Data Preparation

Look at the attribute type; e.g., nominal, ordinal or quantitative.

Categorical/Nominal Attributes:

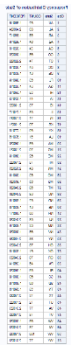
1. State: Customer's state.
2. Area Code: Area code of customer.
3. Phone: Phone number of customer.
4. Intl_Plan: Binary indicator showing whether the customer has an international calling plan (Yes/No).
5. VMail_Plan: Indicator of voicemail plan (Yes/No).
6. Churn: Class attribute with binary values (True for churn and False for not churn).

Numerical/Quantitative Attributes:

1. Account Length: Integer number showing the duration of activity for the customer account.
2. CustServ Calls: The number of calls to customer support service.
3. Day Calls: Discrete attribute indicating the total number of calls during daytime.
4. Day Charge: Charges for using the service during daytime (continuous data type).
5. Day Mins: The number of minutes the customer used the service during daytime (continuous quantitative data type).
6. Eve Calls: The number of calls during evening time.
7. Eve Charge: Charges for using the service during evening time (continuous data type).
8. Eve Mins: The number of minutes the customer used the service during evening time (continuous quantitative data type).
9. Intl Calls: The number of international calls.
10. Intl Charge: Charges for international calls (continuous data type).
11. Intl Mins: The number of minutes the customer used the service to make international calls (continuous quantitative data type).
12. Night Calls: The number of calls during night time.
13. Night Charge: Charges for using the service during night time (continuous data type).
14. Night Mins: Number of minutes the customer used the service during night time (continuous quantitative data type).
15. VMail Message: The number of voicemail messages.

Determine any outlier values (records) for each of the attributes or attributes under consideration (min, max, std. dev, scatter plots, box plots or others can be used).

State – Categorical



Outliers in State Based on Frequency Distribution

Obs	State	COUNT	PERCENT	_TYPE_	_FREQ_	mean_freq	std_freq
1	CA	34	1.02010	0	51	65.3529	11.8014
2	WV	106	3.18032	0	51	65.3529	11.8014

- **Outlier Detection**
  - Outliers are identified using the mean and standard deviation. The typical range for normal data is within:
  - **Mean - 2 × Standard Deviation**
  - **Mean + 2 × Standard Deviation**
  - So, the outlier thresholds are:
  - **Lower Threshold**: Mean - 2 × Std Dev = 65.35 - 2 × 11.80 = 41.75
  - **Upper Threshold**: Mean + 2 × Std Dev = 65.35 + 2 × 11.80 = 89.95
  - Records that fall outside these thresholds are considered outliers.
- **Interpretation of Specific States**
  - **California (CA)**:
    - **Count**: 34
    - **Percentage**: 1.02%
    - **Analysis**:
    - CA has a frequency count of 34, which is below the lower threshold of 41.75.
    - This indicates that CA's count is significantly lower than the average state frequency.
    - **Conclusion**: CA is an outlier with a notably lower frequency compared to other states.
  - **West Virginia (WV)**:
    - **Count**: 106
    - **Percentage**: 3.18%
    - **Analysis**:
    - WV has a frequency count of 106, which is above the upper threshold of 89.95.
    - This indicates that WV's count is significantly higher than the average state frequency.
    - **Conclusion**: WV is an outlier with a notably higher frequency compared to other states

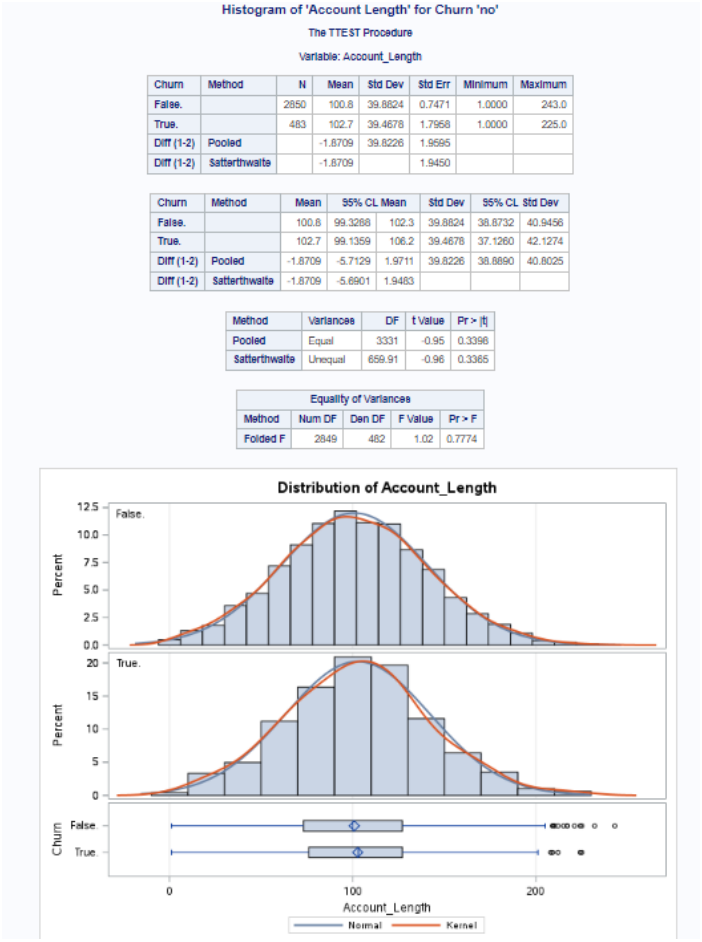
Summary

- **California (CA)** has fewer records than most other states, making it an outlier with a lower frequency.
- **West Virginia (WV)** has more records than most other states, making it an outlier with a higher frequency.

Analyze the distribution of numeric attributes (normal or other). Plot histograms for attributes of concern and analyze whether they have any influence on the class attribute. (NUMERICAL AND CLASS CORRELATIONS)

Account Length and Class Attribute

- Both the pooled and Satterthwaite t-tests show p-values of 0.3398 and 0.3365, respectively. These p-values are greater than the common significance level of 0.05, indicating that the difference in account length between the two groups is not statistically significant.
- The analysis suggests that there is no statistically significant difference in account length between clients who churned and those who did not churn. This means that account length alone is not a strong factor in predicting whether a client will churn.



Histogram of 'Account Length' for Churn 'no'

The TTEST Procedure

Variable: Account_Length

Churn	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
False.		2850	100.8	39.8824	0.7471	1.0000	243.0
True.		483	102.7	39.4678	1.7958	1.0000	225.0
Diff (1-2)	Pooled		-1.8709	39.8226	1.9595		
Diff (1-2)	Satterthwaite		-1.8709		1.9450		

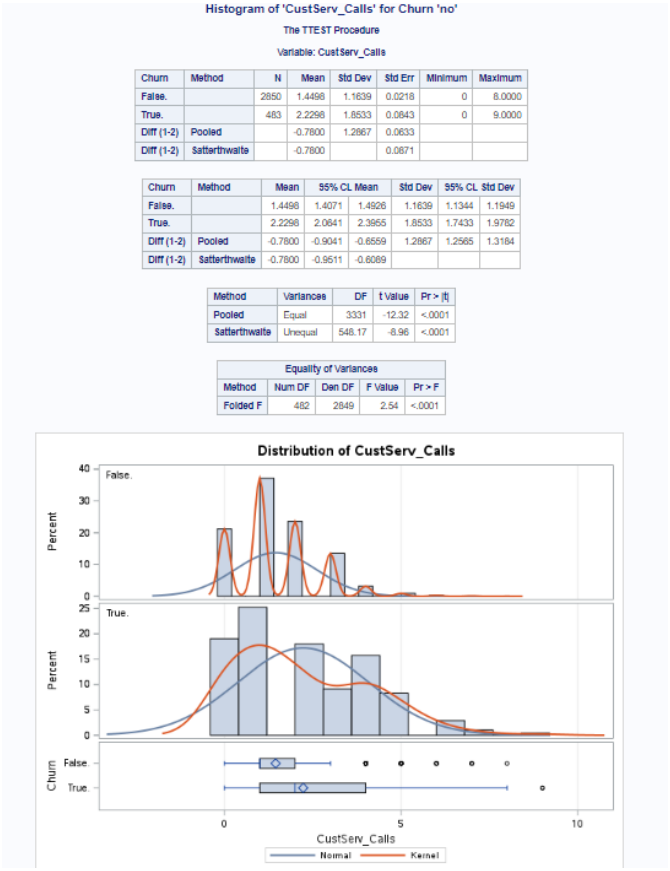
Churn	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
False.		100.8	99.3268	102.3	39.8824	38.8732	40.9456
True.		102.7	99.1359	106.2	39.4678	37.1260	42.1274
Diff (1-2)	Pooled	-1.8709	-5.7129	1.9711	39.8226	38.8690	40.8025
Diff (1-2)	Satterthwaite	-1.8709	-5.6901	1.9483			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	3331	-0.95	0.3398
Satterthwaite	Unequal	659.91	-0.96	0.3365

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	2849	482	1.02	0.7774

CustServ Calls & Class

- The mean number of customer service calls for clients who churned (yes) is higher (2.2298) compared to those who did not churn (no) (1.4498).
    - Both the pooled and Satterthwaite t-tests show highly significant p-values (<.0001), indicating that the difference in the number of customer service calls between the two groups is statistically significant.
    - The analysis reveals a statistically significant difference in the number of customer service calls between clients who have churned and those who have not. Clients who have churned tend to have a higher number of customer service calls compared to those who have not churned. This suggests that customer service calls are an influential factor in determining whether a client churns.



Histogram of 'CustServ_Calls' for Churn 'no'

The TTEST Procedure

Variable: CustServ_Calls

Churn	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
False.		2850	1.4498	1.1639	0.0218	0	8.0000
True.		483	2.2298	1.8533	0.0843	0	9.0000
Diff (1-2)	Pooled		-0.7800	1.2867	0.0633		
Diff (1-2)	Satterthwaite		-0.7800		0.0671		

Churn	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
False.		1.4498	1.4071	1.4926	1.1639	1.1344	1.1949
True.		2.2298	2.0641	2.3955	1.8533	1.7433	1.9782
Diff (1-2)	Pooled	-0.7800	-0.9041	-0.6559	1.2867	1.2565	1.3184
Diff (1-2)	Satterthwaite	-0.7800	-0.9511	-0.6089			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	3331	-12.32	<.0001
Satterthwaite	Unequal	548.17	-8.96	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	482	2849	2.54	<.0001

Day Calls & Class

- The mean number of Day_Calls for clients who churned (yes) is slightly higher (101.3) compared to those who did not churn (no) (100.3).
    - Both the pooled and Satterthwaite t-tests result in p-values of 0.2867 and 0.3165, respectively. These p-values are greater than the conventional significance level of 0.05, indicating that the difference in the number of Day_Calls between the churned and non-churned groups is not statistically significant.
    - The analysis suggests that there is no statistically significant difference in the number of Day_Calls between clients who churned and those who did not. This indicates that the number of Day_Calls may not be a strong factor in predicting churn status.

