83325.38
		Luxury SUV	26	26	16168.61	11739.64	6671.77	50568.26
		SUV	150	150	10883.60	7169.98	2864.82	44795.47
		Sports Car	38	38	10946.38	8764.80	3515.46	39561.08
		Two-Door Car	188	188	6277.78	4489.36	1918.12	29577.28

**** The average customer lifetime value is maximum for females using Large, Luxury SUV Cars and it's 28847.15 . The average customer lifetime value is minimum for males using Large, Two-Door Cars and it's 5853.44.**

<u>3) Do Large cars have a higher lifetime value than medsize cars. Do a ttest and report on your findings.</u>

```
data a2; set carins;if Vehicle_Size="Medsize" or
Vehicle_Size="Large";
proc ttest sides=u;var Customer_Lifetime_Value;class
Vehicle_Size;data a2;run;
```



The TTEST Procedure

Variable: Customer_Lifetime_Value

Vehicle_Size	N	Mean	Std Dev	Std Err	Minimum	Maximum
Large	946	7545.0	6625.4	215.4	1941.0	60556.2
Medsize	6424	8050.7	6833.1	85.2540	1898.0	74228.5
Diff (1-2)		-505.7	6806.8	237.0		

Vehicle_Size	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
Large		7545.0	7122.3	7967.7	6625.4	6339.7	6938.2
Medsize		8050.7	7883.5	8217.8	6833.1	6717.0	6953.4
Diff (1-2)	Pooled	-505.7	-895.6	Infty	6806.8	6698.7	6918.5
Diff (1-2)	Satterthwaite	-505.7	-887.0	Infty			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	7368	-2.13	0.9835
Satterthwaite	Unequal	1259.7	-2.18	0.9854

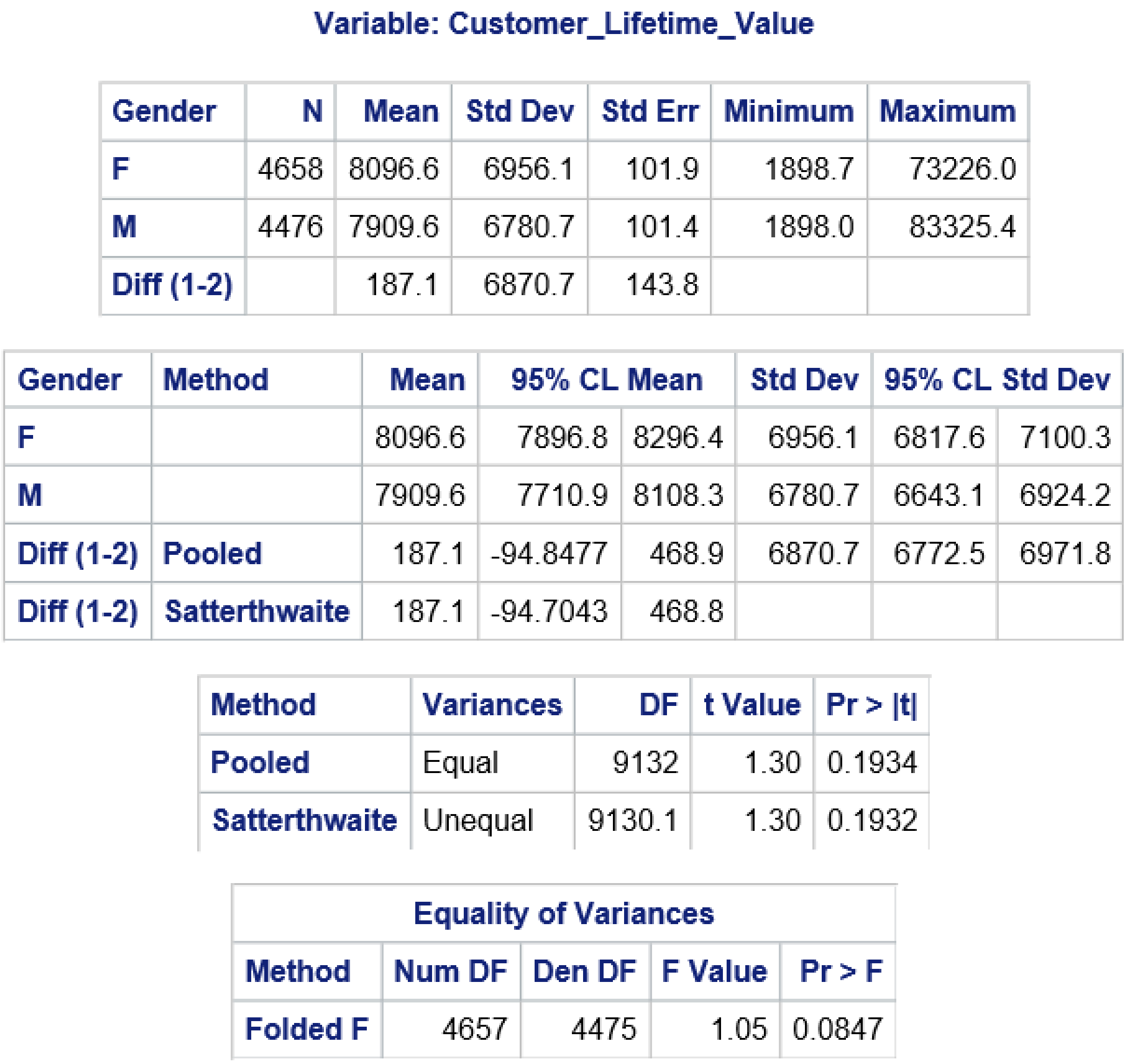
Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	6423	945	1.06	0.2183

**** F- Test result: P value for F-test =0.2183 >0.05, hence we cannot reject null hypothesis of F-test => We have to go ahead with T-test for equal variances.**

P value for the 1-tailed test = 0.9835 > 0.05.
Failing to reject H0, the data suggests that Large cars are not likely to have a higher lifetime value than Medium size cars.

<u>4) Is there a significant difference between men and women in customer lifetime value?</u>

```
data a3; set carins;
proc ttest;var Customer_Lifetime_Value;class Gender;data a3;run;
```



Variable: Customer_Lifetime_Value

Gender	N	Mean	Std Dev	Std Err	Minimum	Maximum
F	4658	8096.6	6956.1	101.9	1898.7	73226.0
M	4476	7909.6	6780.7	101.4	1898.0	83325.4
Diff (1-2)		187.1	6870.7	143.8		

Gender	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
F		8096.6	7896.8	8296.4	6956.1	6817.6	7100.3
M		7909.6	7710.9	8108.3	6780.7	6643.1	6924.2
Diff (1-2)	Pooled	187.1	-94.8477	468.9	6870.7	6772.5	6971.8
Diff (1-2)	Satterthwaite	187.1	-94.7043	468.8			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	9132	1.30	0.1934
Satterthwaite	Unequal	9130.1	1.30	0.1932

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	4657	4475	1.05	0.0847

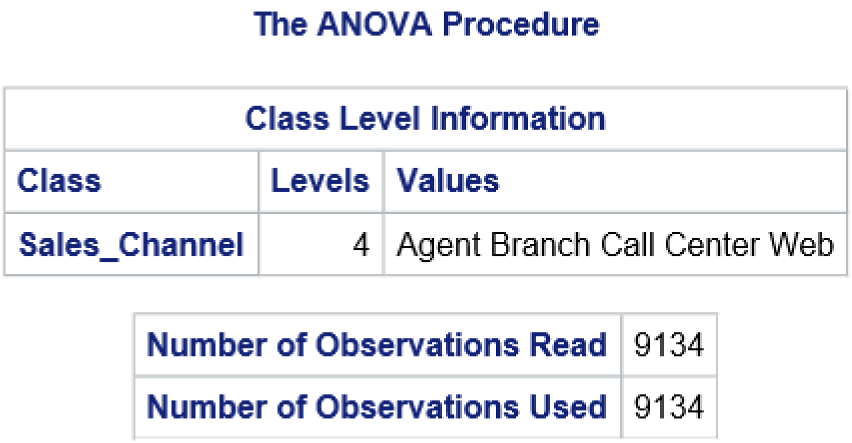
**** Based on the data, we conduct a hypothesis test (with a 0.05 significance level) to see if there is evidence that there is a significant difference between male and Female in customer lifetime value using *α*= 0.05.**

The last part of the result is to check whether the variances of the two groups are equal. Simple rule applied here: Since the probability of F-Test is 0.0847 (greater than 0.05), then the variances are equal and we use the results for the equal variances. The corresponding t-value = 1.30, and the p-value is 0.1934.

The conclusion is that we cannot reject the null hypothesis and don't have enough evidence that the Customer Lifetime Value between male and female are significantly different.

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?

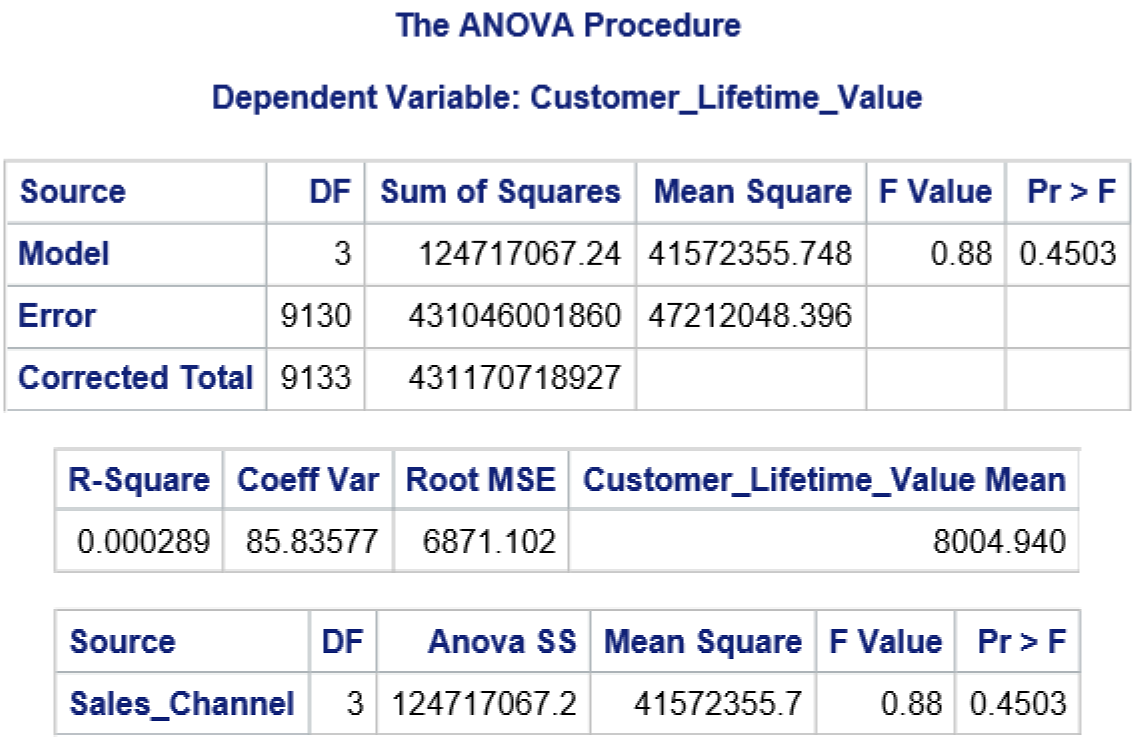
```
proc anova data=carins; class Sales_Channel;model
Customer_Lifetime_Value = Sales_Channel;run;
proc means data=carins;class Sales_Channel;var
Customer_Lifetime_Value;run;
```



The ANOVA Procedure

Class Level Information		
Class	Levels	Values
Sales_Channel	4	Agent Branch Call Center Web

