Advanced Business Analytics

Group X6 (A5)

Ky Anh Le 140044
Giang Nguyen 140043
Mengzhen Shang 140028
Liyuan Cao 140029

Advanced Business Analytics
Professor Adam Korczyński

A. Introduction

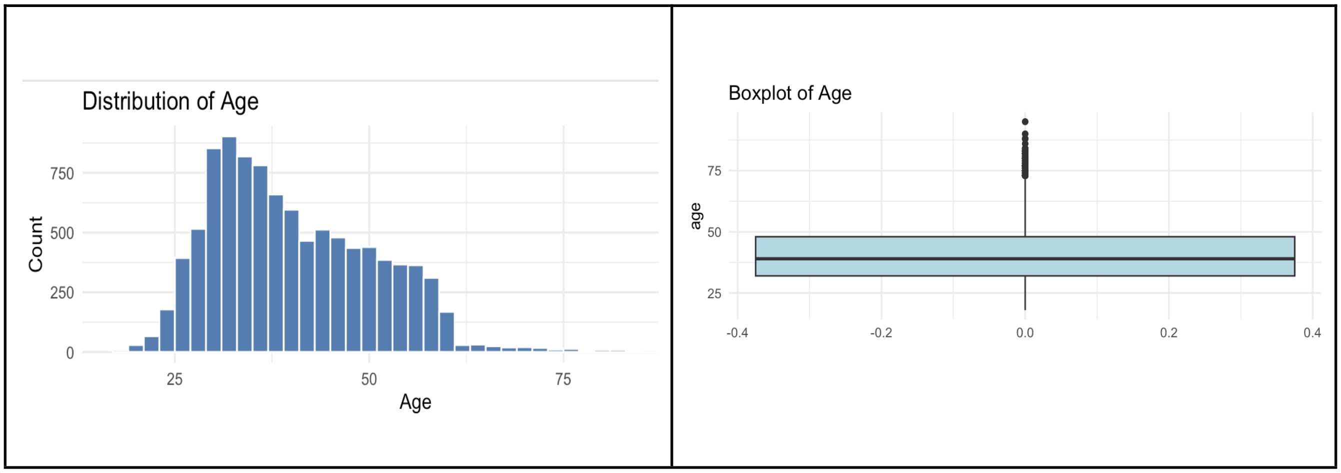
The purpose of this study is to evaluate whether contacting customers through Communication Channel A leads to a higher likelihood of purchasing a new financial product compared to Communication Channel B. Understanding the relative effectiveness of these channels is important for optimizing customer outreach and improving marketing efficiency.
The main research question is: “Does Communication Channel A differ from Channel B in its effectiveness at converting customers?” This leads to the statistical hypotheses:

- Null hypothesis (H_0): Channel A is equally effective as Channel B
- Alternative hypothesis (H_1): Channel A differs in effectiveness

B. Dataset overview

The analysis is based on a dataset of 9,947 observations featuring six key variables.

1. Continuous variable

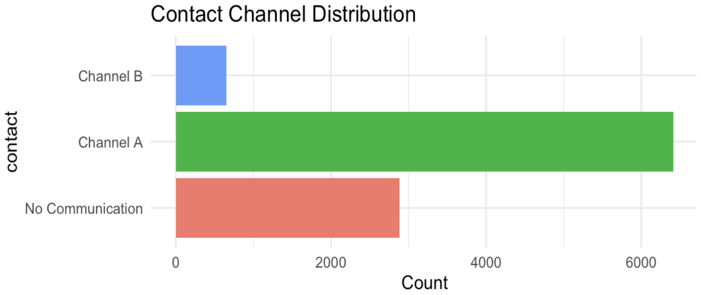
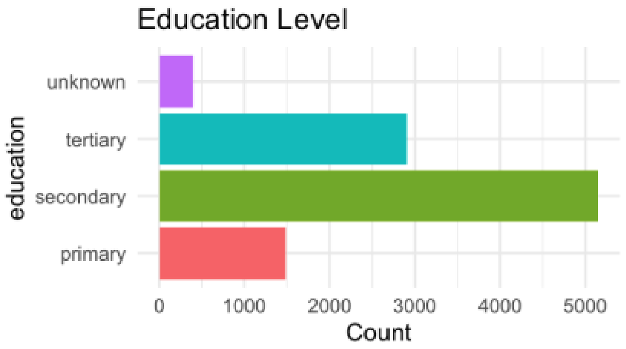
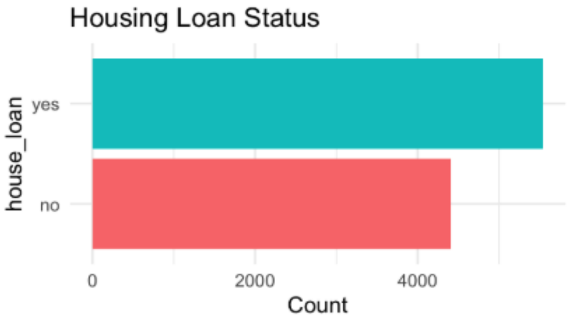
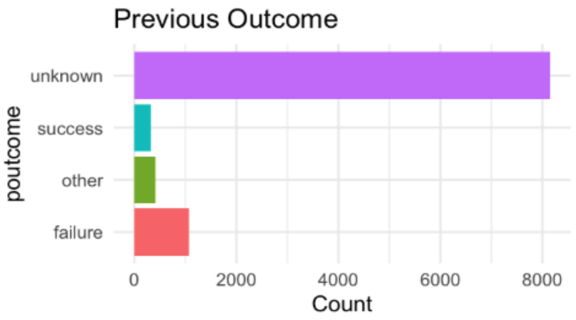