Histogram of 'Day_Calls' for Churn 'no'

The TTEST Procedure

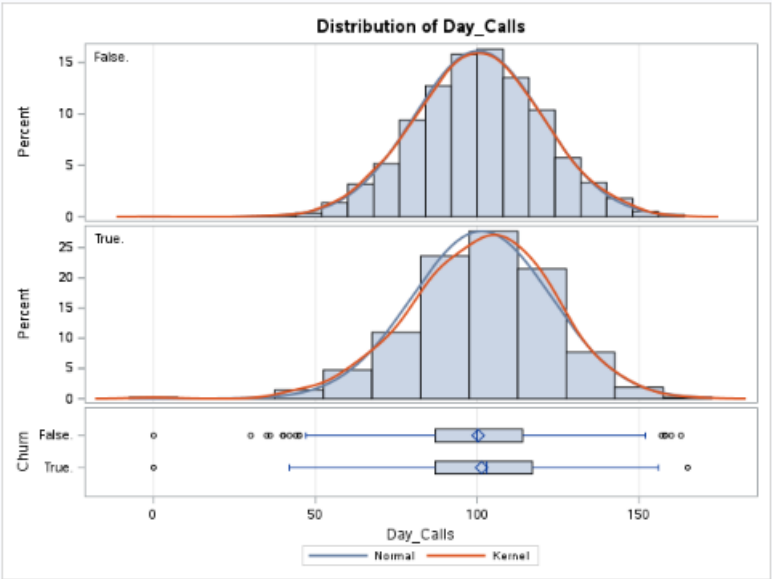
Variable: Day_Calls

Churn	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
False.		2850	100.3	19.8012	0.3709	0	163.0
True.		483	101.3	21.5823	0.9820	0	165.0
Diff (1-2)	Pooled		-1.0522	20.0687	0.9675		
Diff (1-2)	Satterthwaite		-1.0522		1.0497		

Churn	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
False.		100.3	99.5559	101.0	19.8012	19.3001	20.3291
True.		101.3	99.4058	103.3	21.5823	20.3017	23.0368
Diff (1-2)	Pooled	-1.0522	-2.9884	0.8839	20.0687	19.5982	20.5625
Diff (1-2)	Satterthwaite	-1.0522	-3.1137	1.0092			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	3331	-1.07	0.2867
Satterthwaite	Unequal	627.17	-1.00	0.3165

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	482	2849	1.19	0.0111



Day Charge & Class

- The mean Day_Charge for customers who did not churn (no) is lower (29.7804) compared to those who churned (yes) (35.1759).
    - Both the pooled and Satterthwaite t-tests show highly significant p-values (<.0001), indicating that the difference in the Day_Charge between the two groups is statistically significant.
    - The analysis suggests that there is a statistically significant difference in the Day_Charge between customers who churned and those who did not. Customers who churned (yes) tend to have higher Day_Charge compared to those who did not churn (no). This indicates that higher daytime charges may be associated with an increased likelihood of customer churn.

Histogram of 'Day_Charge' for Churn 'no'

The TTEST Procedure

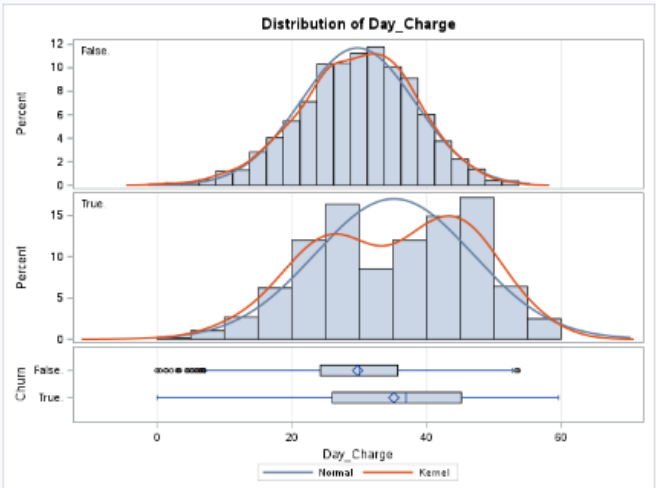
Variable: Day_Charge

Churn	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
False.		2850	29.7804	8.5308	0.1598	0	53.6500
True.		483	35.1759	11.7297	0.5337	0	59.6400
Diff (1-2)	Pooled		-5.3955	9.0639	0.4460		
Diff (1-2)	Satterthwaite		-5.3955		0.5571		

Churn	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
False.		29.7804	29.4671	30.0938	8.5308	8.3150	8.7583
True.		35.1759	34.1272	36.2246	11.7297	11.0337	12.5201
Diff (1-2)	Pooled	-5.3955	-6.2700	-4.5210	9.0639	8.8513	9.2869
Diff (1-2)	Satterthwaite	-5.3955	-6.4898	-4.3012			

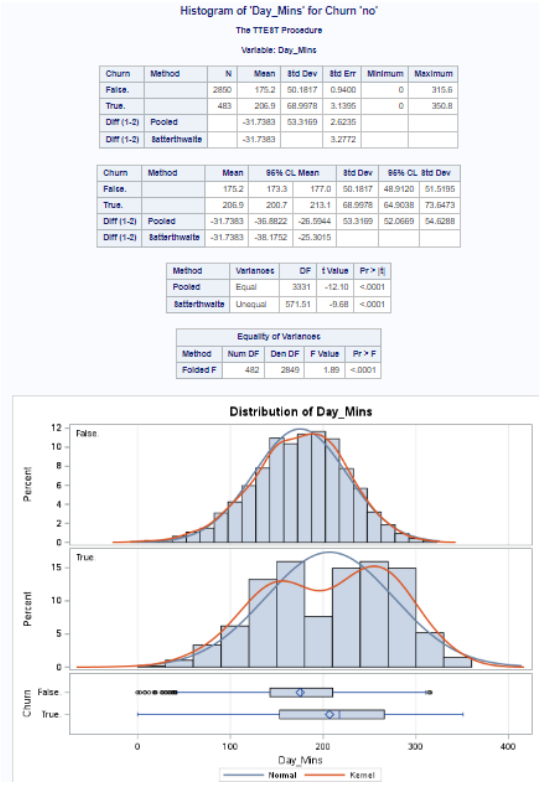
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	3331	-12.10	<.0001
Satterthwaite	Unequal	571.51	-9.68	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	482	2849	1.89	<.0001



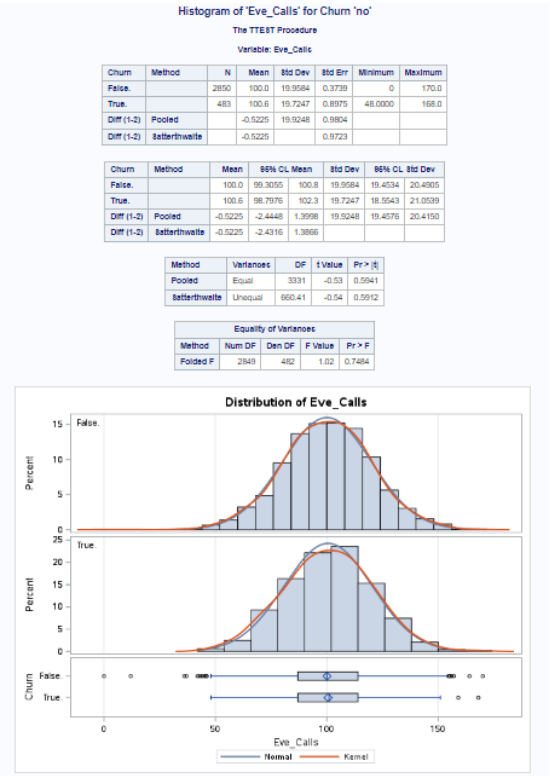
Day Mins & Class

- The mean Day_Mins for customers who churned (yes) is higher (206.9 minutes) compared to those who did not churn (no) (175.2 minutes).
    - Both the pooled and Satterthwaite t-tests show highly significant p-values (<.0001), indicating that the difference in Day_Mins between the two groups is statistically significant.
    - The analysis suggests that there is a statistically significant difference in the number of daytime minutes used between customers who churned and those who did not. Customers who churned tend to have a higher number of daytime minutes, implying that higher usage during the day could be a factor associated with customer churn.



Eve Calls & Class

- The mean number of evening calls is slightly higher for clients who churned (yes) compared to those who did not churn (no), but the difference is very small (-0.5225) and not statistically significant.
    - The p-values from both the pooled and Satterthwaite t-tests are greater than 0.05, indicating that there is no strong evidence to suggest a significant difference in the number of evening calls between the churned and non-churned clients.
    - The analysis implies that the number of evening calls is not a strong predictor of churn status, as there is no significant difference between the two groups in terms of evening calls.



