Analyzing Missing Data in the Palmerpenguins Dataset*

Hechen Zhang

March 5, 2024

Peer reviewed by Dailin Li and Yanyu Wu

Table of contents

*Codes are available at: https://github.com/HechenZ123/Comparative-Analysis-of-Imputed-and-Actual-Value-in-the-Palmer-Penguins-Datase.git

1.0. Introduction

Missingness in data is an essential aspect of data cleaning that is essential and must be dealt with care to tackle the discrepancies in the data. The missing values in the data can lead to statistical results that are prone to biases and significantly influencing the validity and reliability of statistical conclusions. Therefore, the aim of this analysis to to investigate into the missing values in the data set pertaining to penguins from 2007 to 2009 (Horst, Hill, and Gorman 2020). It is essential to handle the missing value in the data set as the preliminary step of data cleaning to avoid skewed research findings based on reduced statistical power (Salgado et al. 2016). The following sections provide an in-depth analysis of the missing values in the data set along with employing simulated methods to understand the various types of missingness in the data.

1.1 Importing Important Packages.

Packages like `palmerpenguins` (Horst, Hill, and Gorman 2020) for accessing the penguins data, `tidyverse` by Wickham et al. (2019) is used for data wrangling, `janitor` package by Firke (2021) is used for data cleaning operations, `knitr` by Xie (2021) for data presentation in data tables.The following code section aims at importing the important packages that are essential for examining the missing values in the data set.

```
library(palmerpenguins)
library(tidyverse)
library(ggplot2)
library(janitor)
library(knitr)
library(lubridate)
library(mice)
```

The following code cell stores the data set in an object called `df` while firstly changing it into a `tibble` structure and using the `clean_names` function from janitors package (Firke 2021). Further, to view if the data is imported in the correct format, the first 6 rows of the data are viewed using the `head` function (2023).

```
df <-
  penguins |>
  as_tibble() |>
  clean_names()

head(df)
```

```
# A tibble: 6 x 8
  species island    bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
  <fct>   <fct>              <dbl>         <dbl>             <int>       <int>
1 Adelie  Torgersen           39.1          18.7               181        3750
2 Adelie  Torgersen           39.5          17.4               186        3800
3 Adelie  Torgersen           40.3          18                 195        3250
4 Adelie  Torgersen           NA            NA                  NA          NA
5 Adelie  Torgersen           36.7          19.3               193        3450
6 Adelie  Torgersen           39.3          20.6               190        3650
# i 2 more variables: sex <fct>, year <int>
```

The data set consists of 344 rows and 8 columns.

```
dim(df)
```

```
[1] 344   8
```

1.2. Data Overview.

In this section, we intend to get an overview of the data to understand the types of the data, each variable belongs to. Based on the output, it can be deduced that there are a few columns like species, islands and sex have *factor* data types instead of *character* type. While, the data in the column years is reported to have *integer* data type. Therefore further steps are important to rectify these data discrepancies.

```
df|>
  glimpse()
```

```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, ~
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186~
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, ~
$ sex               <fct> male, female, female, NA, female, male, female, male~
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

The following code, corrects the data type of species, island and sex variables to *character* data type using the `as.character` function.

```
df$species <- as.character(df$species)
df$island <- as.character(df$island)
df$sex <- as.character(df$sex)
df|>
  str()
```

```
tibble [344 x 8] (S3: tbl_df/tbl/data.frame)
 $ species          : chr [1:344] "Adelie" "Adelie" "Adelie" "Adelie" ...
 $ island           : chr [1:344] "Torgersen" "Torgersen" "Torgersen" "Torgersen" ...
 $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
 $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
 $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
 $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
 $ sex              : chr [1:344] "male" "female" "female" NA ...
 $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

The years column was incorrectly identified as an *integer* type, the `as.Date` function from the lubridate package by Grolemund and Wickham (2011) is employed to change the years to *date* type.

```
df$year <- as.Date(paste(df$year, "-01-01", sep=""));
format(as.Date(paste(df$year, "01", "01", sep="-")), "%Y")
df|>
  str()
```

2.0. Data Sample simulation

Based on the output below, it can be reported that the *bill_length_mm, bill_depth_mm, flipper_length_mm and body_mass_g* columns have 2 missing values each. While there are 11 missing values in the sex column. Therefore, the following section drills deep into the missing values in the *bill_length_mm* column.

```
colSums(is.na(df))|>
  kable()
```

	x
species	0
island	0
bill_length_mm	2
bill_depth_mm	2
flipper_length_mm	2
body_mass_g	2
sex	11
year	0

