RMIT UNIVERSITY

Data Wrangling | <code>

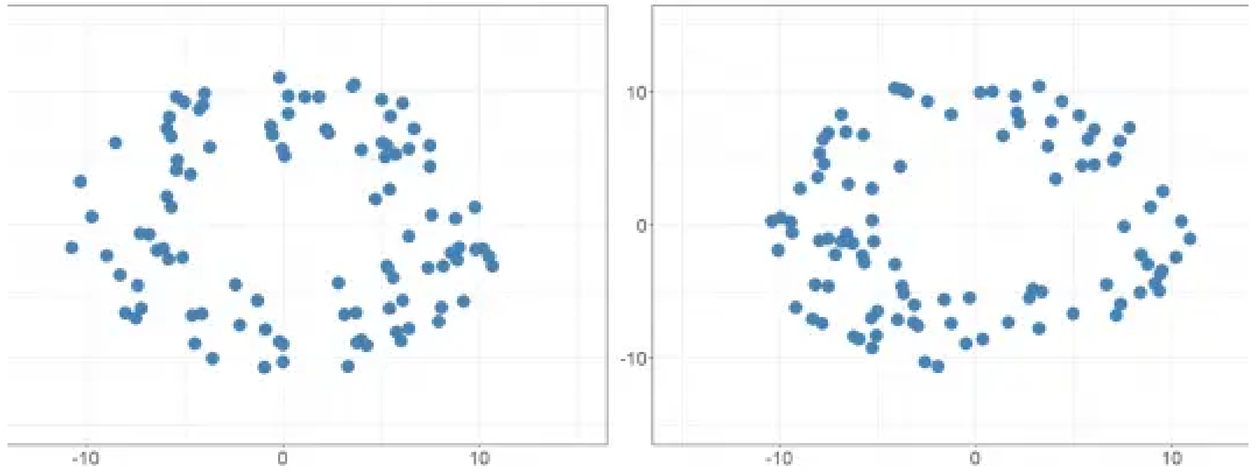
Mid-Semester Assessment | 2024 SEM 1

AIM : To generate synthetic data and analyse the data points

Snigdha Mathur | S4017572

Introduction

Synthetic data generation is a process of **creating data having the characterstics of the real life data**. The synthetically generated datasets further help in **scenario testing**, **development**, and **understand the trends without hampering the actual data points**.



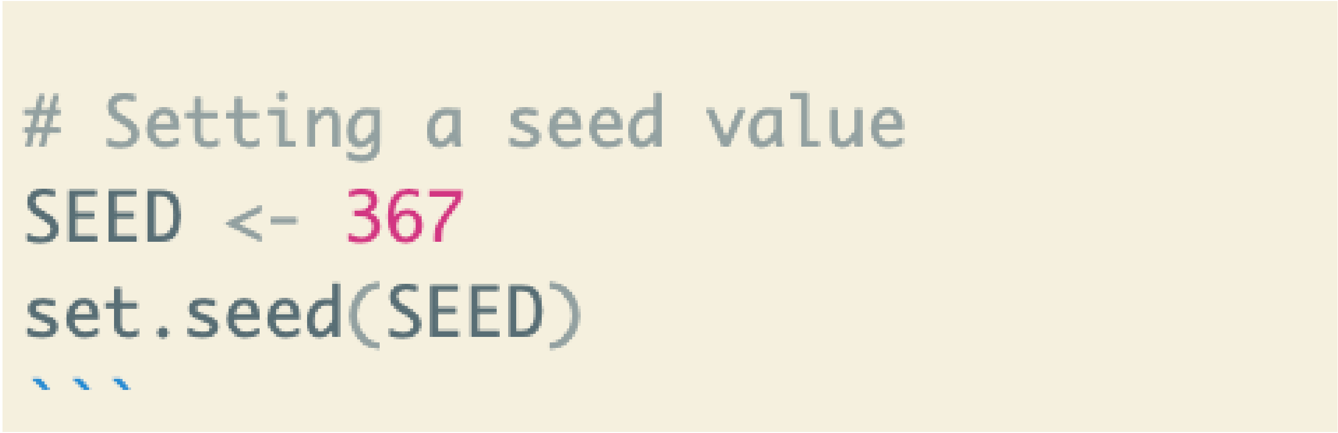
Original data

Synthetic data

The synthetic data retains the structure of the original data but is not the same

For making the data usable and reproducable, we have given a Seed value to randomly generate the data points.