Graph 1&2: Histogram and Boxplot of variable “AGE”

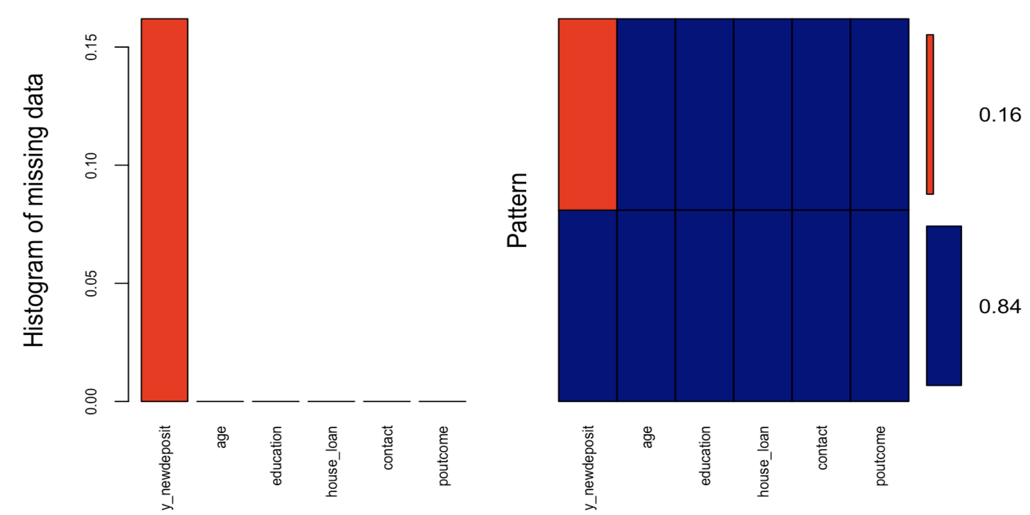
The distribution of customer age is right-skewed, indicating that the client base is predominantly younger, with the highest concentration of customers between 30 and 35 years old. The median age is approximately 39. Although the interquartile range is tight, the boxplot clearly shows the presence of numerous outliers extending to the upper age limit (up to 95 years), representing a small but notable group of older customers.

2. Categorical variables

Table 1: Distribution plot of Contact Channel, Education Level, Housing Loan Status and Previous Outcome Variables