The following code iterates 10 times to randomly select 2 species and ignore them while calculating the average of the *bill_length_mm*. It is therefore, reported based on the results of Table 2 that on excluding the *Chinstrap and Gentoo* species while computing he mean, the mean, for the *bill_length_mm* is the lowest of 38.7.

```
# Initialize an empty tibble to store sample means
sample_means <- tibble(seed = c(), mean = c(), species_ignored = c())

for (i in 1:10) {
  set.seed(i)
  # Sample 2 species to ignore
  dont_get <- sample(x = unique(df$species), size = 2)

  # Average bill length excluding the sampled species
  sample_means <- sample_means |>
    rbind(tibble(
      seed = i,
      mean =
        df |>
        filter(!species %in% dont_get) |>
        summarise(mean = mean(bill_length_mm, na.rm = TRUE))|>
        pull(),
      species_ignored = toString(dont_get)
    ))
}

# Table
sample_means|>
  kable(col.names = c("Iteration", "Mean Bill Length (mm)", "Ignored Species"), digits =2, fo
```

Table 2: Summary Statistics Table

Iteration	Mean Bill Length (mm)	Ignored Species
1	48.83	Adelie, Gentoo
2	47.50	Adelie, Chinstrap
3	48.83	Adelie, Gentoo
4	47.50	Chinstrap, Adelie
5	48.83	Gentoo, Adelie
6	48.83	Adelie, Gentoo
7	48.83	Gentoo, Adelie
8	38.79	Chinstrap, Gentoo
9	47.50	Chinstrap, Adelie
10	47.50	Chinstrap, Adelie

The table below Table 3 summarizes penguin data, indicating key statistics for various attributes: bill length ranges from 32.10 mm to 59.60 mm, with two missing values; bill depth ranges from 13.10 mm to 21.50 mm, also with two missing values. Flipper length ranges from 172.0 mm to 231.0 mm, with two missing values, and body mass ranges from 2700 g to 6300 g, again with two missing values.

```
df|>
  summary()|>
  kable()
```

Table 3: Summary Statistics Table

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
	Length:344	Length:344	Min. :32.10	Min. :13.10	Min. :172.0	Min. :2700	Length:344	Min. :2007
	Class :character	Class :character	1st Qu.:39.23	1st Qu.:15.60	1st Qu.:190.0	1st Qu.:3550	Class :character	1st Qu.:2007
	Mode :character	Mode :character	Median :44.45	Median :17.30	Median :197.0	Median :4050	Mode :character	Median :2008
	NA	NA	Mean :43.92	Mean :17.15	Mean :200.9	Mean :4202	NA	Mean :2008
	NA	NA	3rd Qu.:48.50	3rd Qu.:18.70	3rd Qu.:213.0	3rd Qu.:4750	NA	3rd Qu.:2009

Table 3: Summary Statistics Table

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
NA	NA	Max. :59.60	Max. :21.50	Max. :231.0	Max. :6300	NA	Max. :2009
NA	NA	NA's :2	NA's :2	NA's :2	NA's :2	NA	NA

2.1. Data Missing Completely At Random (MCAR).

This section of the coding exercise, simulates a situation in which generates data that has missing values completely at random (MCAR). Therefore, the data for the *bill_length_mm* is removed at random. Based on the results Table 4, it can be reported that the data for the MCAR simulation has a slightly smaller missing value of 43.91 compared to the actual mean of the original data set.

```
set.seed(1122)

#removing the "bill_length_mm" data for three randomly selected penguins
sample_indices <- sample(x = 1:nrow(df), size = 3, replace = FALSE)

penguins_MCAR <- penguins |>
  mutate(bill_length_mm = if_else(row_number() %in% sample_indices, NA_real_, bill_length_mm)

summary(penguins_MCAR)|>
  kable()
```

Table 4: Summary Statistics Table: (MCAR)

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie :152	Biscoe :168	Min. :32.10	Min. :13.10	Min. :172.0	Min. :2700	female:165	Min. :2007
Chinstrap: 68	Dream :124	1st Qu.:39.20	1st Qu.:15.60	1st Qu.:190.0	1st Qu.:3550	male :168	1st Qu.:2007
Gentoo :124	Torgersen: 52	Median :44.50	Median :17.30	Median :197.0	Median :4050	NA's : 11	Median :2008
NA	NA	Mean :43.91	Mean :17.15	Mean :200.9	Mean :4202	NA	Mean :2008
NA	NA	3rd Qu.:48.50	3rd Qu.:18.70	3rd Qu.:213.0	3rd Qu.:4750	NA	3rd Qu.:2009
NA	NA	Max. :59.60	Max. :21.50	Max. :231.0	Max. :6300	NA	Max. :2009

Table 4: Summary Statistics Table: (MCAR)

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
NA	NA	NA's :5	NA's :2	NA's :2	NA's :2	NA	NA