Number of Observations Read	9134
Number of Observations Used	9134



The ANOVA Procedure

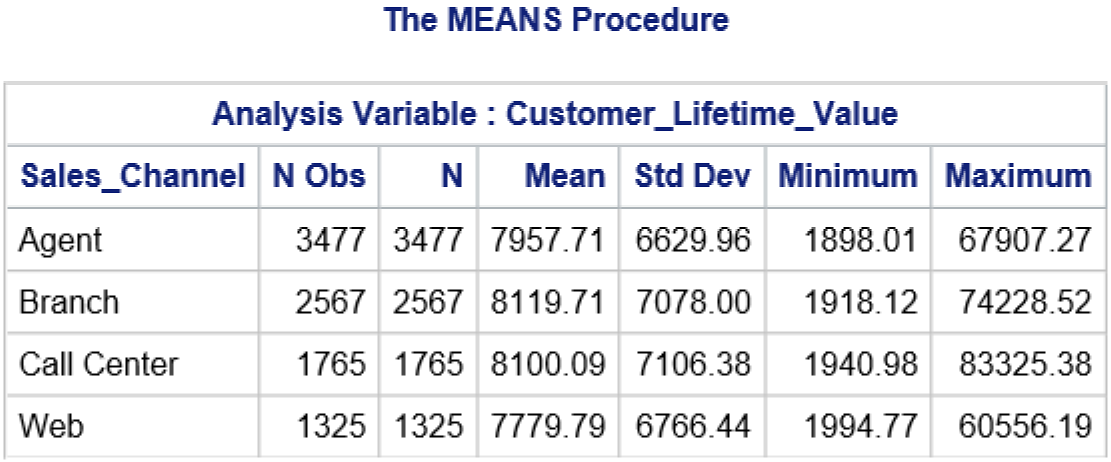
Dependent Variable: Customer_Lifetime_Value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	124717067.24	41572355.748	0.88	0.4503
Error	9130	431046001860	47212048.396		
Corrected Total	9133	431170718927			

R-Square	Coeff Var	Root MSE	Customer_Lifetime_Value Mean
0.000289	85.83577	6871.102	8004.940

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Sales_Channel	3	124717067.2	41572355.7	0.88	0.4503

**** The P-Value of the ANOVA test is 0.4503 (>0.05), which implies that we fail to reject the Null hypothesis. It further implies that mean Customer Lifetime value is same across all Sales Channels.**



The MEANS Procedure

Analysis Variable : Customer_Lifetime_Value

Sales_Channel	N Obs	N	Mean	Std Dev	Minimum	Maximum
Agent	3477	3477	7957.71	6629.96	1898.01	67907.27
Branch	2567	2567	8119.71	7078.00	1918.12	74228.52
Call Center	1765	1765	8100.09	7106.38	1940.98	83325.38
Web	1325	1325	7779.79	6766.44	1994.77	60556.19

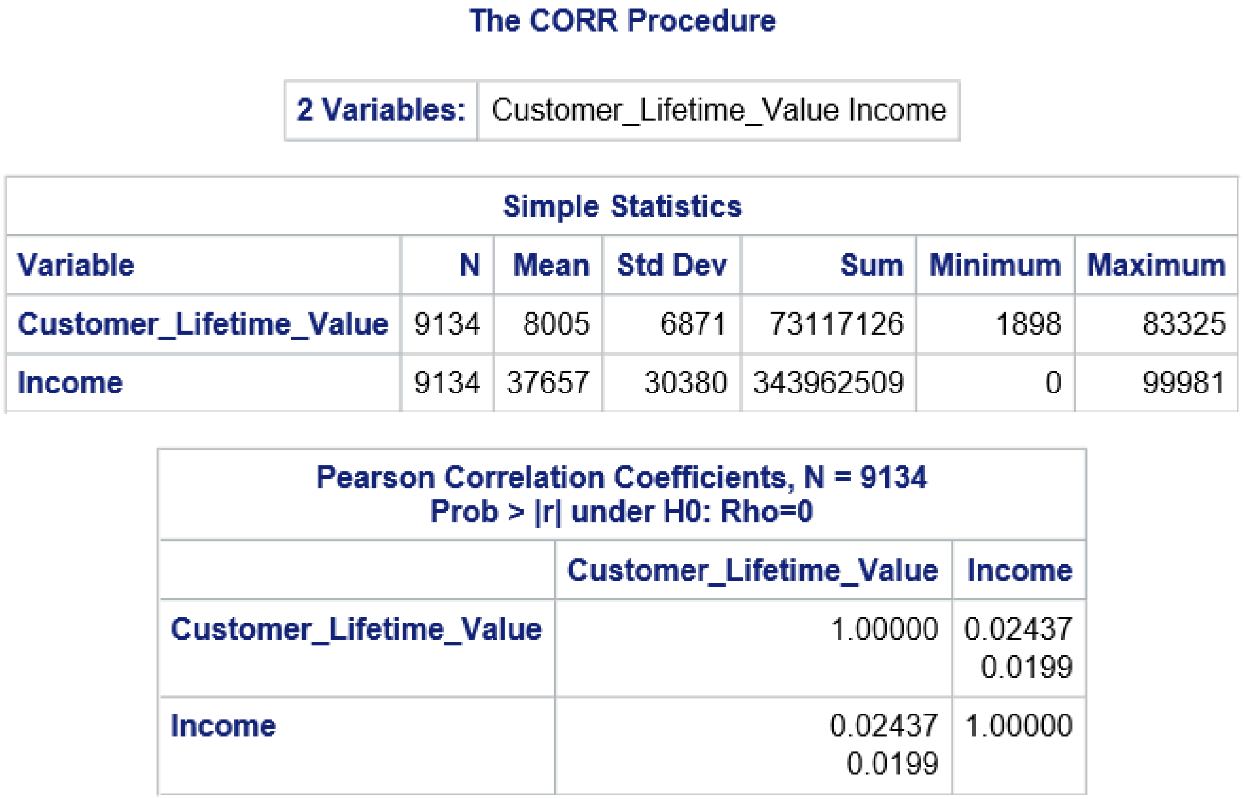
**** Although Call Center records the highest Customer Lifetime Value (83325.38) for one particular case, Agent is the most preferred Sales Channel and is responsible for a major portion of the Customer Lifetime Value among all the other sales Channel.**

This is because the mean Customer Lifetime Value for Agent (7957.71) is closest to the Average Customer Lifetime value of the entire dataset (8004.940). Also, the standard deviation for Agent is least (6629.96) among the Standard deviation of all other Sales Channels.

Moreover, Agent raises the highest Customer Lifetime Value among all the other channels (3477*7957.71).

6) What demographic factors (education, income, marital_status,state,Location_Code) affect customer lifetime value?

```
proc corr data=carins;var Customer_Lifetime_Value Income;run;
proc univariate data=carins;var Customer_Lifetime_Value;run;
data a4;set carins;
if Customer_Lifetime_Value le 4000 then clv=1;
if Customer_Lifetime_Value ge 9000 then clv=3;
if Customer_Lifetime_Value gt 4000 and Customer_Lifetime_Value lt 9000
then clv=2;run;
proc freq data=a4;table Education*clv/CHISQ;run;
proc freq data=a4;table Marital_Status*clv/CHISQ;run;
proc freq data=a4;table State*clv/CHISQ;run;
proc freq data=a4;table Location_Code*clv/CHISQ;run;
proc freq data=a4;table Gender*clv/CHISQ;run;
```



The CORR Procedure

2 Variables:	Customer_Lifetime_Value Income

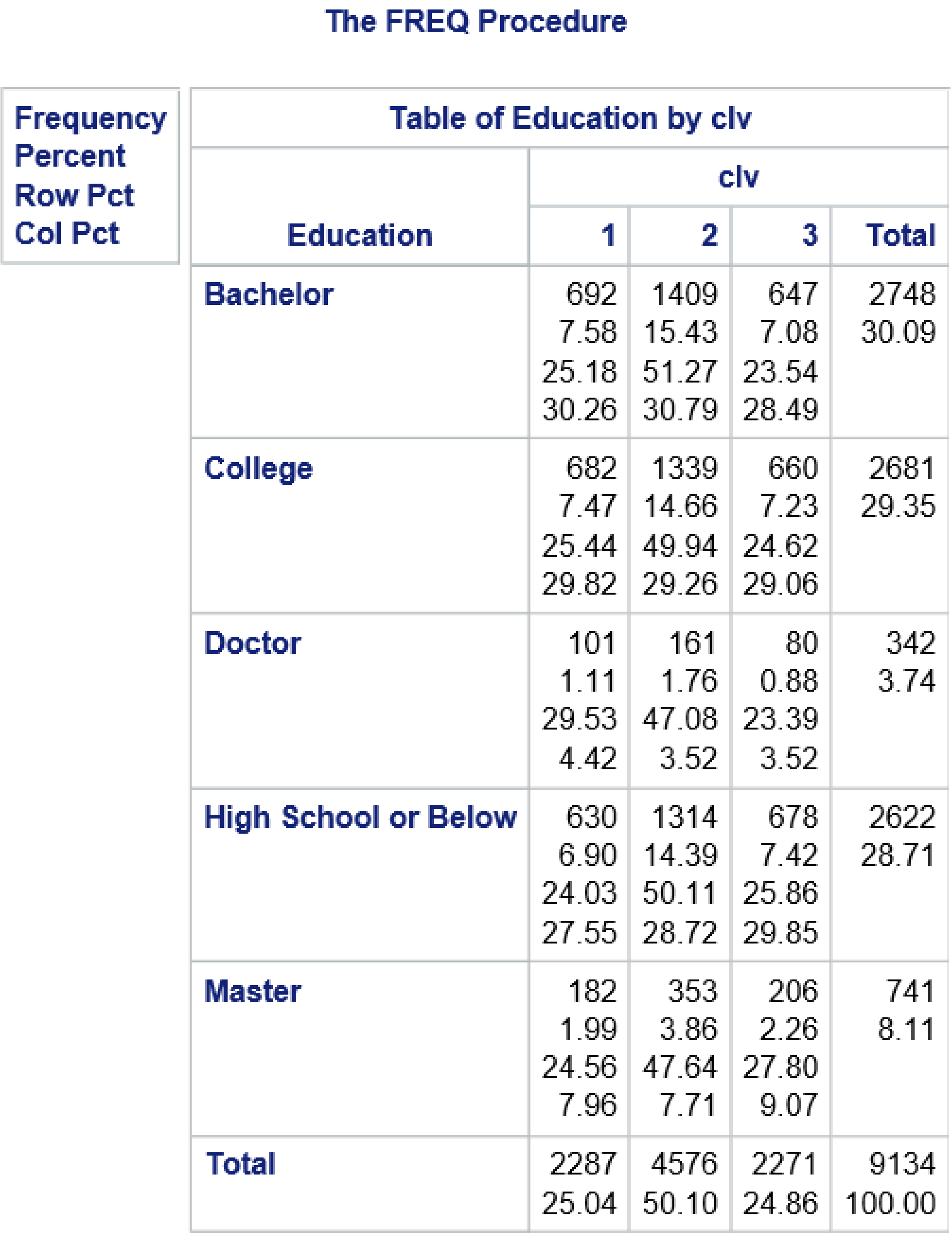
Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Customer_Lifetime_Value	9134	8005	6871	73117126	1898	83325
Income	9134	37657	30380	343962509	0	99981

Pearson Correlation Coefficients, N = 9134
Prob > |r| under H0: Rho=0

	Customer_Lifetime_Value	Income
Customer_Lifetime_Value	1.00000	0.02437 0.0199
Income	0.02437 0.0199	1.00000

