MGMT 571 Homework 1 – Ian Bach

Question 1:

D). None of the Above

Question 2:

<u>Row Totals:</u>

Yes: 60 + 100 + 300 = 190

No: 50 + 90 + 20 = 160

<u>Column Totals:</u>

A: 60 + 50 = 110

B: 100 + 90 = 190

C: 30 + 20 = 50

Total of all: 190 + 160 = 350

<u>Expected Frequencies:</u>

E for Yes in A = (190 x 110) / 350 = 59.71

E for Yes in B = (190 x 190) / 350 = 103.14

E for Yes in C = (190 x 50) / 350 = 27.14

E for No in A = (160 x 110) / 350 = 50.29

E for No in B = (160 x 190) / 350 = 86.86

E for No in C = (160 x 50) / 350 = 22.86

<u>Chi-Square Formula</u>

Yes, A: (60-59.71)^2 / 59.71 = 0.0014

Yes, B: (100 – 103.14)^2 / 103.14 = 0.0956

Yes, C: (30 – 27.14)^2 / 27.14 = 0.3014

No, A: (50-50.29)^2 / 50.29 = 0.0017

No, B: (90-86.86)^2 / 86.86 = 0.1135

No, C: (20 – 22.86)^2 / 22.86 = 0.3578

<u>Sum of Chi-Square Values:</u>

0.0014 + 0.0956 + 0.3014 + 0.0017 + 0.1135 + 0.3578 = 0.8714

Degrees of Freedom:

(2-1) x (3-1) = 2

Question 3:

(d) Determine the variable role and measurement level for each variable.

(none) | not | Equal to

Columns: Label | Mining | Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
AreaCode	Input	Nominal	No		No	.	.
Churn_	Target	Binary	No		No	.	.
CSC_Hi	Input	Interval	No		No	.	.
CustServCalls	Input	Interval	No		No	.	.
DayCalls	Input	Interval	No		No	.	.
DayCharges	Input	Interval	No		No	.	.
DayMins	Input	Interval	No		No	.	.
EveCalls	Input	Interval	No		No	.	.
EveCharges	Input	Interval	No		No	.	.
EveMins	Input	Interval	No		No	.	.
IntlCalls	Input	Interval	No		No	.	.
IntlCharges	Input	Interval	No		No	.	.
IntlMins	Input	Interval	No		No	.	.
IntlPlan	Input	Nominal	No		No	.	.
Length	Input	Interval	No		No	.	.
NightCalls	Input	Interval	No		No	.	.
NightCharges	Input	Interval	No		No	.	.
NightMins	Input	Interval	No		No	.	.
Phone	Input	Nominal	No		Yes	.	.
State	Input	Nominal	No		No	.	.
VMailMessage	Input	Interval	No		No	.	.
VmailPlan	Input	Nominal	No		No	.	.

(e) Compare the Area code and State fields. Discuss any apparent abnormalities.

Sample Statistics

Obs #	Variable ...	Label	Type	Percent ...	Minimum	Maximum	Mean	Number o...	Mode Per...	Mode
1	IntlPlan		CLASS	0	.	.	.	2	90.1	NO
2	Phone		CLASS	0	.	.	.	128+	0.775194	327-3954
3	State		CLASS	0	.	.	.	51	2.65	VT
4	VmailPlan		CLASS	0	.	.	.	2	71.5	NO
5	AreaCode		VAR	0	408	510	438.075	.	.	
6	CSC_Hi		VAR	0	0	1	0.0805	.	.	
7	Churn_		VAR	0	0	1	0.136	.	.	
8	CustServC...		VAR	0	0	9	1.547	.	.	
9	DayCalls		VAR	0	0	165	100.4875	.	.	
10	DayCharges		VAR	0	0	59.64	30.62078	.	.	
11	DayMins		VAR	0	0	350.8	180.1189	.	.	
12	EveCalls		VAR	0	12	168	100.139	.	.	
13	EveCharges		VAR	0	2.65	29.89	17.01906	.	.	
14	EveMins		VAR	0	31.2	351.6	200.2218	.	.	
15	IntlCalls		VAR	0	0	19	4.4825	.	.	
16	IntlCharges		VAR	0	0	5.4	2.75419	.	.	
17	IntlMins		VAR	0	0	20	10.1987	.	.	
18	Length		VAR	0	1	243	99.75	.	.	
19	NightCalls		VAR	0	42	175	100.4155	.	.	
20	NightCharg...		VAR	0	1.04	17.19	9.022145	.	.	
21	NightMins		VAR	0	23.2	381.9	200.4915	.	.	
22	VMailMess...		VAR	0	0	51	8.4235	.	.	

