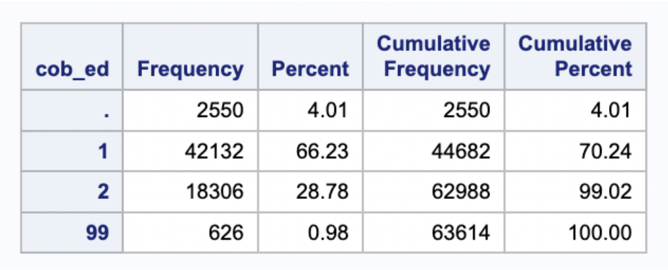
# Data Documentation

## Data Exploration

The original dataset ‘assignmt2a_ed_data.sas7bdat’ contains 15 variables with 63614 observations.

## Data Validation

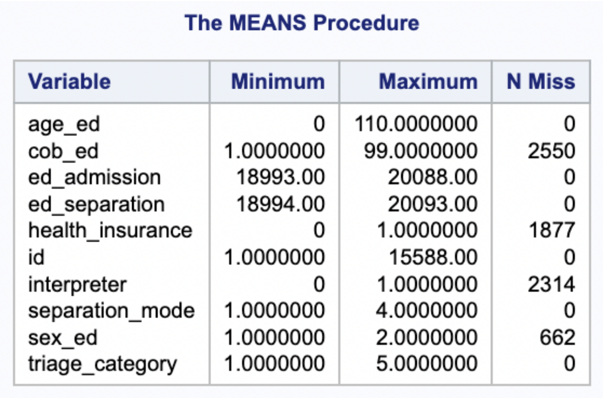
- The ED data dictionary contains 15 variables which corresponds to the 15 variables in the ED dataset.
- The type column of the ED data dictionary requires adjustment as in SAS there are only 2 types of variables ‘numeric’ and ‘character’.
    - The ED data dictionary is adjusted to reflect this, all the ‘number’ and ‘date’ type variables would be converted  to ‘numeric’ variables and ‘character’ variables will remain as ‘character’ variables.
- The maximum value of the ‘cob_ed’ variable is 99 which is not a valid value according to the data dictionary where the ‘cob_ed’ variable only contains 2 levels which are ‘1’ and ‘2’.
    - After investigation, the ‘cob_ed’ variable contains 4 levels which are ‘.’, ‘1’, ‘2’, ‘99’.
        - The ‘cob-ed’ variable is recoded from cob_ed(., 1, 2, 99) to cob_ed(., 1, 2, 3) and the data dictionary is updated where to include ‘3 = unknown’ and ‘. = missing entry’.



| cob_ed | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| . | 2550 | 4.01 | 2550 | 4.01 |
| 1 | 42132 | 66.23 | 44682 | 70.24 |
| 2 | 18306 | 28.78 | 62988 | 99.02 |
| 99 | 626 | 0.98 | 63614 | 100.00 |

Figure 1: Screen Print of cob_ed frequency table

- The minimum value and maximum value of age_ed are 0 and 110, which could be valid since the ED dataset may contain ED admission of infants and very elderly people.
    - As the minimum and maximum values seem plausible, the observations are retained in the dataset.

