Midterm Project

Tina Hajinejad

2023-02-24

Initializing

```
#First Read from csv files
tiketcom = read.csv("tiketcom_bestprice.csv")
distances = read.csv("distance_between_indonesian_airports.csv")

#Now we clean datasets

#Cleaning tiketcom_bestprice.csv
prices_names = str_split_fixed(names(tiketcom), "\\.", 5)
prices_values = vector("character", 5)

for (i in 1:nrow(tiketcom)){
  val = vector("character", 5)
  val = str_split_fixed(tiketcom[i,1], "\\|", 5)
  prices_values <- rbind(prices_values,val)
}
prices = prices_values[!apply(prices_values == "", 1, all),]   #Erasing the first row because it was bl
prices_df <- data.frame(prices)                    #Changing vector to data frame

colnames(prices_df)<-prices_names    #Column names should be varnames(unique variable names)
rownames(prices_df)<-c()
```

```
#Cleaning distance_between_indonesian_airports.csv
distance_names = str_split_fixed(names(distances), "\\.", 4)
distance_values = vector("character", 4)

for (i in 1:nrow(distances)){
  val = vector("character", 4)
  val = str_split_fixed(distances[i,1], "\\|", 4)
  distance_values <- rbind(distance_values,val)
}
distances = distance_values[!apply(distance_values == "", 1, all),]   #Erasing the first row because it
distance_df <- data.frame(distances)                    #Changing vector to data frame

colnames(distance_df)<-distance_names    #Column names should be varnames(unique variable names)
rownames(distance_df)<-c()
```

```
#Changing types:
cols.num <- c("best_price" , "distance_km" , "flight_time_hour")
prices_df[cols.num] <- sapply(prices_df[cols.num],as.numeric)
skim(prices_df)
```

Table 1: Data summary

Name	prices_df
Number of rows	45438
Number of columns	7
Column type frequency:	
character	4
numeric	3
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
extract_timestamp	0	1	26	26	0	15	0
origin	0	1	4	4	0	1	0
destination	0	1	3	3	0	29	0
depart_date	0	1	10	10	0	233	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
best_price	0	1	1491694.30	1087001.09	367200.00	789400.00	1046760.00	1677930.00	5226620.00	
distance_km	0	1	1138.91	948.91	133.25	478.42	853.55	1387.06	3773.77	
flight_time_hour	0	1	2.08	1.06	1.10	1.35	1.62	2.35	5.08	