Loss to Follow-up Project Report

2025-01-10

Contents

Introduction

Project Overview

This report outlines the workflow, analyses, and findings of a biostatistics consulting project focused on longitudinal survey data collected in 2022 and 2024. The project is a collaborative consulting effort, involving data cleaning, imputation, statistical modeling, and exploratory analysis to uncover meaningful insights into survey follow-up patterns and participant characteristics. It aims to provide a structured and reproducible approach to handling longitudinal survey data.

Objectives

The primary objectives of this project are:

1. Analyze survey retention and loss-to-follow-up trends between 2022 and 2024.
2. Investigate predictors of follow-up and non-response using logistic regression models.
3. Explore differences in demographic, socioeconomic, and behavioral characteristics between groups categorized as “lost” or “followed.”
4. Deliver actionable insights and reproducible workflows to inform future survey designs and retention strategies.

Data Handling

Variable Renaming and Selection

In the first step of the data handling process, we renamed several key variables to make them more intuitive and reflective of the actual content.

From the 2022 dataset:

- `sec1_q1` → `dob` (Date of Birth)
- `sec1_q4` → `gender` (Gender)
- `sec1_q5` → `highest_education` (Highest Level of Education)
- `sec1_q6` → `employment_status` (Employment Status)
- `sec1_q7` → `marital_status` (Marital Status)
- `sec1_q8` → `household_income` (Household Income)
- `sec1_q9` → `residence_area` (Residence Area)
- `sec1_q10` → `survey_location` (Survey Location)
- `sec11_start` → `survey_duration` (Survey Duration)
- `sec11_q156` → `religious` (Religious Belief)
- `sec11_q157` → `religion` (Religion)
- `sec11_q157other` → `specified_other_religion` (Specified Other Religion)
- `sec11_q158` → `science_contradict` (Belief Science Contradicts Religion)
- `sec11_q159` → `science_or_religion` (Preference Between Science and Religion)

Additionally, unnecessary variables such as `consent` and `availability` were removed from the dataset to simplify the data structure and focus on relevant variables.

From the 2024 dataset:

- `birthdate` → `dob` (Date of Birth)
- `response_1` → `response_status` (Response Status)
- `response_2` → `response_by` (Response Collected By)
- `educ_level` → `highest_education` (Highest Level of Education)
- `employ_status` → `employment_status` (Employment Status)
- `number_people` → `people_speak_to_daily` (Number of People Spoken to Daily)
- `hh_income` → `household_income` (Household Income)
- `religion_oth` → `specified_other_religion` (Specified Other Religion)
- `call_status` → `call_status` (Call Status)

We selected additional variables relevant to the 2024 data: - `caseid` (Unique Case ID) - `gender` (Gender) - `marital_status` (Marital Status) - `parent_guardian` (Parent/Guardian Status) - `work_industry` (Industry of Work) - `survey_date` (Date of the Survey)

After renaming and selecting these variables, the datasets are now more structured and ready for the next steps in data cleaning and analysis.

Logical Recoding

During the data cleaning process, several categorical variables were logically recoded to simplify the categories and ensure that they were statistically meaningful. Below is a detailed walkthrough of the specific recoding processes applied to key variables for both datasets.

Religion (`religion`)

For `data_2022_cleaned`:

- We noticed that the original `religion` variable in `data_2022_cleaned` contained several categories with very small counts, which would reduce the statistical power of our analysis. Therefore, we grouped

smaller religious affiliations into broader categories:

- **Original categories**: "CCAP," "Traditional African religion," "Seventh Day Adventists," "Pentecostal/Protestant Church," and many smaller specific denominations.
- **Recoding**:
  - Categories such as "CCAP" and "Traditional African religion" were grouped into a broader "Other" category.
  - Smaller Christian denominations such as "Seventh Day Adventists," "Assembly of God Church," "Jehovah's Witness," and others were grouped into the "Other Christian" category.
  - Responses such as "Prefer not to answer" were kept in a standardized format as "Prefer not to answer [do not read aloud]."