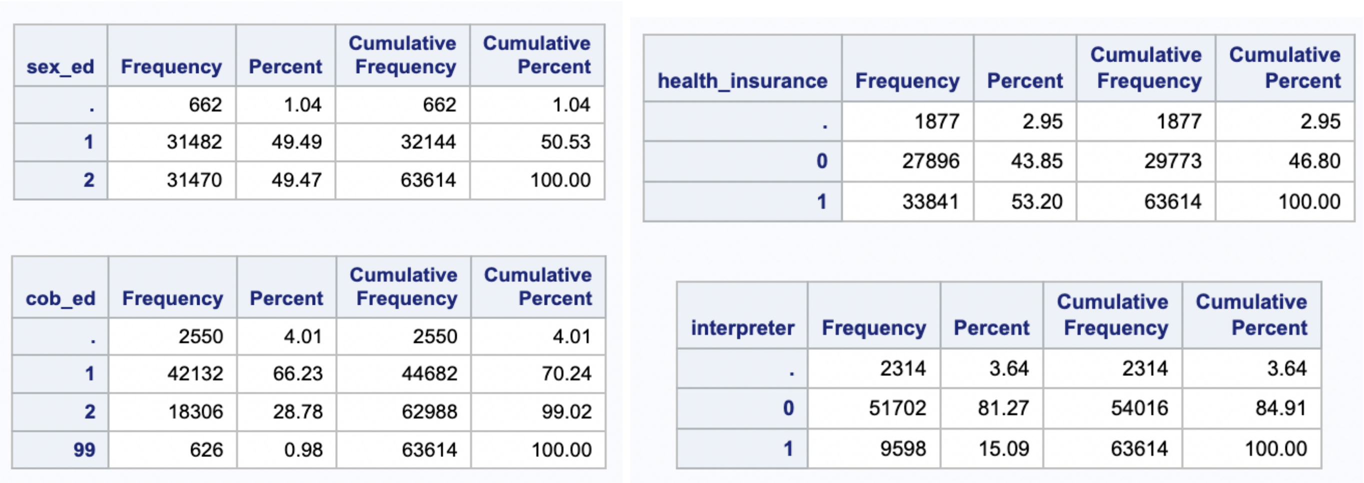


The MEANS Procedure

| Variable | Minimum | Maximum | N Miss |
|---|---|---|---|
| age_ed | 0 | 110.0000000 | 0 |
| cob_ed | 1.0000000 | 99.0000000 | 2550 |
| ed_admission | 18993.00 | 20088.00 | 0 |
| ed_separation | 18994.00 | 20093.00 | 0 |
| health_insurance | 0 | 1.0000000 | 1877 |
| id | 1.0000000 | 15588.00 | 0 |
| interpreter | 0 | 1.0000000 | 2314 |
| separation_mode | 1.0000000 | 4.0000000 | 0 |
| sex_ed | 1.0000000 | 2.0000000 | 662 |
| triage_category | 1.0000000 | 5.0000000 | 0 |

Figure 2: Screen Print Output Table

- There are some missing values in the ED dataset.
    - 2.95% of the observations in ‘health_insurance’ variable is missing.
    - 3.64% of the observations in ‘interpreter’ variable is missing.
    - 1.04% of the observations in ‘sex_ed’ variable is missing.
    - 4.01% of the observations in ‘cob_ed’ variable is missing.



| sex_ed | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| . | 662 | 1.04 | 662 | 1.04 |
| 1 | 31482 | 49.49 | 32144 | 50.53 |
| 2 | 31470 | 49.47 | 63614 | 100.00 |

| health_insurance | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| . | 1877 | 2.95 | 1877 | 2.95 |
| 0 | 27896 | 43.85 | 29773 | 46.80 |
| 1 | 33841 | 53.20 | 63614 | 100.00 |

| cob_ed | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| . | 2550 | 4.01 | 2550 | 4.01 |
| 1 | 42132 | 66.23 | 44682 | 70.24 |
| 2 | 18306 | 28.78 | 62988 | 99.02 |
| 99 | 626 | 0.98 | 63614 | 100.00 |

| interpreter | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| . | 2314 | 3.64 | 2314 | 3.64 |
| 0 | 51702 | 81.27 | 54016 | 84.91 |
| 1 | 9598 | 15.09 | 63614 | 100.00 |

Figure 3: Screen Print of cob_ed, sex_ed, health_insurance and interpreter frequency table

## Data Cleaning

- There are 1047 complete duplicates in the ED dataset.
    - These observations are removed from the final dataset as each patient cannot present to the emergency department twice on the same day.
- Flags were created to inform the removal of partial duplicates, however no partial duplicates were identified in the ED dataset.
- Patterns of missing data was examined through the creation of ad_month and sep_month variables to produce cross-tabular frequency tables across different months against the four variables (sex_ob, health_insuarance, interpreter and cob_ed)
    - No obvious pattern was found. The missing enteries seem to increase as the number of patients increase across the ed_admission and ed_separation variables.
        - Depending on the outcome of interest, these observations may have a relatively minor influence on the interpretation of the result. Therefore, these observations are retained in the dataset.