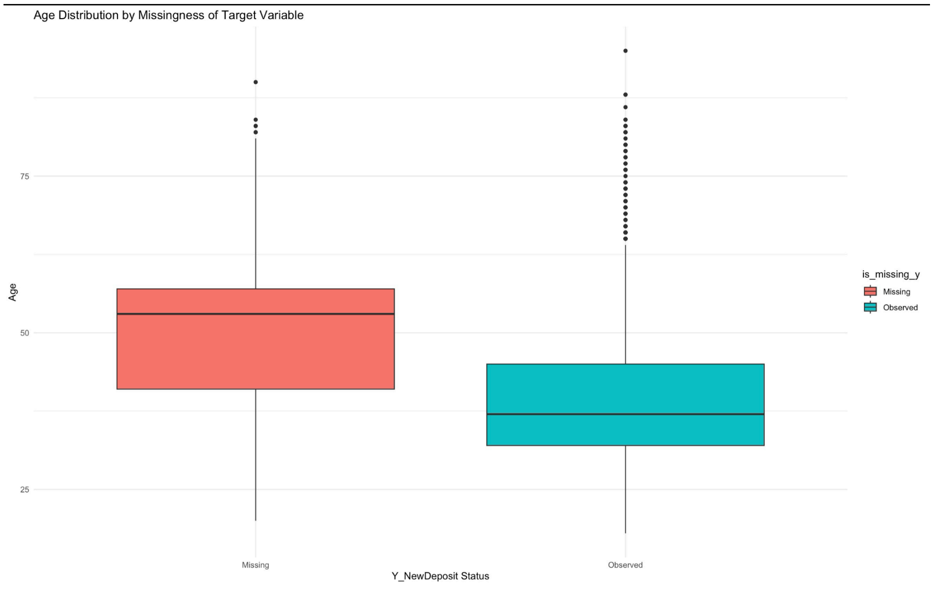
Distribution plot of Variables	Interpretation
	The distribution of the contact channel is highly unbalanced, with channel A being the dominant communication strategy by a significant margin. No communication is the second most frequent status, while channel B represents the smallest fraction of the outreach.
	The Education Level is dominated by customers with secondary education, followed by tertiary education.
	The Housing Loan Status is almost evenly split between customers who have a home loan (yes) and those who do not (no), with a slightly higher count for those with a loan.
	For the Previous Outcome (poutcome), the vast majority of customers have an unknown prior marketing reaction, while failure is the most common known outcome.

C. Missing mechanism analysis



Graph 3: Histogram of missing data in the dataset

In this study, there is only one variable contains missing data, which is Y_NEWDEPOSIT . The proportion of missing data accounts for 16%. In order to access the missing data mechanism, we divide the dataset into 2 groups of observed and missing values and compare the difference between the groups using t-test and chi-square test.

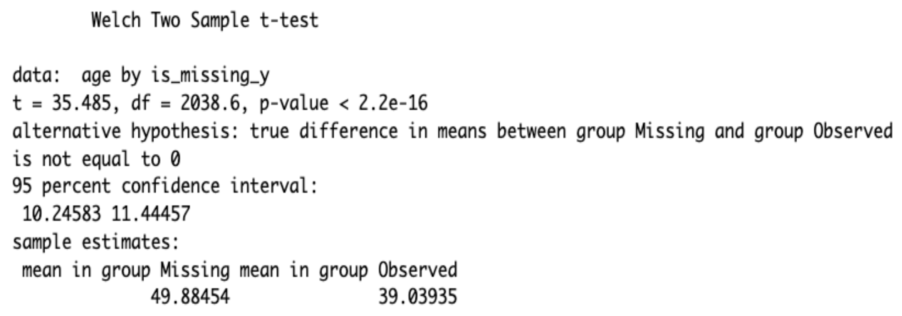


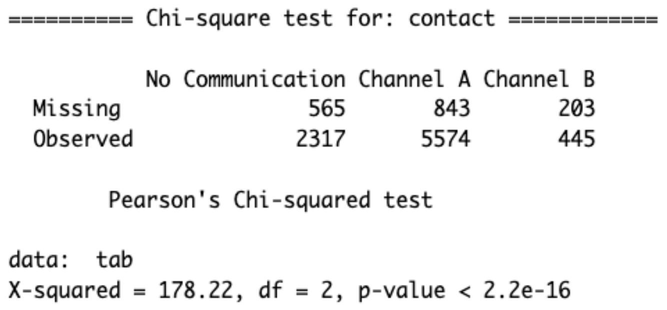
Graph 4: Age Distribution Plot between the 2 groups

