Data Cleaning

Sasit

Data Clening with R Programming

First install package

```
# Project Data Cleaning with R

# Step 1 install packages  and load library
install.packages("tidyverse")
install.packages("skimr")
install.packages("janitor")
install.packages("dplyr")
install.packages("lubridate")
```

And load package by running library

```
library(tidyverse)
library(skimr)
library(janitor)
library(dplyr)
library(lubridate)
```

Improt data

Download data from kaggle Hotel booking demand

```
# import and review data

book_df <- read_csv("hotel_bookings.csv")
```

Explore Data

In the Data Exploration step, we will conduct a general examination of the dataset. This includes checking the number of rows and columns, inspecting data types, identifying missing values in the data, and examining sample data within the table.

```
# view head of data frame
head(book_df)
```

```
## # A tibble: 6 x 32
##   hotel        is_canceled lead_time arrival_date_year arrival_date_month
##   <chr>              <dbl>     <dbl>             <dbl> <chr>
## 1 Resort Hotel           0       342              2015 July
## 2 Resort Hotel           0       737              2015 July
## 3 Resort Hotel           0         7              2015 July
## 4 Resort Hotel           0        13              2015 July
## 5 Resort Hotel           0        14              2015 July
## 6 Resort Hotel           0        14              2015 July
## # i 27 more variables: arrival_date_week_number <dbl>,
## #   arrival_date_day_of_month <dbl>, stays_in_weekend_nights <dbl>,
## #   stays_in_week_nights <dbl>, adults <dbl>, children <dbl>, babies <dbl>,
## #   meal <chr>, country <chr>, market_segment <chr>,
## #   distribution_channel <chr>, is_repeated_guest <dbl>,
## #   previous_cancellations <dbl>, previous_bookings_not_canceled <dbl>,
## #   reserved_room_type <chr>, assigned_room_type <chr>, ...
```

```
# view structure of data frame as number of columns and rows,
# column name, data type and example of data
str(book_df)
```

```
## spc_tbl_ [119,390 x 32] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ hotel                         : chr [1:119390] "Resort Hotel" "Resort Hotel" "Resort Hotel" "Reso
##  $ is_canceled                   : num [1:119390] 0 0 0 0 0 0 0 0 1 1 ...
##  $ lead_time                     : num [1:119390] 342 737 7 13 14 14 0 9 85 75 ...
##  $ arrival_date_year             : num [1:119390] 2015 2015 2015 2015 2015 ...
##  $ arrival_date_month            : chr [1:119390] "July" "July" "July" "July" ...
##  $ arrival_date_week_number      : num [1:119390] 27 27 27 27 27 27 27 27 27 27 ...
##  $ arrival_date_day_of_month     : num [1:119390] 1 1 1 1 1 1 1 1 1 1 ...
##  $ stays_in_weekend_nights       : num [1:119390] 0 0 0 0 0 0 0 0 0 0 ...
##  $ stays_in_week_nights          : num [1:119390] 0 0 1 1 2 2 2 2 3 3 ...
##  $ adults                        : num [1:119390] 2 2 1 1 2 2 2 2 2 2 ...
##  $ children                      : num [1:119390] 0 0 0 0 0 0 0 0 0 0 ...
##  $ babies                        : num [1:119390] 0 0 0 0 0 0 0 0 0 0 ...
##  $ meal                          : chr [1:119390] "BB" "BB" "BB" "BB" ...
##  $ country                       : chr [1:119390] "PRT" "PRT" "GBR" "GBR" ...
##  $ market_segment                : chr [1:119390] "Direct" "Direct" "Direct" "Corporate" ...
##  $ distribution_channel          : chr [1:119390] "Direct" "Direct" "Direct" "Corporate" ...
##  $ is_repeated_guest             : num [1:119390] 0 0 0 0 0 0 0 0 0 0 ...
##  $ previous_cancellations        : num [1:119390] 0 0 0 0 0 0 0 0 0 0 ...
##  $ previous_bookings_not_canceled: num [1:119390] 0 0 0 0 0 0 0 0 0 0 ...
##  $ reserved_room_type            : chr [1:119390] "C" "C" "A" "A" ...
##  $ assigned_room_type            : chr [1:119390] "C" "C" "C" "A" ...
##  $ booking_changes               : num [1:119390] 3 4 0 0 0 0 0 0 0 0 ...
##  $ deposit_type                  : chr [1:119390] "No Deposit" "No Deposit" "No Deposit" "No Deposit
##  $ agent                         : chr [1:119390] "NULL" "NULL" "NULL" "304" ...
##  $ company                       : chr [1:119390] "NULL" "NULL" "NULL" "NULL" ...
##  $ days_in_waiting_list          : num [1:119390] 0 0 0 0 0 0 0 0 0 0 ...
##  $ customer_type                 : chr [1:119390] "Transient" "Transient" "Transient" "Transient" ..
