Data Preperation

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

In the Bank Marketing dataset, the attributes can be categorized into two types: Nominal (Categorical and Quantitative (Numerical). Below is a detailed classification:

1. Nominal (Categorical) Attributes

Nominal attributes represent categories without a specific order. They are qualitative and often describe characteristics or labels. In this dataset, the following attributes are nominal:

- Job: Type of job held by the customer.
- Marital: Marital status of the customer.
- Education: Level of education attained by the customer.
- Default: Indicates whether the customer has any credit in default.
- Housing: Indicates whether the customer has a housing loan.
- Loan: Indicates whether the customer has a personal loan.
- Contact: Method used for the last contact during the marketing campaign (e.g., telephone, cellular).
- Month: The month of the year when the customer was last contacted.
- Poutcome: Outcome of the previous marketing campaign for the customer.
- Y: Class attribute showing whether the customer has subscribed to a term deposit (binary outcome: 'yes' or 'no').

2. Quantitative (Numerical) Attributes

Quantitative attributes involve numerical values that can be measured or counted. They are key for statistical analysis and include:

- Age: Age of the customer (in years).
- Balance: Average yearly balance in the customer's account (in Euros).
- Day: Day of the month when the customer was last contacted.
- Duration: Duration of the last contact with the customer (in seconds).
- Campaign: Number of contacts performed during this marketing campaign.
- Pdays: Number of days since the customer was last contacted from a previous campaign. If the customer was not previously contacted, this value is set to 999.
- Previous: Number of contacts made before this campaign and for this customer.

The CONTENTS Procedure

Data Set Name	WORK.BANK_CSV	Observations	0
Member Type	DATA	Variables	17
Engine	V9	Indexes	0
Created	06/10/2024 15:01:39	Observation Length	136
Last Modified	06/10/2024 15:01:39	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	1
First Data Page	1
Max Obs per Page	962
Obs in First Data Page	0
Number of Data Set Repairs	0
Filename	/saswork/SAS_work16550000F231_odaws01-usw2-2.oda.sas.com/SAS_workB90D0000F231_odaws01-usw2-2.oda.sas.com/bank_csv.sas7bdat
Release Created	9.0401M7
Host Created	Linux
Inode Number	1946157821
Access Permission	rw-r--r--
Owner Name	u63872294
File Size	256KB
File Size (bytes)	262144

Alphabetic List of Variables and Attributes

#	Variable	Type	Len
1	Age	Num	8
6	Balance	Num	8
13	Campaign	Num	8
9	Contact	Char	8
10	Day	Num	8
5	Default	Char	8
12	Duration	Num	8
4	Education	Char	8
7	Housing	Char	8
2	Job	Char	8
8	Loan	Char	8
3	Marital	Char	8
11	Month	Char	8
14	Pdays	Num	8
16	Poutcome	Char	8
15	Previous	Num	8
17	Y	Char	8

Find any missing values

1. Numerical Attributes

- Missing Values: None of the numerical attributes in the dataset have missing values, ensuring that all data points are complete and can be used for analysis without the need for imputation.

The MEANS Procedure

Variable	N Miss
age	0
balance	0
day	0
duration	0
campaign	0
pdays	0
previous	0

2. Categorical Attributes

- Unknown Values: For certain categorical attributes, there are instances where the value is labeled as 'unknown'. This often occurs when the information is not provided or recorded. The attributes with potential unknown values include:
    - Job: Type of job held by the customer.

- Education: Level of education attained by the customer.
- Contact: Method used for the last contact during the marketing campaign.
- Poutcome: Outcome of the previous marketing campaign for the customer.

job	Frequency	Percent	Cumulative Frequency	Cumulative Percent
admin.	478	10.57	478	10.57
blue-collar	946	20.92	1424	31.50
entrepreneur	168	3.72	1592	35.21
housemaid	112	2.48	1704	37.69
management	969	21.43	2673	59.12
retired	230	5.09	2903	64.21
self-employed	183	4.05	3086	68.26
services	417	9.22	3503	77.48
student	84	1.86	3587	79.34
technician	768	16.99	4355	96.33
unemployed	128	2.83	4483	99.16
unknown	38	0.84	4521	100.00

education	Frequency	Percent	Cumulative Frequency	Cumulative Percent
primary	678	15.00	678	15.00
secondary	2306	51.01	2984	66.00
tertiary	1350	29.86	4334	95.86
unknown	187	4.14	4521	100.00

contact	Frequency	Percent	Cumulative Frequency	Cumulative Percent
cellular	2896	64.06	2896	64.06
telephone	301	6.66	3197	70.71
unknown	1324	29.29	4521	100.00

poutcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
failure	490	10.84	490	10.84
other	197	4.36	687	15.20
success	129	2.85	816	18.05
unknown	3705	81.95	4521	100.00

Find max, min, mean and standard deviation of attributes

1. Numerical Attributes

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
age	4521	41.1700951	10.5762110	19.0000000	87.0000000
balance	4521	1422.66	3009.64	-3313.00	71188.00
day	4521	15.9152842	8.2476673	1.0000000	31.0000000
duration	4521	263.9612917	259.8566326	4.0000000	3025.00
campaign	4521	2.7936297	3.1098067	1.0000000	50.0000000
pdays	4521	39.7666445	100.1211244	-1.0000000	871.0000000
previous	4521	0.5425791	1.6935624	0	25.0000000

2. Categorical Attributes

The MEANS Procedure

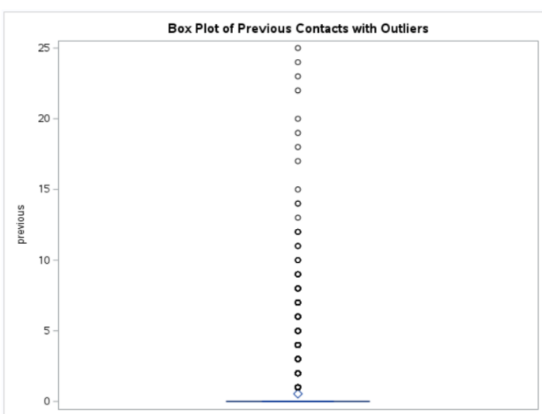
Variable	N	Mean	Std Dev	Minimum	Maximum
Job_num	4521	2.5792966	0.6747373	1.0000000	3.0000000
Marital_num	4521	1.4981199	0.6954711	1.0000000	3.0000000
Education_num	4521	2.2313647	0.7487442	1.0000000	4.0000000
Default_num	4521	0.0168104	0.1285749	0	1.0000000
Housing_num	4521	0.5660252	0.4956763	0	1.0000000
Loan_num	4521	0.1528423	0.3598752	0	1.0000000
Contact_num	3197	1.0941508	0.2920840	1.0000000	2.0000000
Month_num	4521	6.1667773	2.3783802	1.0000000	12.0000000
Poutcome_num	4521	3.6540588	0.7838170	1.0000000	4.0000000
Y_num	4521	0.1152400	0.3193467	0	1.0000000

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)

Below is a summary of the key findings:

1. Numerical Attributes

- Previous Contacts (Previous)
    - Outliers: 816 outliers are identified.
        - Details: Clients with a high number of previous contacts (up to 25 contacts) are considered extreme observations. These could represent persistent attempts to engage certain clients or cases where repeated contacts are necessary.