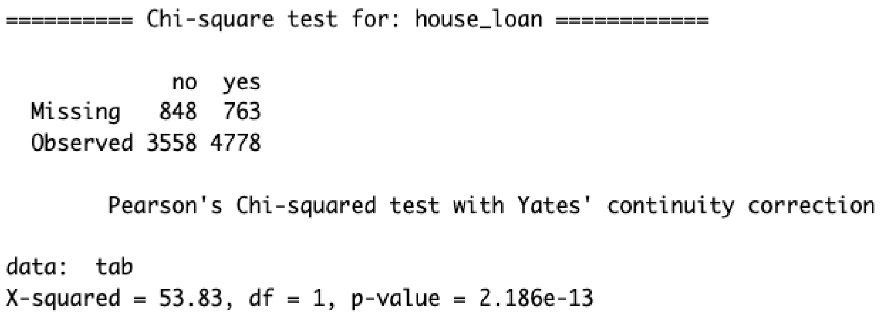
This visual difference in graph 8 suggests that missingness in the target variable is related to age, implying that the missing data mechanism is unlikely to be Missing Completely At Random

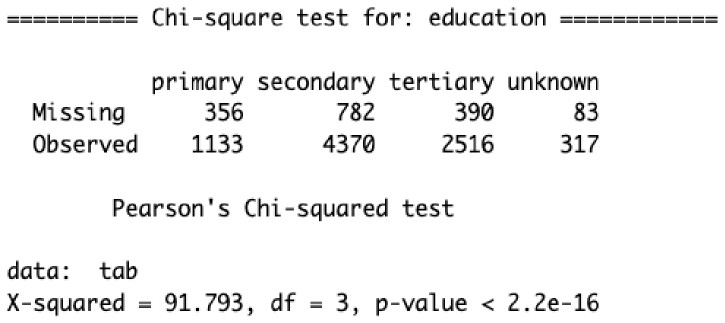
(MCAR). Instead, older customers appear more likely to have missing Y_NEWDEPOSIT values, which supports the assumption of a Missing At Random (MAR) mechanism and justifies the use of Multiple Imputation.

Table 2: Welch Two Sample t-test / Chi-square test results and Interpretation

Welch Two Sample t-test / Chi-square test results	Interpretation
 Welch Two Sample t-test data: age by is_missing_y t = 35.485, df = 2038.6, p-value < 2.2e-16 alternative hypothesis: true difference in means between group Missing and group Observed is not equal to 0 95 percent confidence interval: 10.24583 11.44457 sample estimates: mean in group Missing mean in group Observed 49.88454 39.03935	There is a statistically significant difference in the mean age between the group with missing data and the group with observed data (t = 35.49, p < 0.001). We can see that customers with missing target values are, on average, older (Mean≈49.9 years) compared to those with observed values (Mean≈39.0 years). This suggests that missingness is **not** random with respect to age.


========== Chi-square test for: contact ============
 	
            No Communication Channel A Channel B
  Missing   565   843   203
  Observed   2317   5574   445
 	
        Pearson's Chi-squared test
 	
data: tab
X-squared = 178.22, df = 2, p-value < 2.2e-16
	The chi-square test shows a highly significant association between missingness and the contact category (χ² = 178.22, p < 2.2e−16), indicating that missing values in CONTACT are not missing completely at random.


========== Chi-square test for: house_loan ============
 	
            no yes
  Missing   848 763
  Observed 3558 4778
 	
        Pearson's Chi-squared test with Yates' continuity correction
 	
data: tab
X-squared = 53.83, df = 1, p-value = 2.186e-13
	The chi-square test shows a significant association between missingness and housing loan status (χ² = 53.83, p = 2.19e−13), indicating that missing values in house_loan are not missing completely at random.



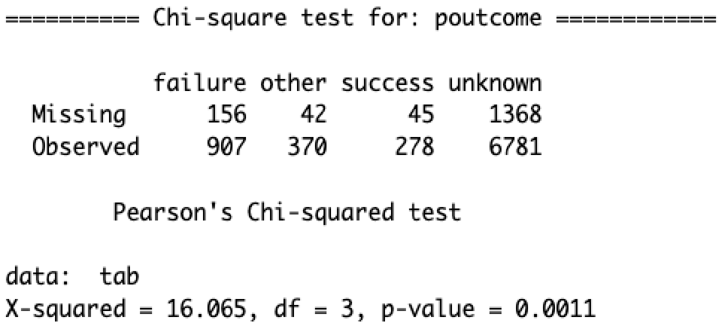
```
========== Chi-square test for: education ============

          primary secondary tertiary unknown
  Missing     356       782      390      83
  Observed   1133      4370     2516     317

       Pearson's Chi-squared test

data:  tab
X-squared = 91.793, df = 3, p-value < 2.2e-16
```

The chi-square test indicates a strong association between missingness and education level (χ² = 91.79, p < 2.2e−16), meaning that missing values in education are not missing completely at random.



```
========== Chi-square test for: poutcome ============

          failure other success unknown
  Missing     156    42      45    1368
  Observed    907   370     278    6781

       Pearson's Chi-squared test

data:  tab
X-squared = 16.065, df = 3, p-value = 0.0011
```

The chi-square test shows a significant association between missingness and previous campaign outcome (χ² = 16.07, p = 0.0011), indicating that missing values in poutcome are not missing completely at random.