2.2. Data Missing At Random (MAR).

This section simulates a scenario, wherein missing values are simulated at random for *bill_length_mm* based on the species. The `mutate()` function is used to assign NA to the *bill_length_mm* columns with the species having the longest average bill length. The objective of this simulation is to replicate missing values using the maximum *bill_length_mm* of the species and to assess the impact of the missing values on results. Therefore, from Table 5, it can be noted that there a significant decline in the average value of the *bill_length_mm* from an original value of 43.92 to a value of 42.70.

```
#species with the highest average bill_length_mm
highest_bill_length_species <- penguins |>
  group_by(species)|>
  summarise(average_bill_length = mean(bill_length_mm, na.rm = TRUE))|>
  slice_max(average_bill_length, n = 1) |>
  pull(species)

# Simulate Missing at Random (MAR) for bill_length_mm based on the selected species
penguins_MAR <- penguins |>
  mutate(bill_length_mm =
           if_else(species %in% highest_bill_length_species, NA_real_, bill_length_mm))

# Summary of the modified dataset
summary(penguins_MAR)|>
  kable()
```

Table 5: Summary Statistics Table:(MAR)

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie :152	Biscoe :168	Min. :32.10	Min. :13.10	Min. :172.0	Min. :2700	female:165	Min. :2007
Chinstrap: 68	Dream :124	1st Qu.:38.35	1st Qu.:15.60	1st Qu.:190.0	1st Qu.:3550	male :168	1st Qu.:2007
Gentoo :124	Torgersen: 52	Median :42.00	Median :17.30	Median :197.0	Median :4050	NA's : 11	Median :2008

Table 5: Summary Statistics Table:(MAR)

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
NA	NA	Mean :42.70	Mean :17.15	Mean :200.9	Mean :4202	NA	Mean :2008
NA	NA	3rd Qu.:46.67	3rd Qu.:18.70	3rd Qu.:213.0	3rd Qu.:4750	NA	3rd Qu.:2009
NA	NA	Max. :59.60	Max. :21.50	Max. :231.0	Max. :6300	NA	Max. :2009
NA	NA	NA's :70	NA's :2	NA's :2	NA's :2	NA	NA

2.3. Data Missing Not At Random (MNAR).

The following simulation assesses the investigation into the potential biases in the data patterns induced due to missing values in the data of *bill_length_mm* due to the species that have an above average bill length. Based on the Table 6, it can be stated that there is a significant decrease in the average *bill_length_mm* of the penguins when the data is missing but not at random.

```r
# Determine the species with the highest average bill length
highest_bill_length_species <- df |>
  group_by(species) |>
  summarise(average_bill_length = mean(bill_length_mm, na.rm = TRUE)) |>
  slice_max(order_by = average_bill_length, n = 2) |>
  pull(species)

# Simulate MNAR for bill_length_mm based on the selected species
penguins_MNAR <- df |>
  mutate(bill_length_mm =
           if_else(species %in% highest_bill_length_species,
                   NA_real_,
                   bill_length_mm))

summary(penguins_MNAR)|>
  kable()
```

Table 6: Summary Statistics Table:(MNAR)

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Length:344	Length:344	Min. :32.10	Min. :13.10	Min. :172.0	Min. :2700	Length:344	Min. :2007
Class :character	Class :character	1st Qu.:36.75	1st Qu.:15.60	1st Qu.:190.0	1st Qu.:3550	Class :character	1st Qu.:2007
Mode :character	Mode :character	Median :38.80	Median :17.30	Median :197.0	Median :4050	Mode :character	Median :2008
NA	NA	Mean :38.79	Mean :17.15	Mean :200.9	Mean :4202	NA	Mean :2008
NA	NA	3rd Qu.:40.75	3rd Qu.:18.70	3rd Qu.:213.0	3rd Qu.:4750	NA	3rd Qu.:2009
NA	NA	Max. :46.00	Max. :21.50	Max. :231.0	Max. :6300	NA	Max. :2009
NA	NA	NA's :193	NA's :2	NA's :2	NA's :2	NA	NA

3.0. Imputing Missing Values.

The missing values in the data set are imputed using the `mice` package by van Buuren and Groothuis-Oudshoorn (2011). The `mice()` function uses the imputation method where missing values are imputed based on the observation values of each variable. Based on Table 7, it can be see that the data set does not have any missing values and that the missingness in the data is dealt with without dropping any observation that could have lead to loss of information.

```r
# multiple imputation
multiple_imputation <- mice(df, m = 5, method = 'pmm', print = FALSE)
```

```
Warning: Number of logged events: 3
```

```r
# dataset after imputing
df_imputed <- complete(multiple_imputation, action = 1)|>
  as_tibble()

# checking for missing values in each column
colSums(is.na(df_imputed))|>
  kable()
```