State Information:

The "State" variable is categorized as a class variable, indicating it represents categorical data (e.g., abbreviations of U.S. states).

The number of distinct states appears to be 51, suggesting it includes all 50 states plus the District of Columbia, which seems correct.

Area Code Information:

The "Area code" variable is a numeric variable, with a minimum value of 408 and a maximum value of 510. These area codes are commonly associated with specific U.S. regions.

The mode of the "Area code" is around 438, suggesting that this is the most frequent area code in the dataset.

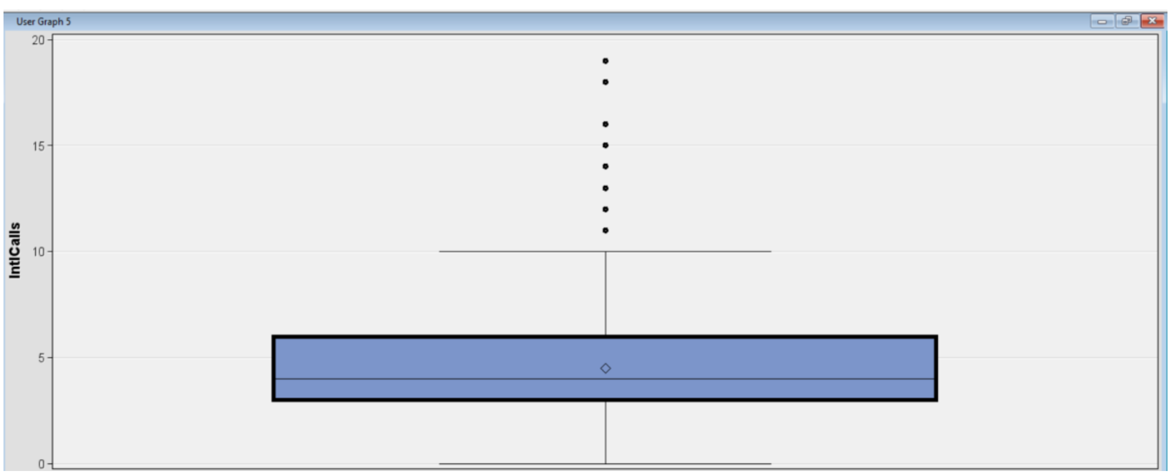
Potential Abnormalities:

Limited Range of Area Codes: The "Area code" values appear to be clustered within a relatively small range (408 to 510). This may indicate that customers in this dataset are primarily located within a particular geographical region. If the dataset claims to cover a wider area (e.g., nationwide), this could be an issue.

Numeric vs. Categorical Representation: "Area code" is treated as a continuous numeric variable. However, it might be better represented as a categorical (class) variable to align it with "State" data and facilitate a more accurate comparison.

Check for Consistency: I need verify if each area code aligns with the corresponding state. For example, area codes 408, 415, and 510 are from California. If these appear in records listed under states like New York or Texas, it would indicate data inconsistencies.

(f) Use a graph to determine visually whether there are any outliers in Total international calls.



Outliers Present:

The plot clearly shows multiple data points above the upper whisker of the box plot, indicating outliers. These are represented by the black dots.

Outliers occur at values higher than around 10 international calls, with the most extreme value close to 20.

Distribution:

The majority of the data is clustered between 0 and 8 international calls, as represented by the interquartile range (the box).