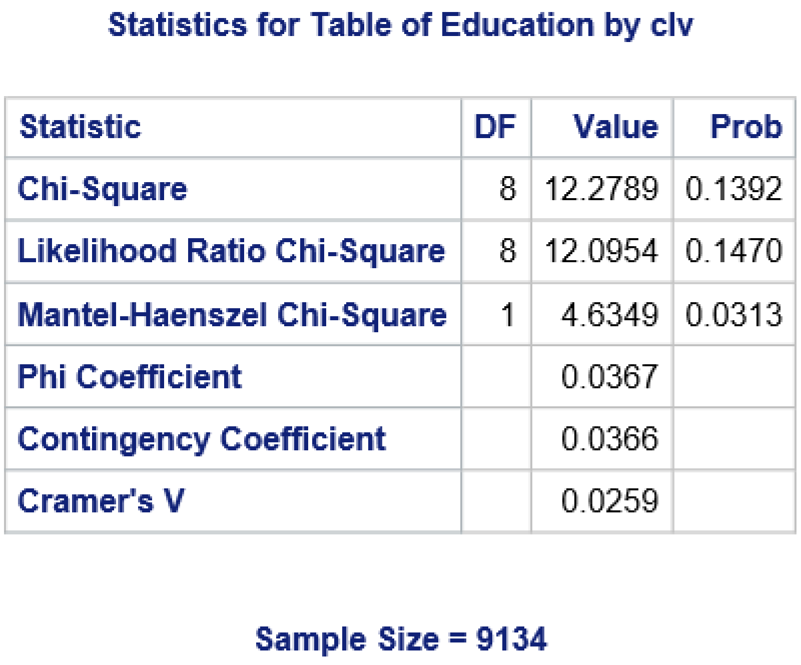
**** There is almost no correlation between Income and Customer Lifetime Value, with a Correlation coefficient of only 0.02437. Also, this coefficient is statistically significant as the p-Value is 0.0199 (< 0.05). Hence, it cannot be said with certainty if income actually has anything to do with the Customer Lifetime Value or not. From the results, it seems more that income does not have any relationship with the Customer Lifetime Value at all.**

Customer_LifeTime_Value Vs Education



The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Education by clv				
		clv			
	Education	1	2	3	Total
	Bachelor	692 7.58 25.18 30.26	1409 15.43 51.27 30.79	647 7.08 23.54 28.49	2748 30.09
	College	682 7.47 25.44 29.82	1339 14.66 49.94 29.26	660 7.23 24.62 29.06	2681 29.35
	Doctor	101 1.11 29.53 4.42	161 1.76 47.08 3.52	80 0.88 23.39 3.52	342 3.74
	High School or Below	630 6.90 24.03 27.55	1314 14.39 50.11 28.72	678 7.42 25.86 29.85	2622 28.71
	Master	182 1.99 24.56 7.96	353 3.86 47.64 7.71	206 2.26 27.80 9.07	741 8.11
	Total	2287 25.04	4576 50.10	2271 24.86	9134 100.00



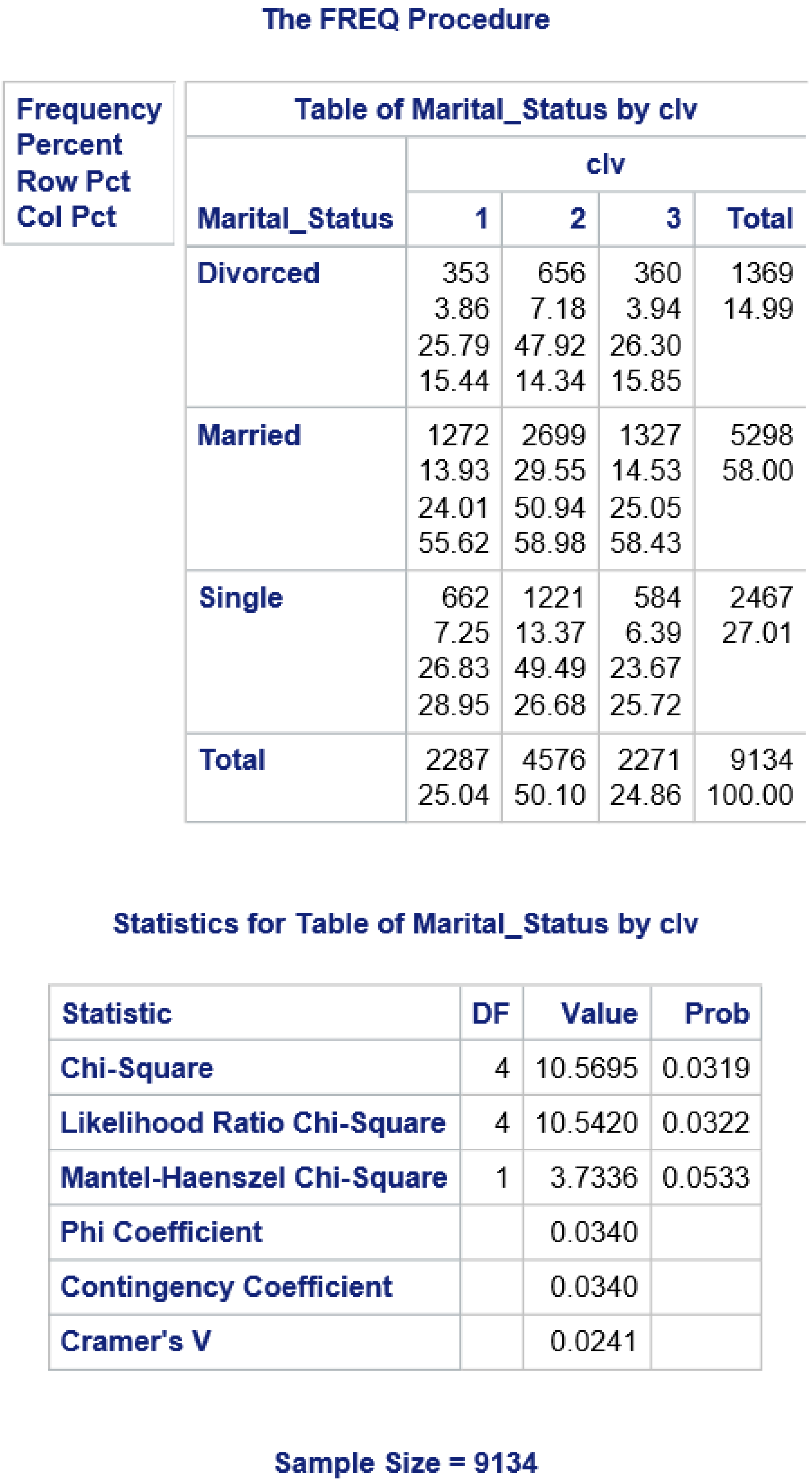
Statistics for Table of Education by clv

Statistic	DF	Value	Prob
Chi-Square	8	12.2789	0.1392
Likelihood Ratio Chi-Square	8	12.0954	0.1470
Mantel-Haenszel Chi-Square	1	4.6349	0.0313
Phi Coefficient		0.0367	
Contingency Coefficient		0.0366	
Cramer's V		0.0259	

Sample Size = 9134

**** From the Chi-Square Test, we see that the p-Value is 0.1392 (>0.05). Hence, we cannot reject the Null hypothesis which imply that there is no relationship exist between Customer_Liftime_Value and Education; they are independent.**

Customer_LifeTime_Value Vs Marital_Status



The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Marital_Status by clv				
	clv			
Marital_Status	1	2	3	Total
Divorced	353 3.86 25.79 15.44	656 7.18 47.92 14.34	360 3.94 26.30 15.85	1369 14.99
Married	1272 13.93 24.01 55.62	2699 29.55 50.94 58.98	1327 14.53 25.05 58.43	5298 58.00
Single	662 7.25 26.83 28.95	1221 13.37 49.49 26.68	584 6.39 23.67 25.72	2467 27.01
Total	2287 25.04	4576 50.10	2271 24.86	9134 100.00

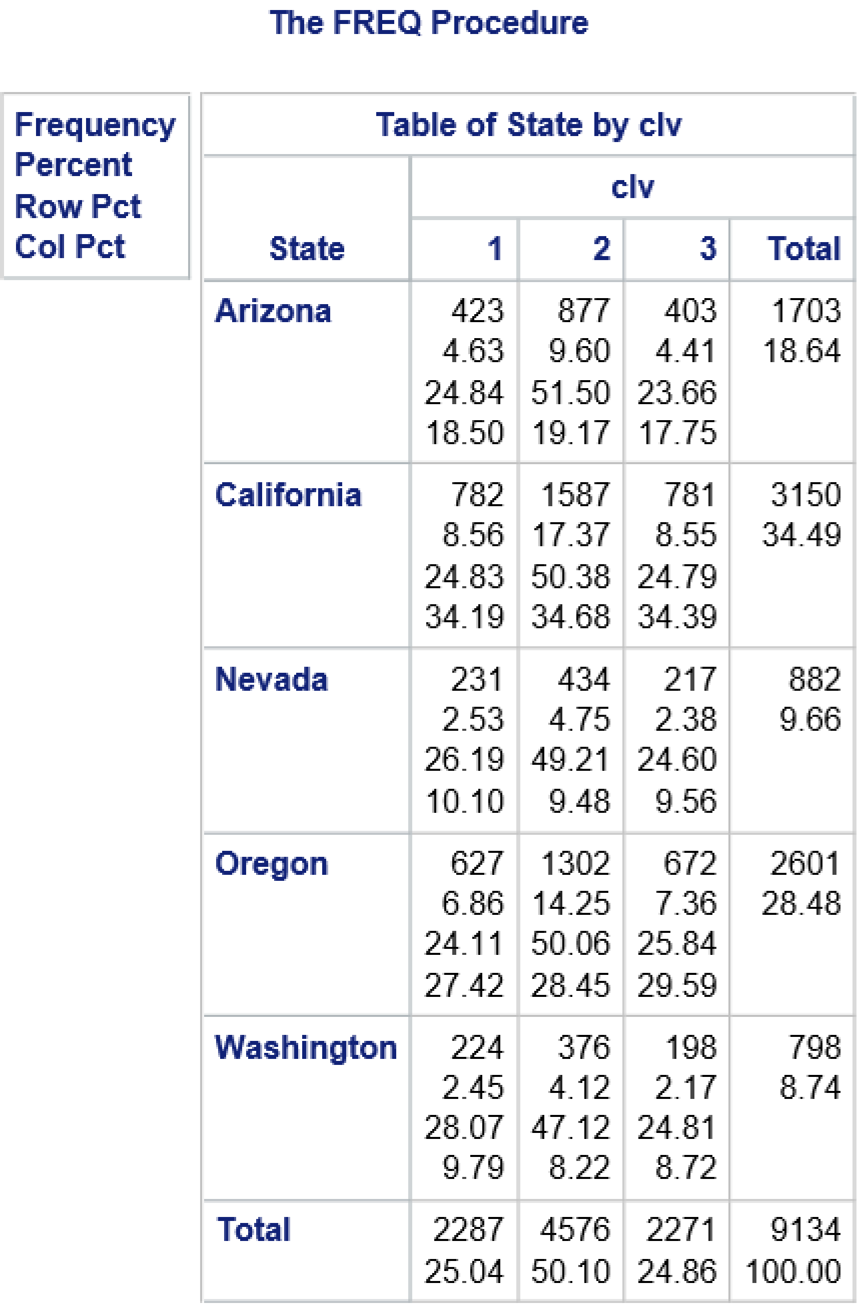
Statistics for Table of Marital_Status by clv

Statistic	DF	Value	Prob
Chi-Square	4	10.5695	0.0319
Likelihood Ratio Chi-Square	4	10.5420	0.0322
Mantel-Haenszel Chi-Square	1	3.7336	0.0533
Phi Coefficient		0.0340	
Contingency Coefficient		0.0340	
Cramer's V		0.0241	

Sample Size = 9134

**** From the above FREQ Procedure for Chi-Square test for Marriage vs Customer Lifetime Value, we see that the p-Value is 0.0319(<0.05). Hence, we can reject the Null hypothesis and we can claim that Customer Lifetime Value is dependent on the Marital Status.**