Histogram of 'Eve_Calls' for Churn 'no'

The TTEST Procedure

Variable: Eve_Calls

Churn	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
False.		2850	100.0	19.9584	0.3739	0	170.0
True.		483	100.6	19.7247	0.8975	48.0000	168.0
Diff (1-2)	Pooled		-0.5225	19.9248	0.9804		
Diff (1-2)	Satterthwaite		-0.5225		0.9723		

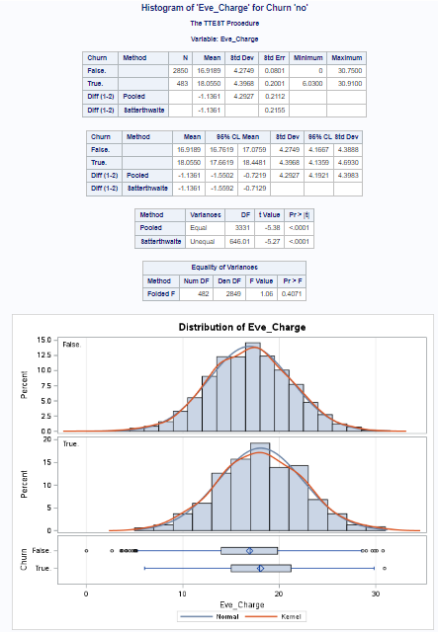
Churn	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
False.		100.0	99.3055	100.8	19.9584	19.4534	20.4905
True.		100.6	98.7976	102.3	19.7247	18.5543	21.0539
Diff (1-2)	Pooled	-0.5225	-2.4448	1.3998	19.9248	19.4576	20.4150
Diff (1-2)	Satterthwaite	-0.5225	-2.4316	1.3866			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	3331	-0.53	0.5941
Satterthwaite	Unequal	660.41	-0.54	0.5912

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	2849	482	1.02	0.7484

Eve Charge & Class

- The mean Eve_Charge is higher for clients who churned (yes) compared to those who did not churn (no). The mean Eve_Charge for churned clients is 18.0550, while for non-churned clients it is 16.9189, resulting in a difference of -1.1361. This difference is statistically significant.
    - The p-values from both the pooled and Satterthwaite t-tests are less than 0.0001, indicating that there is strong evidence to suggest a significant difference in Eve_Charge between churned and non-churned clients.
    - The analysis implies that Eve_Charge is a significant predictor of churn status, with higher Eve_Charge values associated with a higher likelihood of churn.



Eve Mins & Class

- The mean number of evening minutes is higher for clients who churned (yes) compared to those who did not churn (no). Specifically, clients who churned have an average of 212.4 minutes, while non-churned clients have an average of 199.0 minutes. The difference is approximately -13.37 minutes.
    - The p-values from both the pooled and Satterthwaite t-tests are less than 0.0001, indicating that the difference in the number of evening minutes between churned and non-churned clients is statistically significant.
    - The analysis suggests that the number of evening minutes is a significant predictor of churn status. Clients who churn tend to use more evening minutes compared to those who do not churn, which implies that higher evening minutes are associated with a higher likelihood of churn.

Histogram of 'Eve_Mins' for Churn 'no'

The TTEST Procedure

Variable: Eve_Mins

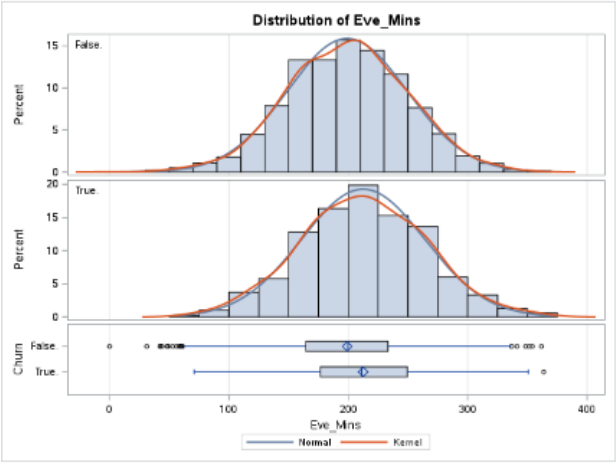
Churn	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
False.		2850	199.0	50.2922	0.9421	0	361.8
True.		483	212.4	51.7289	2.3537	70.9000	363.7
Diff (1-2)	Pooled		-13.3668	50.5026	2.4851		
Diff (1-2)	Satterthwaite		-13.3668		2.5353		

Churn	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
False.		199.0	197.2	200.9	50.2922	49.0197	51.6330
True.		212.4	207.8	217.0	51.7289	48.6596	55.2147
Diff (1-2)	Pooled	-13.3668	-18.2392	-8.4945	50.5026	49.3186	51.7453
Diff (1-2)	Satterthwaite	-13.3668	-18.3452	-8.3885			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	3331	-5.38	<.0001
Satterthwaite	Unequal	645.99	-5.27	<.0001

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	482	2849	1.06	0.4062



Intl Calls & Class

- The mean number of international calls is slightly higher for clients who did not churn (no) compared to those who churned (yes), with a difference of 0.3694. This difference is statistically significant.
    - The p-values from both the pooled t-test (0.0023) and the Satterthwaite t-test (0.0032) are less than 0.05, indicating that there is a significant difference in the number of international calls between churned and non-churned clients.
    - The analysis suggests that international calls do have a significant relationship with churn status. Non-churned clients tend to make slightly more international calls compared to churned clients. This difference, though statistically significant, is relatively small, implying that while there is a difference, it may not be a strong predictor of churn on its own.

Histogram of 'Intl_Calls' for Churn 'no'

The TTEST Procedure

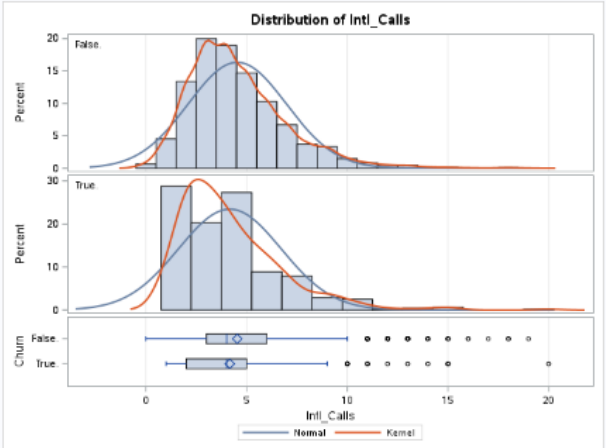
Variable: Intl_Calls

Churn	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
False.		2850	4.5330	2.4420	0.0457	0	19.0000
True.		483	4.1636	2.5516	0.1161	1.0000	20.0000
Diff (1-2)	Pooled		0.3694	2.4581	0.1210		
Diff (1-2)	Satterthwaite		0.3694		0.1248		

Churn	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
False.		4.5330	4.4433	4.6227	2.4420	2.3802	2.5071
True.		4.1636	3.9354	4.3917	2.5516	2.4002	2.7235
Diff (1-2)	Pooled	0.3694	0.1323	0.6066	2.4581	2.4005	2.5186
Diff (1-2)	Satterthwaite	0.3694	0.1244	0.6145			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	3331	3.05	0.0023
Satterthwaite	Unequal	640.64	2.96	0.0032

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	482	2849	1.09	0.1971



Intl Charge & Class

- The mean Intl_Charge is slightly higher for clients who churned (yes) compared to those who did not churn (no). The mean Intl_Charge is 2.8895 for churned clients and 2.7434 for non-churned clients, with a difference of -0.1461.
    - The p-values from both the pooled and Satterthwaite t-tests are less than 0.05 (0.0001), indicating strong evidence to suggest a significant difference in the Intl_Charge between churned and non-churned clients.
    - The analysis implies that Intl_Charge is a significant predictor of churn status, as there is a statistically significant difference between the two groups in terms of Intl_Charge. Specifically, churned clients tend to have a slightly higher average Intl_Charge compared to non-churned clients.

Histogram of 'Intl_Charge' for Churn 'no'

The TTEST Procedure

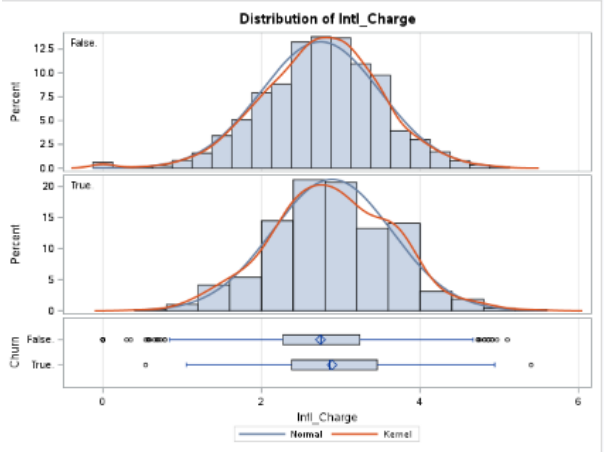
Variable: Intl_Charge

Churn	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
False.		2850	2.7434	0.7518	0.0141	0	5.1000
True.		483	2.8895	0.7542	0.0343	0.5400	5.4000
Diff (1-2)	Pooled		-0.1461	0.7521	0.0370		
Diff (1-2)	Satterthwaite		-0.1461		0.0371		

Churn	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
False.		2.7434	2.7158	2.7710	0.7518	0.7328	0.7718
True.		2.8895	2.8221	2.9570	0.7542	0.7094	0.8050
Diff (1-2)	Pooled	-0.1461	-0.2187	-0.0736	0.7521	0.7345	0.7706
Diff (1-2)	Satterthwaite	-0.1461	-0.2190	-0.0733			

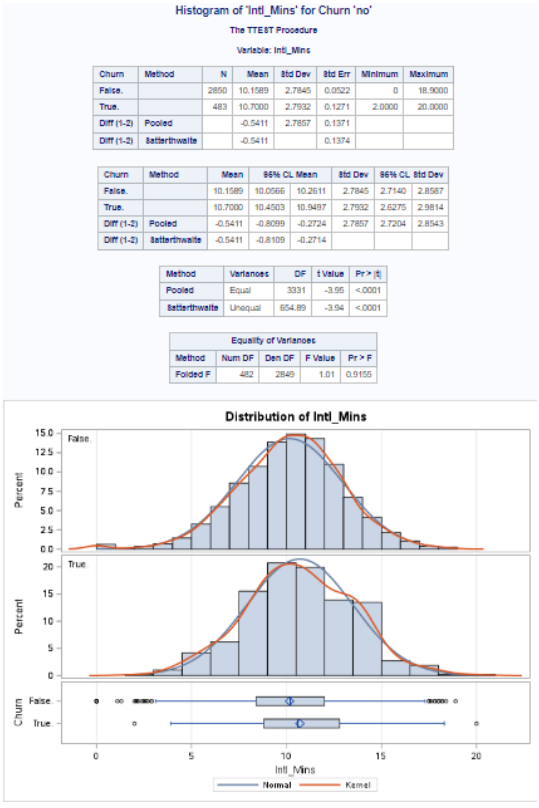
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	3331	-3.95	<.0001
Satterthwaite	Unequal	654.88	-3.94	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	482	2849	1.01	0.9149



Intl Mins & Class

- The mean number of international minutes is slightly higher for clients who churned (yes) compared to those who did not churn (no), but the difference is statistically significant.
- The mean Intl_Mins for clients who churned is 10.70, compared to 10.16 for those who did not churn. The difference in means is -0.5411.
    - The p-values from both the pooled and Satterthwaite t-tests are less than 0.05 (0.0001), indicating a significant difference in the number of international minutes between churned and non-churned clients.
    - The analysis suggests that Intl_Mins does have a statistically significant relationship with churn status, implying that the number of international minutes could be a relevant factor in predicting customer churn.