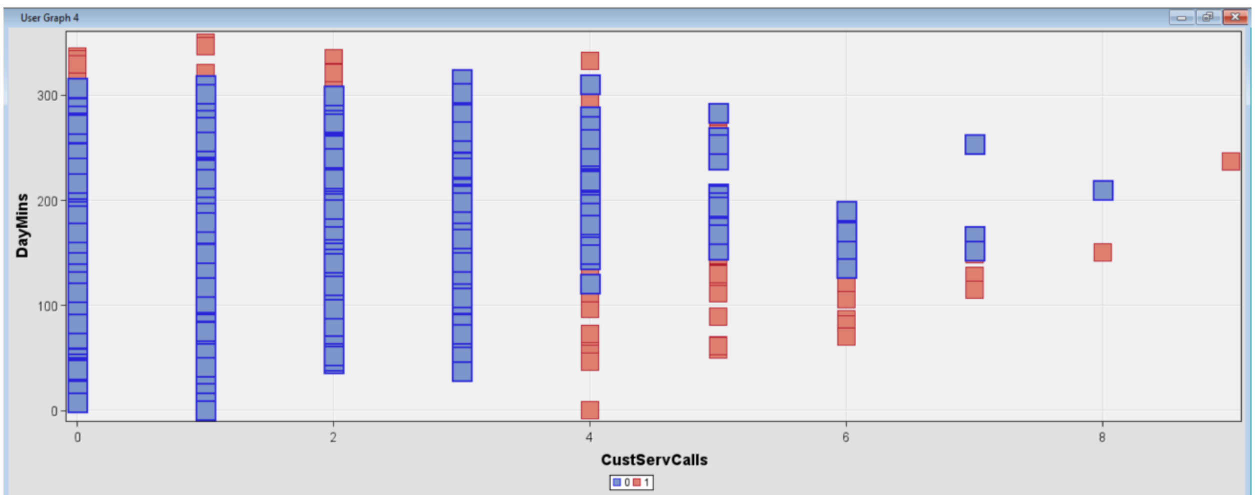
The box plot's central line (median) indicates that the typical customer makes around 4-6 international calls.

Conclusion:

The box plot identifies several customers who make significantly more international calls than the average, marking them as outliers. These outliers could represent heavy international users, and their behavior might be worth analyzing separately, especially if there is a correlation between high international call usage and churn.

(g) Does a 2D scatter plot between Number of calls to customer service and

Total day minutes (group by the target variable) reveal any information?



High Number of Customer Service Calls:

There appears to be a trend where customers with more calls to customer service (values of 4 and above on the x-axis) show more instances of churn (red markers). This suggests that frequent contact with customer service might be linked to dissatisfaction, potentially leading to churn.

Distribution Across DayMins:

Customers with varying DayMins usage (from low to high) can be seen throughout the plot. There doesn't seem to be a clear trend indicating that higher or lower DayMins alone directly correlates with churn.

However, when combined with higher CustServCalls, there seems to be a stronger indication of churn. This suggests an interaction effect where a higher number of customer service calls may lead to churn regardless of the level of day usage.

Potential Groupings:

Consider segmenting customers based on CustServCalls into groups (e.g., low, medium, and high contact) and further analyze their likelihood to churn based on other variables.

(h) Utilize the Chi-square table with a significance level of 0.05 to determine which,

out of the 20 predictor variables, are useful in predicting customer churn.