For our analysis, a seed value of 367 has been set as indicated -->



~~~
# Setting a seed value
SEED <- 367
set.seed(SEED)
```
~~~

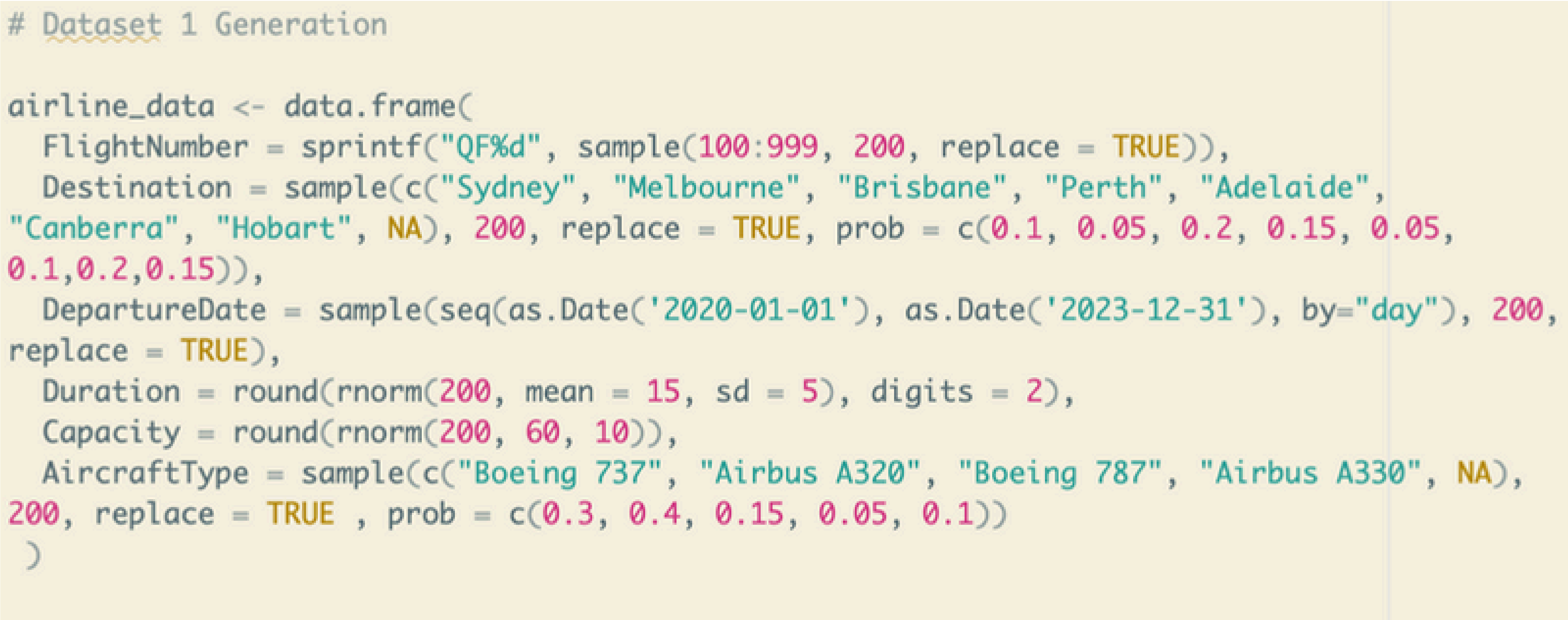
Brand Chosen : Qantas Airlines

Synthetic data generation is a process of **creating data having the characterstics of the real life data**. The synthetically generated datasets further help in **scenario testing**, **development**, and **understand the trends without hampering the actual data points**.

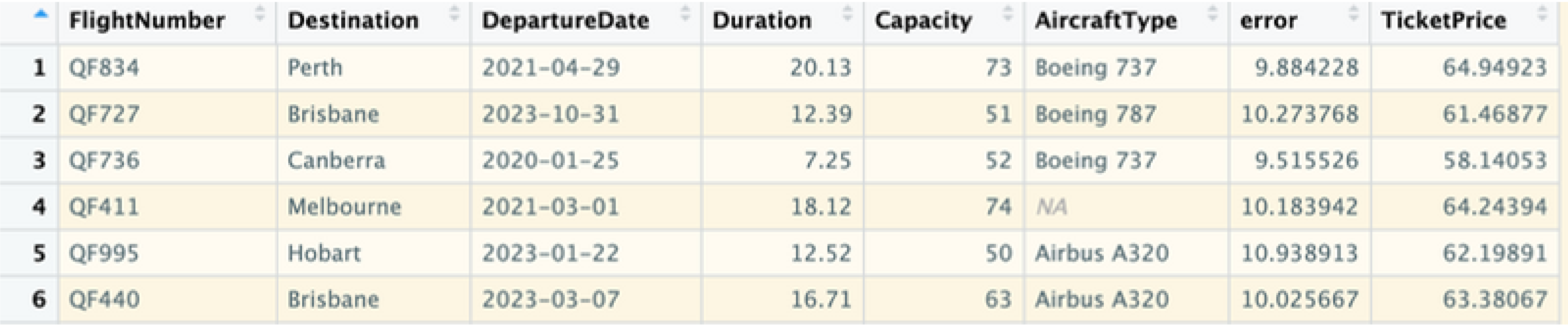
Synthetic Data Generation

We have generated 3 datasets for Qantas airlines : **Airline** , Passenger and Customer Feedback **datasets**

Airline Dataset