- **Customer_LifeTime_Value Vs State**

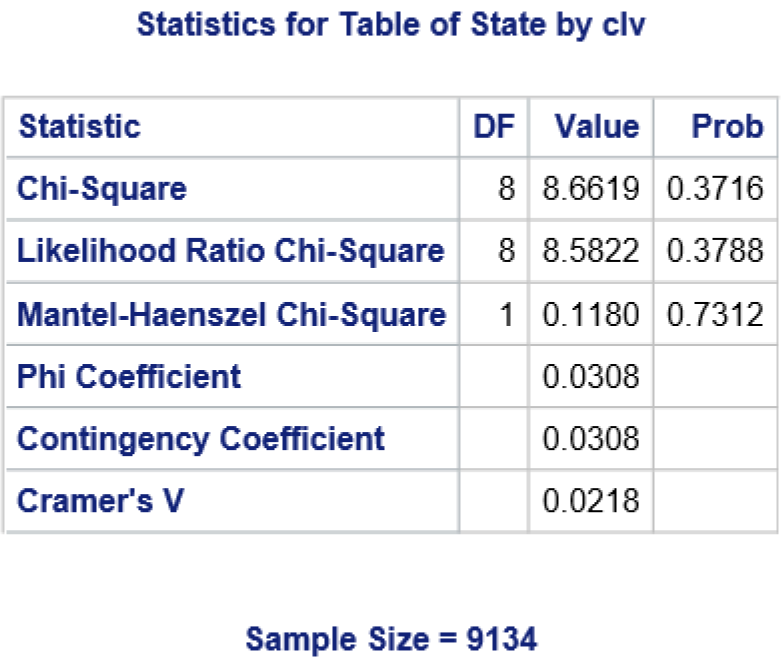


The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of State by clv

	clv			
State	1	2	3	Total
Arizona	423 4.63 24.84 18.50	877 9.60 51.50 19.17	403 4.41 23.66 17.75	1703 18.64
California	782 8.56 24.83 34.19	1587 17.37 50.38 34.68	781 8.55 24.79 34.39	3150 34.49
Nevada	231 2.53 26.19 10.10	434 4.75 49.21 9.48	217 2.38 24.60 9.56	882 9.66
Oregon	627 6.86 24.11 27.42	1302 14.25 50.06 28.45	672 7.36 25.84 29.59	2601 28.48
Washington	224 2.45 28.07 9.79	376 4.12 47.12 8.22	198 2.17 24.81 8.72	798 8.74
Total	2287 25.04	4576 50.10	2271 24.86	9134 100.00



Statistics for Table of State by clv

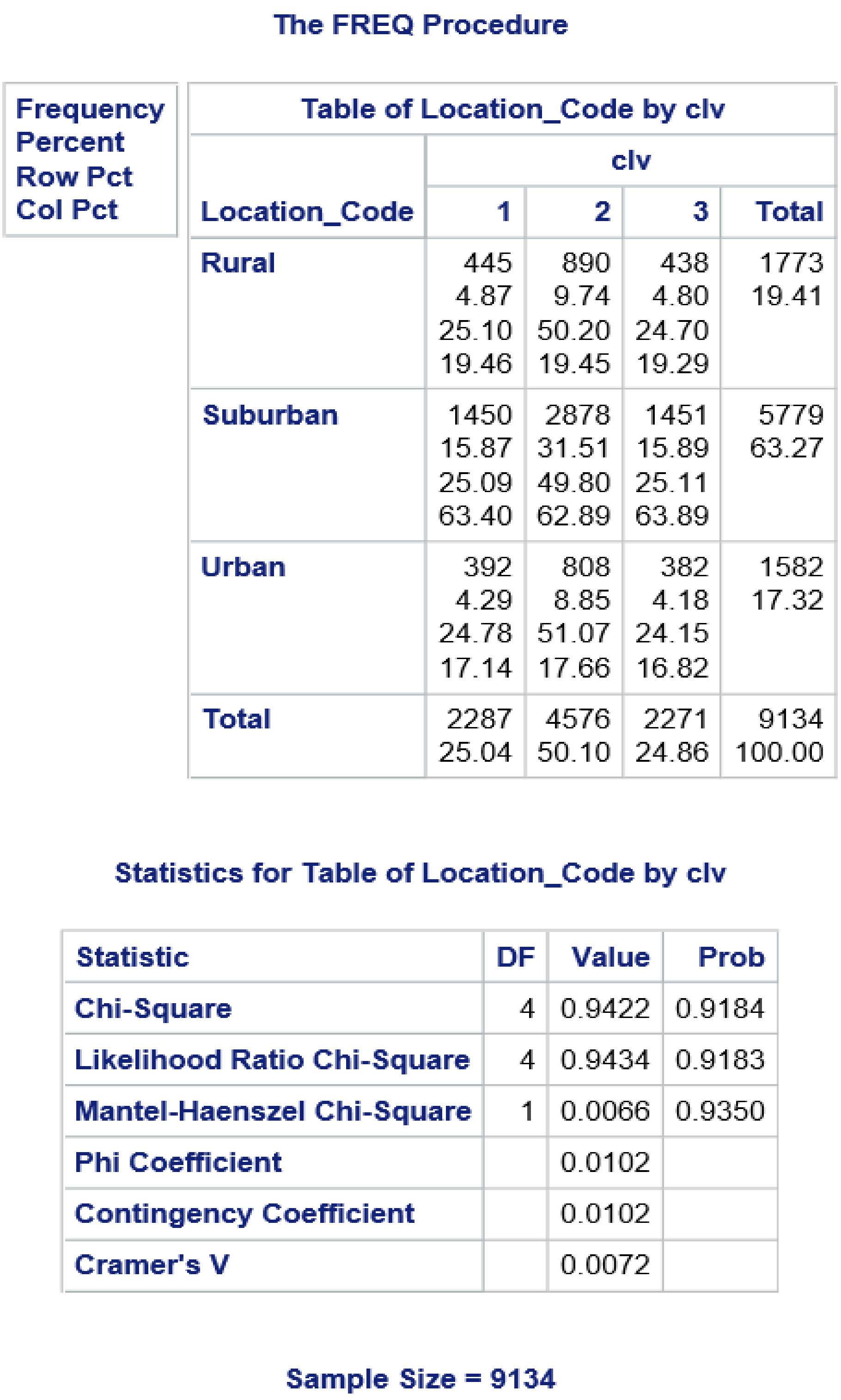
Statistic	DF	Value	Prob
Chi-Square	8	8.6619	0.3716
Likelihood Ratio Chi-Square	8	8.5822	0.3788
Mantel-Haenszel Chi-Square	1	0.1180	0.7312
Phi Coefficient		0.0308	
Contingency Coefficient		0.0308	
Cramer's V		0.0218	

Sample Size = 9134

**** The Chi-Square test of State vs Customer Lifetime Value yields a p-Value of 0.3716 (>0.05). Hence, here we cannot reject the Null hypothesis which further imply that there is no relationship exist between Customer_Liftime_Value and State; they are independent.**

.

- **Customer_LifeTime_Value** Vs **Location_Code**



The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of Location_Code by clv

Location_Code	clv 1	2	3	Total
Rural	445 4.87 25.10 19.46	890 9.74 50.20 19.45	438 4.80 24.70 19.29	1773 19.41
Suburban	1450 15.87 25.09 63.40	2878 31.51 49.80 62.89	1451 15.89 25.11 63.89	5779 63.27
Urban	392 4.29 24.78 17.14	808 8.85 51.07 17.66	382 4.18 24.15 16.82	1582 17.32
Total	2287 25.04	4576 50.10	2271 24.86	9134 100.00

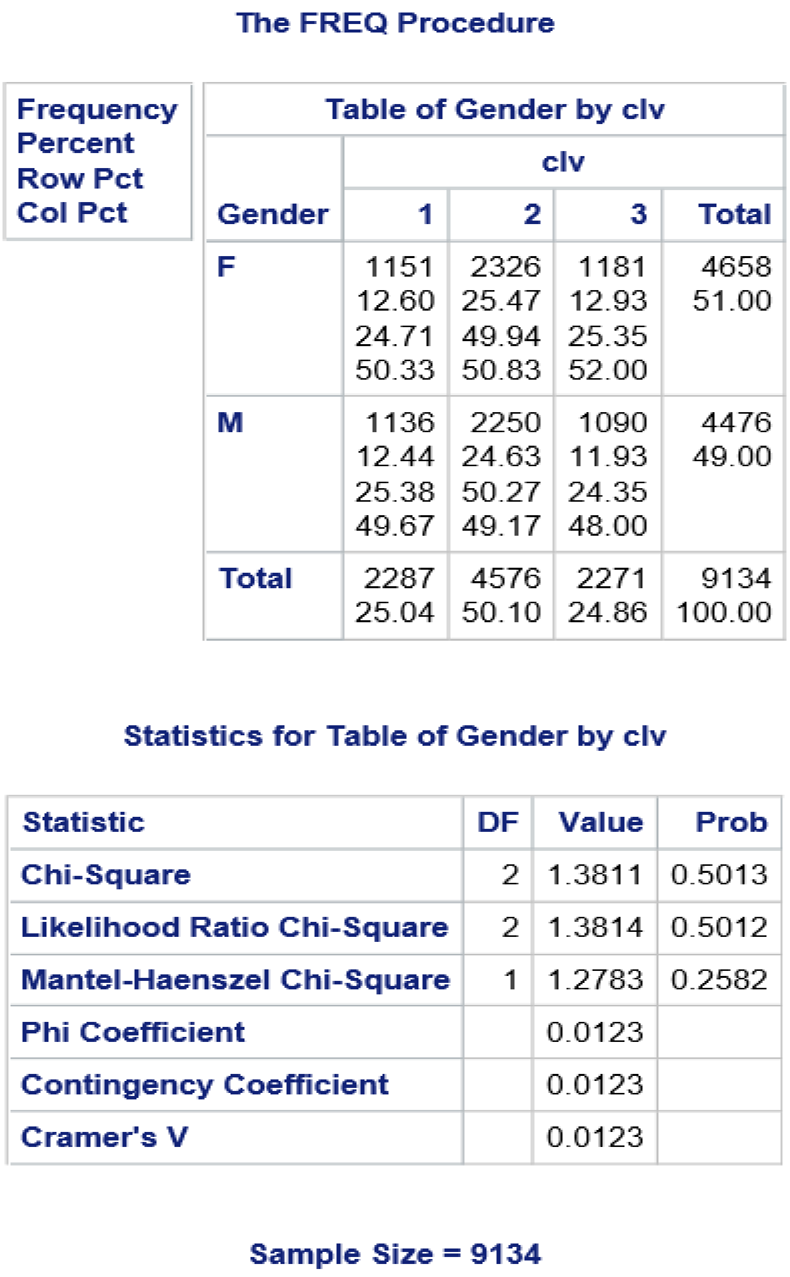
Statistics for Table of Location_Code by clv

Statistic	DF	Value	Prob
Chi-Square	4	0.9422	0.9184
Likelihood Ratio Chi-Square	4	0.9434	0.9183
Mantel-Haenszel Chi-Square	1	0.0066	0.9350
Phi Coefficient		0.0102	
Contingency Coefficient		0.0102	
Cramer's V		0.0072	

Sample Size = 9134

**** From the above FREQ Procedure for Chi-Square test for Location Code vs Customer Lifetime Value, we see that the p-Value is 0.9184(>0.05). Hence, we cannot reject the Null hypothesis and we can say that there is no relationship exist between Customer_Liftime_Value and Location_Code; they are independent.**

- **Customer_LifeTime_Value Vs Gender**



The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Gender by clv				
	clv			
Gender	1	2	3	Total
F	1151 12.60 24.71 50.33	2326 25.47 49.94 50.83	1181 12.93 25.35 52.00	4658 51.00
M	1136 12.44 25.38 49.67	2250 24.63 50.27 49.17	1090 11.93 24.35 48.00	4476 49.00
Total	2287 25.04	4576 50.10	2271 24.86	9134 100.00

Statistics for Table of Gender by clv

Statistic	DF	Value	Prob
Chi-Square	2	1.3811	0.5013
Likelihood Ratio Chi-Square	2	1.3814	0.5012
Mantel-Haenszel Chi-Square	1	1.2783	0.2582
Phi Coefficient		0.0123	
Contingency Coefficient		0.0123	
Cramer's V		0.0123	

Sample Size = 9134

**** The p-Value for Chi-Square test of Gender vs Customer Lifetime Value is 0.5013 (>0.05). Hene, we cannot reject the Null hypothesis. Thus, Customer Lifetime Value might not have any dependency on Gender.**

**** From the analysis of all the demographic attributes above, we can collaboratively conclude that Customer Lifetime value is dependent on Marital status only and might have no dependency at all on other demographic attributes like- Gender, State, Location Code, Income, Education etc.**

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

```
PROC FREQ data=carins ;TABLE
Renew_Offer_Type*Response/CHISQ;RUN;
```

**** The Chi-Squared Test has a result of p-Value <0.0001 (<0.05). Hence, we can reject the Null hypothesis for Chi-Squared test and conclude that there is a relationship between the Renew Offer type and the Response.**