Histogram of 'Intl_Mins' for Churn 'no'

The TTEST Procedure

Variable: Intl_Mins

Churn	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
False.		2850	10.1589	2.7845	0.0522	0	18.9000
True.		483	10.7000	2.7932	0.1271	2.0000	20.0000
Diff (1-2)	Pooled		-0.5411	2.7857	0.1371		
Diff (1-2)	Satterthwaite		-0.5411		0.1374		

Churn	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
False.		10.1589	10.0566	10.2611	2.7845	2.7140	2.8587
True.		10.7000	10.4503	10.9497	2.7932	2.6275	2.9814
Diff (1-2)	Pooled	-0.5411	-0.8099	-0.2724	2.7857	2.7204	2.8543
Diff (1-2)	Satterthwaite	-0.5411	-0.8109	-0.2714			

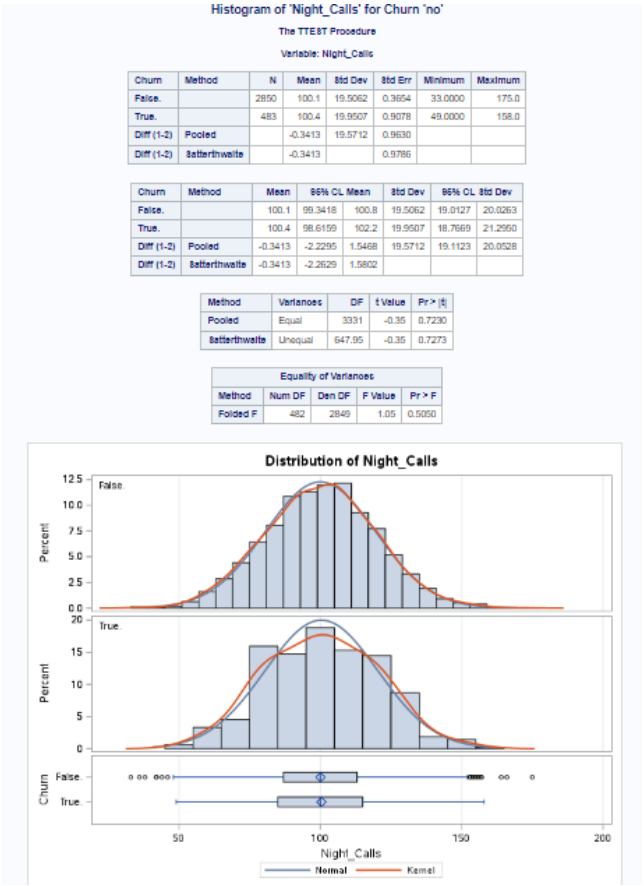
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	3331	-3.95	<.0001
Satterthwaite	Unequal	654.89	-3.94	<.0001

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	482	2849	1.01	0.9155

Night Calls & Class

- Mean Difference: The mean number of night calls is slightly higher for clients who churned (yes) compared to those who did not churn (no). However, the difference in means is very small, with a difference of -0.3413.
    - Statistical Significance:
        - The p-values from both the pooled t-test (p = 0.7230) and the Satterthwaite t-test (p = 0.7273) are greater than 0.05. This indicates that there is no strong evidence to suggest a significant difference in the number of night calls between clients who churned and those who did not.
        - The analysis implies that the number of night calls is not a strong predictor of churn status, as the observed differences in night calls between the two groups are not statistically significant.



Histogram of 'Night_Calls' for Churn 'no'

The TTEST Procedure

Variable: Night_Calls

Churn	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
False.		2850	100.1	19.5062	0.3654	33.0000	175.0
True.		483	100.4	19.9507	0.9078	49.0000	158.0
Diff (1-2)	Pooled		-0.3413	19.5712	0.9630		
Diff (1-2)	Satterthwaite		-0.3413		0.9786		

Churn	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
False.		100.1	99.3418	100.8	19.5062	19.0127	20.0263
True.		100.4	98.6159	102.2	19.9507	18.7669	21.2950
Diff (1-2)	Pooled	-0.3413	-2.2295	1.5468	19.5712	19.1123	20.0528
Diff (1-2)	Satterthwaite	-0.3413	-2.2629	1.5802			

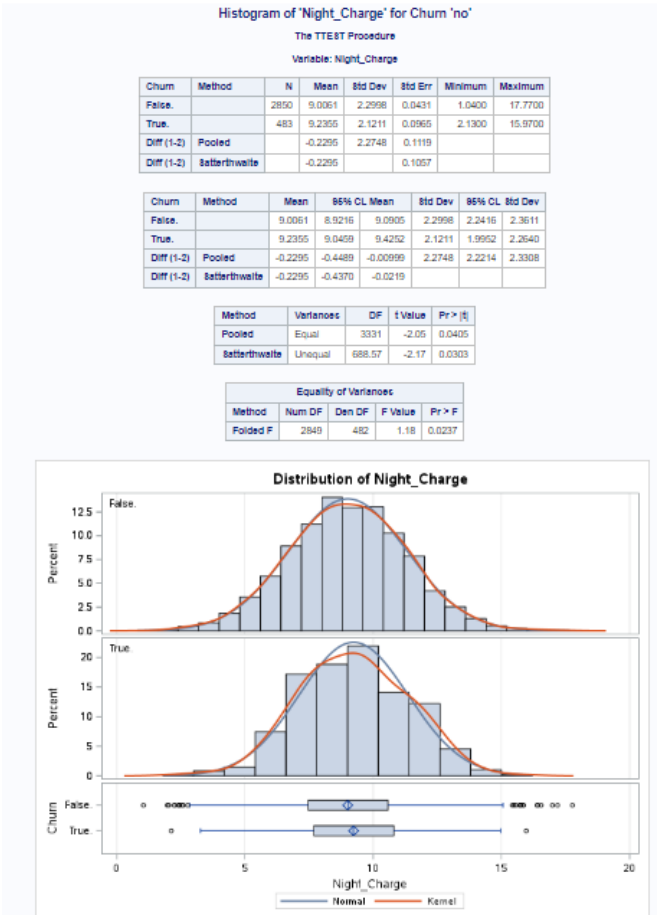
Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	3331	-0.35	0.7230
Satterthwaite	Unequal	647.95	-0.35	0.7273

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	482	2849	1.05	0.5050

Night Charge & Class

- The mean night charge is slightly higher for clients who churned (9.24) compared to those who did not churn (9.01). The difference is small (-0.2295) but statistically significant at the 0.05 level.
    - The p-values from both the pooled (0.0405) and Satterthwaite (0.0303) t-tests are less than 0.05, indicating that there is evidence to suggest a significant difference in the night charge between the churned and non-churned clients.
    - Although the difference in night charge between the two groups is statistically significant, the magnitude of the difference is small, which may limit its practical significance as a predictor of churn status.



Night Mins & Class

- The mean number of night minutes is slightly higher for clients who churned (205.2) compared to those who did not churn (200.1), but the difference is small (-5.0985) and statistically significant.
    - The p-values from both the pooled (p = 0.0405) and Satterthwaite (p = 0.0303) t-tests are less than 0.05, indicating that there is strong evidence to suggest a significant difference in the number of night minutes between the churned and non-churned clients.
    - The analysis implies that the number of night minutes might be a predictor of churn status, as there is a statistically significant difference between the two groups in terms of night minutes. However, the practical significance of this difference should be considered, as the actual difference in means is relatively small.

Histogram of 'Night_Mins' for Churn 'no'

The TTEST Procedure

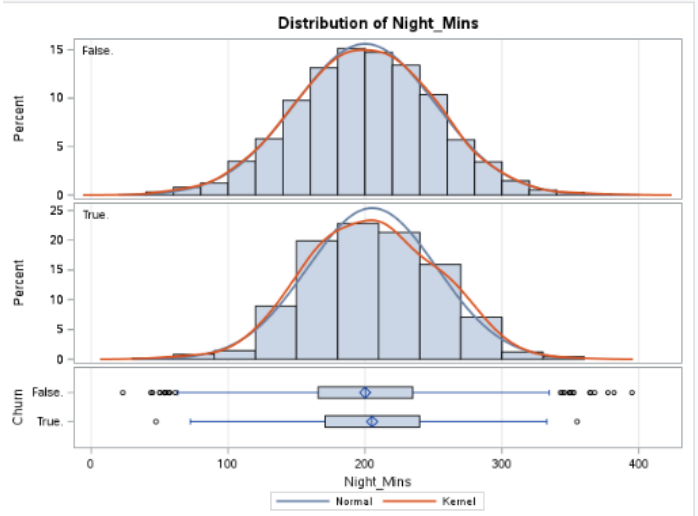
Variable: Night_Mins

Churn	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
False.		2850	200.1	51.1050	0.9573	23.2000	395.0
True.		483	205.2	47.1328	2.1446	47.4000	354.9
Diff (1-2)	Pooled		-5.0985	50.5496	2.4874		
Diff (1-2)	Satterthwaite		-5.0985		2.3486		

Churn	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
False.		200.1	198.3	202.0	51.1050	49.8120	52.4675
True.		205.2	201.0	209.4	47.1328	44.3362	50.3089
Diff (1-2)	Pooled	-5.0985	-9.9754	-0.2216	50.5496	49.3644	51.7934
Diff (1-2)	Satterthwaite	-5.0985	-9.7097	-0.4873			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	3331	-2.05	0.0405
Satterthwaite	Unequal	688.58	-2.17	0.0303

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	2849	482	1.18	0.0237



Vmail Message & Class

- The mean number of voicemail messages is significantly higher for clients who did not churn (no) compared to those who did churn (yes), with a mean difference of 3.4886. This difference is statistically significant.
    - The p-values from both the pooled and Satterthwaite t-tests are less than 0.05, indicating strong evidence of a significant difference in the number of voicemail messages between churned and non-churned clients.
    - The analysis suggests that the number of voicemail messages is a significant predictor of churn status, as there is a substantial and statistically significant difference between the two groups in terms of voicemail messages.