```
# Dataset 1 Generation

airline_data <- data.frame(
  FlightNumber = sprintf("QF%d", sample(100:999, 200, replace = TRUE)),
  Destination = sample(c("Sydney", "Melbourne", "Brisbane", "Perth", "Adelaide",
"Canberra", "Hobart", NA), 200, replace = TRUE, prob = c(0.1, 0.05, 0.2, 0.15, 0.05,
0.1,0.2,0.15)),
  DepartureDate = sample(seq(as.Date('2020-01-01'), as.Date('2023-12-31'), by="day"), 200,
replace = TRUE),
  Duration = round(rnorm(200, mean = 15, sd = 5), digits = 2),
  Capacity = round(rnorm(200, 60, 10)),
  AircraftType = sample(c("Boeing 737", "Airbus A320", "Boeing 787", "Airbus A330", NA),
200, replace = TRUE , prob = c(0.3, 0.4, 0.15, 0.05, 0.1))
 )
```

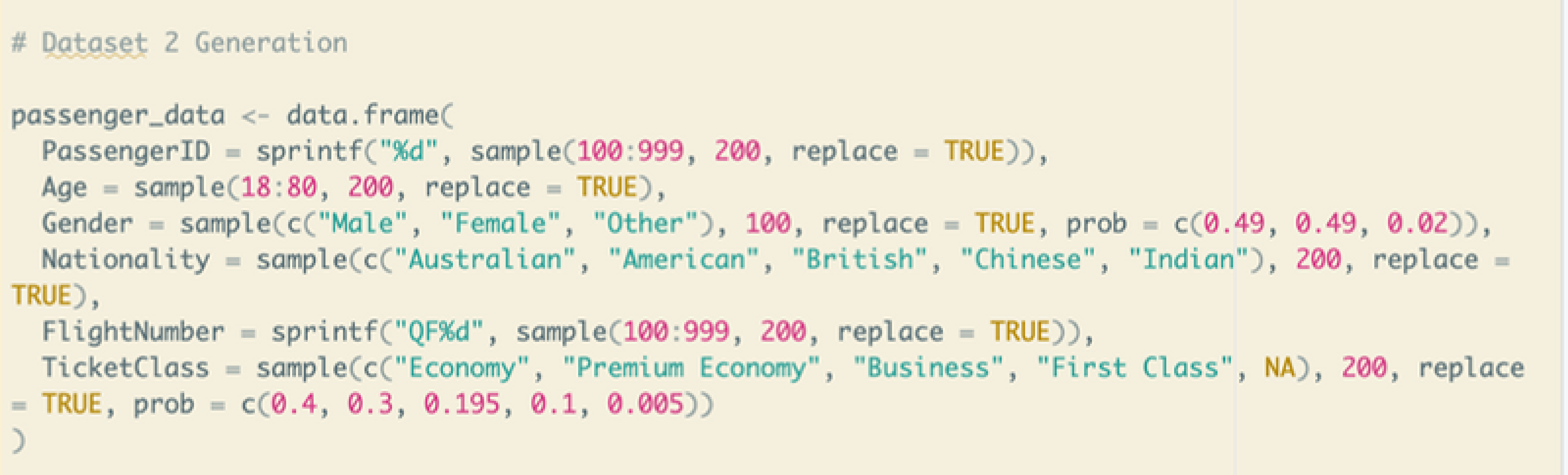


	FlightNumber	Destination	DepartureDate	Duration	Capacity	AircraftType	error	TicketPrice
1	QF834	Perth	2021-04-29	20.13	73	Boeing 737	9.884228	64.94923
2	QF727	Brisbane	2023-10-31	12.39	51	Boeing 787	10.273768	61.46877
3	QF736	Canberra	2020-01-25	7.25	52	Boeing 737	9.515526	58.14053
4	QF411	Melbourne	2021-03-01	18.12	74	NA	10.183942	64.24394
5	QF995	Hobart	2023-01-22	12.52	50	Airbus A320	10.938913	62.19891
6	QF440	Brisbane	2023-03-07	16.71	63	Airbus A320	10.025667	63.38067

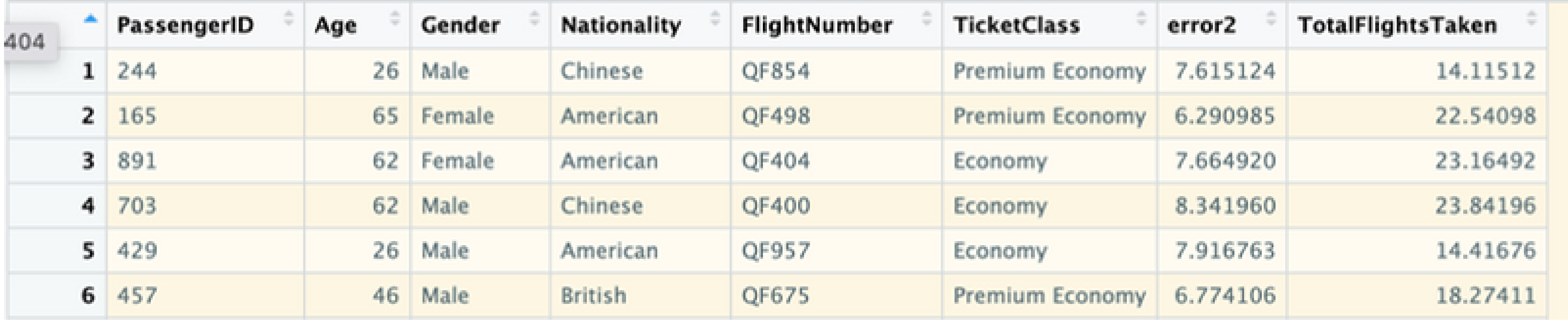
Synthetic Data Generation

We have generated 3 datasets for Qantas airlines : Airline , **Passenger** and Customer Feedback **datasets**

Passenger Dataset