The UNIVARIATE Procedure
Variable: previous

Moments			
N	4521	Sum Weights	4521
Mean	0.54257908	Sum Observations	2453
Std Deviation	1.69356235	Variance	2.86815344
Skewness	5.87525868	Kurtosis	51.9952115
Uncorrected SS	14295	Corrected SS	12964.0535
Coeff Variation	312.131895	Std Error Mean	0.02518743

Basic Statistical Measures			
Location		Variability	
Mean	0.542579	Std Deviation	1.69356
Median	0.000000	Variance	2.86815
Mode	0.000000	Range	25.00000
		Interquartile Range	0

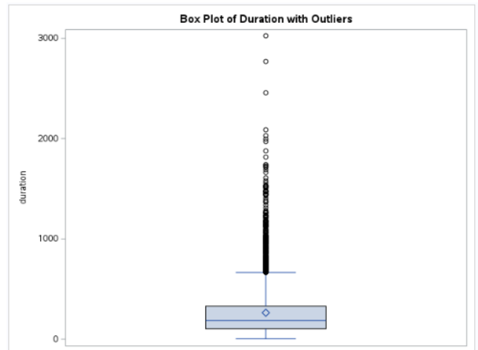
Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	21.54166	Pr > |t|	<.0001
Sign	M	408	Pr >= |M|	<.0001
Signed Rank	S	166668	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	25
99%	8
95%	3
90%	2
75% Q3	0
50% Median	0
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0

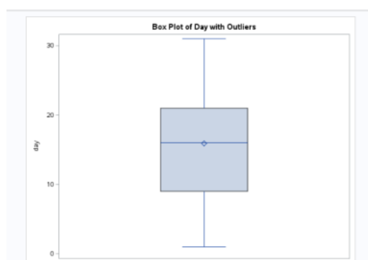
Extreme Observations					
Lowest			Highest		
Value	previous	Obs	Value	previous	Obs
0	0	4519	20	20	98
0	0	4518	22	22	3683
0	0	4517	23	23	4081
0	0	4516	24	24	3427
0	0	4514	25	25	4398

- Duration Since Last Contact (Pdays)
    - Outliers: 816 outliers are identified.

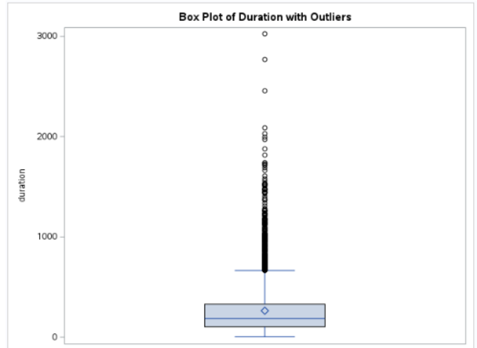
- Details: Extreme observations include values like -1 (likely a placeholder for missing or undefined data) and values up to 871 days. These extreme values could indicate significant delays between contacts or data entry issues.



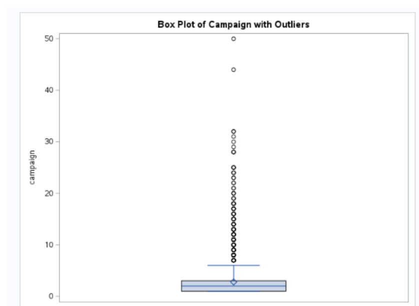
- Day of Month Contact was Made (Day)
  - Outliers: None detected.
    - Details: This attribute is well-distributed without any extreme observations, suggesting that contact days are evenly spread across the month.



- Duration of Last Contact (Duration)
  - Outliers: 330 outliers are identified from a total of 4521 observations.
    - Details: These longer durations might indicate successful engagements where clients were interested or required more time to discuss the term deposit offer. Notably, durations as long as 3025 seconds could highlight successful interactions, potentially leading to higher subscription rates.



- Number of Contacts During This Campaign (Campaign)
    - Outliers: 318 outliers are identified from a total of 4521 observations.
        - Details: These extreme values suggest instances where the number of contacts performed during the campaign is significantly higher than the majority of observations. Such cases could potentially skew the analysis or indicate specific conditions that warrant further investigation.



2. Categorical Attributes

- Outcome of Previous Marketing Campaign (Poutcome)
    - Unknown Values: 81.9509% of the values are labeled as 'unknown'.
        - Details: A high proportion of unknown outcomes can impact the analysis, as it limits the ability to draw meaningful insights from past campaign results.

The FREQ Procedure

poutcome	Frequency	Percent	Cumulative Frequency	Cumulative Percent
failure	490	10.84	490	10.84
other	197	4.36	687	15.20
success	129	2.85	816	18.05
unknown	3705	81.95	4521	100.00

Frequency Distribution of Poutcome

Obs	poutcome	COUNT	PERCENT
1	failure	490	10.8383
2	other	197	4.3574
3	success	129	2.8534
4	unknown	3705	81.9509

- Month of Last Contact (Month)
  - Outliers: One significant outlier.
    - Details: The month of May accounts for 30.9224% of all contacts, making it a potential outlier. This suggests that a disproportionate number of contacts were made in May.

The FREQ Procedure

month	Frequency	Percent	Cumulative Frequency	Cumulative Percent
apr	293	6.48	293	6.48
aug	633	14.00	926	20.48
dec	20	0.44	946	20.92
feb	222	4.91	1168	25.83
jan	148	3.27	1316	29.11
jul	706	15.62	2022	44.72
jun	531	11.75	2553	56.47
mar	49	1.08	2602	57.55
may	1398	30.92	4000	88.48
nov	389	8.60	4389	97.08
oct	80	1.77	4469	98.85
sep	52	1.15	4521	100.00

Frequency Distribution of Month

Obs	month	COUNT	PERCENT
1	apr	293	6.4809
2	aug	633	14.0013
3	dec	20	0.4424
4	feb	222	4.9104
5	jan	148	3.2736
6	jul	706	15.6160
7	jun	531	11.7452
8	mar	49	1.0838
9	may	1398	30.9224
10	nov	389	8.6043
11	oct	80	1.7695
12	sep	52	1.1502

Outliers in Month Based on Frequency Distribution

Obs	month	COUNT	PERCENT	_TYPE_	_FREQ_	mean_freq	std_freq
1	may	1398	30.9224	0	12	376.75	399.143

- Subscribed to Term Deposit (Y)
  - Class Imbalance: 88.4760% of the records are classified as "no".
    - Details: The target variable is highly imbalanced, with the majority of clients not subscribing to the term deposit. This imbalance may affect the performance of predictive models and should be addressed, potentially through resampling techniques.