Histogram of 'Vmail_Message' for Churn 'no'

The TTEST Procedure

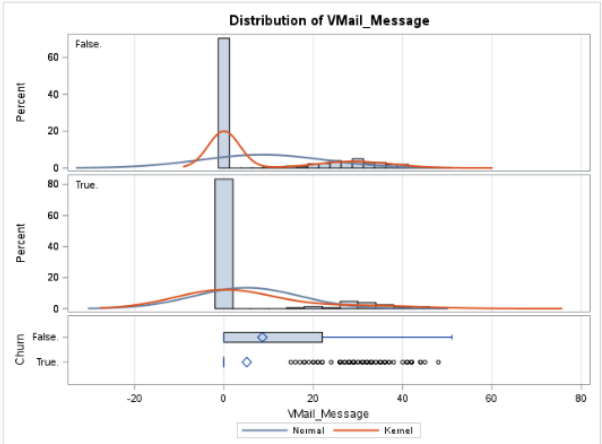
Variable: VMail_Message

Churn	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
False.		2850	8.6046	13.9131	0.2606	0	51.0000
True.		483	5.1159	11.8601	0.5397	0	48.0000
Diff (1-2)	Pooled		3.4886	13.6352	0.6709		
Diff (1-2)	Satterthwaite		3.4886		0.5993		

Churn	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
False.		8.6046	8.0935	9.1156	13.9131	13.5611	14.2841
True.		5.1159	4.0556	6.1763	11.8601	11.1564	12.6593
Diff (1-2)	Pooled	3.4886	2.1731	4.8041	13.6352	13.3155	13.9707
Diff (1-2)	Satterthwaite	3.4886	2.3121	4.6652			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	3331	5.20	<.0001
Satterthwaite	Unequal	726.36	5.82	<.0001

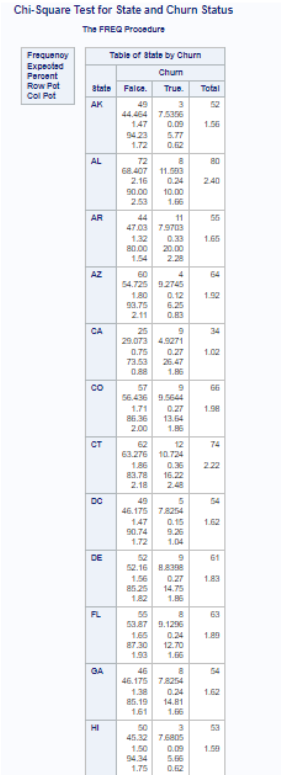
Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	2849	482	1.38	<.0001



(CATEGORICAL AND CLASS CORRELATIONS)

State & Class

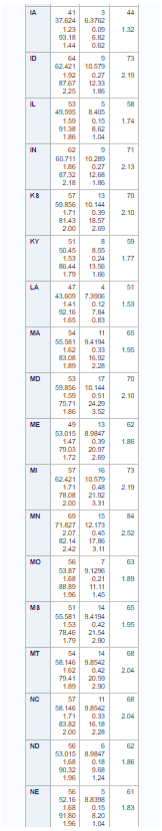
- The p-value of 0.0023 indicates a significant association between the state and whether a client churned. This suggests that the likelihood of churn is not independent of the state a client is from.
- The Phi Coefficient and Cramer's V values are both 0.1578, which indicates a weak to moderate association between State and Churn. This implies that while there is a significant association, the strength of this relationship is not very strong.
- States with Higher Churn Rates:
    - West Virginia (WV): Has the highest proportion of churn with 14.49% of clients churning. This state shows a higher likelihood of churn compared to other states.
    - Mississippi (MS): This state has a relatively high churn rate at 21.54%, suggesting a greater likelihood of churn.
- States with Lower Churn Rates:
    - North Dakota (ND): Has the lowest proportion of churn with only 9.68% of clients churning. This state exhibits a lower likelihood of churn.
    - Nebraska (NE): Also has a lower churn rate at 8.20%, indicating a reduced likelihood of churn compared to other states.



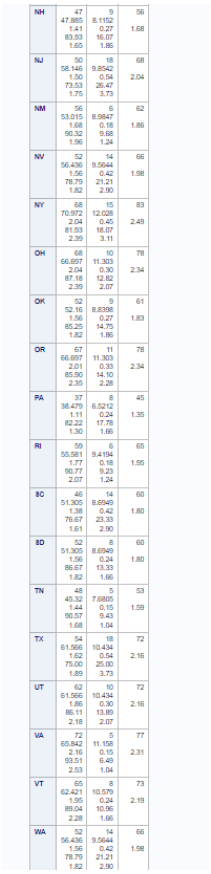
Chi-Square Test for State and Churn Status

The FREQ Procedure

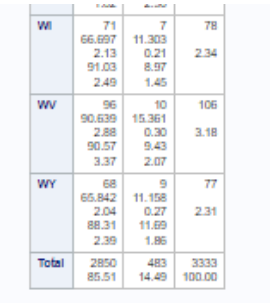
Frequency Expected Percent Row Pct Col Pct	Table of State by Churn		
	Churn		
State	False.	True.	Total
AK	49 44.464 1.47 94.23 1.72	3 7.5356 0.09 5.77 0.62	52 1.56
AL	72 68.407 2.16 90.00 2.53	8 11.593 0.24 10.00 1.66	80 2.40
AR	44 47.03 1.32 80.00 1.54	11 7.9703 0.33 20.00 2.28	55 1.65
AZ	60 54.725 1.80 93.75 2.11	4 9.2745 0.12 6.25 0.83	64 1.92
CA	25 29.073 0.75 73.53 0.88	9 4.9271 0.27 26.47 1.86	34 1.02
CO	57 56.436 1.71 86.36 2.00	9 9.5644 0.27 13.64 1.86	66 1.98
CT	62 63.276 1.86 83.78 2.18	12 10.724 0.36 16.22 2.48	74 2.22
DC	49 46.175 1.47 90.74 1.72	5 7.8254 0.15 9.26 1.04	54 1.62
DE	52 52.16 1.56 85.25 1.82	9 8.8398 0.27 14.75 1.86	61 1.83
FL	55 53.87 1.65 87.30 1.93	8 9.1296 0.24 12.70 1.66	63 1.89
GA	46 46.175 1.38 85.19 1.61	8 7.8254 0.24 14.81 1.66	54 1.62
HI	50 45.32 1.50 94.34 1.75	3 7.6805 0.09 5.66 0.62	53 1.59



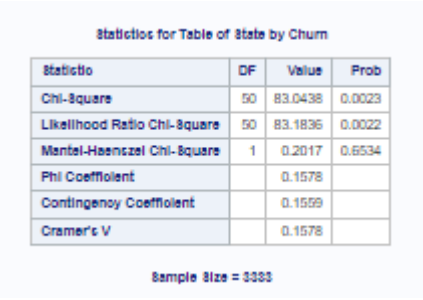
IA	41 37.624 1.23 93.18 1.44	3 6.3762 0.09 6.82 0.62	44 1.32
ID	64 62.421 1.92 87.67 2.25	9 10.579 0.27 12.33 1.86	73 2.19
IL	53 49.595 1.59 91.38 1.86	5 8.405 0.15 8.62 1.04	58 1.74
IN	62 60.711 1.86 87.32 2.18	9 10.289 0.27 12.68 1.86	71 2.13
KS	57 59.856 1.71 81.43 2.00	13 10.144 0.39 18.57 2.69	70 2.10
KY	51 50.45 1.53 86.44 1.79	8 8.55 0.24 13.56 1.66	59 1.77
LA	47 43.609 1.41 92.16 1.65	4 7.3906 0.12 7.84 0.83	51 1.53
MA	54 55.581 1.62 83.08 1.89	11 9.4194 0.33 16.92 2.28	65 1.95
MD	53 59.856 1.59 75.71 1.86	17 10.144 0.51 24.29 3.52	70 2.10
ME	49 53.015 1.47 79.03 1.72	13 8.9847 0.39 20.97 2.69	62 1.86
MI	57 62.421 1.71 78.08 2.00	16 10.579 0.48 21.92 3.31	73 2.19
MN	69 71.827 2.07 82.14 2.42	15 12.173 0.45 17.86 3.11	84 2.52
MO	56 53.87 1.68 88.89 1.96	7 9.1296 0.21 11.11 1.45	63 1.89
MS	51 55.581 1.53 78.46 1.79	14 9.4194 0.42 21.54 2.90	65 1.95
MT	54 58.146 1.62 79.41 1.89	14 9.8542 0.42 20.59 2.90	68 2.04
NC	57 58.146 1.71 83.82 2.00	11 9.8542 0.33 16.18 2.28	68 2.04
ND	56 53.015 1.68 90.32 1.96	6 8.9847 0.18 9.68 1.24	62 1.86
NE	56 52.16 1.68 91.80 1.96	5 8.8398 0.15 8.20 1.04	61 1.83



NH	47 47.885 1.41 83.93 1.65	9 8.1152 0.27 16.07 1.86	56 1.68
NJ	50 58.146 1.50 73.53 1.75	18 9.8542 0.54 26.47 3.73	68 2.04
NM	56 53.015 1.68 90.32 1.96	6 8.9847 0.18 9.68 1.24	62 1.86
NV	52 56.436 1.56 78.79 1.82	14 9.5644 0.42 21.21 2.90	66 1.98
NY	68 70.972 2.04 81.93 2.39	15 12.028 0.45 18.07 3.11	83 2.49
OH	68 66.697 2.04 87.18 2.39	10 11.303 0.30 12.82 2.07	78 2.34
OK	52 52.16 1.56 85.25 1.82	9 8.8398 0.27 14.75 1.86	61 1.83
OR	67 66.697 2.01 85.90 2.35	11 11.303 0.33 14.10 2.28	78 2.34
PA	37 38.479 1.11 82.22 1.30	8 6.5212 0.24 17.78 1.66	45 1.35
RI	59 55.581 1.77 90.77 2.07	6 9.4194 0.18 9.23 1.24	65 1.95
SC	46 51.305 1.38 76.67 1.61	14 8.6949 0.42 23.33 2.90	60 1.80
SD	52 51.305 1.56 86.67 1.82	8 8.6949 0.24 13.33 1.66	60 1.80
TN	48 45.32 1.44 90.57 1.68	5 7.6805 0.15 9.43 1.04	53 1.59
TX	54 61.566 1.62 75.00 1.89	18 10.434 0.54 25.00 3.73	72 2.16
UT	62 61.566 1.86 86.11 2.18	10 10.434 0.30 13.89 2.07	72 2.16
VA	72 65.842 2.16 93.51 2.53	5 11.158 0.15 6.49 1.04	77 2.31
VT	65 62.421 1.95 89.04 2.28	8 10.579 0.24 10.96 1.66	73 2.19
WA	52 56.436 1.56 78.79 1.82	14 9.5644 0.42 21.21 2.90	66 1.98



