BUAN 6337 Predictive Analytics using SAS

Homework Assignment 1

Group 11

Prithvi Shivashankar

Soham Bhalerao

Harkirat kaur

Arjitkumar Sureshkumar

Anjana Sebastian

Sai Pratheek Banda

Q1. What is the distribution of gender, vehicle size, and vehicle class?

The SAS System

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

Distribution of Gender: There is not much variation in terms of gender distribution for the given data.

Distribution of Vehicle Size: Medium-sized vehicles constitute 70% of the data.

Distribution of Vehicle Class: The vehicles are distributed widely. The top 3 dominant vehicle classes in the given data are Four-door car, two-door car and SUV respectively.

Q2. What is the average customer life time value of each level of gender, vehicle size, and vehicle class?

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	3810.24	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

For both males and females,
Luxury SUV has higher Lifetime value under Large and Medium vehicle size and Luxury car has higher lifetime value under small vehicle size on an average.

Q3. Do Large cars have a higher lifetime value than medsize cars. Do a ttest and report on your findings.

The SAS System

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	-970.3	-40.9917	6806.8	6698.7	6918.5
Diff (1-2)	Satterthwaite	-505.7	-960.2	-51.1690			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	7368	-2.13	0.0329
Satterthwaite	Unequal	1259.7	-2.18	0.0292

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

Average vehicle size for medium cars – $\mu 1$
Average vehicle size for large cars – $\mu 2$

Initial test for equality of variance:

- H0: $\mu 1 = \mu 2$
- H1: $\mu 1 \neq \mu 2$

From The equality of variances test we conclude that the variances are equal as the p-value is 0.2183.

We can hence conclude at 5% level of significance that the null hypothesis that the variances are equal cannot be rejected.

H0: Large cars do not have higher lifetime value than medium size car (Medium size car>=Large size car)
H1: Large cars have higher lifetime value than medium size car (Medium size car<Large size car)

The T-value for the case of equality of variance (as tested above) is -2.13 and since this is a left tail test, the t-critical value is -1.64. As the calculated t-value is lesser than the t-critical value, we can reject the null hypothesis in favor of the alternative. Hence, we can conclude that large cars have higher lifetime value than medium sized cars.

Q4. Is there a significant difference between men and women in customer life time value?

The SAS System

The TTEST Procedure

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

Average Customer value for men – μ1
Average Customer value for women – μ2

Initial test for equality of variance:

- H0: μ1= μ2
- H1: μ1≠ μ2

From The equality of variances test we conclude that the variances are equal as the pvalue is 0.0847. We can hence conclude at 5% level of significance that the null hypothesis that the variances are equal cannot be rejected.

H0: Customer life time value of men and women is not different.
H1: Customer life time value of men and women is different.
The p value for equal variances is 0.1934, so we do not reject the null hypothesis at 5% significance level and we don't have enough evidence to claim that customer lifetime value for men and women is different.

Q5. Use ANOVA to test whether there is difference in customer lifetime value across different sales channels. Which sales channel generates the highest lifetime value?

The SAS System

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 SAS System

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 SAS System

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

H0: Customer lifetime value across different sales channels is not different.

H1: Atleast customer lifetime value across one channel is different.

In The above ANOVA test, it is observed that the p value is 0.4503, Hence we do not reject the null hypothesis. Therefore, the data explains that customer lifetime value across different channels is not different.

As per the findings, it is observed that as a total (N * Mean), Agent sales channel generates the highest customer lifetime value.

Q6. What demographic factors (education, income, marital_status) affect customer lifetime value?

- **INCOME and customer life time value:**

The SAS System

The CORR Procedure

2 Variables:	Income Customer_Lifetime_Value

Simple Statistics

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

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

	Income	Customer_Lifetime_Value
Income	1.00000	0.02437 0.0199
Customer_Lifetime_Value	0.02437 0.0199	1.00000

We calculated the correlation to find how income affects customer life time value. Correlation of 0.0243 suggests that there's almost no correlation between income and CLV and since, the p-value is 0.0199 , we conclude that the correlation is not zero. Hence the data explains Income isn't significantly affecting customer life time value.

- **Education and customer life time value:**

The FREQ Procedure

Table of clv by Education

Frequency Percent Row Pct Col Pct	Education					
clv	Bachelor	College	Doctor	High School or	Master	Total
1	692 7.58 30.30 25.18	680 7.44 29.77 25.36	101 1.11 4.42 29.53	629 6.89 27.54 23.99	182 1.99 7.97 24.56	2284 25.01
2	1408 15.41 30.83 51.24	1339 14.66 29.32 49.94	160 1.75 3.50 46.78	1308 14.32 28.64 49.89	352 3.85 7.71 47.50	4567 50.00
3	648 7.09 28.38 23.58	662 7.25 29.00 24.69	81 0.89 3.55 23.68	685 7.50 30.00 26.13	207 2.27 9.07 27.94	2283 24.99
Total	2748 30.09	2681 29.35	342 3.74	2622 28.71	741 8.11	9134 100.00

Statistics for Table of clv by Education

Statistic	DF	Value	Prob
Chi-Square	8	13.0316	0.1108
Likelihood Ratio Chi-Square	8	12.8480	0.1172
Mantel-Haenszel Chi-Square	1	5.1311	0.0235
Phi Coefficient		0.0378	
Contingency Coefficient		0.0377	
Cramer's V		0.0267	

Since Chi square is a procedure for testing if two categorical variables are related or not, we divided Customer Lifetime Value into three levels:
if Customer Lifetime Value < 3994 then clv=1
if 8963 < Customer Lifetime Value < 3994 then clv= 2
if Customer Lifetime Value >= 8963 then clv=3

Chi square test:
H0: Education and Customer life time value are independent. No relationship exists.
H1: Education and Customer life time value are dependent

As you can see, the p value of the chi square test is 0.1108, hence we fail to reject the null hypothesis. Therefore, there is no relationship between education and customer life time value.