In addition, to rigorously evaluate the simultaneous effect of all variables on the probability of missingness, we fitted a logistic regression model with the target as the probability of missingness.

Table 3: Predictors of Missingness (Logistic Regression Results)

Variable	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.621263	0.217940	-25.793	<2e-16
age	0.099346	0.003182	31.221	<2e-16
contactChannel A	-0.576512	0.070523	-8.175	2.96e-16
contactChannel B	-0.002247	0.113734	-0.020	0.9842
educationsecondary	0.033044	0.081211	0.407	0.6841
educationtertiary	0.005670	0.092189	0.062	0.9510
educationunknown	-0.149386	0.152067	-0.982	0.3259
house_loanyes	-0.048506	0.064311	-0.754	0.4507
poutcomeother	-0.443663	0.204733	-2.167	0.0302
poutcomesuccess	-0.460500	0.211814	-2.174	0.0297
poutcomeunknown	-0.033386	0.105563	-0.316	0.7518

We can see that some variables are statistically significant to the probability of missing data in the target variable such as Age, Contact Channel A, previous campaign outcome.

The analysis provides strong evidence to reject the Missing Completely at Random (MCAR) assumption as the probability of missing data is significantly dependent on observed variables, specifically Age and Communication Channel. According to the above t-test on continuous variable and chi-square test on categorical variables, we also see that there is difference in the two groups observed and missing value. Since the missingness can be explained by observed data, we conclude that the data follows a Missing at Random (MAR) mechanism.

Therefore, to obtain unbiased estimates of the marketing campaign's effect, we cannot simply discard incomplete records. We can make adjustments to the estimates using data imputation. The imputation model must explicitly include Age and Contact (along with other covariates) as predictors. By conditioning on these variables during imputation, we satisfy the MAR assumption and correct for the selection bias identified in the missingness analysis.

D. Model building under MCAR and MAR assumptions

1. MCAR assumption

The Logit model was initially evaluated under the assumption of Missing Completely At Random using Complete-Case Analysis as a base to compare.

Table 4: Model Result under Complete-Case Analysis

Variable	Estimate	Std. Error	z value	p-value	Significance
(Intercept)	-2.272352	0.269877	-8.420	< 2e-16	***
age	-0.008825	0.004083	-2.162	0.03065	*
contactChannel A	0.930243	0.119617	7.777	7.44e-15	***
contactChannel B	1.033828	0.182317	5.670	1.42e-08	***
house_loanyes	-0.650713	0.076526	-8.503	< 2e-16	***
poutcomeother	0.069776	0.176917	0.394	0.69329	ns
poutcomesuccess	2.208930	0.160980	13.722	< 2e-16	***
poutcomeunknown	-0.368564	0.110044	-3.349	0.00081	***
educationsecondary	0.232742	0.130399	1.785	0.07429	. (marginal)
educationtertiary	0.424294	0.135193	3.138	0.00170	**
educationunknown	0.118018	0.232193	0.508	0.61126	ns

In the complete-case logistic regression model, several predictors show statistically significant associations with the probability of purchasing a new product.

Contact Channel A has a strong and significant positive effect (Estimate = 0.930, p < 0.001), indicating that customers contacted through Channel A have substantially higher odds of purchasing compared to those with no contact (the reference group). Similarly, Contact Channel

B is also significantly positive (Estimate = 1.034, p < 0.001), suggesting an even stronger increase in purchase likelihood relative to the reference category. However, the model does not directly compare Channel A versus Channel B.

Age has a small but statistically significant negative effect (Estimate = -0.009, p = 0.031), implying that older customers are slightly less likely to purchase.

Customers with a housing loan exhibit significantly lower purchase probability (Estimate = -0.650, p < 0.001).

Regarding previous campaign outcomes, poutcome = success has a very strong and highly significant positive effect (Estimate = 2.209, p < 0.001), indicating a large increase in purchase odds.
In contrast, poutcome = other is not significant (p = 0.694), while poutcome = unknown shows a moderate but significant negative effect (Estimate = –0.369, p < 0.001).