##  $ adr                           : num [1:119390] 0 0 75 75 98 ...
##  $ required_car_parking_spaces   : num [1:119390] 0 0 0 0 0 0 0 0 0 0 ...
##  $ total_of_special_requests     : num [1:119390] 0 0 0 0 1 1 0 1 1 0 ...
##  $ reservation_status            : chr [1:119390] "Check-Out" "Check-Out" "Check-Out" "Check-Out" ..
##  $ reservation_status_date       : Date[1:119390], format: "2015-07-01" "2015-07-01" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   hotel = col_character(),
##   ..   is_canceled = col_double(),
##   ..   lead_time = col_double(),
##   ..   arrival_date_year = col_double(),
##   ..   arrival_date_month = col_character(),
##   ..   arrival_date_week_number = col_double(),
##   ..   arrival_date_day_of_month = col_double(),
##   ..   stays_in_weekend_nights = col_double(),
##   ..   stays_in_week_nights = col_double(),
##   ..   adults = col_double(),
##   ..   children = col_double(),
##   ..   babies = col_double(),
##   ..   meal = col_character(),
##   ..   country = col_character(),
##   ..   market_segment = col_character(),
##   ..   distribution_channel = col_character(),
##   ..   is_repeated_guest = col_double(),
##   ..   previous_cancellations = col_double(),
##   ..   previous_bookings_not_canceled = col_double(),
##   ..   reserved_room_type = col_character(),
##   ..   assigned_room_type = col_character(),
##   ..   booking_changes = col_double(),
##   ..   deposit_type = col_character(),
##   ..   agent = col_character(),
##   ..   company = col_character(),
##   ..   days_in_waiting_list = col_double(),
##   ..   customer_type = col_character(),
##   ..   adr = col_double(),
##   ..   required_car_parking_spaces = col_double(),
##   ..   total_of_special_requests = col_double(),
##   ..   reservation_status = col_character(),
##   ..   reservation_status_date = col_date(format = "")
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
glimpse(book_df)
```

```
## Rows: 119,390
## Columns: 32
## $ hotel                          <chr> "Resort Hotel", "Resort Hotel", "Resort~
## $ is_canceled                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, ~
## $ lead_time                      <dbl> 342, 737, 7, 13, 14, 14, 0, 9, 85, 75, ~
## $ arrival_date_year              <dbl> 2015, 2015, 2015, 2015, 2015, 2015, 201~
## $ arrival_date_month             <chr> "July", "July", "July", "July", "July",~
## $ arrival_date_week_number       <dbl> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,~
## $ arrival_date_day_of_month      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ stays_in_weekend_nights        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ stays_in_week_nights           <dbl> 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, ~
## $ adults                         <dbl> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ children                       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ babies                         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ meal                           <chr> "BB", "BB", "BB", "BB", "BB", "BB", "BB~
## $ country                        <chr> "PRT", "PRT", "GBR", "GBR", "GBR", "GBR~
## $ market_segment                 <chr> "Direct", "Direct", "Direct", "Corporat~
## $ distribution_channel           <chr> "Direct", "Direct", "Direct", "Corporat~
## $ is_repeated_guest              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ previous_cancellations         <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ previous_bookings_not_canceled <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ reserved_room_type             <chr> "C", "C", "A", "A", "A", "A", "C", "C",~
## $ assigned_room_type             <chr> "C", "C", "C", "A", "A", "A", "C", "C",~
## $ booking_changes                <dbl> 3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ deposit_type                   <chr> "No Deposit", "No Deposit", "No Deposit~
## $ agent                          <chr> "NULL", "NULL", "NULL", "304", "240", "~
## $ company                        <chr> "NULL", "NULL", "NULL", "NULL", "NULL",~
## $ days_in_waiting_list           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ customer_type                  <chr> "Transient", "Transient", "Transient", ~
## $ adr                            <dbl> 0.00, 0.00, 75.00, 75.00, 98.00, 98.00,~
## $ required_car_parking_spaces    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ total_of_special_requests      <dbl> 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 3, ~
## $ reservation_status             <chr> "Check-Out", "Check-Out", "Check-Out", ~
## $ reservation_status_date        <date> 2015-07-01, 2015-07-01, 2015-07-02, 20~
```