- **Marital status and customer life time value**:

Frequency Percent Row Pct Col Pct	Table of clv by Marital_Status			
	Marital_Status			
clv	Divorced	Married	Single	Total
1	353 3.86 15.46 25.79	1271 13.92 55.65 23.99	660 7.23 28.90 26.75	2284 25.01
2	654 7.16 14.32 47.77	2693 29.48 58.97 50.83	1220 13.36 26.71 49.45	4567 50.00
3	362 3.96 15.86 26.44	1334 14.60 58.43 25.18	587 6.43 25.71 23.79	2283 24.99
Total	1369 14.99	5298 58.00	2467 27.01	9134 100.00

Statistics for Table of clv by Marital_Status

Statistic	DF	Value	Prob
Chi-Square	4	10.4112	0.0340
Likelihood Ratio Chi-Square	4	10.3889	0.0344
Mantel-Haenszel Chi-Square	1	3.6196	0.0571
Phi Coefficient		0.0338	
Contingency Coefficient		0.0337	
Cramer's V		0.0239	

Chi square test:
H0: Marital status and Customer life time value are independent. No relationship exists.
H1: Marital status and Customer life time value are dependent.

The p value here is 0.0340, hence we reject the null hypothesis at 5% significance level. Therefore, we conclude that marital status does affect customer lifetime value.

Q7. Is there a relationship between renew_offer_type and response (use Chi-sq test)? Which offer type generates the highest response rate?

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	

H0: Renew Offer Type and Response rate are independent.

H1: Renew Offer Type and Response rate are dependent.

The p value here is less than 0.0001, hence we reject the null hypothesis at 5% significance level. Therefore, we conclude that renew offer type does affect response rate.

Offer type 1 generates the highest response rate.

Q8. Do different renew_offer_types have different lifetime values? Which offer type is the best?

The SAS System

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 SAS System

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

The SAS System

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

H0: Customer lifetime value across different renew offer types is not different.

H1: Customer lifetime value across different renew offer types is different.

The p value here is less than 0.0001, hence we reject the null hypothesis at 5% significance level. Therefore, we conclude that customer lifetime value across different renew offer types is different.

Offer 1 type is the best.

Q9. Is the effectiveness of renew_offer_type different across different states with respect to lifetime value?

Class Level Information		
Class	Levels	Values
State	5	Arizona California Nevada Oregon Washington
Renew_Offer_Type	4	Offer1 Offer2 Offer3 Offer4

Number of Observations Read	9134
Number of Observations Used	9134

The SAS System

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*Renew_Offer_Ty	19	4079881684	214730615	4.58	<.0001

ANOVA test:
Ho: Mean of Renew offer type with respect to lifetime value is equal across different states.
H1: Mean of at least one Renew offer type across different states with respect to lifetime value is different.

The p value here is <0.0001, hence we reject the null hypothesis. Therefore, we conclude that Renew offer type with respect to lifetime value is different across different states at 5% significance level.

Q10. 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.

Insight 1: According to marketing theory, it is always wise to invest more in customers with high customer lifetime value. From the data, we can find out that premium renew type has more CLV than basic type. If we can prove that there is a significant difference between Premium Renew type and Basic

Renew Type with respect to CLV, which can be done using ANOVA, then we can recommend upgrading customers who are likely to churn from basic to premium for free because the cost is minuscule for the company and we can retain those customers.

Insight 2: People who work in urban areas tend to stay in suburban areas because of high cost of housing in the urban areas and factors such as lesser noise and stress. Such people tend to drive from the suburban areas to the urban areas and there is a higher probability that such people could encounter accidents on the road. We could first check the average claim amount for the three different location types and then run an ANOVA to check if there is a significant difference between the three groups and based on that, we would be able to see whether the people from suburban locations have a higher claim amount and then in such a case, we could charge a higher premium.

The SAS System

The MEANS Procedure

Analysis Variable : Total_Claim_Amount						
Location_Code	N Obs	N	Mean	Std Dev	Minimum	Maximum
Rural	1773	1773	109.9050952	76.8463187	0.0990070	562.0885870
Suburban	5779	5779	562.1598701	275.1666070	292.8000000	2893.24
Urban	1582	1582	329.5723289	124.1755121	156.9212470	1065.05

Based on the above table, we can see that people from suburban areas have much higher claim amounts on average and then by running an ANOVA, we can check whether they are significantly different and then based on that, we can decide to charge a higher premium for the suburban people.

Insight 3: If we plot a trend line between customer lifetime value and Months since the last claim, we can find that there is inverse relationship i.e. as the number of months since the last claim increases CLV tends to decrease over time. If we can prove this by running a regression and get that the relationship is significant then we can get a segment of the customer, we can focus on.