**** The highest response rate is generated by Offer2 (23.38 %).**

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of Renew_Offer_Type by Response			
	Response		
Renew_Offer_Type	No	Yes	Total
Offer1	3158 34.57 84.17 40.35	594 6.50 15.83 45.41	3752 41.08
Offer2	2242 24.55 76.62 28.65	684 7.49 23.38 52.29	2926 32.03
Offer3	1402 15.35 97.91 17.91	30 0.33 2.09 2.29	1432 15.68
Offer4	1024 11.21 100.00 13.08	0 0.00 0.00 0.00	1024 11.21
Total	7826 85.68	1308 14.32	9134 100.00

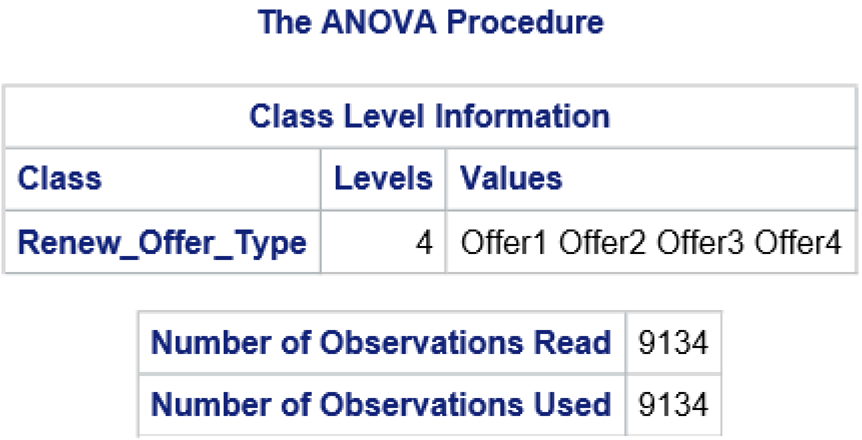
Statistics for Table of Renew_Offer_Type by Response

Statistic	DF	Value	Prob
Chi-Square	3	548.1645	<.0001
Likelihood Ratio Chi-Square	3	751.4675	<.0001
Mantel-Haenszel Chi-Square	1	242.3027	<.0001
Phi Coefficient		0.2450	
Contingency Coefficient		0.2379	
Cramer's V		0.2450	

Sample Size = 9134

8) Do different renew_offer_types have different lifetime values? Which offer type is the best?

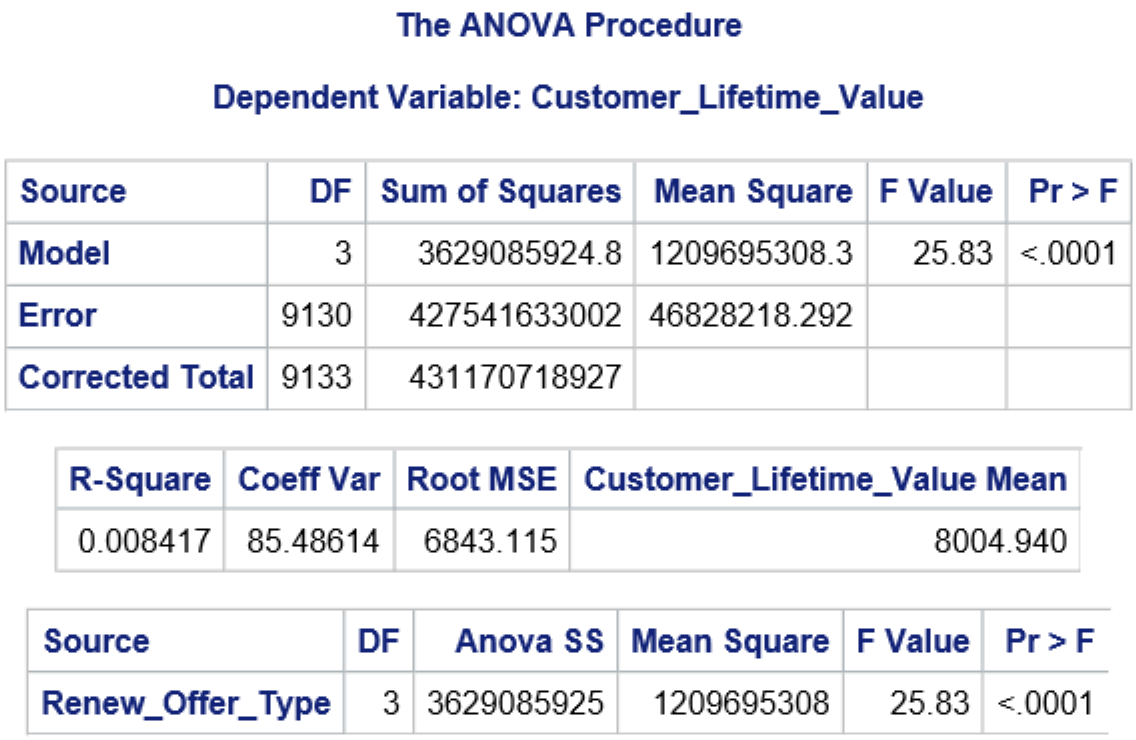
```
proc anova data=carins; class Renew_Offer_Type;model
Customer_Lifetime_Value = Renew_Offer_Type;run;
proc means data=carins;class Renew_Offer_Type;var
Customer_Lifetime_Value;run;
```



The ANOVA Procedure

Class Level Information		
Class	Levels	Values
Renew_Offer_Type	4	Offer1 Offer2 Offer3 Offer4

Number of Observations Read	9134
Number of Observations Used	9134



The ANOVA Procedure

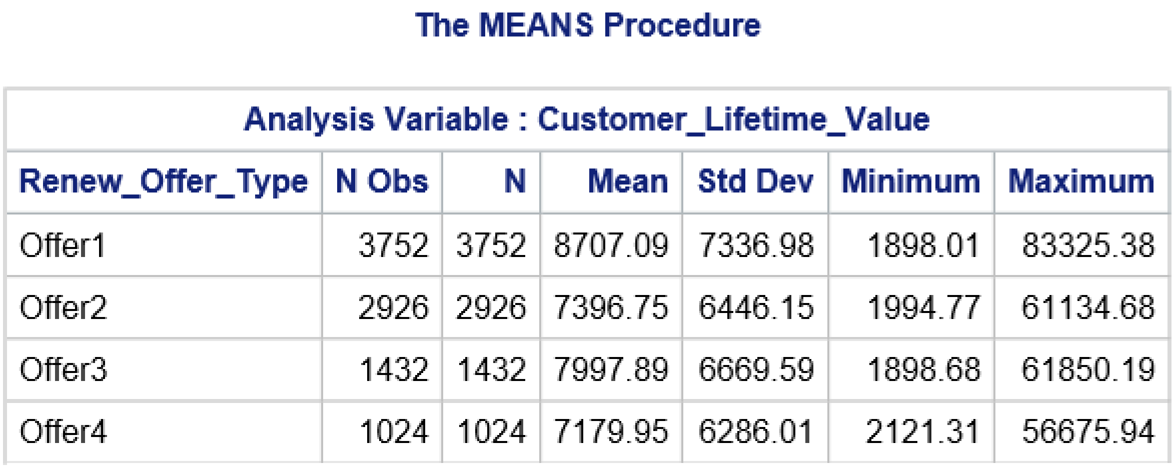
Dependent Variable: Customer_Lifetime_Value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3629085924.8	1209695308.3	25.83	<.0001
Error	9130	427541633002	46828218.292		
Corrected Total	9133	431170718927			

R-Square	Coeff Var	Root MSE	Customer_Lifetime_Value Mean
0.008417	85.48614	6843.115	8004.940

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Renew_Offer_Type	3	3629085925	1209695308	25.83	<.0001

**** From ANOVA, the F Test has a P-Value of <0.0001 (<0.05). Hence, the means are different which imply that different renew offer types have different average Customer Lifetime Values.**



The MEANS Procedure

Analysis Variable : Customer_Lifetime_Value						
Renew_Offer_Type	N Obs	N	Mean	Std Dev	Minimum	Maximum
Offer1	3752	3752	8707.09	7336.98	1898.01	83325.38
Offer2	2926	2926	7396.75	6446.15	1994.77	61134.68
Offer3	1432	1432	7997.89	6669.59	1898.68	61850.19
Offer4	1024	1024	7179.95	6286.01	2121.31	56675.94

**** An offer is best if it yields Customer Lifetime Value with optimal mean and less variability in the long run. In this context, Offer3 has an average Customer Lifetime Value of 7997.89 which is pretty close to the Average Customer Lifetime Value of the entire dataset (8004.90). Also, Offer3 has a**

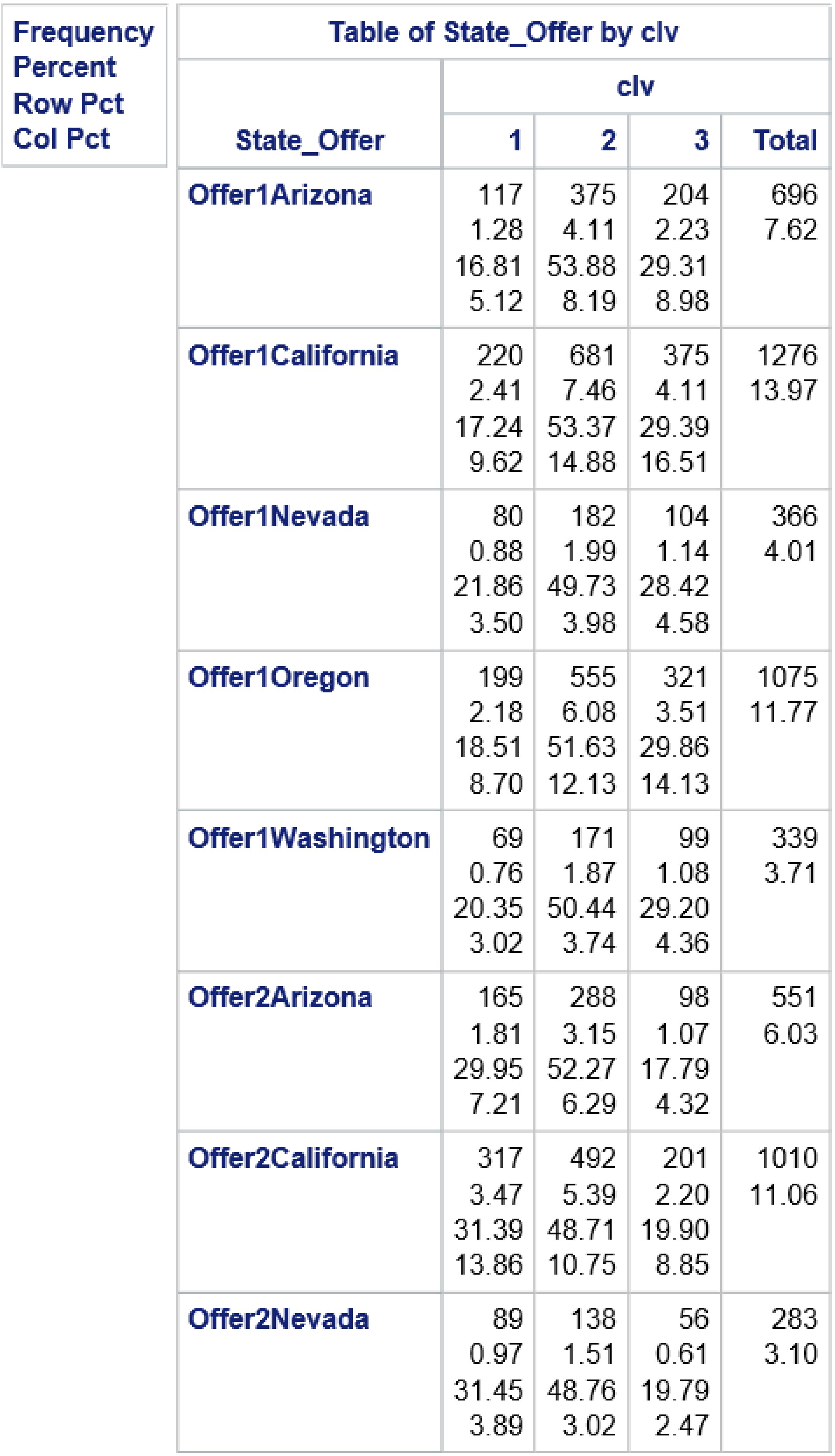
Standard deviation of 6669.59 only. Hence, Offer3 is undoubtedly the best renew offer.