The FREQ Procedure

y	Frequency	Percent	Cumulative Frequency	Cumulative Percent
no	4000	88.48	4000	88.48
yes	521	11.52	4521	100.00

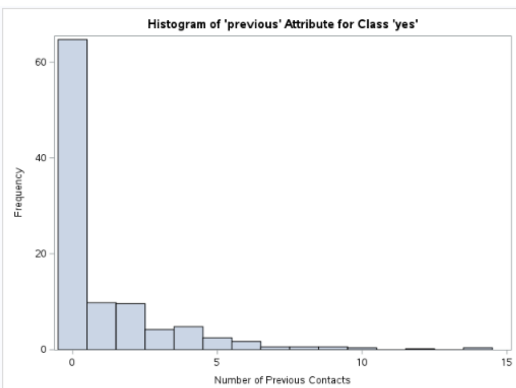
Frequency Distribution of Subscribed (y)

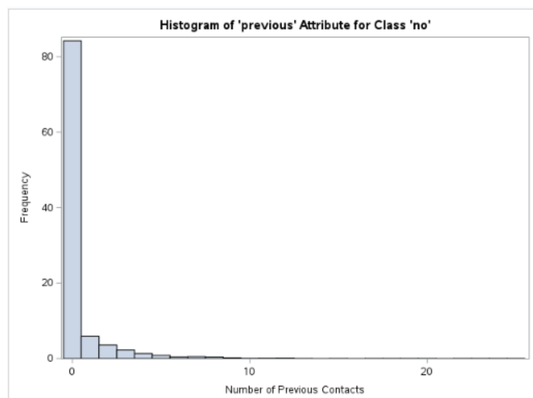
Obs	y	COUNT	PERCENT
1	no	4000	88.4760
2	yes	521	11.5240

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.

Below is an analysis of key numeric attributes, along with histograms and insights on how they might influence the likelihood of subscription:

- Previous Contacts (Previous) – Numerical
  - Distribution: The mean number of previous contacts for clients who subscribed (mean = 1.0902) is higher than for those who did not subscribe (mean = 0.4713).
  - Statistical Significance: Both the pooled and Satterthwaite t-tests show highly significant p-values (<.0001), indicating a statistically significant difference between the two groups.
  - Influence on Subscription:
    - Clients with more previous contacts are more likely to subscribe to a term deposit.
    - Marketing Strategy Insight: Previous contacts are an influential factor. Engaging clients multiple times may increase the likelihood of subscription.





Histogram of 'previous' Attribute for Class 'no'

The TTEST Procedure

Variable: previous

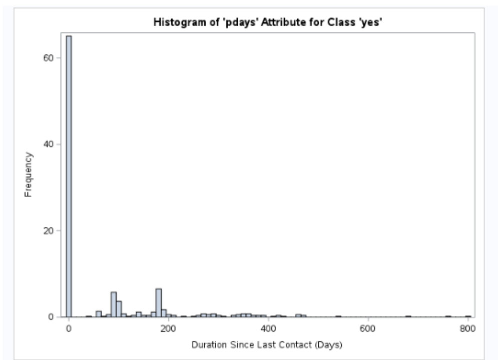
y	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
no		4000	0.4713	1.6274	0.0257	0	25.0000
yes		521	1.0902	2.0554	0.0900	0	14.0000
Diff (1-2)	Pooled		-0.6190	1.6822	0.0784		
Diff (1-2)	Satterthwaite		-0.6190		0.0937		

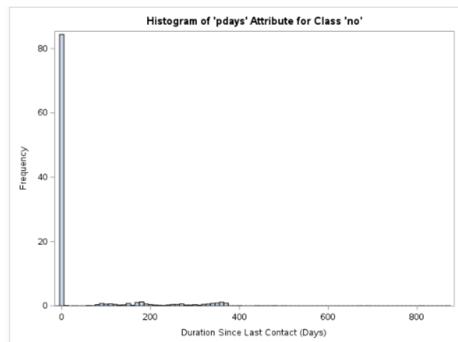
y	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
no		0.4713	0.4208	0.5217	1.6274	1.5925	1.6638
yes		1.0902	0.9133	1.2671	2.0554	1.9377	2.1884
Diff (1-2)	Pooled	-0.6190	-0.7726	-0.4654	1.6822	1.6482	1.7176
Diff (1-2)	Satterthwaite	-0.6190	-0.8029	-0.4350			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	4519	-7.90	<.0001
Satterthwaite	Unequal	607.86	-6.61	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	520	3999	1.60	<.0001

- Duration Since Last Contact (Pdays) – Numerical
    - Distribution: The Pdays variable indicates a significant difference between clients who subscribed and those who did not. Higher Pdays (indicating longer time since last contact) is associated with a higher likelihood of subscription.
    - Statistical Significance: The difference is statistically significant, suggesting that the timing of the previous contact plays a role in the likelihood of subscription.
    - Influence on Subscription:
        - Clients contacted after a longer period are more likely to subscribe.
        - Marketing Strategy Insight: Consider the behavior of Pdays post-subscription. Adjusting the time between contacts might increase receptiveness to the marketing campaign.





Histogram of 'pdays' Attribute for Class 'no'

The TTEST Procedure

Variable: pdays

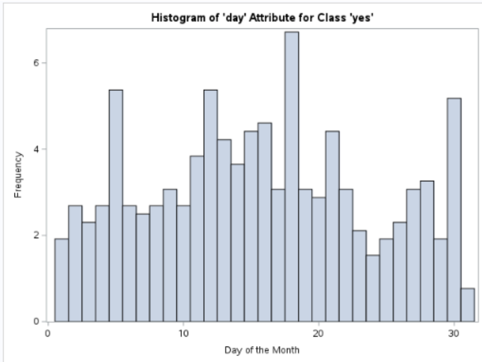
y	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
no		4000	36.0060	96.2977	1.5226	-1.0000	871.0
yes		521	68.6392	122.0	5.3433	-1.0000	804.0
Diff (1-2)	Pooled		-32.6332	99.5883	4.6385		
Diff (1-2)	Satterthwaite		-32.6332		5.5560		