Table 7: Missing Values aafter imutation.

	x
species	0
island	0
bill_length_mm	0
bill_depth_mm	0
flipper_length_mm	0
body_mass_g	0
sex	11
year	0

This section of the code assesses if the mean imputation for the missing values in the *bill_length_mm* variable for all three species as suggested to me by **Dailin Li**. It can be stated, based on Table 8 that the imputed values closely resemble the the actual values and therefore the method of mean imputation is valid for this data set. I further deduced based on the suggested method by **Yanyu Wu** that since the number of missing values were significantly lower, there would be lower probability of biases in the data due to mean imputation as there are no outliers, that would skew the mean.

```
# Calculating the mean for input replacement (simple mean imputation)
mean_bill_length <- mean(df$bill_length_mm, na.rm = TRUE)

# Replacing NA in bill_length_mm with mean_bill_length for simple mean imputation
df_input_mean <- df
df_input_mean$bill_length_mm[is.na(df$bill_length_mm)] <- mean_bill_length

# Actual mean by species
actual_by_species <- df |>
  group_by(species) |>
  summarize(Actual = mean(bill_length_mm, na.rm = TRUE))

# Input mean by species
input_mean_by_species <- df_input_mean|>
  group_by(species)|>
  summarize(Input_mean = mean(bill_length_mm))

# Multiple imputation mean by species from df_imputed
multi_imp_by_species <- df_imputed |>
  group_by(species) |>
  summarize(Multiple_imputation = mean(bill_length_mm))
```

```
# Merging the tables together
comparison_table <- reduce(list(actual_by_species, input_mean_by_species, multi_imp_by_speci
comparison_table|>
  kable()
```

Table 8: Comparive Analysis Mean Imputation.

species	Actual	Input_mean	Multiple_imputation
Adelie	38.79139	38.82514	38.82632
Chinstrap	48.83382	48.83382	48.83382
Gentoo	47.50488	47.47598	47.49839

4.0. Comparative Mean

A comparison of missing data handling strategies in the penguin data set offers insights about their impact on calculating overall bill length based on Table 9. The *Drop_missing* and *Input_mean* techniques produce similar overall mean bill lengths of roughly 43.92 mm, indicating the elimination of missing data and simple mean imputation. Multiple imputation produces a slightly higher mean of approximately 43.93 mm, demonstrating its capacity to capture uncertainty. The *Actual_mean*, determined from the original data set, closely matches the means obtained from the drop_missing and input_mean approaches, showing the importance of evaluating the true data distribution. These findings illustrate the trade-offs between simplicity and accuracy in managing missing data, emphasizing the benefits of approaches such as multiple imputation in capturing variability.

```
# Drop missing observations for bill_length_mm to calculate the mean
mean_bill_length_drop_missing <- mean(df$bill_length_mm, na.rm = TRUE)
# Input mean
df_input_mean <- df
df_input_mean$bill_length_mm[is.na(df_input_mean$bill_length_mm)] <- mean_bill_length_drop_mi

# multiple imutation mean
multiple_imputation <- mice(df, m = 5, method = 'pmm', maxit = 5, print = FALSE, seed = 123)
```

```
Warning: Number of logged events: 3
```

```
df_imputed <- complete(multiple_imputation, action = 1) |>
  as_tibble()
```

```r
#  overall mean for each method
overall_mean_drop_missing <- mean_bill_length_drop_missing
overall_mean_input_mean <- mean(df_input_mean$bill_length_mm)
overall_mean_multiple_imputation <- mean(df_imputed$bill_length_mm)
overall_mean_actual <- mean(df$bill_length_mm, na.rm = TRUE)

# comparison table
comparison_table <- data.frame(
  Observation = c("Overall"),
  Drop_missing = overall_mean_drop_missing,
  Input_mean = overall_mean_input_mean,
  Multiple_imputation = overall_mean_multiple_imputation,
  Actual = overall_mean_actual
)

# Printing the comparison table
comparison_table|>
  kable()
```

Table 9: Comparetive Analysis.

Observation	Drop_missing	Input_mean	Multiple_imputation	Actual
Overall	43.92193	43.92193	43.91541	43.92193

References:

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://doi.org/10.5281/zenodo.3960218.

R Core Team. 2023. "R: A Language and Environment for Statistical Computing." https://www.R-project.org/.

Salgado, Cátia M., Carlos Azevedo, Hugo Proença, and Susana M. Vieira. 2016. "Missing Data." Springer, Cham (CH). http://europepmc.org/books/NBK543620.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in r." *Journal of Statistical Software* 45 (3): 1–67. https://doi.org/10.18637/jss.v045.i03.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://yihui.org/knitr/.