WI	71 66.697 2.13 91.03 2.49	7 11.303 0.21 8.97 1.45	78 2.34
WV	96 90.639 2.88 90.57 3.37	10 15.361 0.30 9.43 2.07	106 3.18
WY	68 65.842 2.04 88.31 2.39	9 11.158 0.27 11.69 1.86	77 2.31
Total	2850 85.51	483 14.49	3333 100.00



Statistics for Table of State by Churn

Statistic	DF	Value	Prob
Chi-Square	50	83.0438	0.0023
Likelihood Ratio Chi-Square	50	83.1836	0.0022
Mantel-Haenszel Chi-Square	1	0.2017	0.6534
Phi Coefficient		0.1578	
Contingency Coefficient		0.1559	
Cramer's V		0.1578	

Sample Size = 3333

Area_Code & Class

- The p-value of 0.9151 indicates no significant association between Area_Code and whether a client churned. This suggests that the likelihood of churn is independent of the area code a client is from.
- The Phi Coefficient and Cramer's V values are both 0.0073, indicating a very weak association between Area_Code and Churn. This implies that while there may be some relationship, it is not strong.
- Area Codes with Higher Churn Rates:
    - Area Code 415: This area code has the highest number of churned clients with 236 out of 1655 clients, indicating a higher likelihood of churn compared to other area codes.
    - Area Code 510: This area code has 125 churned clients out of 840, suggesting a notable churn rate compared to the other area codes.
- Area Codes with Lower Churn Rates:
    - Area Code 408: This area code has 122 churned clients out of 1226, indicating a somewhat lower likelihood of churn compared to the higher churn rate area codes.
    - Area Code 838: This area code has 83 churned clients out of 331, showing a similar level of likelihood for churn compared to the other lower churn rate area codes.

Contingency Table for Area_Code and Churn Status

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of Area_Code by Churn		
	Churn		
Area_Code	False.	True.	Total
408	716 21.48 85.44 25.12	122 3.66 14.56 25.26	838 25.14
415	1419 42.57 85.74 49.79	236 7.08 14.26 48.86	1655 49.65
510	715 21.45 85.12 25.09	125 3.75 14.88 25.88	840 25.20
Total	2850 85.51	483 14.49	3333 100.00

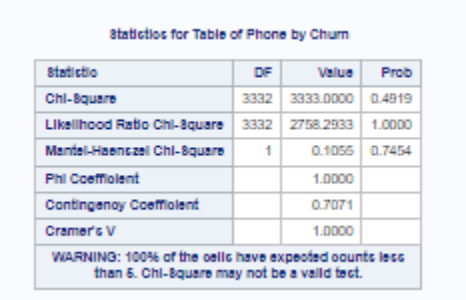
Statistics for Table of Area_Code by Churn

Statistic	DF	Value	Prob
Chi-Square	2	0.1775	0.9151
Likelihood Ratio Chi-Square	2	0.1771	0.9153
Mantel-Haenszel Chi-Square	1	0.1270	0.7215
Phi Coefficient		0.0073	
Contingency Coefficient		0.0073	
Cramer's V		0.0073	

Sample Size = 3333

Phone & Class

- Chi-Square Test:
  - p-value: 0.4919
  - The p-value indicates that there is no significant association between Phone and Churn at the conventional significance level (e.g., 0.05). A p-value of 0.4919 suggests that the relationship between Phone and Churn is likely due to chance.
- Phi Coefficient/Cramer's V:
  - Phi Coefficient: 1.0000
  - Cramer's V: 1.0000
  - Both the Phi Coefficient and Cramer's V are at their maximum values. However, these values are not meaningful here due to the warning about expected counts. When 100% of cells have expected counts less than 5, these measures may not accurately represent the strength of association.



Statistics for Table of Phone by Churn

Statistic	DF	Value	Prob
Chi-Square	3332	3333.0000	0.4919
Likelihood Ratio Chi-Square	3332	2758.2933	1.0000
Mantel-Haenszel Chi-Square	1	0.1055	0.7454
Phi Coefficient		1.0000	
Contingency Coefficient		0.7071	
Cramer's V		1.0000	

WARNING: 100% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Intl_Plan & Class

- The p-value of <0.0001 indicates a highly significant association between the Intl_Plan status and whether a client churned. This suggests that the likelihood of churn is not independent of whether a client has an international plan.
- The Phi Coefficient and Cramer's V values are both 0.2599, which indicates a moderate association between Intl_Plan and Churn. This implies that while there is a significant relationship between having an international plan and the likelihood of churn, the strength of this relationship is moderate.
- Clients with Higher Churn Rates:
    - Clients with an international plan (Intl_Plan = yes):
        - 137 out of 323 clients with an international plan churned, which is a higher churn rate compared to those without an international plan.
- Clients with Lower Churn Rates:
    - Clients without an international plan (Intl_Plan = no):
        - 346 out of 3010 clients without an international plan churned, which is a lower churn rate compared to those with an international plan.

Contingency Table for Intl_Plan and Churn Status

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of Intl_Plan by Churn

Intl_Plan	Churn: False.	Churn: True.	Total
no	2664 79.93 88.50 93.47	346 10.38 11.50 71.64	3010 90.31
yes	186 5.58 57.59 6.53	137 4.11 42.41 28.36	323 9.69
Total	2850 85.51	483 14.49	3333 100.00

Statistics for Table of Intl_Plan by Churn

Statistic	DF	Value	Prob
Chi-Square	1	225.0541	<.0001
Likelihood Ratio Chi-Square	1	170.3998	<.0001
Continuity Adj. Chi-Square	1	222.5658	<.0001
Mantel-Haenszel Chi-Square	1	224.9866	<.0001
Phi Coefficient		0.2599	
Contingency Coefficient		0.2515	
Cramer's V		0.2599	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	2664
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	<.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Sample Size = 3333

Vmail_Plan & Class

- The p-value of <0.0001 indicates a highly significant association between the Vmail_Plan status and whether a client churned. This suggests that the likelihood of churn is not independent of whether a client has a voicemail plan.
- The Phi Coefficient and Cramer's V values are both -0.1021, which indicates a weak association between Vmail_Plan and Churn. This implies that while there is a significant relationship between having a voicemail plan and the likelihood of churn, the strength of this relationship is weak.
- Clients with Higher Churn Rates:
    - Clients without a voicemail plan (VMail_Plan = no):
        - Out of 2411 clients without a voicemail plan, 403 churned, resulting in a churn rate of approximately 16.73%.
- Clients with Lower Churn Rates:
    - Clients with a voicemail plan (VMail_Plan = yes):
        - Out of 922 clients with a voicemail plan, 80 churned, resulting in a churn rate of approximately 8.68%.
- This corrected analysis indicates that clients without a voicemail plan have a higher churn rate compared to those with a voicemail plan, suggesting that having a voicemail plan might be associated with lower churn

Contingency Table for Vmail_Plan and Churn Status

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of VMail_Plan by Churn		
	Churn		
VMail_Plan	False.	True.	Total
no	2008 60.25 83.28 70.46	403 12.09 16.72 83.44	2411 72.34
yes	842 25.26 91.32 29.54	80 2.40 8.68 16.56	922 27.66
Total	2850 85.51	483 14.49	3333 100.00

Statistics for Table of VMail_Plan by Churn

Statistic	DF	Value	Prob
Chi-Square	1	34.7773	<.0001
Likelihood Ratio Chi-Square	1	37.9643	<.0001
Continuity Adj. Chi-Square	1	34.1317	<.0001
Mantel-Haenszel Chi-Square	1	34.7669	<.0001
Phi Coefficient		-0.1021	
Contingency Coefficient		0.1016	
Cramer's V		-0.1021	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	2008
Left-sided Pr <= F	<.0001
Right-sided Pr >= F	1.0000
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Sample Size = 3333

(NUMERICAL CORRELATIONS)