For `data_2024_cleaned`:

- Similar to `data_2022_cleaned`, the `religion` variable in `data_2024_cleaned` was recoded in a consistent manner:
  - **Original categories**: Similar religious affiliations as in the 2022 data.
  - **Recoding**:
    - Categories like "Seventh Day Adventists," "Pentecostal/Protestant Church," and other smaller denominations were grouped into the "Other Christian" category.
    - Categories such as "Akorino" and "Traditional African religion" were grouped into the "Other" category.
    - Any "Prefer not to answer" responses were relabeled for consistency.

Employment Status (`employment_status`)

For `data_2022_cleaned`:

- The `employment_status` variable in `data_2022_cleaned` initially contained a large number of specific employment types. We recoded these categories to create fewer, more meaningful groupings:
  - **Original categories**: Categories such as "Self-employed (includes agribusiness)," "Peasant farmer," "Government employee," and "Prefer not to answer" were initially present.
  - **Recoding**:
    - Categories like "Self-employed (includes agribusiness)" and "Peasant farmer" were combined into a new "Self-employed" category, reducing the number of categories and focusing on the nature of employment.
    - Responses such as "Prefer not to answer" were retained and standardized.

For `data_2024_cleaned`:

- The `employment_status` variable in `data_2024_cleaned` was also recoded similarly:
  - **Recoding**:
    - Similar employment types, such as "Self-employed" and "Peasant farmer," were grouped into the "Self-employed" category.
    - Missing employment data were recoded as "Missing" to ensure completeness in the dataset.

Other Categorical Variables

For `data_2022_cleaned`:

- Other categorical variables such as `marital_status`, `parent_guardian`, and `household_income` were standardized:

  - For any missing values in `marital_status`, we used a "Missing" category to ensure data completeness.
  - The `parent_guardian` variable was recoded to include only relevant responses, with any unclear or missing data labeled as "Missing."

For `data_2024_cleaned`:

- Similar recoding was applied to `marital_status`, `work_industry`, and other key variables in `data_2024_cleaned`:
  - Missing values were treated uniformly across all categorical variables, with "Missing" used as a placeholder where necessary.

Standardization Across Datasets

For both datasets (`data_2022_cleaned` and `data_2024_cleaned`), we ensured that categorical variables were standardized, particularly for `religion`, `employment_status`, and other demographics, so that the datasets could be merged seamlessly. This standardization allows for a more consistent and reliable comparison across both time periods.

Data Merging and Imputation

We merged the 2022 and 2024 datasets using a common unique identifier (`caseid`). During this process, missing values in the 2024 dataset were imputed using data from the 2022 dataset. For instance, if gender or date of birth information was missing in the 2024 data, it was filled using the corresponding values from the 2022 dataset. This ensured that our data was complete and minimized the bias introduced by missing data.

- **Imputation of Missing Values**: If a key variable (such as gender or date of birth) was missing in the 2024 data, we used the values from the 2022 dataset to fill in the gaps. In cases where both datasets were missing a particular value, we coded the value as "Unknown."

Exploratory Analysis

Summary Statistics

We start by providing summary statistics for key variables across both datasets (2022 and 2024) to give an overview of the data. The file is attached as seperate summary pdf.

Missing Data Analysis

We investigated the percentage of missing data for key variables in both datasets. Below are bar charts illustrating the missing data percentages:

Pie Charts

- The 2024 dataset shows significantly more missing data compared to 2022. This may hinder our ability to fully account for the changes between the two time points. The missing data in key variables such as employment status, income, and education is particularly problematic for the analysis.

Missing Data Imputation Using MICE

The missing values in the 2022 dataset were imputed using the `mice` package in R to ensure data completeness and consistency. The imputation process followed the methodology outlined below:

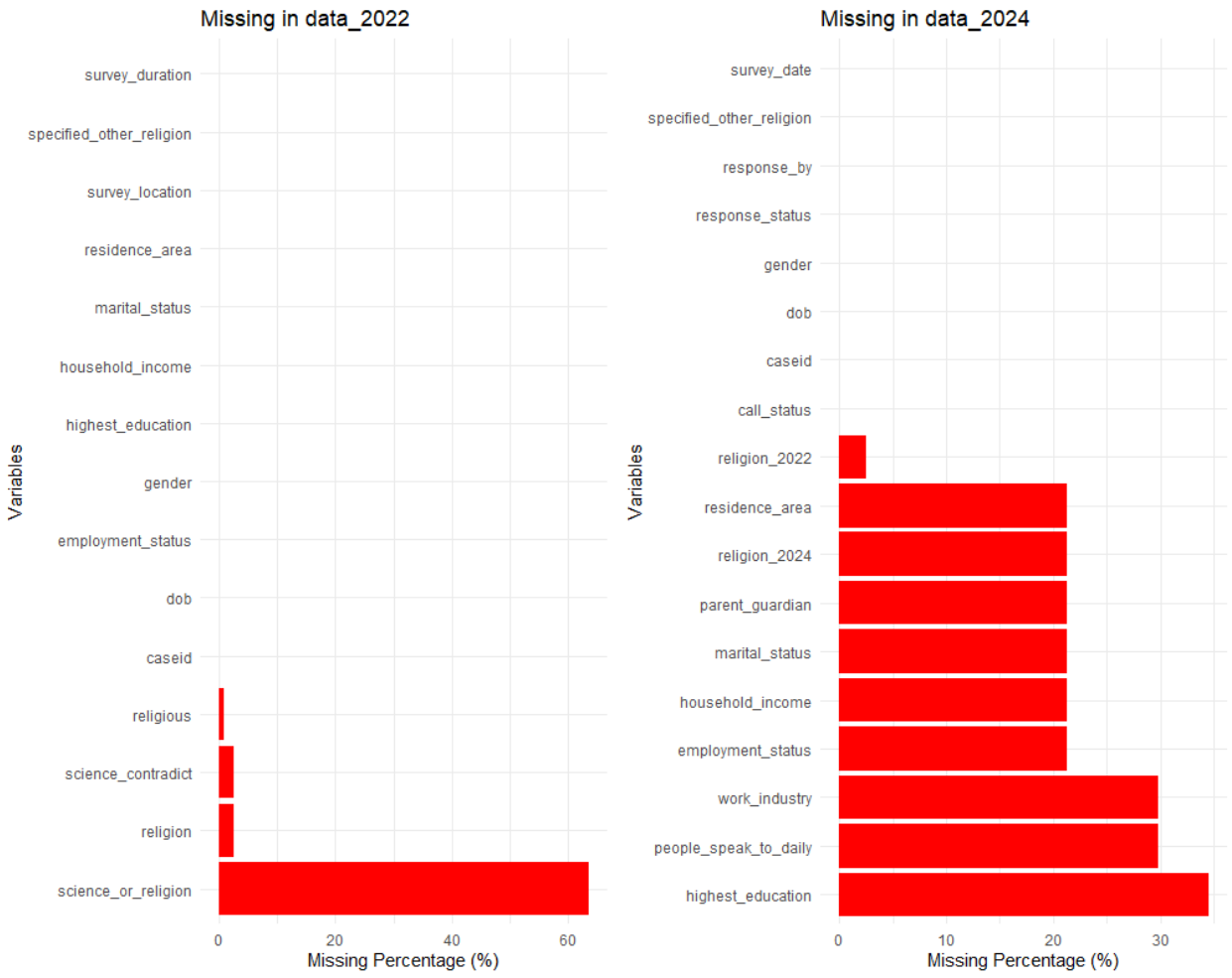


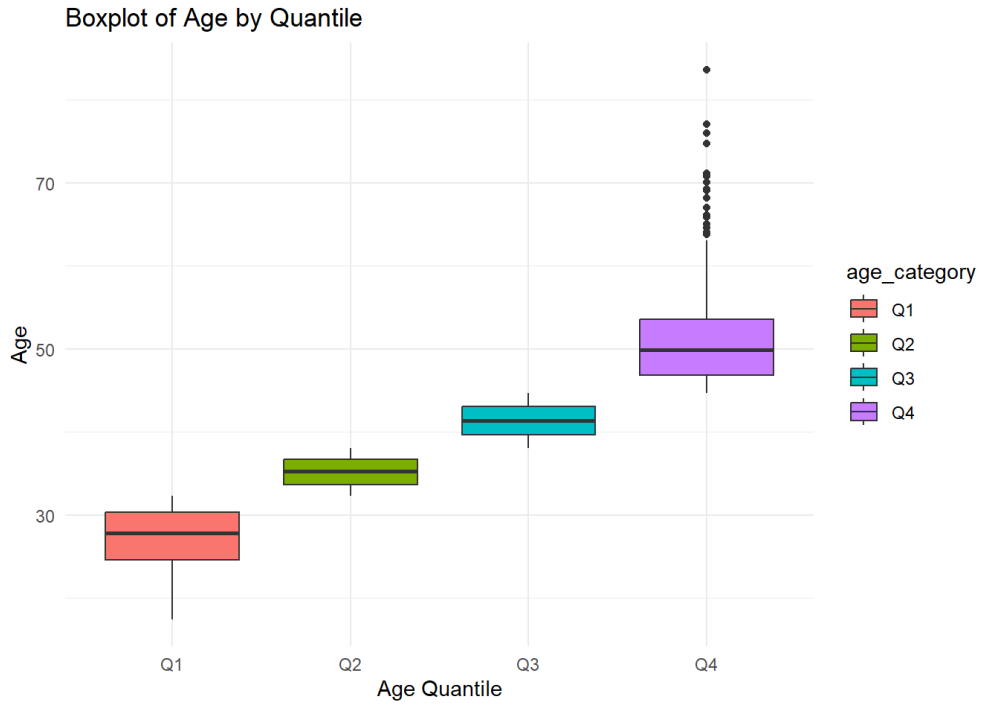
Figure 1: Missing Values

(a) Changes in Lost Group (b) Changes in Followed Group

Figure 2: Comparing Changes in Social-economic Variables across Different Groups

Creating Age Categories

Age was calculated using the Date of Birth (`dob`) and the survey date. The resulting age variable was then categorized into four quartiles to create a categorical variable for further analysis.



(a) Age Quantiles

Modeling

Definitions of Loss and Followed

- **If_follow**: Participants who both answered correctly and completed the survey in 2024.
- **If_follow2**: Participants who either "Completed," "Answered, but not completed/Appointment," or "Refused."

Logistic Regression Models

Two logistic regression models were applied to examine the relationship between the defined follow-up categories (`If_follow` and `If_follow2`) and the selected variables of interest. The models included covariates from the **2022 dataset**:

- Gender
- Employment status
- Highest education level
- Household income
- Marital status
- religion
- age (quantile)

The models are as followed:

If_follow:

logistic_followup_model1 <- glm(if_follow ~ highest_education + employment_status + household_income + residence_area + gender + marital_status + religion + age_category, data = completed_data_2022, family = binomial)

If_follow2:

logistic_followup_model2 <- glm(if_follow2 ~ highest_education + employment_status + household_income + residence_area + gender + marital_status + religion + age_category, data = completed_data_2022, family = binomial)

Model Summaries

1. For those who answered the phone and are correct respondent (Refer to Table 1 & Figure 4)