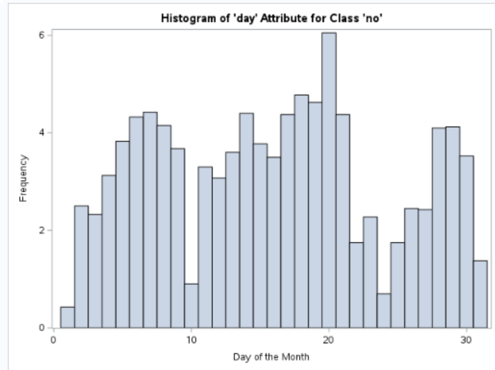
y	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
no		36.0060	33.0209	38.9911	96.2977	94.2329	98.4556
yes		68.6392	58.1420	79.1363	122.0	115.0	129.9
Diff (1-2)	Pooled	-32.6332	-41.7269	-23.5394	99.5883	97.5770	101.7
Diff (1-2)	Satterthwaite	-32.6332	-43.5445	-21.7218			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	4519	-7.04	<.0001
Satterthwaite	Unequal	607.36	-5.87	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	520	3999	1.60	<.0001

- Day of Month Contact was Made (Day) – Numerical
    - Distribution: The `Day` variable shows no significant difference between those who subscribed and those who did not.
    - Statistical Significance: P-values for both the pooled and Satterthwaite methods are approximately 0.45, much higher than the typical significance level of 0.05.
    - Influence on Subscription:
        - The day of the month when contact was made does not significantly influence the likelihood of subscription.





The TTEST Procedure

Variable: day

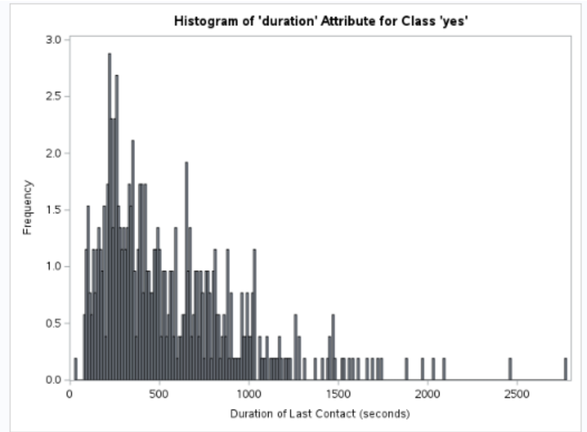
y	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
no		4000	15.9488	8.2497	0.1304	1.0000	31.0000
yes		521	15.6583	8.2351	0.3608	1.0000	31.0000
Diff (1-2)	Pooled		0.2904	8.2481	0.3842		
Diff (1-2)	Satterthwaite		0.2904		0.3836		

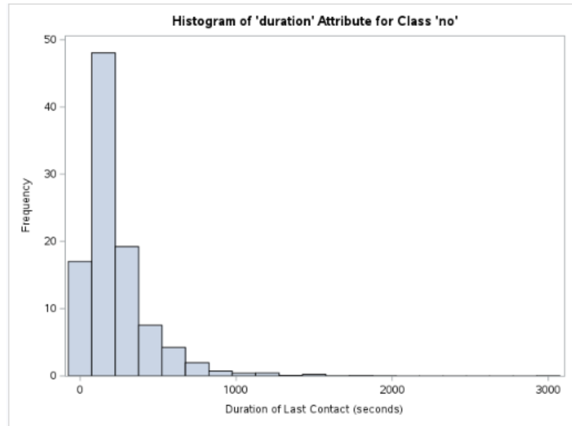
y	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
no		15.9488	15.6930	16.2045	8.2497	8.0728	8.4346
yes		15.6583	14.9496	16.3671	8.2351	7.7636	8.7681
Diff (1-2)	Pooled	0.2904	-0.4628	1.0436	8.2481	8.0815	8.4217
Diff (1-2)	Satterthwaite	0.2904	-0.4629	1.0437			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	4519	0.76	0.4497
Satterthwaite	Unequal	663.35	0.76	0.4493

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	3999	520	1.00	0.9707

- Duration of Last Contact (Duration) – Numerical
  - Distribution: The mean call duration is significantly longer for clients who subscribed (mean = 552.7 seconds) compared to those who did not subscribe (mean = 226.3 seconds).
  - Statistical Significance: P-values for both the pooled and Satterthwaite methods are less than 0.0001, indicating a statistically significant difference between the two groups.
  - Influence on Subscription:
    - Longer call durations are associated with a higher likelihood of subscription.
    - Marketing Strategy Insight: Emphasize strategies that promote longer and more engaging interactions with potential customers.





The TTEST Procedure

Variable: duration

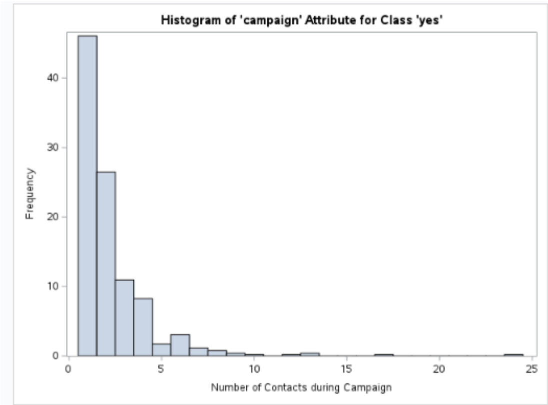
y	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
no		4000	226.3	210.3	3.3254	4.0000	3025.0
yes		521	552.7	390.3	17.1005	30.0000	2769.0
Diff (1-2)	Pooled		-326.4	238.1	11.0881		
Diff (1-2)	Satterthwaite		-326.4		17.4208		

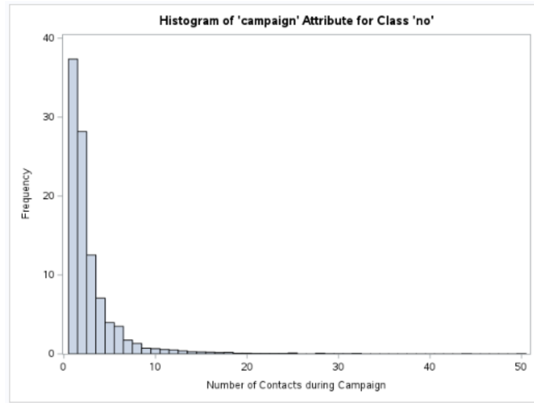
y	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
no		226.3	219.8	232.9	210.3	205.8	215.0
yes		552.7	519.1	586.3	390.3	368.0	415.6
Diff (1-2)	Pooled	-326.4	-348.1	-304.7	238.1	233.3	243.1
Diff (1-2)	Satterthwaite	-326.4	-360.6	-292.2			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	4519	-29.44	<.0001
Satterthwaite	Unequal	559.97	-18.74	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	520	3999	3.44	<.0001

- Number of Contacts During This Campaign (Campaign) – Numerical
    - Distribution: Clients who subscribed had a lower mean number of contacts compared to those who did not subscribe.
    - Statistical Significance: Both the Pooled and Satterthwaite methods show highly significant p-values (<0.0001).
    - Influence on Subscription:
        - Fewer contacts during the campaign are associated with higher subscription rates.
        - Marketing Strategy Insight: Focus on a strategic number of contacts. Excessive contacts may not lead to higher subscription rates, and tailoring the number of contacts could improve outcomes.