Pearson Correlation Coefficients, N = 3333
Prob > |r| under H0: Rho=0

	Account_Length	CustServ_Calls	Day_Calls	Day_Charge	Day_Mins	Eve_Calls	Eve_Charge	Eve_Mins	Intl_Calls	Intl_Charge	Intl_Mins	Night_Calls	Night_Charge	Night_Mins	VMail_Message
Account_Length	1.00000	-0.00380 0.8266	0.03847 0.0264	0.00621 0.7199	0.00622 0.7198	0.01926 0.2663	-0.00675 0.6971	-0.00676 0.6966	0.02066 0.2331	0.00955 0.5817	0.00951 0.5830	-0.01318 0.4470	-0.00896 0.6051	-0.00896 0.6053	-0.00463 0.7894
CustServ_Calls	-0.00380 0.8266	1.00000	-0.01894 0.2743	-0.01343 0.4384	-0.01342 0.4385	0.00242 0.8888	-0.01299 0.4535	-0.01298 0.4536	-0.01756 0.3108	-0.00967 0.5766	-0.00964 0.5780	-0.01280 0.4600	-0.00928 0.5924	-0.00929 0.5920	-0.01326 0.4440
Day_Calls	0.03847 0.0264	-0.01894 0.2743	1.00000	0.00675 0.6967	0.00675 0.6969	0.00646 0.7092	-0.02145 0.2157	-0.02145 0.2157	0.00457 0.7918	0.02167 0.2111	0.02156 0.2133	-0.01956 0.2590	0.02293 0.1857	0.02294 0.1855	-0.00955 0.5816
Day_Charge	0.00621 0.7199	-0.01343 0.4384	0.00675 0.6967	1.00000	1.00000 <.0001	0.01577 0.3628	0.00704 0.6847	0.00705 0.6841	0.00803 0.6430	-0.01009 0.5602	-0.01016 0.5578	0.02297 0.1849	0.00430 0.8040	0.00432 0.8029	0.00078 0.9643
Day_Mins	0.00622 0.7198	-0.01342 0.4385	0.00675 0.6969	1.00000 <.0001	1.00000	0.01577 0.3628	0.00703 0.6850	0.00704 0.6844	0.00803 0.6429	-0.01009 0.5603	-0.01015 0.5578	0.02297 0.1849	0.00430 0.8040	0.00432 0.8030	0.00078 0.9642
Eve_Calls	0.01926 0.2663	0.00242 0.8888	0.00646 0.7092	0.01577 0.3628	0.01577 0.3628	1.00000	-0.01142 0.5097	-0.01143 0.5095	0.01743 0.3143	0.00867 0.6167	0.00870 0.6155	0.00771 0.6564	-0.00206 0.9056	-0.00209 0.9039	-0.00586 0.7350
Eve_Charge	-0.00675 0.6971	-0.01299 0.4535	-0.02145 0.2157	0.00704 0.6847	0.00703 0.6850	-0.01142 0.5097	1.00000	1.00000 <.0001	0.00254 0.8834	-0.01107 0.5227	-0.01104 0.5239	0.00760 0.6611	-0.01260 0.4671	-0.01259 0.4674	0.01758 0.3103
Eve_Mins	-0.00676 0.6966	-0.01298 0.4536	-0.02145 0.2157	0.00705 0.6841	0.00704 0.6844	-0.01143 0.5095	1.00000 <.0001	1.00000	0.00254 0.8834	-0.01107 0.5230	-0.01103 0.5242	0.00759 0.6615	-0.01259 0.4674	-0.01258 0.4677	0.01756 0.3108
Intl_Calls	0.02066 0.2331	-0.01756 0.3108	0.00457 0.7918	0.00803 0.6430	0.00803 0.6429	0.01743 0.3143	0.00254 0.8834	0.00254 0.8834	1.00000	0.03237 0.0617	0.03230 0.0622	0.00030 0.9860	-0.01233 0.4767	-0.01235 0.4759	0.01396 0.4205
Intl_Charge	0.00955 0.5817	-0.00967 0.5766	0.02167 0.2111	-0.01009 0.5602	-0.01009 0.5603	0.00867 0.6167	-0.01107 0.5227	-0.01107 0.5230	0.03237 0.0617	1.00000	0.99999 <.0001	-0.01363 0.4315	-0.01519 0.3808	-0.01518 0.3810	0.00288 0.8678
Intl_Mins	0.00951 0.5830	-0.00964 0.5780	0.02156 0.2133	-0.01016 0.5578	-0.01015 0.5578	0.00870 0.6155	-0.01104 0.5239	-0.01103 0.5242	0.03230 0.0622	0.99999 <.0001	1.00000	-0.01360 0.4323	-0.01521 0.3799	-0.01521 0.3801	0.00286 0.8691
Night_Calls	-0.01318 0.4470	-0.01280 0.4600	-0.01956 0.2590	0.02297 0.1849	0.02297 0.1849	0.00771 0.6564	0.00760 0.6611	0.00759 0.6615	0.00030 0.9860	-0.01363 0.4315	-0.01360 0.4323	1.00000	0.01119 0.5185	0.01120 0.5179	0.00712 0.6810
Night_Charge	-0.00896 0.6051	-0.00928 0.5924	0.02293 0.1857	0.00430 0.8040	0.00430 0.8040	-0.00206 0.9056	-0.01260 0.4671	-0.01259 0.4674	-0.01233 0.4767	-0.01519 0.3808	-0.01521 0.3799	0.01119 0.5185	1.00000	1.00000 <.0001	0.00766 0.6583
Night_Mins	-0.00896 0.6053	-0.00929 0.5920	0.02294 0.1855	0.00432 0.8029	0.00432 0.8030	-0.00209 0.9039	-0.01259 0.4674	-0.01258 0.4677	-0.01235 0.4759	-0.01518 0.3810	-0.01521 0.3801	0.01120 0.5179	1.00000 <.0001	1.00000	0.00768 0.6576
VMail_Message	-0.00463 0.7894	-0.01326 0.4440	-0.00955 0.5816	0.00078 0.9643	0.00078 0.9642	-0.00586 0.7350	0.01758 0.3103	0.01756 0.3108	0.01396 0.4205	0.00288 0.8678	0.00286 0.8691	0.00712 0.6810	0.00766 0.6583	0.00768 0.6576	1.00000

(CATEGORICAL CORRELATIONS)

>Only showing Correlation = Yes

Area_Code & Intl_Plan

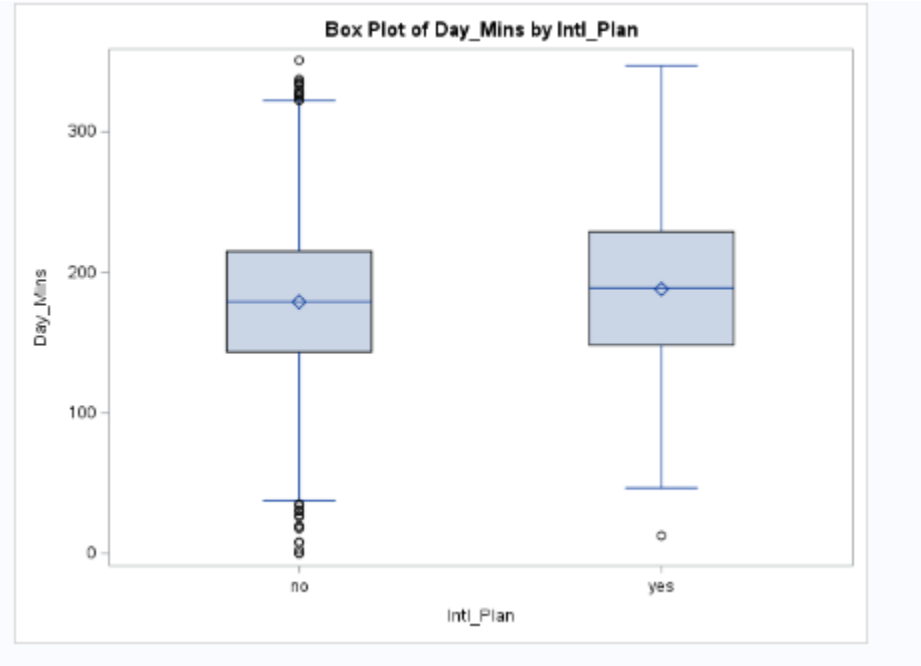
- p-value: The p-value is 0.0189, which is less than 0.05. This indicates a significant association between Area_Code and Intl_Plan. This suggests that the likelihood of having an international plan is not independent of the area code.
- Phi Coefficient and Cramer's V: Both values are 0.0488, which indicates a very weak association between Area_Code and Intl_Plan. This implies that while there is a statistically significant relationship, the strength of the association is quite weak.
- Frequency Distribution:
    - Clients with Intl_Plan = yes:
        - The percentage of clients with an international plan varies by area code:
            - Area Code 408: 8.47%
            - Area Code 415: 9.06%
            - Area Code 510: 12.14%
    - Clients with Intl_Plan = no:
        - The percentage of clients without an international plan:
            - Area Code 408: 91.53%
            - Area Code 415: 90.94%
            - Area Code 510: 87.86%

(CATEGORICAL and NUMERICAL CORRELATIONS)

>Only showing Correlation = Yes

Day_Mins & Intl_Plan

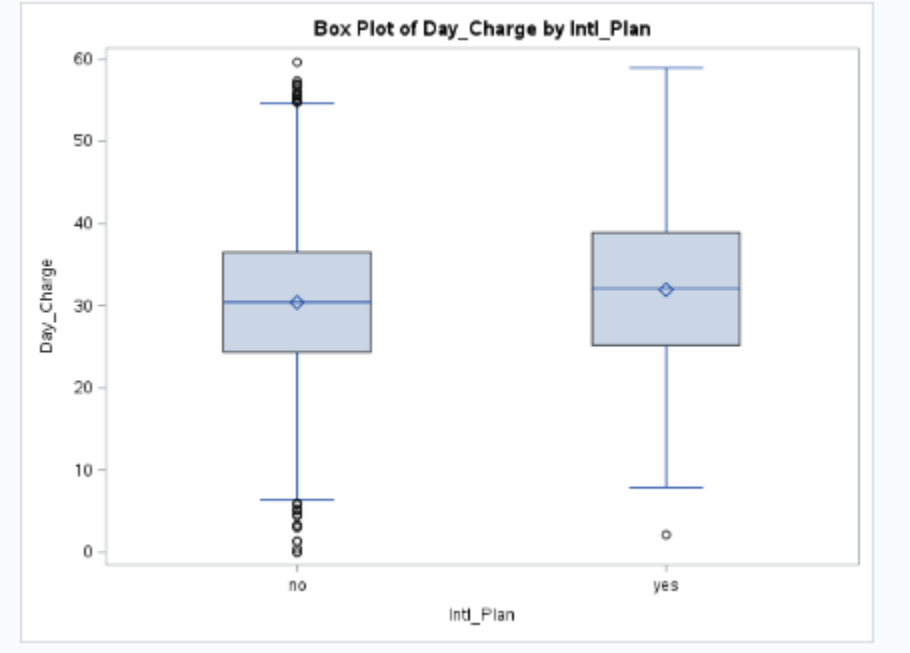
- The mean number of Day_Mins is slightly higher for clients with an international plan (Intl_Plan = yes) compared to those without (Intl_Plan = no), but the difference is relatively small and statistically significant.
    - The F-value is 8.15 with a p-value of 0.0043, indicating a statistically significant difference in Day_Mins between clients with and without an international plan.
    - The R-squared value is 0.002440, suggesting that the model explains a very small proportion of the variance in Day_Mins.
    - The p-value from the ANOVA is less than 0.05, which means there is sufficient evidence to suggest a significant difference in Day_Mins based on the Intl_Plan status.
    - The analysis indicates that while there is a significant difference in Day_Mins related to having an international plan, the practical significance might be limited due to the small R-squared value.



Day_Charge & Intl_Plan

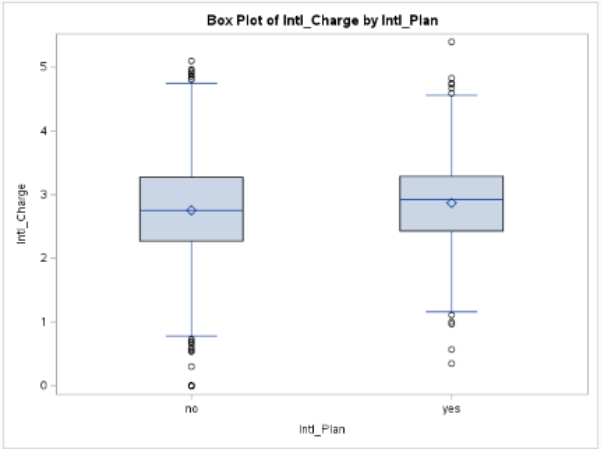
- The mean Day_Charge is slightly higher for clients with an international plan (Intl_Plan = yes) compared to those without (Intl_Plan = no), and this difference is statistically significant.
    - The F-value is 8.15 with a p-value of 0.0043, indicating a statistically significant difference in Day_Charge between clients with and without an international plan.
    - The R-squared value is 0.002440, which shows that the model explains a very small proportion of the variance in Day_Charge.