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

```
proc sql;
create table a5 as select *,cats(Renew_Offer_Type,State) as State_Offer from carins;
quit;
data a6;set a5;
if Customer_Lifetime_Value le 4000 then clv=1;
if Customer_Lifetime_Value ge 9000 then clv=3;
if Customer_Lifetime_Value gt 4000 and Customer_Lifetime_Value lt 9000 then clv=2;run;
proc freq data=a6;table State_Offer*clv/CHISQ;run;
proc anova data=a5; class State_Offer;model Customer_Lifetime_Value = State_Offer;run;
```

**** In order to see whether different renew offer types across different states are impacting the Customer Lifetime Value, we stacked the 2 categorical variables -State and Renew Offer Type, and performed both Chi-Square Test as well as the ANOVA.**

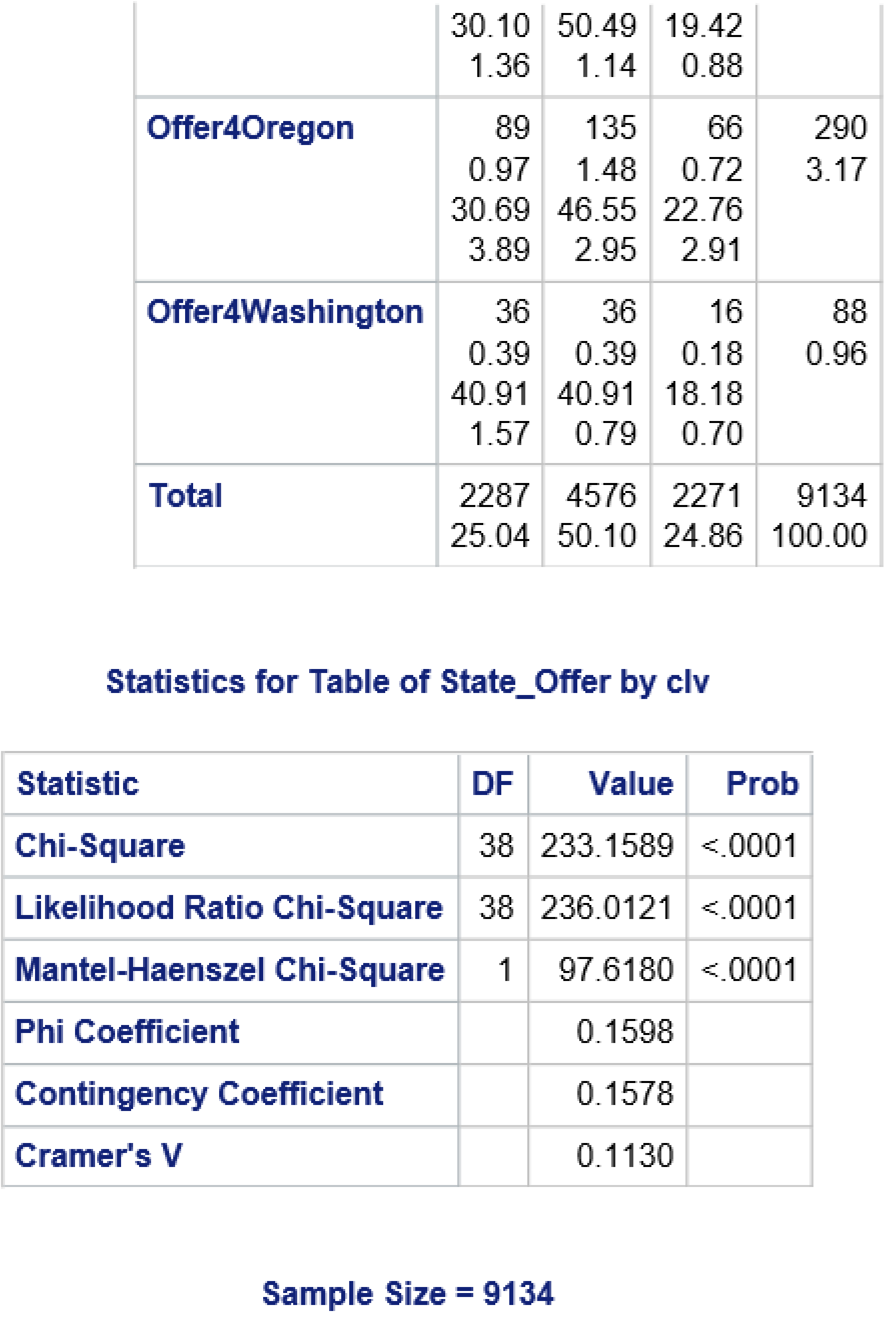
The FREQ Procedure



Frequency
Percent
Row Pct
Col Pct

Table of State_Offer by clv				
	clv			
State_Offer	1	2	3	Total
Offer1Arizona	117 1.28 16.81 5.12	375 4.11 53.88 8.19	204 2.23 29.31 8.98	696 7.62
Offer1California	220 2.41 17.24 9.62	681 7.46 53.37 14.88	375 4.11 29.39 16.51	1276 13.97
Offer1Nevada	80 0.88 21.86 3.50	182 1.99 49.73 3.98	104 1.14 28.42 4.58	366 4.01
Offer1Oregon	199 2.18 18.51 8.70	555 6.08 51.63 12.13	321 3.51 29.86 14.13	1075 11.77
Offer1Washington	69 0.76 20.35 3.02	171 1.87 50.44 3.74	99 1.08 29.20 4.36	339 3.71
Offer2Arizona	165 1.81 29.95 7.21	288 3.15 52.27 6.29	98 1.07 17.79 4.32	551 6.03
Offer2California	317 3.47 31.39 13.86	492 5.39 48.71 10.75	201 2.20 19.90 8.85	1010 11.06
Offer2Nevada	89 0.97 31.45 3.89	138 1.51 48.76 3.02	56 0.61 19.79 2.47	283 3.10

Offer2Oregon	237 2.59 28.42 10.36	406 4.44 48.68 8.87	191 2.09 22.90 8.41	834 9.13
Offer2Washington	85 0.93 34.27 3.72	111 1.22 44.76 2.43	52 0.57 20.97 2.29	248 2.72
Offer3Arizona	72 0.79 27.17 3.15	124 1.36 46.79 2.71	69 0.76 26.04 3.04	265 2.90
Offer3California	122 1.34 23.83 5.33	258 2.82 50.39 5.64	132 1.45 25.78 5.81	512 5.61
Offer3Nevada	31 0.34 23.85 1.36	62 0.68 47.69 1.35	37 0.41 28.46 1.63	130 1.42
Offer3Oregon	102 1.12 25.37 4.46	206 2.26 51.24 4.50	94 1.03 23.38 4.14	402 4.40
Offer3Washington	34 0.37 27.64 1.49	58 0.63 47.15 1.27	31 0.34 25.20 1.37	123 1.35
Offer4Arizona	69 0.76 36.13 3.02	90 0.99 47.12 1.97	32 0.35 16.75 1.41	191 2.09
Offer4California	123 1.35 34.94 5.38	156 1.71 44.32 3.41	73 0.80 20.74 3.21	352 3.85
Offer4Nevada	31 0.34	52 0.57	20 0.22	103 1.13



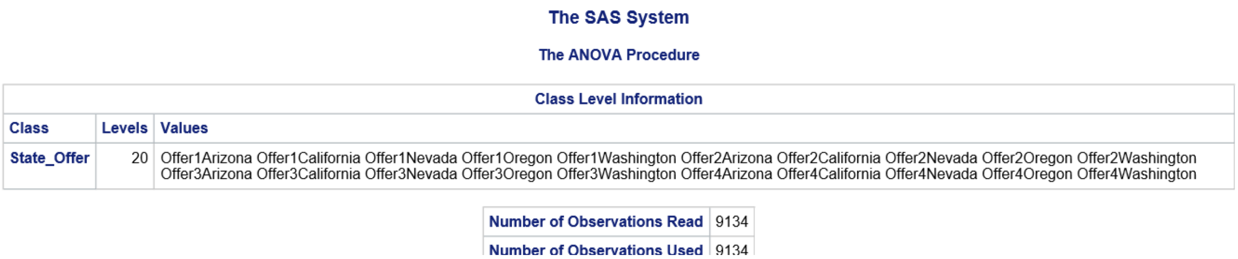
	30.10 1.36	50.49 1.14	19.42 0.88	
Offer4Oregon	89 0.97 30.69 3.89	135 1.48 46.55 2.95	66 0.72 22.76 2.91	290 3.17
Offer4Washington	36 0.39 40.91 1.57	36 0.39 40.91 0.79	16 0.18 18.18 0.70	88 0.96
Total	2287 25.04	4576 50.10	2271 24.86	9134 100.00

Statistics for Table of State_Offer by clv

Statistic	DF	Value	Prob
Chi-Square	38	233.1589	<.0001
Likelihood Ratio Chi-Square	38	236.0121	<.0001
Mantel-Haenszel Chi-Square	1	97.6180	<.0001
Phi Coefficient		0.1598	
Contingency Coefficient		0.1578	
Cramer's V		0.1130	

Sample Size = 9134

**** The p-value of Chi-Squared test is <0.0001(<0.05). Hence, we can reject the Null hypothesis and conclude that Customer Lifetime value has dependency on a combination of State and Renew Offer Type.**

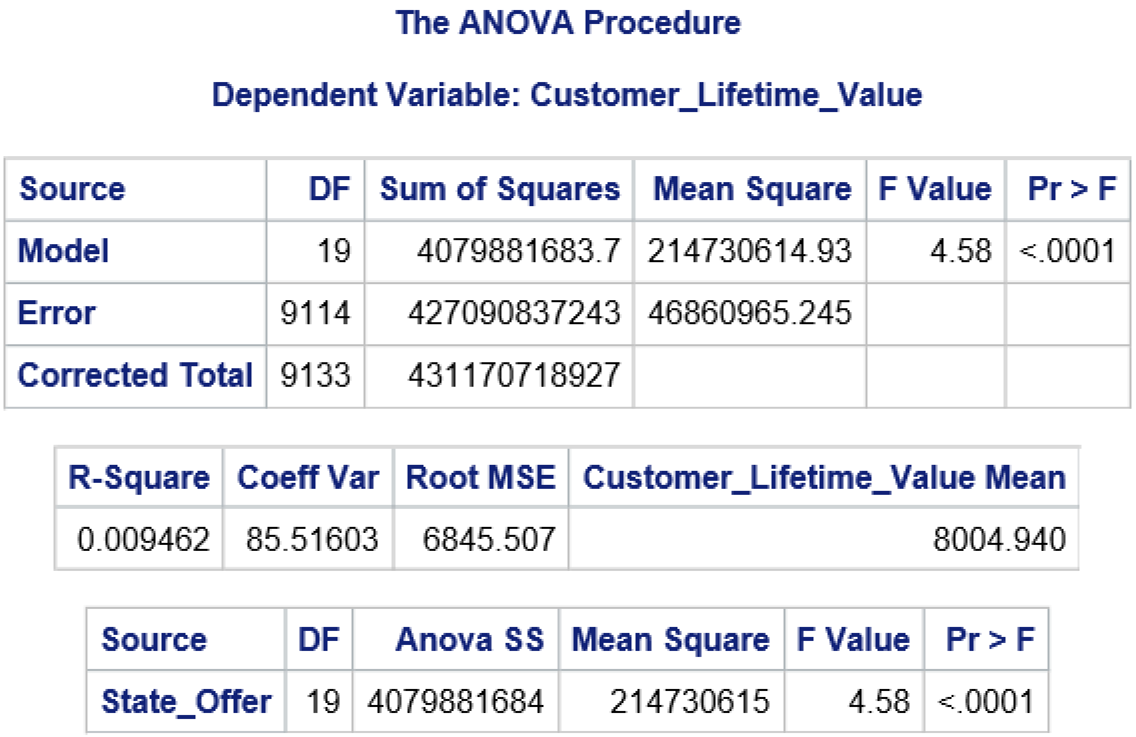


The SAS System

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
State_Offer	20	Offer1Arizona Offer1California Offer1Nevada Offer1Oregon Offer1Washington Offer2Arizona Offer2California Offer2Nevada Offer2Oregon Offer2Washington Offer3Arizona Offer3California Offer3Nevada Offer3Oregon Offer3Washington Offer4Arizona Offer4California Offer4Nevada Offer4Oregon Offer4Washington