```
# Dataset 2 Generation

passenger_data <- data.frame(
  PassengerID = sprintf("%d", sample(100:999, 200, replace = TRUE)),
  Age = sample(18:80, 200, replace = TRUE),
  Gender = sample(c("Male", "Female", "Other"), 100, replace = TRUE, prob = c(0.49, 0.49, 0.02)),
  Nationality = sample(c("Australian", "American", "British", "Chinese", "Indian"), 200, replace =
TRUE),
  FlightNumber = sprintf("QF%d", sample(100:999, 200, replace = TRUE)),
  TicketClass = sample(c("Economy", "Premium Economy", "Business", "First Class", NA), 200, replace
= TRUE, prob = c(0.4, 0.3, 0.195, 0.1, 0.005))
)
```

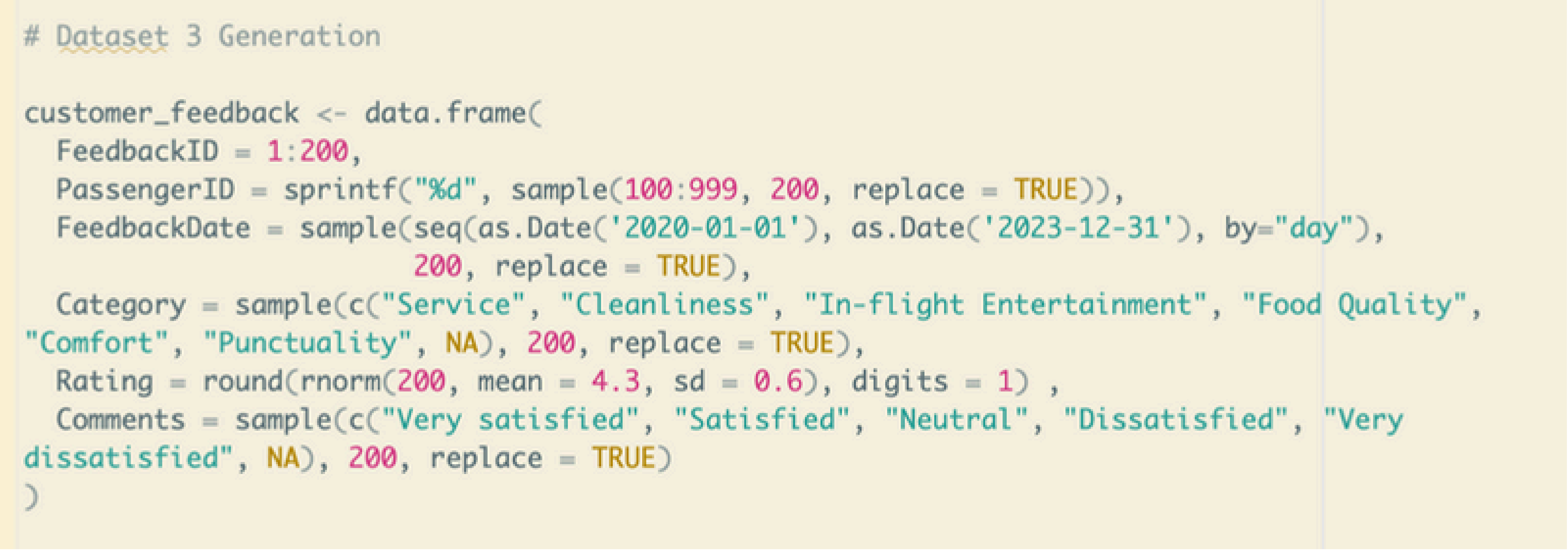


