Data Engineering Project with R

Dataset: INSERT DATASET NAME with LINK

INSERT YOUR NAME HERE

2023-12-21

Table of contents

List of Figures

List of Tables

List of Listings

Instructions

- Use this template for your project report.
- In the YAML field `author:` insert your name.
- In the YAML field `subtitle:` insert the name of your dataset a hyper-linked it to your data source.
- In the YAML fields `output:` for HTML and PDF insert the correct name for the output files (`name-of-dataset`*`your-surname`*`your-first-name`) including correct extensions `.html` `.pdf`.
- Do not remove import YAML setting like: `embed-resources: true`
- Remove all `callout-caution ### Instructions` Instructions before submission from your final report.
- Render your report into HTML and PDF format.
- Submit in Moodle the following three files:

  1. Rendered HTML report
  2. Rendered PDF report
  3. dataset file

Instructions

Your report must be of high quality, meaning that your report:

- is visually and textually pleasing of
- does not look/read/feel like a draft instead of a finished analysis
- explains/discusses your findings and results in the main text, e.g., explain/discuss all figures/table in the main text
- is representable such that it can show to any interested third party

- uses figure/table captions/linking/reference (see example further down)
- Do not show any standard printout of R-code, use for data.frame/tibbles `knitr::kable()` printing.
- Do not simply print datasets (too many lines) use instead `rmarkdown::paged_table()`

1 Introduction

Instructions

Write some text that introduces the reader to your report. Give a brief overview of what is shown in the sections of your report, like the data handling, descriptive statistics, data visualisations, etc.

1.1 Libraries

```
library <- function(...) {suppressPackageStartupMessages(base::library(...))}
library(tidyverse)
```

2 Data

2.1 Data source

Instructions

Show your data source and provide appropriate links to your data including data descriptive files if available.

2.2 Data import

Instructions

Download and import your dataset. Assign each variable appropriate/correct data types. Discuss/explain your choice for each variable (groups) in the main text.

2.3 Data dictionary

Instructions

Make a data dictionary (table) that describes all variables in the dataset (shows what they mean/measure).

3 Summary statistic tables

Instructions

Make summary statistics of your data.

3.1 Numeric iris

Table 1 shows for the numerical variables in the iris dataset some summary statistics.

```
iris |>
  janitor::clean_names() |>
  mutate(row = row_number() |> factor()) |>
  pivot_longer(cols = where(is.numeric)) |>
  group_by(name) |>
  summarize(N = n(),
            min = min(value),
            mean = mean(value),
            median = median(value),
            max = max(value),
            st.dev = sd(value)
            ) |>
  knitr::kable(digits = 2)
```

Table 1: Summary statistics of numerical variables in `datasets::iris` with tidyverse (ungrouped)

name	N	min	mean	median	max	st.dev
petal_length	150	1.0	3.76	4.35	6.9	1.77
petal_width	150	0.1	1.20	1.30	2.5	0.76
sepal_length	150	4.3	5.84	5.80	7.9	0.83
sepal_width	150	2.0	3.06	3.00	4.4	0.44

3.2 Numeric iris grouped

Table 2 shows for the numerical variables in the iris dataset grouped summary statistics for different Species.

```r
iris |>
  janitor::clean_names() |>
  mutate(row = row_number() |> as.character()) |>
  pivot_longer(cols = where(is.numeric)) |>
  group_by(name, species) |>
  summarize(N = n(),
            min = min(value),
            mean = mean(value),
            median = median(value),
            max = max(value),
            st.dev = sd(value)
            ) |>
  knitr::kable(digits = 2)
```

Table 2: Summary statistics of numerical variables in `datasets::iris` with tidyverse grouped by Species

name	species	N	min	mean	median	max	st.dev
petal_length	setosa	50	1.0	1.46	1.50	1.9	0.17
petal_length	versicolor	50	3.0	4.26	4.35	5.1	0.47
petal_length	virginica	50	4.5	5.55	5.55	6.9	0.55
petal_width	setosa	50	0.1	0.25	0.20	0.6	0.11
petal_width	versicolor	50	1.0	1.33	1.30	1.8	0.20
petal_width	virginica	50	1.4	2.03	2.00	2.5	0.27
sepal_length	setosa	50	4.3	5.01	5.00	5.8	0.35
sepal_length	versicolor	50	4.9	5.94	5.90	7.0	0.52
sepal_length	virginica	50	4.9	6.59	6.50	7.9	0.64
sepal_width	setosa	50	2.3	3.43	3.40	4.4	0.38
sepal_width	versicolor	50	2.0	2.77	2.80	3.4	0.31
sepal_width	virginica	50	2.2	2.97	3.00	3.8	0.32

3.3 Nominal iris

Table 3 shows summary statistics for the iris factor variables.

```
iris |>
  janitor::clean_names() |>
  mutate(row = row_number() |> as.character()) |>
  select(where(is.factor)) |>
  pivot_longer(cols = where(is.factor)) |>
  group_by(name) |>
  count(value) |>
  ungroup() |>
  arrange(desc(name), n) |>
  knitr::kable(digits = 2)
```