Number of Observations Read	9134
Number of Observations Used	9134



The ANOVA Procedure

Dependent Variable: Customer_Lifetime_Value

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	19	4079881683.7	214730614.93	4.58	<.0001
Error	9114	427090837243	46860965.245		
Corrected Total	9133	431170718927			

R-Square	Coeff Var	Root MSE	Customer_Lifetime_Value Mean
0.009462	85.51603	6845.507	8004.940

Source	DF	Anova SS	Mean Square	F Value	Pr > F
State_Offer	19	4079881684	214730615	4.58	<.0001

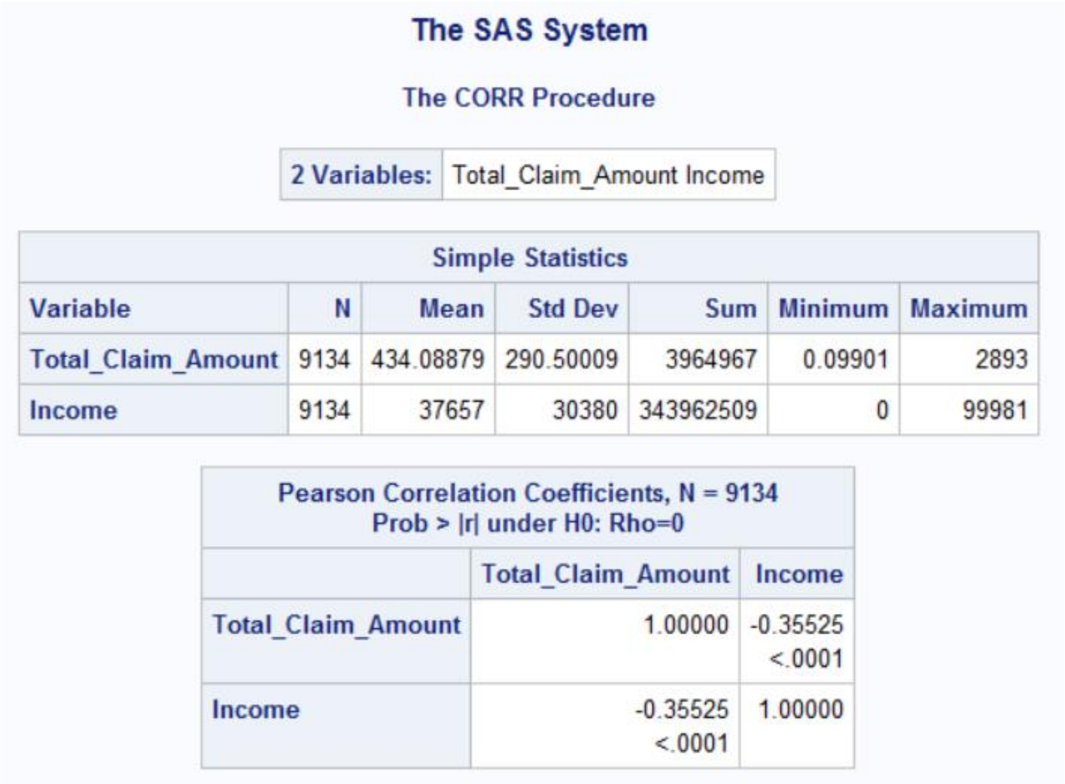
**** From ANOVA, the p-value is <0.0001 (<0.05). Hence, it can be easily said that for different state and renew offer type combinations, we have different average Customer Lifetime Values.**

**** Combining the results from both the Chi-Squared Test and ANOVA, we can certainly conclude that the effectiveness of renew offer types is different across different states with respect to Customer Lifetime Value.**

10) What other interesting insights that are useful to the company in terms of action can be obtained from the data? Write any 3 and indicate which type of analysis is appropriate.

1. How Income and Total_claim_amount is related to each other.?

```
proc corr data=carins;var income total_claim_amount;run;
proc sgplot data = carins;
scatter x = income y = total_claim_amount;
run;
```



The SAS System

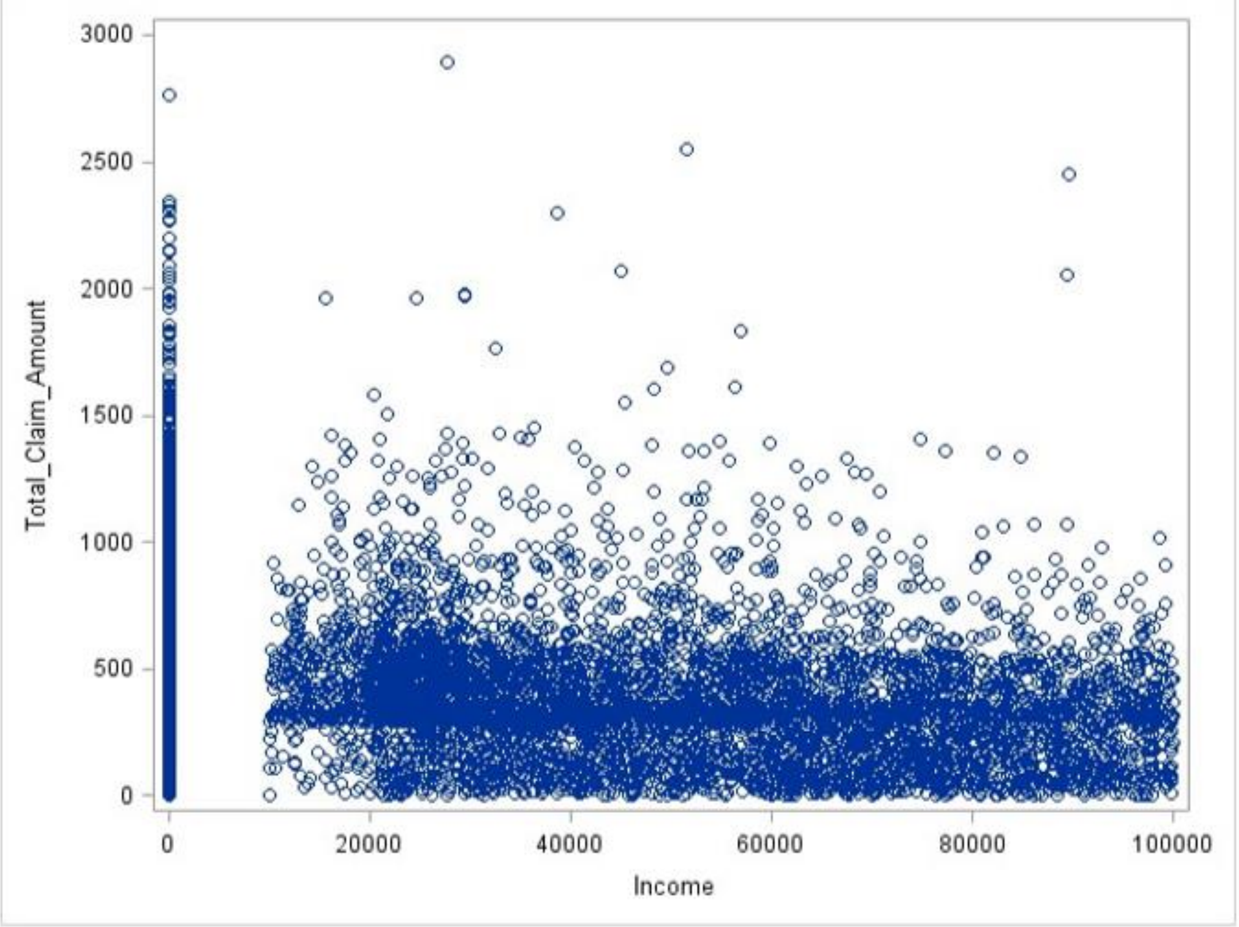
The CORR Procedure

2 Variables:	Total_Claim_Amount Income

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Total_Claim_Amount	9134	434.08879	290.50009	3964967	0.09901	2893
Income	9134	37657	30380	343962509	0	99981

Pearson Correlation Coefficients, N = 9134
Prob > |r| under H0: Rho=0

	Total_Claim_Amount	Income
Total_Claim_Amount	1.00000	-0.35525 <.0001
Income	-0.35525 <.0001	1.00000

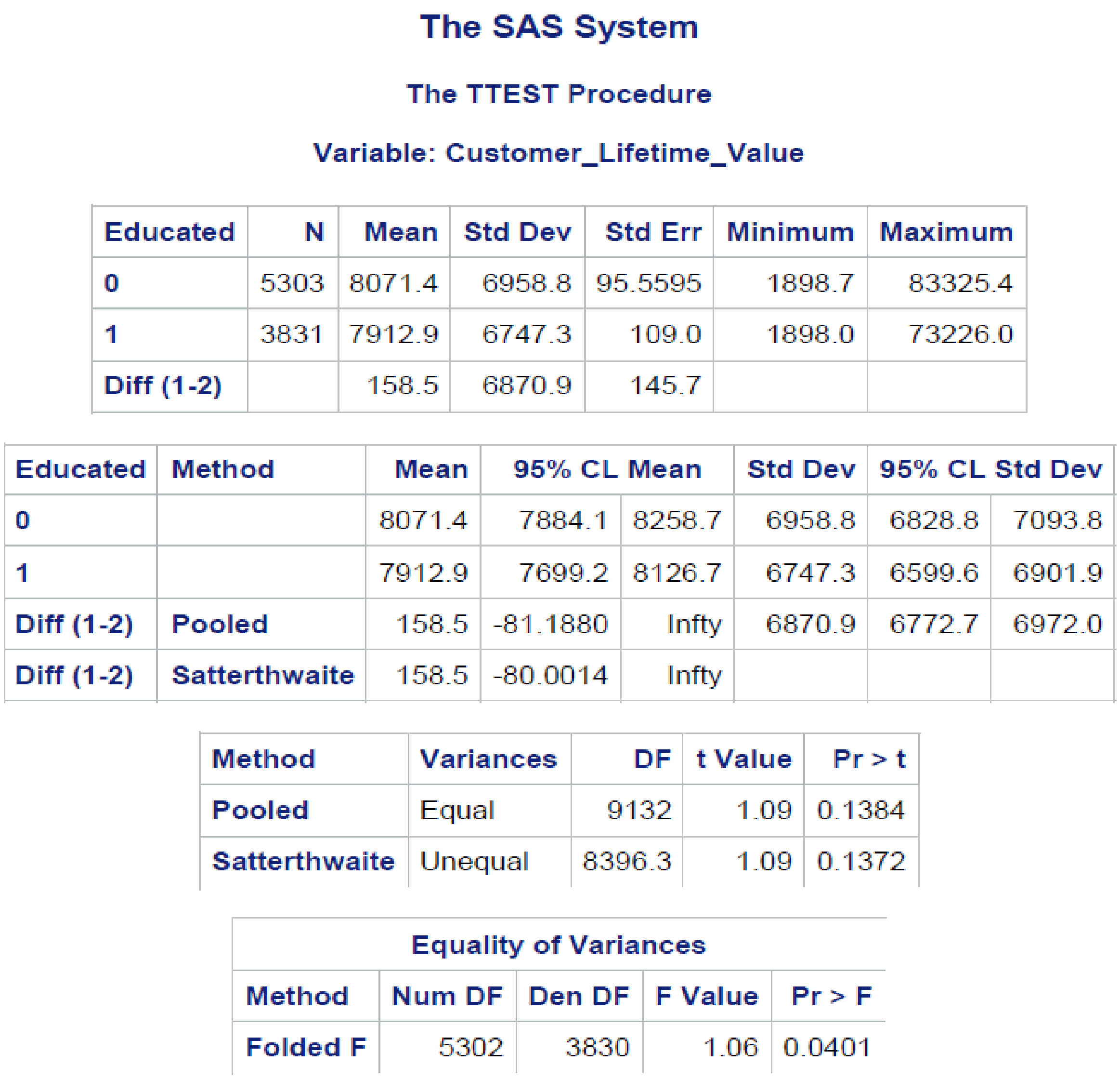


**** There is moderate to weak negative linear relationship between Income and Total_claim_amount, with a Correlation coefficient of -0.3552. Also, this coefficient is statistically significant as the p-Value is < 0.0001 Hence, it cannot be said with certainty if income actually has anything to do with the Total_claim_amount or not. From the results, it seems more that income does not have any relationship with the Total_claim_amount at all.**