404	PassengerID	Age	Gender	Nationality	FlightNumber	TicketClass	error2	TotalFlightsTaken
1	244	26	Male	Chinese	QF854	Premium Economy	7.615124	14.11512
2	165	65	Female	American	QF498	Premium Economy	6.290985	22.54098
3	891	62	Female	American	QF404	Economy	7.664920	23.16492
4	703	62	Male	Chinese	QF400	Economy	8.341960	23.84196
5	429	26	Male	American	QF957	Economy	7.916763	14.41676
6	457	46	Male	British	QF675	Premium Economy	6.774106	18.27411

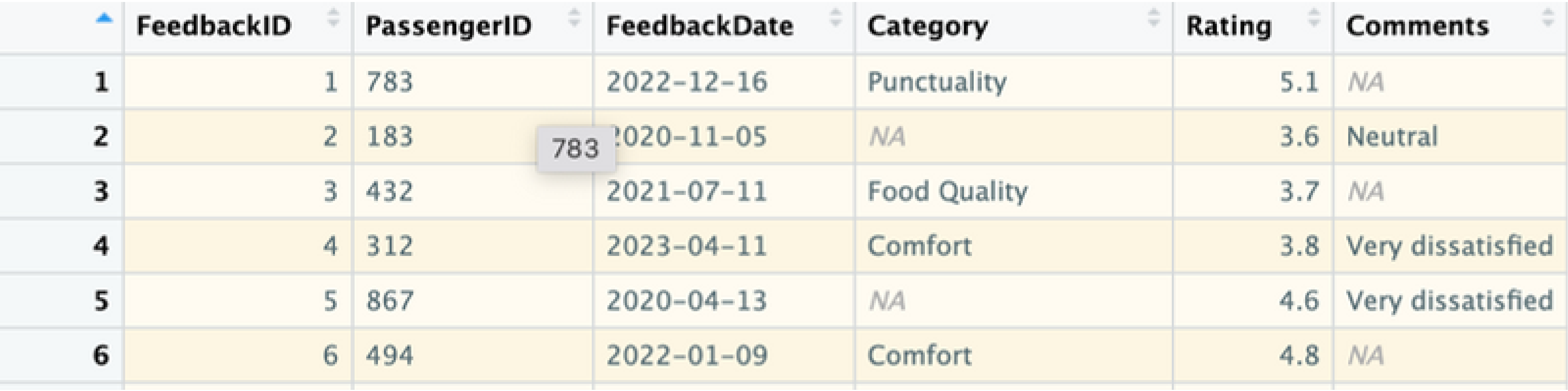
Synthetic Data Generation

We have generated 3 datasets for Qantas airlines : Airline , Passenger and **Customer Feedback datasets**

Customer Feedback Dataset



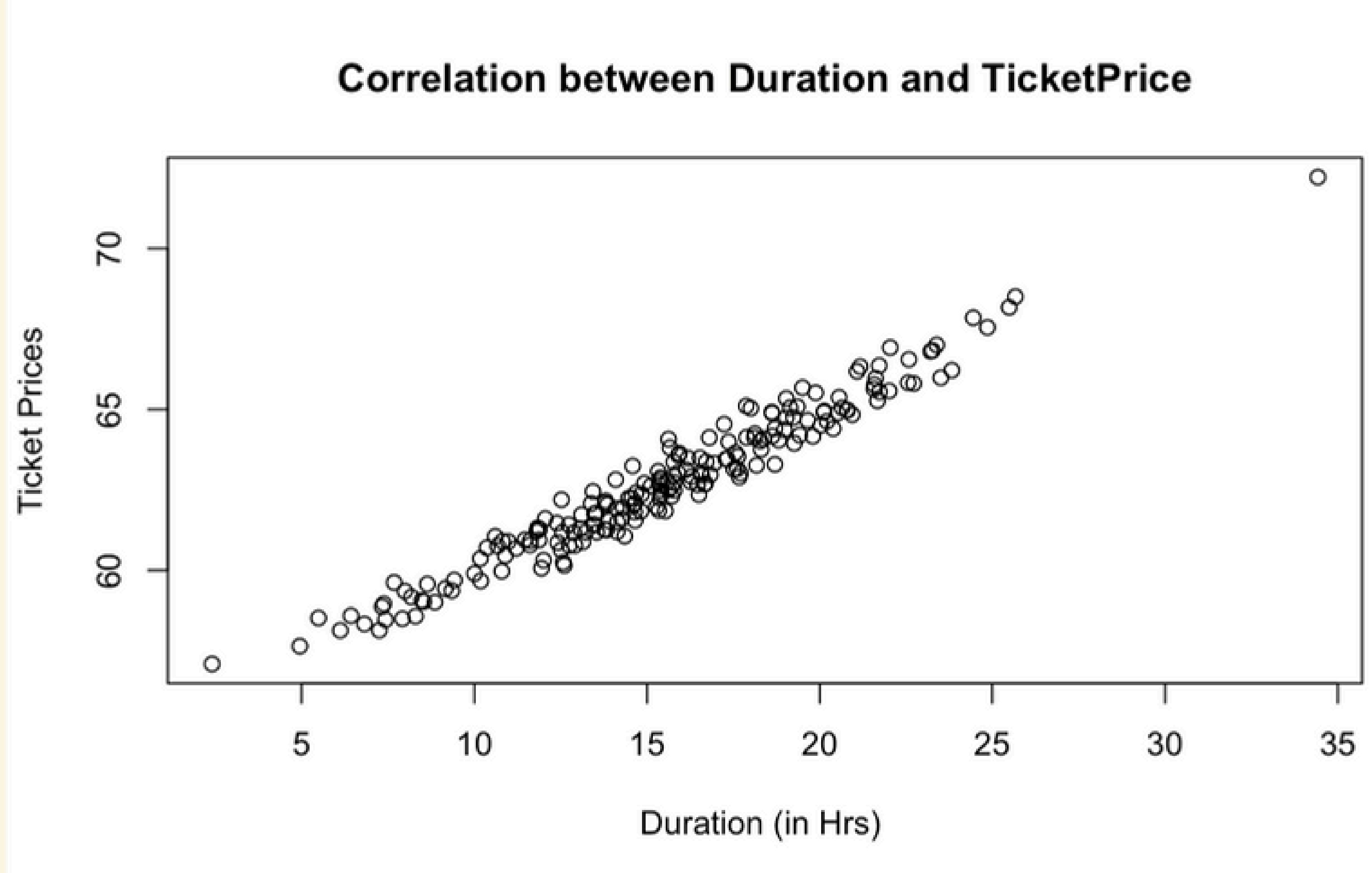
```
# Dataset 3 Generation