The TTEST Procedure

Variable: campaign

y	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
no		4000	2.8623	3.2126	0.0508	1.0000	50.0000
yes		521	2.2668	2.0921	0.0917	1.0000	24.0000
Diff (1-2)	Pooled		0.5955	3.1043	0.1446		
Diff (1-2)	Satterthwaite		0.5955		0.1048		

y	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
no		2.8623	2.7627	2.9618	3.2126	3.1437	3.2846
yes		2.2668	2.0867	2.4469	2.0921	1.9723	2.2275
Diff (1-2)	Pooled	0.5955	0.3120	0.8789	3.1043	3.0416	3.1697
Diff (1-2)	Satterthwaite	0.5955	0.3898	0.8011			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	4519	4.12	<.0001
Satterthwaite	Unequal	877.72	5.68	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	3999	520	2.36	<.0001

Which attributes seem to be correlated? Which attributes seem to be most linked to the class attribute?

Previous Contacts (Previous) vs. Duration Since Last Contact (Pdays)

- Correlation Coefficient: 0.57756 (Moderate positive linear relationship)
- p-value: < 0.0001 (Statistically significant)

Interpretation:

- There is a moderate positive correlation between the number of previous contacts (Previous) and the number of days since the last contact (Pdays).
- As the number of previous contacts increases, the duration since the last contact tends to increase.
- The statistically significant p-value suggests that this relationship is not due to random chance. It implies that clients who were contacted more frequently in the past tend to have longer periods since their last contact.

The CORR Procedure

2 Variables:	previous pdays

Pearson Correlation Coefficients, N = 4521
Prob > |r| under H0: Rho=0

	previous	pdays
previous	1.00000	0.57756 <.0001
pdays	0.57756 <.0001	1.00000

Outcome of Previous Marketing Campaign (Poutcome) vs. Subscribed

- p-value: < 0.0001 (Statistically significant)

Interpretation:

- There is a significant association between the outcome of the previous campaign (Poutcome) and whether the client subscribed to a term deposit (Subscribed).
- Clients with a 'success' outcome in the previous campaign have a much higher likelihood of subscribing (64.34%) compared to those with other outcomes. Conversely, clients with 'unknown' or 'failure' outcomes are less likely to subscribe.
- This suggests that the outcome of the previous campaign is a strong predictor of the likelihood of a client subscribing to a term deposit, with a 'success' outcome being particularly indicative of a positive response.

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of poutcome by y

	y		
poutcome	no	yes	Total
failure	427 9.44 87.14 10.68	63 1.39 12.86 12.09	490 10.84
other	159 3.52 80.71 3.98	38 0.84 19.29 7.29	197 4.36
success	46 1.02 35.66 1.15	83 1.84 64.34 15.93	129 2.85
unknown	3368 74.50 90.90 84.20	337 7.45 9.10 64.68	3705 81.95
Total	4000 88.48	521 11.52	4521 100.00

Statistics for Table of poutcome by y

Statistic	DF	Value	Prob
Chi-Square	3	386.8774	<.0001
Likelihood Ratio Chi-Square	3	235.5376	<.0001
Mantel-Haenszel Chi-Square	1	30.8628	<.0001
Phi Coefficient		0.2925	
Contingency Coefficient		0.2808	
Cramer's V		0.2925	

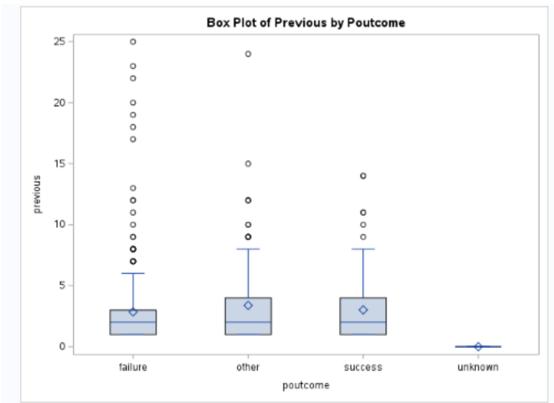
Sample Size = 4521

Number of Previous Contacts (Previous) vs. Outcome of Previous Marketing Campaign (Poutcome)

- p-value: < 0.0001 (Statistically significant)

Interpretation:

- The ANOVA results indicate a strong relationship between the number of previous contacts (Previous) and the outcome of the previous marketing campaign (Poutcome).
- The significant p-value suggests that the differences in mean Previous values across different Poutcome categories are unlikely due to random chance.
- The high R-Square value indicates that Poutcome plays a crucial role in predicting the number of previous contacts made with clients.



ANOVA for Previous by Poutcome

The GLM Procedure

Dependent Variable: previous

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	6083.27984	2027.75995	1331.16	<.0001
Error	4517	6880.77369	1.52331		
Corrected Total	4520	12964.05353			

R-Square	Coeff Var	Root MSE	previous Mean
0.469242	227.4734	1.234223	0.542579

Source	DF	Type I SS	Mean Square	F Value	Pr > F
poutcome	3	6083.279843	2027.759948	1331.16	<.0001

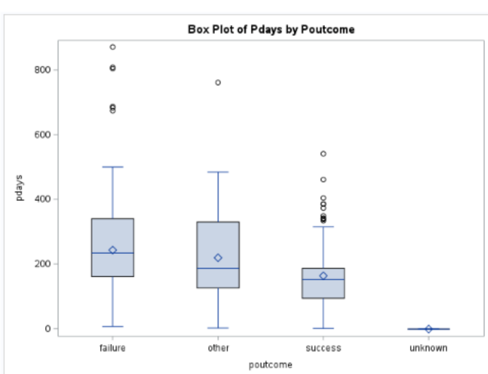
Source	DF	Type III SS	Mean Square	F Value	Pr > F
poutcome	3	6083.279843	2027.759948	1331.16	<.0001

<u>Duration Since Last Contact (Pdays) vs. Outcome of Previous Marketing Campaign (Poutcome)</u>

- p-value: < 0.0001 (Statistically significant)

Interpretation:

- There is a significant association between Poutcome and Pdays, suggesting that knowing the outcome of the previous marketing campaign allows for a high degree of accuracy in predicting the number of days since the last contact.
- Poutcome is a strong predictor of Pdays, indicating that the success or failure of a previous campaign has a substantial impact on the duration since the last contact.