Table 3: Summary statistics of factor variables in `datasets::iris` with tidyverse.

name	value	n
species	setosa	50
species	versicolor	50
species	virginica	50

3.4 Nominal penguins

Table 4 shows summary statistics for the iris factor variables.s and penguins factor variables.

```
data("penguins", package="palmerpenguins")
penguins |>
  janitor::clean_names() |>
  mutate(row = row_number() |> as.character()) |>
  select(where(is.factor)) |>
  pivot_longer(cols = where(is.factor)) |>
  group_by(name) |>
  count(value) |>
  ungroup() |>
  arrange(desc(name), n) |> #dput()
  knitr::kable(digits = 2)
```

Table 4: Summary statistics of factor variables in `palmerpenguins::penguins` with tidyverse.

name	value	n
species	Chinstrap	68
species	Gentoo	124
species	Adelie	152
sex	NA	11
sex	female	165
sex	male	168
island	Torgersen	52
island	Dream	124
island	Biscoe	168

3.5 All variable statistics

Table 5 shows summary statistics applicable to different data type.

```r
# specify full dataset
data_all <- iris
data_all <- palmerpenguins::penguins

## numerical data
data_num <- data_all[sapply(data_all, is.numeric)]
data_num <- subset(data_all, select = sapply(data_all,is.numeric))
data_num <- data_all |> select(where(is.numeric))

## nominal data
data_nom <- data_all[!sapply(data_all, is.numeric)]
data_chr <- data_all[sapply(data_all, is.character)]
data_lgl <- data_all[sapply(data_all, is.logical)]
data_fct <- data_all[sapply(data_all, is.factor)]
```

```r
data.frame(
  #var = names(data_all),
  n_obs = sapply(data_all, function(.) length(na.omit(.)))
  ,n_all = sapply(na.omit(data_all), length)
  ,n_missing = sapply(data_all, function(.) sum(is.na(.)))
  ,mode = data_all |> sapply(mode)
  ,class = data_all |> sapply(class)
) |> #as_tibble() |>
  knitr::kable()
```

Table 5: Base R statistics applicable to all variables

	n_obs	n_all	n_missing	mode	class
species	344	333	0	numeric	factor
island	344	333	0	numeric	factor
bill_length_mm	342	333	2	numeric	numeric
bill_depth_mm	342	333	2	numeric	numeric
flipper_length_mm	342	333	2	numeric	integer
body_mass_g	342	333	2	numeric	integer
sex	333	333	11	numeric	factor
year	344	333	0	numeric	integer

3.6 Numerical data

```r
library(tidyverse)
## n_obs individual by variables
## n_all all variables have same n observations

library(tidyverse)
data.frame(
  var = names(data_num),
  n_obs = sapply(data_num, function(.) length(na.omit(.)))
  ,n_all = sapply(na.omit(data_num), length)
  ,min = sapply(data_num, function(.) min(na.omit(.)))
  ,min_all = sapply(na.omit(data_num), min)
  ,mean = sapply(data_num, function(.) mean(na.omit(.)))
  ,mean_all = sapply(na.omit(data_num), mean)
  ,median = sapply(data_num, function(.) median(na.omit(.)))
  ,median_all = sapply(na.omit(data_num), median)
  ,max = sapply(data_num, function(.) max(na.omit(.)))
  ,max_all = sapply(na.omit(data_num), max)
  ,sd = sapply(data_num, function(.) sd(na.omit(.)))
  ,sd_all = sapply(na.omit(data_num), sd)
) |>
  as_tibble() |>
  select(!contains("_all")) |>
  knitr::kable(digits = 2)
```

var	n_obs	min	mean	median	max	sd
bill_length_mm	342	32.1	43.92	44.45	59.6	5.46
bill_depth_mm	342	13.1	17.15	17.30	21.5	1.97
flipper_length_mm	342	172.0	200.92	197.00	231.0	14.06
body_mass_g	342	2700.0	4201.75	4050.00	6300.0	801.95
year	344	2007.0	2008.03	2008.00	2009.0	0.82

3.7 print

```r
rmarkdown::paged_table(iris)
```

```
# A tibble: 150 x 5
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
          <dbl>       <dbl>        <dbl>       <dbl> <fct>
 1          5.1         3.5          1.4         0.2 setosa
 2          4.9         3            1.4         0.2 setosa
 3          4.7         3.2          1.3         0.2 setosa
 4          4.6         3.1          1.5         0.2 setosa
 5          5           3.6          1.4         0.2 setosa
 6          5.4         3.9          1.7         0.4 setosa
 7          4.6         3.4          1.4         0.3 setosa
 8          5           3.4          1.5         0.2 setosa
 9          4.4         2.9          1.4         0.2 setosa
10          4.9         3.1          1.5         0.1 setosa
# i 140 more rows
```

4 Data visualisations

Instructions

Visually explore your data.

5 Summary

Instructions

Summarise your finding.