customer_feedback <- data.frame(
  FeedbackID = 1:200,
  PassengerID = sprintf("%d", sample(100:999, 200, replace = TRUE)),
  FeedbackDate = sample(seq(as.Date('2020-01-01'), as.Date('2023-12-31'), by="day"),
                      200, replace = TRUE),
  Category = sample(c("Service", "Cleanliness", "In-flight Entertainment", "Food Quality",
"Comfort", "Punctuality", NA), 200, replace = TRUE),
  Rating = round(rnorm(200, mean = 4.3, sd = 0.6), digits = 1) ,
  Comments = sample(c("Very satisfied", "Satisfied", "Neutral", "Dissatisfied", "Very
dissatisfied", NA), 200, replace = TRUE)
)
```



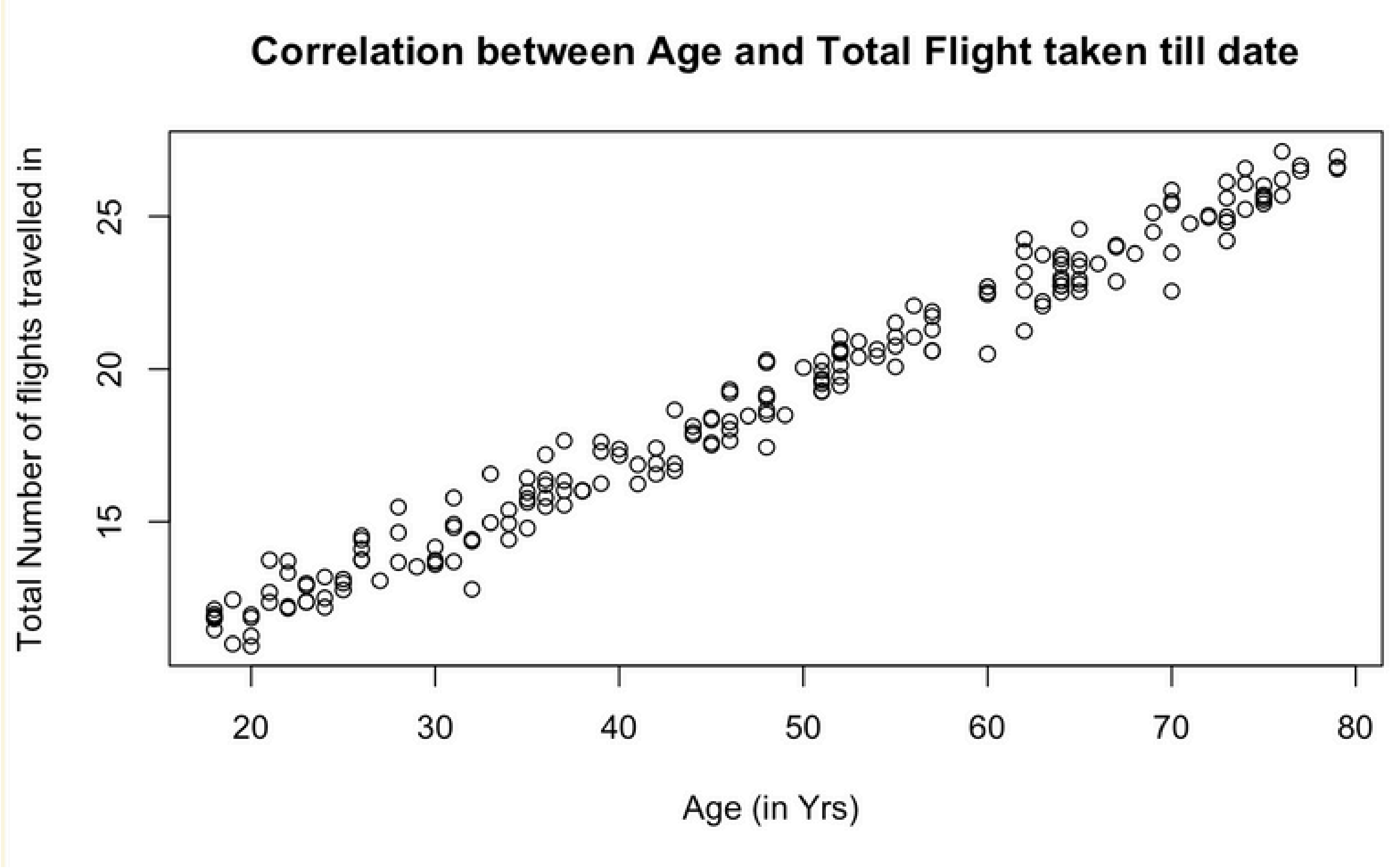
	FeedbackID	PassengerID	FeedbackDate	Category	Rating	Comments
1	1	783	2022-12-16	Punctuality	5.1	NA
2	2	183	2020-11-05	NA	3.6	Neutral
3	3	432	2021-07-11	Food Quality	3.7	NA
4	4	312	2023-04-11	Comfort	3.8	Very dissatisfied
5	5	867	2020-04-13	NA	4.6	Very dissatisfied
6	6	494	2022-01-09	Comfort	4.8	NA

Correlation variable

A correlation between variables : Duration of flight and Ticket prices has been given. The **longer the duration of flight the higher will be the prices of flight**.



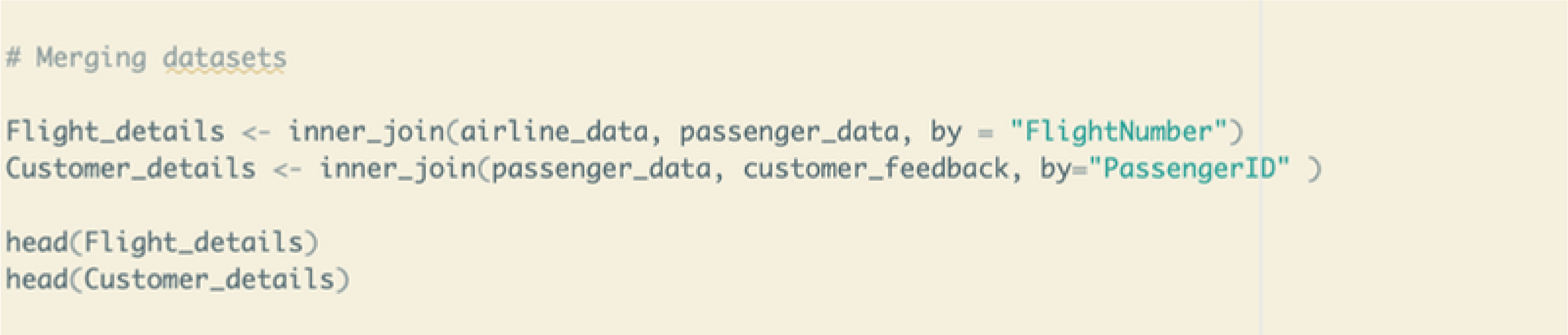
Age and the total number of flights travelled in till date have been assigned a **positive relationship** younger people tend to travel less, with age the number of flght trips increase



Merging datasets

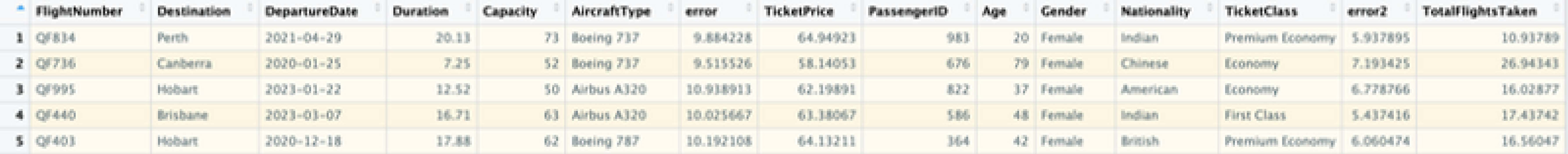
In our analysis, we have merged the following:

1. Airline and Passenger dataset on the variable “**Flight Number** “
2. Passenger and Customer feedback dataset on “**PassengerID**”



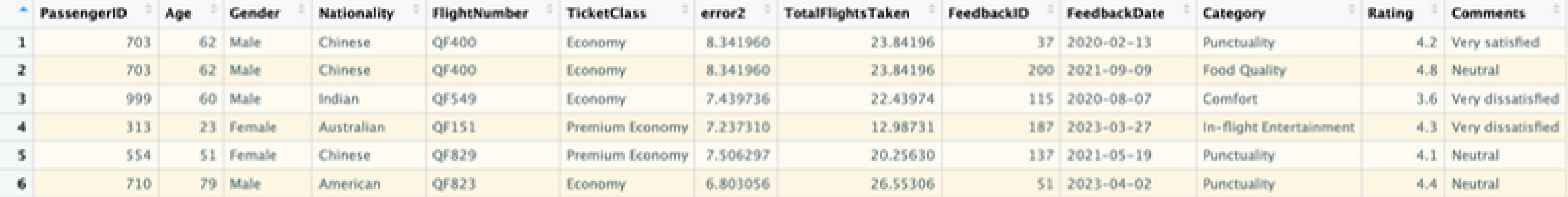
```
# Merging datasets