The analysis revealed that Age Category Q4 (= 1.00, p < 0.001) was a strong predictor, showing that individuals in this group were much more likely to exhibit the outcome. Self-employment (= 0.32, p = 0.04) was also positively associated, while being employed full-time (= 0.50, p = 0.07) and being in Age Category Q3 (= 0.34, p = 0.09) showed some evidence of association but fell short of statistical significance. A notable finding was the borderline negative association for individuals identifying as Muslim (= -0.67, p = 0.05). Other factors, such as household income, marital status, and other religious affiliations, did not show significant effects. Overall, age and employment status emerged as the most relevant predictors in this analysis.

2. For those who just answered the phone (Refer to Table 2 & Figure 5)

The analysis for Model 2 identifies Age Category Q4 (= 1.01, p < 0.001) as the most significant variable, strongly predicting the outcome with a positive association. Employment Status: Employed Full-Time (= 0.58, p = 0.04) and Employment Status: Not Employed but Looking for Work (= 0.62, p = 0.03) were also significant, suggesting their relevance in the model. Additionally, Gender: Male (= 0.32, p = 0.04) showed a significant positive effect. Religion: Muslim (= -0.57, p = 0.10) and some other variables demonstrated trends toward significance but did not meet the threshold. Most other predictors, including household income, marital status, and residence area, did not significantly affect the outcome, with wide confidence intervals overlapping zero. After applying the Benjamini-Hochberg method to control for false positives, only Age Category Q4 remained significant, indicating its robustness as a key predictor.