```r
# view only column name
colnames(book_df)
```

```
##  [1] "hotel"                          "is_canceled"
##  [3] "lead_time"                      "arrival_date_year"
##  [5] "arrival_date_month"             "arrival_date_week_number"
##  [7] "arrival_date_day_of_month"      "stays_in_weekend_nights"
##  [9] "stays_in_week_nights"           "adults"
## [11] "children"                       "babies"
## [13] "meal"                           "country"
## [15] "market_segment"                 "distribution_channel"
## [17] "is_repeated_guest"              "previous_cancellations"
## [19] "previous_bookings_not_canceled" "reserved_room_type"
## [21] "assigned_room_type"             "booking_changes"
## [23] "deposit_type"                   "agent"
## [25] "company"                        "days_in_waiting_list"
## [27] "customer_type"                  "adr"
## [29] "required_car_parking_spaces"    "total_of_special_requests"
## [31] "reservation_status"             "reservation_status_date"
```

```r
# show summary of data
summary(book_df)
```

```
##     hotel            is_canceled       lead_time   arrival_date_year
##  Length:119390      Min.   :0.0000   Min.   :  0   Min.   :2015
##  Class :character   1st Qu.:0.0000   1st Qu.: 18   1st Qu.:2016
##  Mode  :character   Median :0.0000   Median : 69   Median :2016
##                     Mean   :0.3704   Mean   :104   Mean   :2016
##                     3rd Qu.:1.0000   3rd Qu.:160   3rd Qu.:2017
##                     Max.   :1.0000   Max.   :737   Max.   :2017
##
##  arrival_date_month arrival_date_week_number arrival_date_day_of_month
##  Length:119390      Min.   : 1.00            Min.   : 1.0
##  Class :character   1st Qu.:16.00            1st Qu.: 8.0
##  Mode  :character   Median :28.00            Median :16.0
##                     Mean   :27.17            Mean   :15.8
##                     3rd Qu.:38.00            3rd Qu.:23.0
##                     Max.   :53.00            Max.   :31.0
##
##  stays_in_weekend_nights stays_in_week_nights     adults
##  Min.   : 0.0000         Min.   : 0.0         Min.   : 0.000
##  1st Qu.: 0.0000         1st Qu.: 1.0         1st Qu.: 2.000
##  Median : 1.0000         Median : 2.0         Median : 2.000
##  Mean   : 0.9276         Mean   : 2.5         Mean   : 1.856
##  3rd Qu.: 2.0000         3rd Qu.: 3.0         3rd Qu.: 2.000
##  Max.   :19.0000         Max.   :50.0         Max.   :55.000
##
##     children           babies              meal            country
##  Min.   : 0.0000   Min.   : 0.000000   Length:119390      Length:119390
##  1st Qu.: 0.0000   1st Qu.: 0.000000   Class :character   Class :character
##  Median : 0.0000   Median : 0.000000   Mode  :character   Mode  :character
##  Mean   : 0.1039   Mean   : 0.007949
##  3rd Qu.: 0.0000   3rd Qu.: 0.000000
##  Max.   :10.0000   Max.   :10.000000
##  NA's   :4
##  market_segment     distribution_channel is_repeated_guest
##  Length:119390      Length:119390        Min.   :0.00000
##  Class :character   Class :character     1st Qu.:0.00000
##  Mode  :character   Mode  :character     Median :0.00000
##                                          Mean   :0.03191
##                                          3rd Qu.:0.00000
##                                          Max.   :1.00000
##
##  previous_cancellations previous_bookings_not_canceled reserved_room_type
##  Min.   : 0.00000       Min.   : 0.0000                Length:119390
##  1st Qu.: 0.00000       1st Qu.: 0.0000                Class :character
##  Median : 0.00000       Median : 0.0000                Mode  :character
##  Mean   : 0.08712       Mean   : 0.1371
##  3rd Qu.: 0.00000       3rd Qu.: 0.0000
##  Max.   :26.00000       Max.   :72.0000
##
##  assigned_room_type booking_changes   deposit_type          agent
##  Length:119390      Min.   : 0.0000   Length:119390      Length:119390
##  Class :character   1st Qu.: 0.0000   Class :character   Class :character
##  Mode  :character   Median : 0.0000   Mode  :character   Mode  :character
##                     Mean   : 0.2211
##                     3rd Qu.: 0.0000
##                     Max.   :21.0000
##
##    company           days_in_waiting_list customer_type           adr
##  Length:119390      Min.   :  0.000      Length:119390      Min.   :  -6.38
##  Class :character   1st Qu.:  0.000      Class :character   1st Qu.:  69.29
##  Mode  :character   Median :  0.000      Mode  :character   Median :  94.58
##                     Mean   :  2.321                         Mean   : 101.83
##                     3rd Qu.:  0.000                         3rd Qu.: 126.00
##                     Max.   :391.000                         Max.   :5400.00
##
##  required_car_parking_spaces total_of_special_requests reservation_status
##  Min.   :0.00000             Min.   :0.0000             Length:119390
##  1st Qu.:0.00000             1st Qu.:0.0000             Class :character
##  Median :0.00000             Median :0.0000             Mode  :character
##  Mean   :0.06252             Mean   :0.5714
##  3rd Qu.:0.00000             3rd Qu.:1.0000
##  Max.   :8.00000             Max.   :5.0000
##
##  reservation_status_date
##  Min.   :2014-10-17
##  1st Qu.:2016-02-01
##  Median :2016-08-07
##  Mean   :2016-07-30
##  3rd Qu.:2017-02-08
##  Max.   :2017-09-14
##
```

```
# show summary of NA in each column
summary(is.na(book_df)) # column children have 4 NA
```

```
##    hotel          is_canceled     lead_time       arrival_date_year
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:119390    FALSE:119390    FALSE:119390    FALSE:119390
##
##  arrival_date_month arrival_date_week_number arrival_date_day_of_month
##  Mode :logical      Mode :logical            Mode :logical
##  FALSE:119390       FALSE:119390             FALSE:119390
##
##  stays_in_weekend_nights stays_in_week_nights   adults         children
##  Mode :logical          Mode :logical        Mode :logical   Mode :logical
##  FALSE:119390           FALSE:119390         FALSE:119390    FALSE:119386
##                                                               TRUE :4
##    babies           meal           country        market_segment
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:119390    FALSE:119390    FALSE:119390    FALSE:119390
##
##  distribution_channel is_repeated_guest previous_cancellations
##  Mode :logical        Mode :logical     Mode :logical
##  FALSE:119390         FALSE:119390      FALSE:119390
##
##  previous_bookings_not_canceled reserved_room_type assigned_room_type
##  Mode :logical                  Mode :logical      Mode :logical
##  FALSE:119390                   FALSE:119390       FALSE:119390
##
##  booking_changes deposit_type      agent          company
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:119390    FALSE:119390    FALSE:119390    FALSE:119390
##
##  days_in_waiting_list customer_type      adr
##  Mode :logical        Mode :logical   Mode :logical
##  FALSE:119390         FALSE:119390    FALSE:119390
##
##  required_car_parking_spaces total_of_special_requests reservation_status
##  Mode :logical               Mode :logical             Mode :logical
##  FALSE:119390                FALSE:119390              FALSE:119390
##
##  reservation_status_date
##  Mode :logical
##  FALSE:119390
##
```

```
# show only rows have NA
subset(book_df, is.na(children))
```

```
## # A tibble: 4 x 32
##   hotel      is_canceled lead_time arrival_date_year arrival_date_month
##   <chr>            <dbl>     <dbl>             <dbl> <chr>
## 1 City Hotel           1         2              2015 August
## 2 City Hotel           1         1              2015 August
## 3 City Hotel           1         1              2015 August
## 4 City Hotel           1         8              2015 August
## # i 27 more variables: arrival_date_week_number <dbl>,
## #   arrival_date_day_of_month <dbl>, stays_in_weekend_nights <dbl>,
## #   stays_in_week_nights <dbl>, adults <dbl>, children <dbl>, babies <dbl>,
## #   meal <chr>, country <chr>, market_segment <chr>,
## #   distribution_channel <chr>, is_repeated_guest <dbl>,
## #   previous_cancellations <dbl>, previous_bookings_not_canceled <dbl>,
## #   reserved_room_type <chr>, assigned_room_type <chr>, ...
```