2. Educated customers (with a bachelors or equivalent/more degree) are more valuable than others ?

```
data a7; set carins;
if Education="Bachelor" or Education="Master" or Education="Doctor"
then Educated=1;
```

```
if Education="College" or Education="High School or Below" then
Educated=0;run;
proc ttest sides=u;var Customer_Lifetime_Value;class Educated;data a7;run;
```



The SAS System

The TTEST Procedure

Variable: Customer_Lifetime_Value

Educated	N	Mean	Std Dev	Std Err	Minimum	Maximum
0	5303	8071.4	6958.8	95.5595	1898.7	83325.4
1	3831	7912.9	6747.3	109.0	1898.0	73226.0
Diff (1-2)		158.5	6870.9	145.7		

Educated	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		8071.4	7884.1	8258.7	6958.8	6828.8	7093.8
1		7912.9	7699.2	8126.7	6747.3	6599.6	6901.9
Diff (1-2)	Pooled	158.5	-81.1880	Infty	6870.9	6772.7	6972.0
Diff (1-2)	Satterthwaite	158.5	-80.0014	Infty			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	9132	1.09	0.1384
Satterthwaite	Unequal	8396.3	1.09	0.1372

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	5302	3830	1.06	0.0401

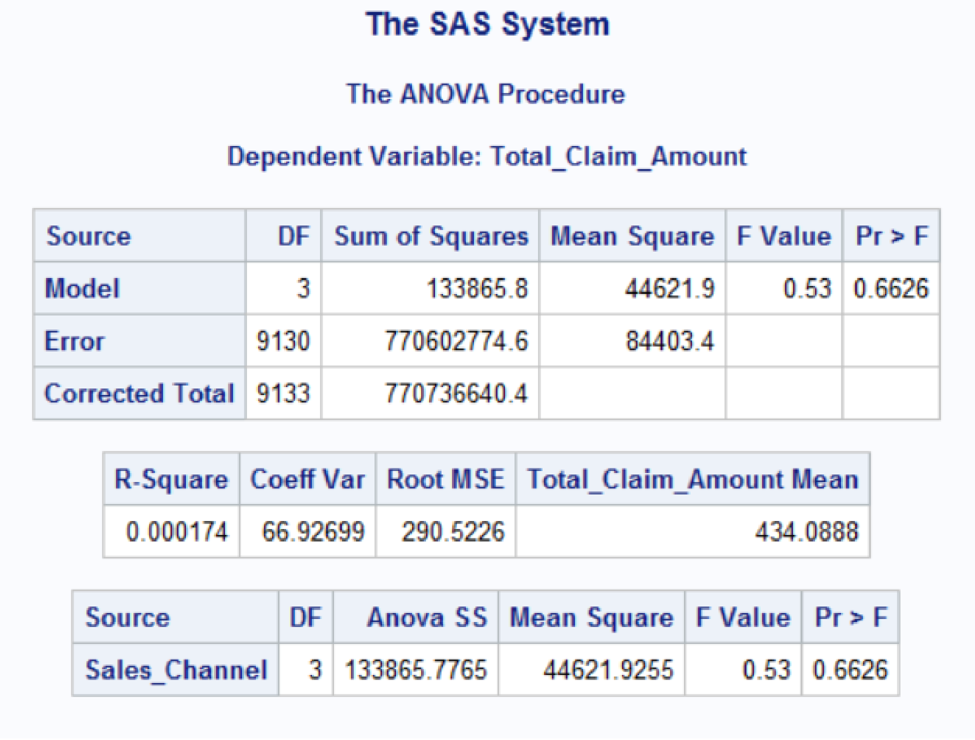
**** F- Test result: P value for F-test =0.0401 < 0.05, hence we can reject null hypothesis of F-test => We have to go ahead with T-test for unequal variances.**

P value for the 1-tailed test = 0.1372 > 0.05.

Failing to reject H0, We don't have enough evidence to claim that Educated customers (with a bachelors or equivalent/more degree) are more valuable than others.

3. The distribution of Total Claim Amount, Monthly Premium Auto and Number of Policies are not significantly different across all the sales channels. ANOVA is the appropriate analysis to check this insight.

```
proc anova data=carins;class Sales_Channel;
model Total_Claim_Amount=Sales_Channel;run;
proc anova data=carins;class Sales_Channel;
model Number_of_Policies=Sales_Channel;run;
proc anova data=carins;class Sales_Channel;
model Monthly_Premium_Auto=Sales_Channel;run;
```



The SAS System

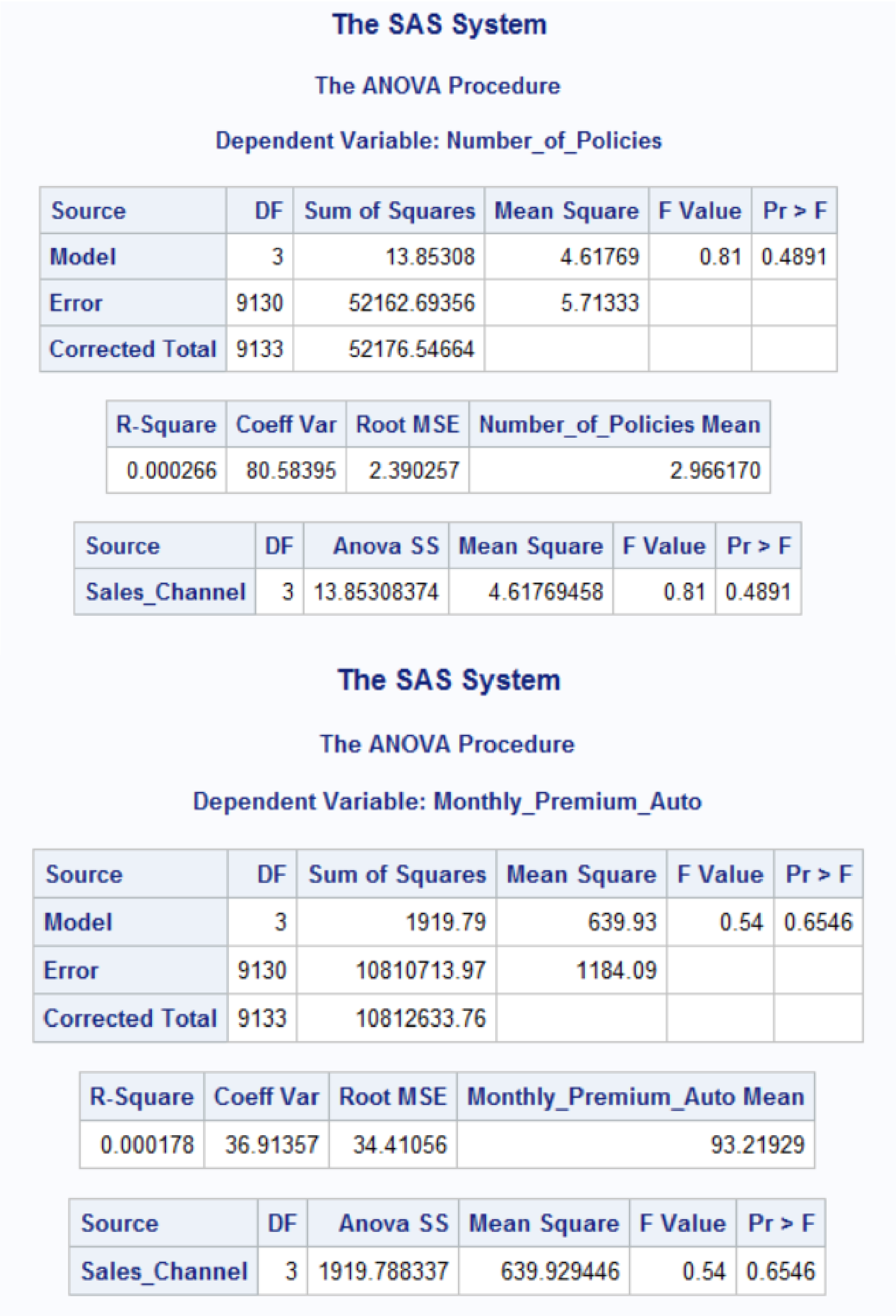
The ANOVA Procedure

Dependent Variable: Total_Claim_Amount

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	133865.8	44621.9	0.53	0.6626
Error	9130	770602774.6	84403.4		
Corrected Total	9133	770736640.4			

R-Square	Coeff Var	Root MSE	Total_Claim_Amount Mean
0.000174	66.92699	290.5226	434.0888

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Sales_Channel	3	133865.7765	44621.9255	0.53	0.6626



The SAS System

The ANOVA Procedure

Dependent Variable: Number_of_Policies

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	13.85308	4.61769	0.81	0.4891
Error	9130	52162.69356	5.71333		
Corrected Total	9133	52176.54664			

R-Square	Coeff Var	Root MSE	Number_of_Policies Mean
0.000266	80.58395	2.390257	2.966170

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Sales_Channel	3	13.85308374	4.61769458	0.81	0.4891

The SAS System

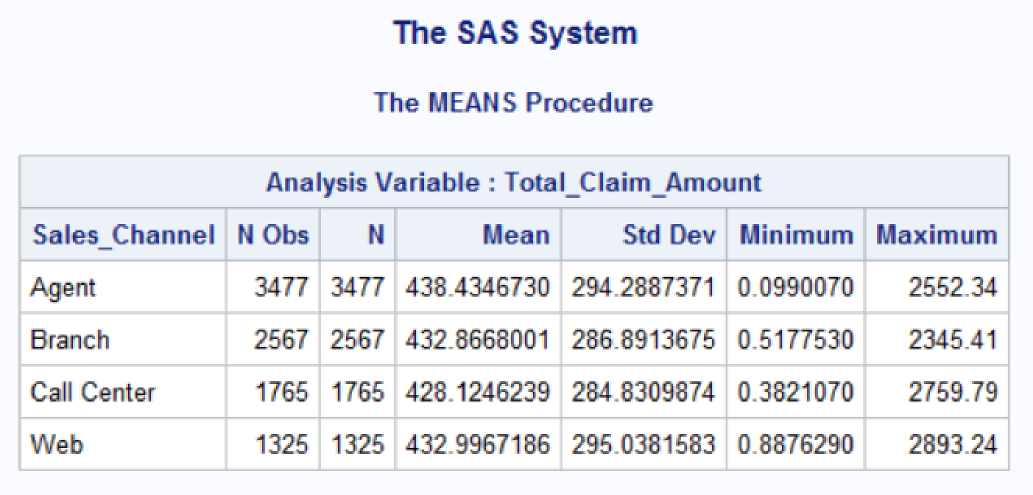
The ANOVA Procedure

Dependent Variable: Monthly_Premium_Auto

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1919.79	639.93	0.54	0.6546
Error	9130	10810713.97	1184.09		
Corrected Total	9133	10812633.76			

R-Square	Coeff Var	Root MSE	Monthly_Premium_Auto Mean
0.000178	36.91357	34.41056	93.21929

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Sales_Channel	3	1919.788337	639.929446	0.54	0.6546



The SAS System

The MEANS Procedure

Analysis Variable : Total_Claim_Amount						
Sales_Channel	N Obs	N	Mean	Std Dev	Minimum	Maximum
Agent	3477	3477	438.4346730	294.2887371	0.0990070	2552.34
Branch	2567	2567	432.8668001	286.8913675	0.5177530	2345.41
Call Center	1765	1765	428.1246239	284.8309874	0.3821070	2759.79
Web	1325	1325	432.9967186	295.0381583	0.8876290	2893.24