ANOVA for Pdays by Poutcome

The GLM Procedure

Dependent Variable: pdays

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	34767227.46	11589075.82	4965.49	<.0001
Error	4517	10542335.35	2333.92		
Corrected Total	4520	45309562.81			

R-Square	Coeff Var	Root MSE	pdays Mean
0.767326	121.4855	48.31070	39.76664

Source	DF	Type I SS	Mean Square	F Value	Pr > F
poutcome	3	34767227.46	11589075.82	4965.49	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
poutcome	3	34767227.46	11589075.82	4965.49	<.0001

Day of Month Contact was Made (Day) vs. Duration of Last Contact (Duration)

- Correlation Coefficient: -0.02463 (Very weak negative correlation)
- p-value: 0.0978 (Not statistically significant)

Interpretation:

- There is a very weak negative correlation between the day of the month (Day) and the duration of the call (Duration), which is not statistically significant. This suggests that the timing of the contact during the month has no meaningful relationship with the duration of the calls.

The CORR Procedure

3 Variables:	day duration campaign

Pearson Correlation Coefficients, N = 4521
Prob > |r| under H0: Rho=0

	day	duration	campaign
day	1.00000	-0.02463 0.0978	0.16071 <.0001
duration	-0.02463 0.0978	1.00000	-0.06838 <.0001
campaign	0.16071 <.0001	-0.06838 <.0001	1.00000

Day of Month Contact was Made (Day) vs. Number of Contacts (Campaign)

- Correlation Coefficient: 0.16071 (Weak positive correlation)
- p-value: < 0.0001 (Statistically significant)

Interpretation:

- There is a weak positive correlation between the day of the month (Day) and the number of contacts (Campaign), which is statistically significant. This indicates a slight tendency for the day of the month to be associated with the number of contacts made to the customer.

The CORR Procedure

3 Variables:	day duration campaign

Pearson Correlation Coefficients, N = 4521 Prob > |r| under H0: Rho=0			
	day	duration	campaign
day	1.00000	-0.02463 0.0978	0.16071 <.0001
duration	-0.02463 0.0978	1.00000	-0.06838 <.0001
campaign	0.16071 <.0001	-0.06838 <.0001	1.00000

<u>Duration of Last Contact (Duration) vs. Number of Contacts (Campaign)</u>

- Correlation Coefficient: -0.06838 (Weak negative correlation)
- p-value: < 0.0001 (Statistically significant)

Interpretation:

- There is a weak negative correlation between the duration of the call (Duration) and the number of contacts (Campaign), which is statistically significant. This suggests that as the number of contacts increases, the duration of each individual contact tends to slightly decrease.

The CORR Procedure

3 Variables:	day duration campaign

Pearson Correlation Coefficients, N = 4521 Prob > |r| under H0: Rho=0			
	day	duration	campaign
day	1.00000	-0.02463 0.0978	0.16071 <.0001
duration	-0.02463 0.0978	1.00000	-0.06838 <.0001
campaign	0.16071 <.0001	-0.06838 <.0001	1.00000

<u>Month of Contact vs. Outcome of Previous Marketing Campaign (Poutcome)</u>

- p-value: < 0.0001 (Statistically significant)

Interpretation:

- There is a significant association between the month of contact (Month) and the outcome of the previous marketing campaign (Poutcome). This suggests that the outcomes of previous campaigns are not independent of the month in which contacts are made.
- The month of contact influences the likelihood of different outcomes, indicating potential seasonal trends in marketing campaign outcomes.

Contingency Table for Month and Poutcome

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of month by poutcome

month	poutcome: failure	other	success	unknown	Total
apr	72 1.59 24.57 14.69	28 0.62 9.56 14.21	15 0.33 5.12 11.63	178 3.94 60.75 4.80	293 6.48
aug	22 0.49 3.48 4.49	11 0.24 1.74 5.58	16 0.35 2.53 12.40	584 12.92 92.26 15.76	633 14.00
dec	4 0.09 20.00 0.82	3 0.07 15.00 1.52	7 0.15 35.00 5.43	6 0.13 30.00 0.16	20 0.44
feb	55 1.22 24.77 11.22	14 0.31 6.31 7.11	8 0.18 3.60 6.20	145 3.21 65.32 3.91	222 4.91
jan	31 0.69 20.95 6.33	20 0.44 13.51 10.15	9 0.20 6.08 6.98	88 1.95 59.46 2.38	148 3.27
jul	14 0.31 1.98 2.86	3 0.07 0.42 1.52	9 0.20 1.27 6.98	680 15.04 96.32 18.35	706 15.62
jun	11 0.24 2.07 2.24	11 0.24 2.07 5.58	6 0.13 1.13 4.65	503 11.13 94.73 13.58	531 11.75
mar	5 0.11 10.20 1.02	3 0.07 6.12 1.52	7 0.15 14.29 5.43	34 0.75 69.39 0.92	49 1.08
may	177 3.92 12.66 36.12	65 1.44 4.65 32.99	20 0.44 1.43 15.50	1136 25.13 81.26 30.66	1398 30.92
nov	74 1.64 19.02 15.10	28 0.62 7.20 14.21	10 0.22 2.57 7.75	277 6.13 71.21 7.48	389 8.60
oct	18 0.40 22.50 3.67	4 0.09 5.00 2.03	11 0.24 13.75 8.53	47 1.04 58.75 1.27	80 1.77
sep	7 0.15 13.46 1.43	7 0.15 13.46 3.55	11 0.24 21.15 8.53	27 0.60 51.92 0.73	52 1.15
Total	490 10.84	197 4.36	129 2.85	3705 81.95	4521 100.00

Statistics for Table of month by poutcome

Statistic	DF	Value	Prob
Chi-Square	33	693.9689	<.0001
Likelihood Ratio Chi-Square	33	610.0894	<.0001
Mantel-Haenszel Chi-Square	1	4.1195	0.0424
Phi Coefficient		0.3918	
Contingency Coefficient		0.3648	
Cramer's V		0.2262	
WARNING: 21% of the cells have expected counts less than 5. Chi-Square may not be a valid test.			

Sample Size = 4521

<u>Month of Contact vs. Subscribed</u>

- p-value: < 0.0001 (Statistically significant)

Interpretation:

- There is a significant association between the month of contact (Month) and whether the client subscribed to a term deposit (Subscribed). Certain months, such as May and August, have higher subscription rates, suggesting optimal timing for contacting potential clients.