Flight_details <- inner_join(airline_data, passenger_data, by = "FlightNumber")
Customer_details <- inner_join(passenger_data, customer_feedback, by="PassengerID" )

head(Flight_details)
head(Customer_details)
```

Flight_Details



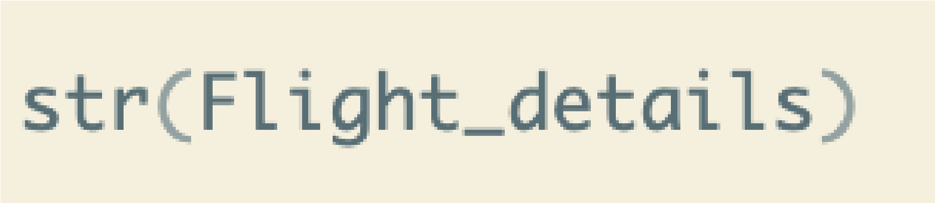
	FlightNumber	Destination	DepartureDate	Duration	Capacity	AircraftType	error	TicketPrice	PassengerID	Age	Gender	Nationality	TicketClass	error2	TotalFlightsTaken
1	QF834	Perth	2021-04-29	20.13	73	Boeing 737	9.884228	64.94923	983	20	Female	Indian	Premium Economy	5.937895	10.93789
2	QF736	Canberra	2020-01-25	7.25	52	Boeing 737	9.515526	58.14053	676	79	Female	Chinese	Economy	7.193425	26.94343
3	QF995	Hobart	2023-01-22	12.52	50	Airbus A320	10.938913	62.19891	822	37	Female	American	Economy	6.778766	16.02877
4	QF440	Brisbane	2023-03-07	16.71	63	Airbus A320	10.025667	63.38067	586	48	Female	Indian	First Class	5.437416	17.43742
5	QF403	Hobart	2020-12-18	17.88	62	Boeing 787	10.192108	64.13211	364	42	Female	British	Premium Economy	6.060474	16.56047

Customer_Details

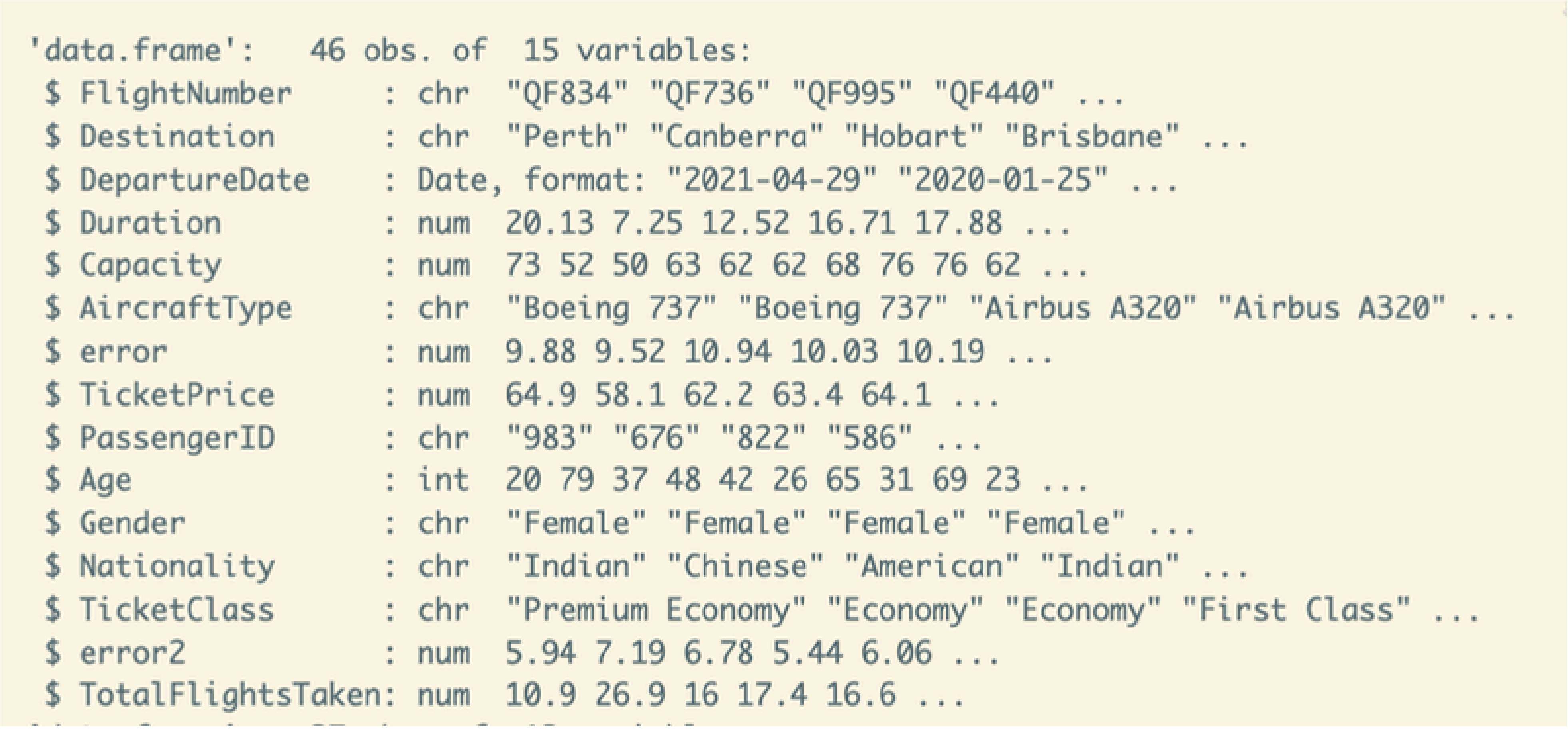


	PassengerID	Age	Gender	Nationality	FlightNumber	TicketClass	error2	TotalFlightsTaken	FeedbackID	FeedbackDate	Category	Rating	Comments