For education, tertiary education is significantly associated with higher purchase likelihood (Estimate = 0.423, p = 0.002), while secondary education is marginal (p ≈ 0.074), and unknown education is not significant (p = 0.612).

Overall, the complete-case model identifies strong and statistically significant predictors, especially contact channel, campaign success, housing loan status, and education level. These effects are notable both in magnitude and statistical precision.

2. MAR assumption

We performed Multiple Imputation using the MICE (Multivariate Imputation by Chained Equations) algorithm with the following specification:

- Imputation Method: Logistic Regression for the binary target variable.
- Number of Imputations: m = 100 datasets were generated.
- Predictors: The imputation model included all available covariates to preserve the multivariate structure:

$$Y_{newdeposit} \sim Age + Contact + Education + House\_Loan + Poutcome$$

To assess the difference in effectiveness between the two contact channels, Channel B is set as the reference category, such that the coefficient associated with Channel A directly measures the difference in the log-odds of purchase between the two channels. The logistic regression model was fitted on each of the 100 imputed datasets, and the results were pooled using Rubin's Rules.

Table 5: Model Result after Multiple Imputation

Variable	Estimate	Std. Error	Statistic (z)	p-value	Significance
(Intercept)	-1.230989	0.290646	-4.235	2.48e-05	***
age	-0.008897	0.004112	-2.182	2.95e-02	*
No contact	-1.030618	0.186269	-5.533	4.07e-08	***
ContactChannel A	-0.101143	0.152430	-0.663	5.07e-01	ns
house_loan = yes	-0.650462	0.077220	-8.423	5.41e-17	***
education = secondary	0.232670	0.130287	1.786	7.43e-02	.
education = tertiary	0.422864	0.133439	3.170	1.55e-03	**
education = unknown	0.106608	0.237400	0.450	6.53e-01	ns
poutcome = other	0.070781	0.176328	0.401	6.88e-01	ns
poutcome = success	2.200036	0.162193	13.564	1.25e-40	***
poutcome = unknown	-0.371711	0.109363	-3.399	6.84e-04	***

The logistic regression model estimated on the multiply imputed datasets reveals several statistically significant predictors of purchase probability. Regarding the main variable of interest, communication channel, the coefficient for Channel A is negative and statistically insignificant (Estimate = -0.101, p = 0.449). This indicates that Channel A does not differ significantly from Channel B, the reference category, in terms of effectiveness.

In contrast, the coefficient for No Communication is strongly negative and highly significant (Estimate = -1.031, p < 0.001), showing that not contacting customers substantially reduces the likelihood of purchase relative to Channel B. Channel A vs. Channel B have OR ≈ 0.90, indicating that the odds of purchasing after Channel A contact are about 10% lower, though this effect is not statistically significant. In practice, this means Channel A performs equivalently to Channel B. No Communication vs. Channel B have OR ≈ 0.36, showing that not contacting customers reduces purchase odds by approximately 64%, highlighting the importance of proactive outreach.

Among demographic and financial variables, age has a small but statistically significant negative effect (Estimate = –0.009, p = 0.029), suggesting that older customers are slightly less likely to buy.

Customers with a housing loan exhibit significantly lower purchase probability (Estimate = –0.650, p < 0.001), consistent with results from the complete-case analysis. Housing loan variable (yes) have OR ≈ 0.52, meaning these customers have about half the odds of purchasing compared to those without a housing loan.

Education effects show mixed patterns. Tertiary education has a significant positive association with purchase probability (Estimate = 0.423, p = 0.002), while secondary education is only marginally significant (p ≈ 0.072), and unknown education remains non-significant. Education tertiary with OR ≈ 1.53, indicating a 53% increase in purchase odds relative to the reference education level.

Campaign history is a strong predictor: customers who previously experienced a successful outcome are much more likely to purchase again (Estimate = 2.200, p < 0.001).
The “other” outcome category is not significant, whereas unknown outcomes have a significant negative effect (Estimate = –0.372, p < 0.001). Previous success with OR ≈ 9.03, meaning these customers are over nine times more likely to purchase than those with a failed previous campaign, by far the strongest predictor in the model.

Overall, the multiply imputed model indicates that Channel A performs similarly to Channel B, and the strongest predictors of purchase likelihood are previous campaign success, housing loan status, age, and education level.