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of month by y

month	y: no	y: yes	Total
apr	237 5.24 80.89 5.93	56 1.24 19.11 10.75	293 6.48
aug	554 12.25 87.52 13.85	79 1.75 12.48 15.16	633 14.00
dec	11 0.24 55.00 0.28	9 0.20 45.00 1.73	20 0.44
feb	184 4.07 82.88 4.60	38 0.84 17.12 7.29	222 4.91
jan	132 2.92 89.19 3.30	16 0.35 10.81 3.07	148 3.27
jul	645 14.27 91.36 16.13	61 1.35 8.64 11.71	706 15.62
jun	476 10.53 89.64 11.90	55 1.22 10.36 10.56	531 11.75
mar	28 0.62 57.14 0.70	21 0.46 42.86 4.03	49 1.08
may	1305 28.87 93.35 32.63	93 2.06 6.65 17.85	1398 30.92
nov	350 7.74 89.97 8.75	39 0.86 10.03 7.49	389 8.60
oct	43 0.95 53.75 1.08	37 0.82 46.25 7.10	80 1.77
sep	35 0.77 67.31 0.88	17 0.38 32.69 3.26	52 1.15
Total	4000 88.48	521 11.52	4521 100.00

Statistics for Table of month by y

Statistic	DF	Value	Prob
Chi-Square	11	250.5001	<.0001
Likelihood Ratio Chi-Square	11	187.4051	<.0001
Mantel-Haenszel Chi-Square	1	7.5732	0.0059
Phi Coefficient		0.2354	
Contingency Coefficient		0.2291	
Cramer's V		0.2354	

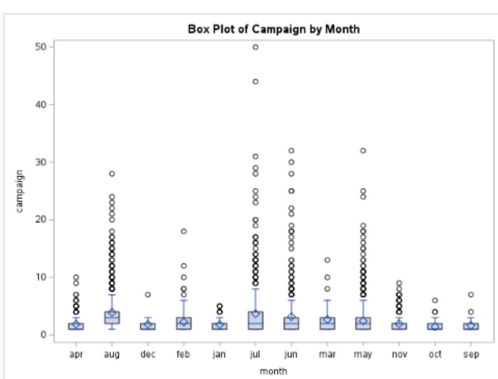
Sample Size = 4521

<u>Month of Contact vs. Number of Contacts (Campaign)</u>

- p-value: < 0.0001 (Statistically significant)

Interpretation:

- The timing of the campaign within the year has a statistically significant effect on the number of contacts made. However, the R-Square value indicates that the month alone explains a small portion of the variability in Campaign, suggesting other factors also play a role.



ANOVA for Campaign by Month

The GLM Procedure

Dependent Variable: campaign

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	2621.08164	238.28015	26.15	<.0001
Error	4509	41091.37489	9.11319		
Corrected Total	4520	43712.45654			

R-Square	Coeff Var	Root MSE	campaign Mean
0.059962	108.0604	3.018806	2.793630

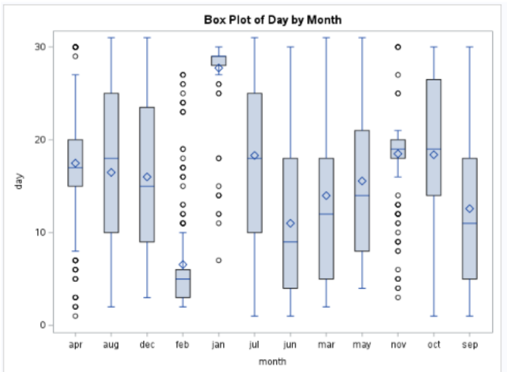
Source	DF	Type I SS	Mean Square	F Value	Pr > F
month	11	2621.081643	238.280149	26.15	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
month	11	2621.081643	238.280149	26.15	<.0001

<u>Month vs. Day of Contact</u>

- p-value: < 0.0001 (Statistically significant)

Interpretation: There is a statistically significant relationship between the month in which the contact was made and the day of the month when the contact occurred. The R-Square value indicates that approximately 20.16% of the variance in the day of contact can be explained by differences between month categories. This suggests that the month has a notable impact on the specific day within the month when contacts are made.



ANOVA for Day by Month

The GLM Procedure

Dependent Variable: day

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	61997.1668	5636.1061	103.53	<.0001
Error	4509	245471.3870	54.4403		
Corrected Total	4520	307468.5539			

R-Square	Coeff Var	Root MSE	day Mean
0.201637	46.36027	7.378368	15.91528

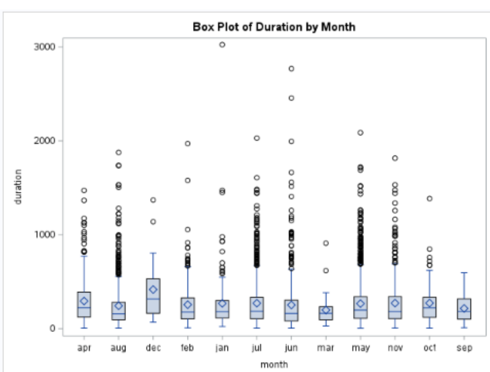
Source	DF	Type I SS	Mean Square	F Value	Pr > F
month	11	61997.16683	5636.10608	103.53	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
month	11	61997.16683	5636.10608	103.53	<.0001

<u>Month vs. Duration of Contact</u>

- p-value: < 0.0001 (Statistically significant)

Interpretation: There is a statistically significant relationship between the month in which the contact was made and the duration of the contact. The R-Square value of 15.08% indicates that the month of contact can explain a moderate portion of the variance in contact duration, suggesting that the month does influence how long a contact lasts.



ANOVA for Duration by Month

The GLM Procedure

Dependent Variable: duration

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	1500647.7	136422.5	2.03	0.0225
Error	4509	303714474.5	67357.4		
Corrected Total	4520	305215122.2			

R-Square	Coeff Var	Root MSE	duration Mean
0.004917	98.32238	259.5330	263.9613

Source	DF	Type I SS	Mean Square	F Value	Pr > F
month	11	1500647.709	136422.519	2.03	0.0225

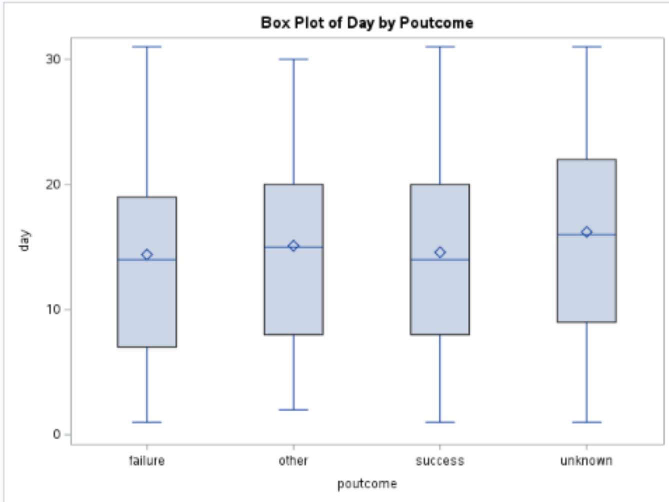
Source	DF	Type III SS	Mean Square	F Value	Pr > F
month	11	1500647.709	136422.519	2.03	0.0225

<u>Outcome of Previous Marketing Campaign (Poutcome) vs. Day of Contact</u>

- p-value: < 0.0001 (Statistically significant)

Interpretation:

- Although Poutcome has a statistically significant effect on the day of the month (Day), the effect size is very small, indicating limited practical significance.