Table: Chi-Square Plot

Data Role	Segment	Segment Id	Segment Name:Value	Target	Input	Cramer's V	Prob	Chi-Square	Df	Role	Label	Ordered Inputs	Group	Plot
TRAIN			_OVERALL_	Churn_	CustServCalls	0.314044	<.0001	328.7132	4	INPUT	CustServCalls	1	1	1
TRAIN			_OVERALL_	Churn_	CSC_Hi	0.311804	<.0001	324.0392	1	INPUT	CSC_Hi	2	2	1
TRAIN			_OVERALL_	Churn_	DayCharges	0.306068	<.0001	312.2281	4	INPUT	DayCharges	3	3	1
TRAIN			_OVERALL_	Churn_	DayMins	0.306068	<.0001	312.2281	4	INPUT	DayMins	4	4	1
TRAIN			_OVERALL_	Churn_	IntlPlan	0.259852	<.0001	225.0541	1	INPUT	IntlPlan	5	5	1
TRAIN			_OVERALL_	Churn_	State	0.157847	0.0023	83.0438	50	INPUT	State	6	6	1
TRAIN			_OVERALL_	Churn_	VMailMessage	0.10734	<.0001	38.4021	4	INPUT	VMailMessage	7	7	1
TRAIN			_OVERALL_	Churn_	VmailPlan	0.102148	<.0001	34.7773	1	INPUT	VmailPlan	8	8	1
TRAIN			_OVERALL_	Churn_	EveMins	0.084761	<.0001	23.9455	4	INPUT	EveMins	9	9	1
TRAIN			_OVERALL_	Churn_	EveCharges	0.083749	0.0001	23.3770	4	INPUT	EveCharges	10	10	1
TRAIN			_OVERALL_	Churn_	IntlCharges	0.071407	0.0019	16.9946	4	INPUT	IntlCharges	11	11	1
TRAIN			_OVERALL_	Churn_	IntlMins	0.071407	0.0019	16.9946	4	INPUT	IntlMins	12	12	1
TRAIN			_OVERALL_	Churn_	IntlCalls	0.059672	0.0184	11.8680	4	INPUT	IntlCalls	13	13	1
TRAIN			_OVERALL_	Churn_	DayCalls	0.047048	0.1172	7.3776	4	INPUT	DayCalls	14	14	1
TRAIN			_OVERALL_	Churn_	NightMins	0.042402	0.1997	5.9926	4	INPUT	NightMins	15	15	1
TRAIN			_OVERALL_	Churn_	NightCharges	0.04196	0.2092	5.8681	4	INPUT	NightCharges	16	16	1
TRAIN			_OVERALL_	Churn_	NightCalls	0.033148	0.4536	3.6622	4	INPUT	NightCalls	17	17	1
TRAIN			_OVERALL_	Churn_	Length	0.024653	0.7310	2.0256	4	INPUT	Length	18	18	1
TRAIN			_OVERALL_	Churn_	EveCalls	0.017894	0.8994	1.0672	4	INPUT	EveCalls	19	19	1
TRAIN			_OVERALL_	Churn_	AreaCode	0.007298	0.9151	0.1775	2	INPUT	AreaCode	20	20	1

Key Variables with Significant Association ($p < 0.05$):

CustServCalls: High Cramer's V value (0.314), indicating a strong association with churn. This suggests that the number of calls to customer service is an important predictor.

CSC_H: Significant with a strong association (Cramer's V = 0.311), likely indicating customer service issues or high engagement in this category.

DayCharges & DayMins: Both have high Cramer's V values (0.307 and 0.306), suggesting a strong correlation. Since they are highly correlated, you may choose one to avoid redundancy.

IntlPlan: Moderate association (0.259), indicating customers on international plans are more likely to churn.

State: Moderate significance (Cramer's V = 0.158), which could be useful for geographic segmentation.

Variables with Lower Predictive Value:

AreaCode: Very low Cramer's V (0.007), and a high p-value, suggesting it has little to no association with churn. This variable can be rejected.

Length (Account length): Low association (Cramer's V = 0.025), meaning it may not be as useful in predicting churn.

EveCalls, NightCalls, and NightCharges: Lower Cramer's V values (0.033 and below), suggesting less predictive strength.

Recommendations:

Focus on Key Variables: Prioritize high Cramer's V variables such as CustServCalls, DayCharges, IntlPlan, and CSC_H.

Consider Rejecting Low-Predictive Variables: Consider rejecting variables like AreaCode, Length, and those with low Cramer's V values (e.g., NightCharges, EveCalls).