Data Cleaning

- Begin by deleting rows with missing values.
- Next, rename the 'ard' column to 'average_daily_rate' for better clarity.
- Change the data type of the 'arrival_date_month' column from Char to int.
- Combine the values of 'arrival_date_year,' 'arrival_date_month,' and 'arrival_date_day_of_month' into 'arrival_date' and format it as YYYY-MM-DD.

```
# Data Cleaning
book_df_clean <- book_df %>%
  na.omit() %>% # Note Delete rows have missing value
  rename(average_daily_rate = adr) %>% # Note Change column name adr to Average Daily Rate
  mutate(
    arrival_date_month = as.integer(factor(arrival_date_month, levels = month.name)),
    arrival_date = paste(arrival_date_year, sprintf("%02d", arrival_date_month),
                         sprintf("%02d", arrival_date_day_of_month), sep = "-"))
# Note change data type in column arrival_date_month from char to int
    # Note combine date of arrival_date_year, arrival_date_month and arrival_date_day_of_month

#View(book_df_clean)

#glimpse(book_df_clean)
```

- Combine the number of guests from the 'adults,' 'children,' and 'babies' columns into a new column named 'number_of_guests.'
- Sum the number of nights spent on weekends ('stays_in_weekend_nights') and weekdays ('stays_in_week_nights') to create a new column called 'day_of_stays' representing the total length of stay.
- Select columns to exclude from the dataset.

```
# sum number of adults, children and babies and create new column
# sum number of stays_in_week_nights and stays_in_weekend_nights in new column day_of_stays
book_df_clean2 <- book_df_clean %>%
  mutate(number_of_guests = adults + children + babies ) %>%
  mutate(day_of_stays = stays_in_weekend_nights + stays_in_week_nights) %>%
  select(-arrival_date_day_of_month, -arrival_date_month, -arrival_date_year,
         -arrival_date_week_number, -adults, -children, -babies,
         -stays_in_weekend_nights, -stays_in_week_nights)

#View(book_df_clean2)
```

After completing the Data cleaning process, save the file in CSV format using the command.

```
# save file csv
write.csv(book_df_clean2, file = "hotel_bookings_clean")
```