ANOVA for Day by Poutcome

The GLM Procedure

Dependent Variable: day

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1804.1278	601.3759	8.89	<.0001
Error	4517	305664.4261	67.6698		
Corrected Total	4520	307468.5539			

R-Square	Coeff Var	Root MSE	day Mean
0.005868	51.68720	8.226165	15.91528

Source	DF	Type I SS	Mean Square	F Value	Pr > F
poutcome	3	1804.127765	601.375922	8.89	<.0001

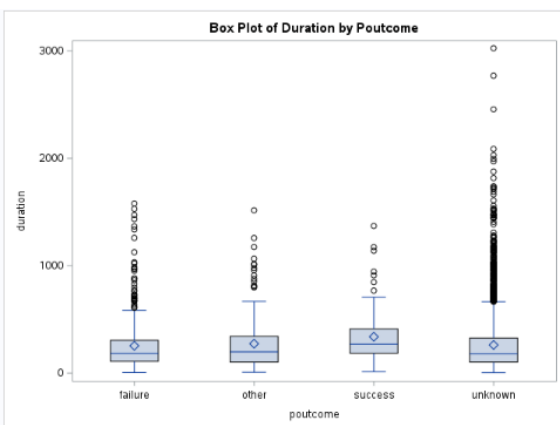
Source	DF	Type III SS	Mean Square	F Value	Pr > F
poutcome	3	1804.127765	601.375922	8.89	<.0001

<u>Outcome of Previous Marketing Campaign (Poutcome) vs. Duration of Last Contact</u>

- p-value: 0.0081 (Statistically significant)

Interpretation:

- The relationship between Poutcome and Duration is statistically significant, but the small R-Square value indicates that Poutcome explains only a very small proportion of the variance in Duration.



ANOVA for Duration by Poutcome

The GLM Procedure

Dependent Variable: duration

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	796274.4	265424.8	3.94	0.0081
Error	4517	304418847.8	67394.0		
Corrected Total	4520	305215122.2			

R-Square	Coeff Var	Root MSE	duration Mean
0.002609	98.34912	259.6036	263.9613

Source	DF	Type I SS	Mean Square	F Value	Pr > F
poutcome	3	796274.3943	265424.7981	3.94	0.0081

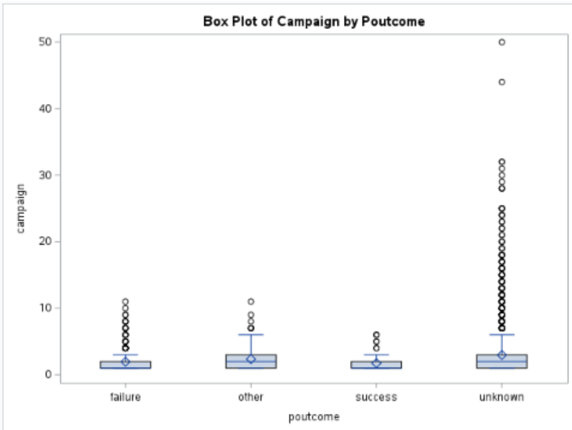
Source	DF	Type III SS	Mean Square	F Value	Pr > F
poutcome	3	796274.3943	265424.7981	3.94	0.0081

<u>Outcome of Previous Marketing Campaign (Poutcome) vs. Number of Contacts (Campaign)</u>

- p-value: < 0.0001 (Statistically significant)

Interpretation:

- There is a statistically significant association between Poutcome and Campaign, but the low R-Square value suggests that the relationship is not strong in terms of explaining the variation in Campaign.



ANOVA for Campaign by Poutcome

The GLM Procedure

Dependent Variable: campaign

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	636.13445	212.04482	22.24	<.0001
Error	4517	43076.32209	9.53649		
Corrected Total	4520	43712.45654			

R-Square	Coeff Var	Root MSE	campaign Mean
0.014553	110.5415	3.088121	2.793630

Source	DF	Type I SS	Mean Square	F Value	Pr > F
poutcome	3	636.1344478	212.0448159	22.24	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
poutcome	3	636.1344478	212.0448159	22.24	<.0001

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 Decision Tree and Naïve Bayes report

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.

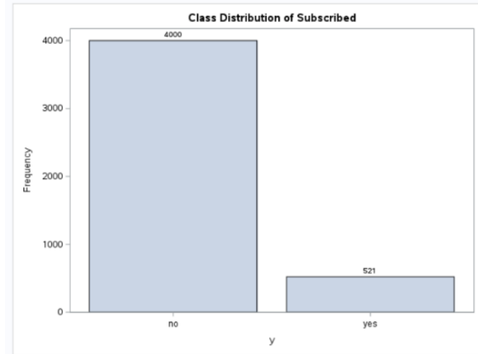
The analysis of the class distribution for the target variable "y" (representing whether clients subscribed to a product or not) reveals a significant imbalance:

- Yes (Subscribed): 521 out of 4521 records (11.5%)
- No (Not Subscribed): 4000 out of 4521 records (88.5%)

Class Distribution of Subscribed

The FREQ Procedure

y	Frequency
no	4000
yes	521



Interpretation:

- The "Yes" class, which represents clients who subscribed, constitutes only 11.5% of the total dataset, while the "No" class, representing clients who did not subscribe, makes up a substantial 88.5%. This significant disparity indicates that the dataset is imbalanced, with the "Yes" class being underrepresented.

In the predictive modeling section of this assignment, we have balanced the class attribute "y" (which represents whether clients subscribed or not) accordingly.

Reason for Balancing:

- Given the significant class imbalance in the dataset—where the "Yes" class (subscribed) constitutes only 11.5% of the data, while the "No" class (not subscribed) makes up 88.5%—balancing the dataset was necessary. This imbalance could lead to a model that is biased towards the majority class, resulting in poor performance when predicting the minority class.

Impact of Balancing:

- Balancing the dataset helps the model to better learn and predict the minority class ("Yes"), improving key performance metrics such as precision, recall, and F1-score for the subscribed class. This approach ensures that the model is more reliable and effective in making accurate predictions across all classes, particularly in identifying clients who are likely to subscribe.