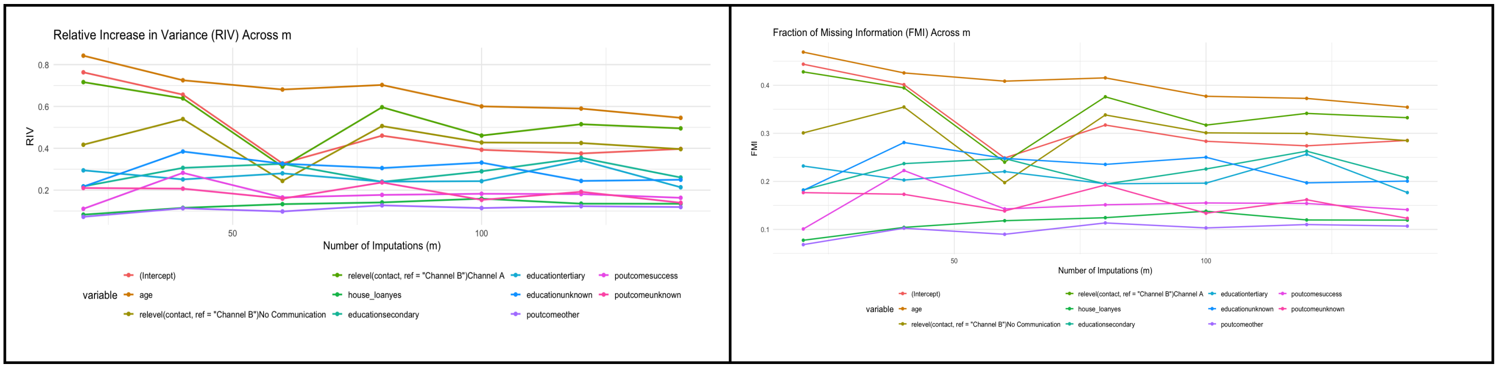
3. Models comparison

The complete-case model, which removes all observations with missing data, produces stronger effects and more statistically significant coefficients. This is expected, as dropping cases reduces sample size and can introduce bias when missingness is related to key variables.

In contrast, the multiple imputation (MI) model uses all available information and incorporates uncertainty from missing data. As a result, MI estimates are more conservative with larger standard errors. Importantly, while the complete-case model suggested a strong positive effect for Channel A, the MI results show no significant difference between Channels A and B, indicating that the earlier result was likely influenced by biased case deletion. Key predictors such as age, housing loan status, and previous success remain significant in both models, though their effects are slightly attenuated under MI. Overall, MI provides more reliable and less biased inference, leading to more cautious and statistically defensible conclusions about communication channel effectiveness.

E. Imputation diagnostic

To evaluate the quality and stability of the multiple imputation procedure, we examined two standard diagnostic measures proposed by Rubin (1987): the Relative Increase in Variance (RIV) and the Fraction of Missing Information (FMI). These metrics quantify how much additional uncertainty is introduced into the parameter estimates due to missing data.
In our analysis, we compared RIV and FMI across different numbers of imputations (m=20, 40, 60, 80, 100, 120, 140).



Graph 5 & 6: Diagnostic plots for Multiple Imputation (RIV and FMI across m)

Based on the diagnostic plots of RIV and FMI across different numbers of imputations, we observe that both measures stabilize when m ≥ 100. This indicates that the uncertainty introduced by missing data is adequately captured from this point, and that additional imputations yield diminishing returns. Therefore, using 100 imputations provides a reliable balance between computational efficiency and statistical precision in this analysis.

F. Conclusion

This study examined whether Communication Channel A differs from Channel B in converting customers to purchase a new product. After handling missing data using multiple imputation and estimating a logistic regression model with Channel B as the reference, the effect of Channel A was found to be small and statistically non-significant (Estimate = -0.101, p-value = 0.508). This shows that Channel A does not perform differently from Channel B in terms of purchase probability.

Several control variables, however, displayed strong and significant effects. A previous successful campaign outcome greatly increases purchase likelihood, while having a housing loan or an unknown prior outcome significantly decreases it. Age and tertiary education also exhibit modest significant effects.

In summary, once missing-data uncertainty is properly accounted for, the key finding is that Channels A and B are equally effective, and choosing one over the other does not meaningfully change customers’ likelihood of purchasing the product.