Conclusion

The analysis highlights Age Category Q4 as the most consistent and robust predictor across both models, demonstrating a strong positive association with the outcome (Model 1: = 1.00, p < 0.001; Model 2: = 1.01, p < 0.001). Employment status also emerged as an important factor, with Self-employment (Model 1: = 0.32, p = 0.04), Employed Full-Time (Model 2: = 0.58, p = 0.04), and Not Employed but Looking for Work (Model 2: = 0.62, p = 0.03) showing significant positive associations. Additionally, Gender: Male (Model 2: = 0.32, p = 0.04) was identified as a significant predictor.

While Religion: Muslim demonstrated a borderline negative association in both models, it did not reach statistical significance. Other variables, such as household income, marital status, and residence area, were not significant predictors, with wide confidence intervals suggesting limited influence on the outcome. After controlling for multiple comparisons using the Benjamini-Hochberg method, Age Category Q4 remained the sole significant predictor, underscoring its central role in this analysis.

In summary, age and employment status are the most relevant predictors, with Age Category Q4 standing out as the strongest and most consistent factor influencing the outcome.

Table 1: Model 1 Analysis Results

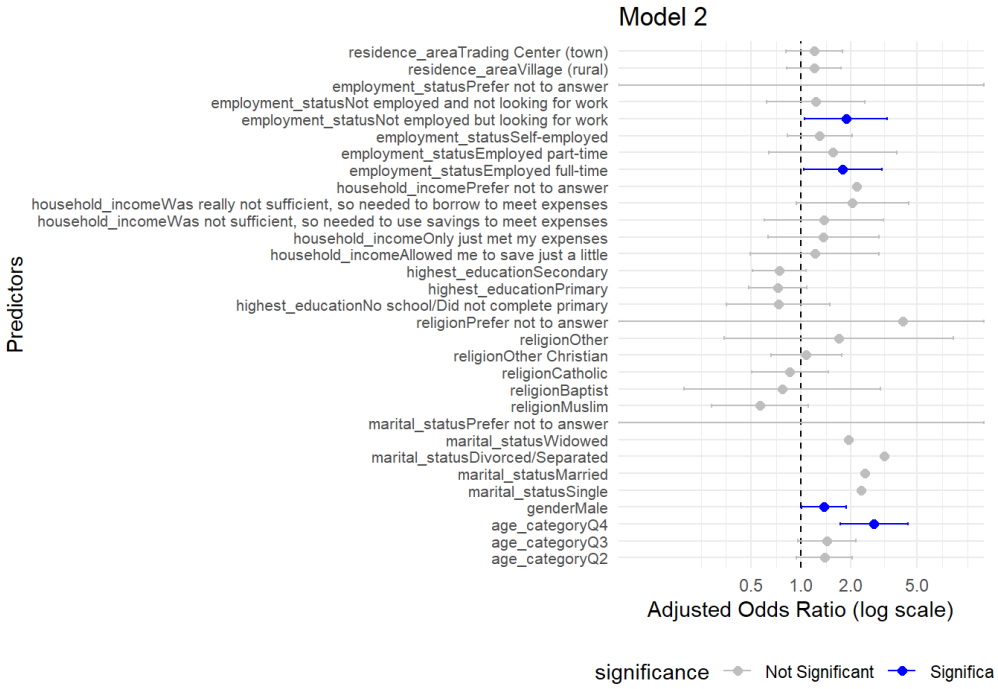
Variable	Coefficient	p-value	CI Lower	CI Upper
Employ_Status				
No school/Did not complete primary	-0.23	0.53	-0.95	0.49
Primary	-0.25	0.22	-0.65	0.15
Secondary	-0.27	0.15	-0.63	0.10
Employed full-time	0.50	0.07	-0.04	1.04
Employed part-time	0.29	0.51	-0.57	1.15
Not employed and not looking for work	0.09	0.80	-0.58	0.76
Not employed but looking for work	0.44	0.12	-0.12	1.00
Prefer not to answer	28.34	0.97	-1435.07	1491.76
Self-employed	0.22	0.33	-0.23	0.68
Household_Income				
Allowed me to save just a little	0.02	0.96	-0.85	0.90
Only just met my expenses	0.26	0.50	-0.50	1.03
Prefer not to answer	0.73	0.52	-1.51	2.97
Was not sufficient, so needed to use savings to meet expenses	0.17	0.69	-0.65	0.99
Was really not sufficient, so needed to borrow to meet expenses	0.60	0.13	-0.18	1.38
Residence_Area				
Trading Center (Town)	0.16	0.41	-0.22	0.54
Village (rural)	0.22	0.24	-0.15	0.59
Gender				
Male	0.32	0.04	0.01	0.62
Marital_Status				
Divorced/Separated	1.02	0.33	-1.02	3.05
Married	0.86	0.38	-1.05	2.76
Prefer not to answer	-14.58	0.98	-1087.70	1058.54
Single	0.78	0.43	-1.14	2.71
Widowed	0.67	0.52	-1.40	2.74
Religion				
Baptist	-0.35	0.61	-1.72	1.01
Catholic	-0.24	0.37	-0.78	0.29
Muslim	-0.67	0.05	-1.34	0.00
Other	0.40	0.62	-1.19	1.99
Other Christian	-0.05	0.83	-0.55	0.44
Prefer not to answer	1.24	1.00	-2239.78	2242.25
Age_Group				
age_categoryQ2 (25% to 50% age range in all respondents)	0.33	0.09	-0.06	0.72
age_categoryQ3 (50% to 75% age range in all respondents)	0.34	0.09	-0.06	0.73
age_categoryQ4 (75% to 100% age range in all respondents)	1.00	0.00	0.54	1.46