- The p-value from the ANOVA is less than 0.05, providing strong evidence of a significant difference in Day_Charge based on the Intl_Plan status.
- Despite the statistical significance, the small R-squared value suggests that the practical impact of Intl_Plan on Day_Charge might be limited, as the model does not explain much of the variability in the dependent variable.



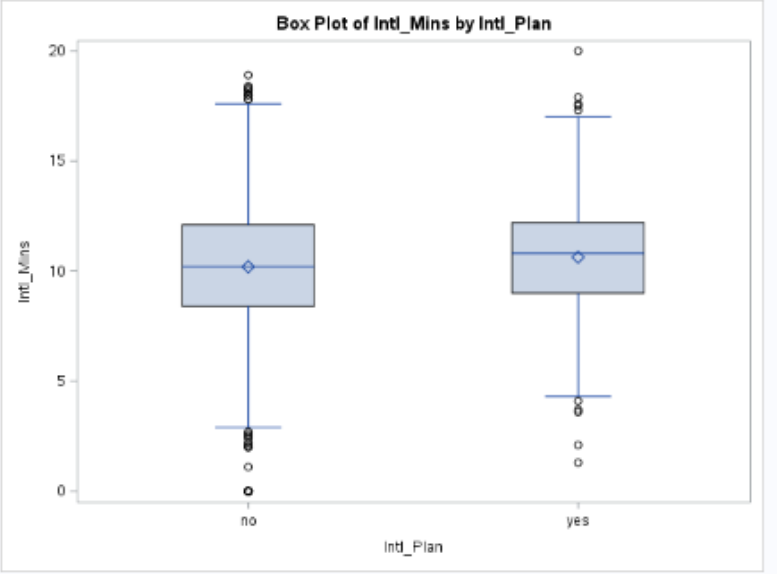
Intl_Charge & Intl_Plan

- The mean Intl_Charge is slightly higher for clients with an international plan (Intl_Plan = yes) compared to those without (Intl_Plan = no), and this difference is statistically significant.
  - The F-value is 7.00 with a p-value of 0.0082, indicating a statistically significant difference in Intl_Charge between clients with and without an international plan.
  - The R-squared value is 0.002096, which shows that the model explains a very small proportion of the variance in Intl_Charge.
  - The p-value from the ANOVA is less than 0.05, providing strong evidence of a significant difference in Intl_Charge based on the Intl_Plan status.
  - Despite the statistical significance, the small R-squared value suggests that the practical impact of Intl_Plan on Intl_Charge might be limited, as the model does not explain much of the variability in the dependent variable.



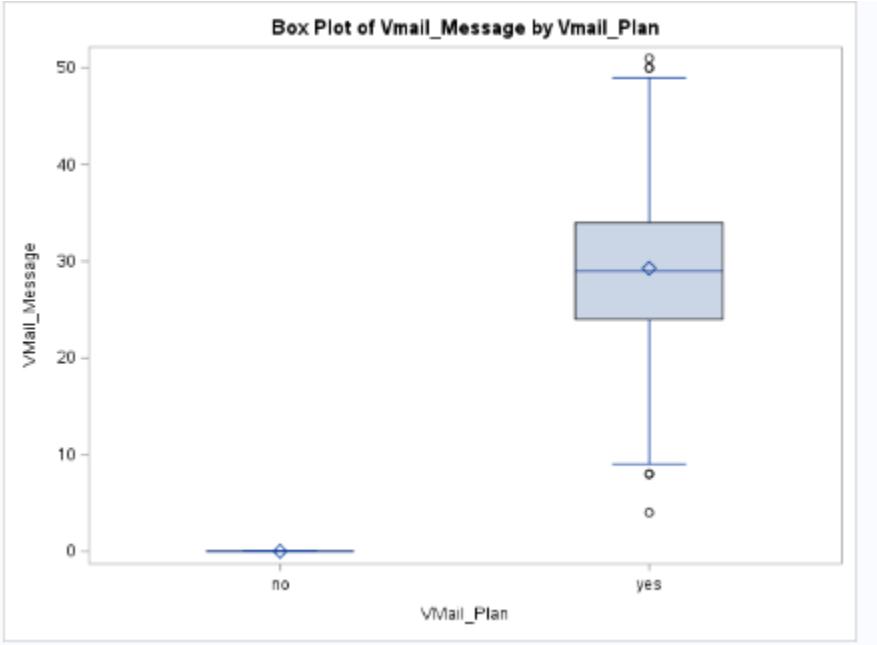
Intl_Mins & Intl_Plan

- The mean Intl_Mins is slightly higher for clients with an international plan (Intl_Plan = yes) compared to those without (Intl_Plan = no), and this difference is statistically significant.
    - The F-value is 7.02 with a p-value of 0.0081, indicating a statistically significant difference in Intl_Mins between clients with and without an international plan.
    - The R-squared value is 0.002104, which shows that the model explains a very small proportion of the variance in Intl_Mins.
    - The p-value from the ANOVA is less than 0.05, providing strong evidence of a significant difference in Intl_Mins based on the Intl_Plan status.
    - Despite the statistical significance, the small R-squared value suggests that the practical impact of Intl_Plan on Intl_Mins might be limited, as the model does not explain much of the variability in the dependent variable.



Vmail_Message & Vmail_Plan

- The mean number of Vmail_Message is significantly different between clients with a voicemail plan (Vmail_Plan = yes) and those without (Vmail_Plan = no), and this difference is highly statistically significant.
    - The F-value is 36186.7 with a p-value of <0.0001, indicating a highly significant difference in Vmail_Message between clients with and without a voicemail plan.
    - The R-squared value is 0.9157, which shows that the model explains a substantial proportion of the variance in Vmail_Message.
    - The p-value from the ANOVA is less than 0.05, providing very strong evidence of a significant difference in Vmail_Message based on the Vmail_Plan status.
    - The large R-squared value suggests that the presence of a voicemail plan has a considerable impact on the number of voicemail messages received, explaining a significant portion of the variability in Vmail_Message.



Which attributes do you think can be eliminated or included in the analysis? This can be a subjective decision or an objective decision based on statistical method

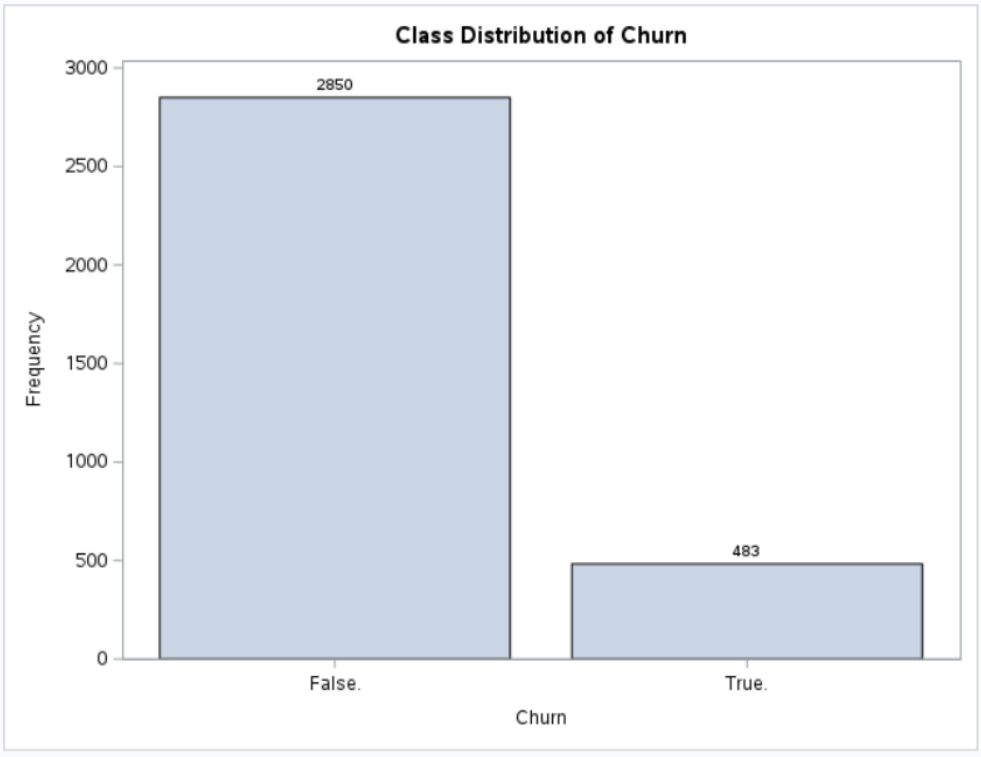
Selection Strategy included in Naïve Bayes and Decision Tree model

Determine whether the dataset has an imbalanced class distribution (same proportion of records of different types or not) and do you need to balance the dataset.

Class Distribution of Churn

The FREQ Procedure

Churn	Frequency
False.	2850
True.	483



- The class distribution of the variable Churn (which represents whether clients have churned or not) is imbalanced:
    - The True class (churned) constitutes 483 records, which is approximately 14.6% of the total dataset.
    - The False class (not churned) constitutes 2850 records, which is approximately 85.4% of the total dataset.
- The True class is significantly smaller compared to the False class, indicating an imbalance. The percentage of the True class is under the 20-30% threshold, which highlights the imbalance in the dataset.

Balancing the Class Attribute:

In the predictive modeling efforts, I have addressed the class imbalance in the dataset, specifically focusing on the Churn attribute, which indicates whether clients have churned or not.

Reason for Balancing:

- The dataset exhibits a notable class imbalance, with the True class (churned) representing only 14.6% of the data, while the False class (not churned) accounts for 85.4%. This imbalance could result in a model that is biased towards predicting the majority class (False), which may impair its ability to accurately identify clients who have actually churned (True).

Impact of Balancing:

- Balancing the dataset is crucial as it allows the model to more effectively learn and predict the minority class (True). This approach enhances key performance metrics such as precision, recall, and F1-score for the churned class. By addressing the imbalance, the model becomes more robust and reliable in making accurate predictions, particularly improving its ability to identify clients who are at risk of churning.