Table 2: Model 2 Analysis Results

Variable	Coefficient	p-value	CI Lower	CI Upper
Employ_Status				
No school/Did not complete primary	-0.31	0.39	-1.03	0.40
Primary	-0.32	0.12	-0.73	0.08
Secondary	-0.30	0.11	-0.67	0.07
Employed full-time	0.58	0.04	0.04	1.12
Employed part-time	0.44	0.33	-0.45	1.32
Not employed and not looking for work	0.20	0.56	-0.48	0.88
Not employed but looking for work	0.62	0.03	0.05	1.19
Prefer not to answer	28.54	0.97	-1422.58	1479.66
Self-employed	0.26	0.26	-0.19	0.71
Household_Income				
Allowed me to save just a little	0.19	0.68	-0.70	1.08
Only just met my expenses	0.31	0.43	-0.46	1.08
Prefer not to answer	0.77	0.50	-1.47	3.00
Was not sufficient, so needed to use savings to meet expenses	0.32	0.45	-0.51	1.14
Was really not sufficient, so needed to borrow to meet expenses	0.71	0.07	-0.07	1.50
Residence_Area				
Trading Center (Town)	0.18	0.36	-0.21	0.57
Village (rural)	0.18	0.35	-0.20	0.56
Gender				
Male	0.32	0.04	0.01	0.63
Marital_Status				
Divorced/Separated	1.16	0.27	-0.90	3.21
Married	0.88	0.36	-1.03	2.80
Prefer not to answer	-14.67	0.98	-1080.81	1051.47
Single	0.84	0.40	-1.10	2.77
Widowed	0.66	0.53	-1.42	2.74
Religion				
Baptist	-0.26	0.71	-1.63	1.10
Catholic	-0.16	0.56	-0.69	0.37
Muslim	-0.57	0.10	-1.24	0.10
Other	0.52	0.52	-1.07	2.11
Other Christian	0.07	0.78	-0.42	0.56
Prefer not to answer	1.41	1.00	-2237.21	2240.03
Age_Group				
age_categoryQ2 (25% to 50% age range in all respondents)	0.32	0.11	-0.07	0.71
age_categoryQ3 (50% to 75% age range in all respondents)	0.36	0.08	-0.04	0.76
age_categoryQ4 (75% to 100% age range in all respondents)	1.01	0.00	0.55	1.48

(a) Model 1

Figure 4: Model 1 Odds Ratio Report



(a) Model 2

Figure 5: Model 2 Odds Ratio Report