1	703	62	Male	Chinese	QF400	Economy	8.341960	23.84196	37	2020-02-13	Punctuality	4.2	Very satisfied
2	703	62	Male	Chinese	QF400	Economy	8.341960	23.84196	200	2021-09-09	Food Quality	4.8	Neutral
3	999	60	Male	Indian	QF549	Economy	7.439736	22.43974	115	2020-08-07	Comfort	3.6	Very dissatisfied
4	313	23	Female	Australian	QF151	Premium Economy	7.237310	12.98731	187	2023-03-27	In-flight Entertainment	4.3	Very dissatisfied
5	554	51	Female	Chinese	QF829	Premium Economy	7.506297	20.25630	137	2021-05-19	Punctuality	4.1	Neutral
6	710	79	Male	American	QF823	Economy	6.803056	26.55306	51	2023-04-02	Punctuality	4.4	Neutral

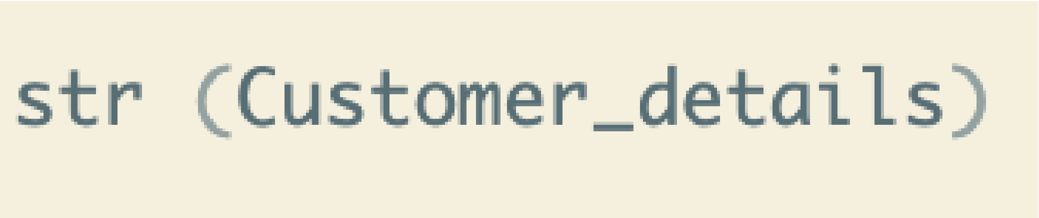
Checking structure of dataset



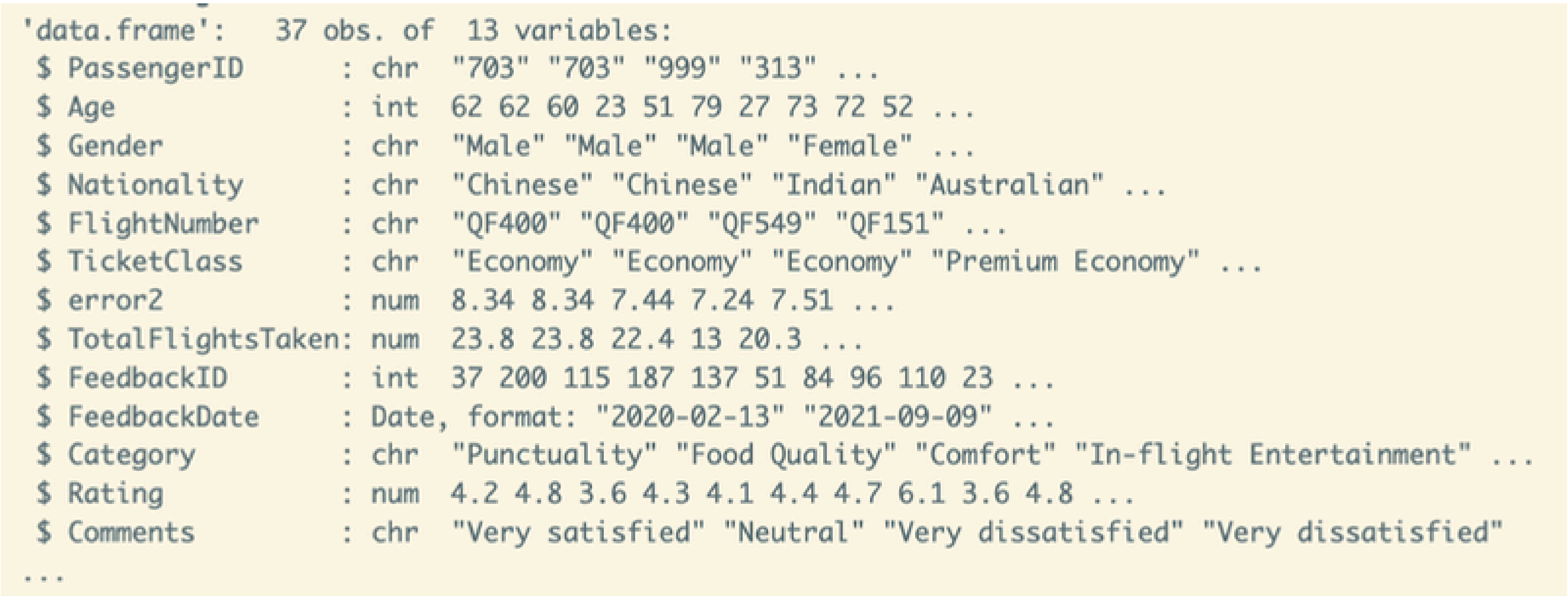
```
str(Flight_details)
```



```
'data.frame':	46 obs. of  15 variables:
 $ FlightNumber     : chr  "QF834" "QF736" "QF995" "QF440" ...
 $ Destination      : chr  "Perth" "Canberra" "Hobart" "Brisbane" ...
 $ DepartureDate    : Date, format: "2021-04-29" "2020-01-25" ...
 $ Duration         : num  20.13 7.25 12.52 16.71 17.88 ...
 $ Capacity         : num  73 52 50 63 62 62 68 76 76 62 ...
 $ AircraftType     : chr  "Boeing 737" "Boeing 737" "Airbus A320" "Airbus A320" ...
 $ error            : num  9.88 9.52 10.94 10.03 10.19 ...
 $ TicketPrice      : num  64.9 58.1 62.2 63.4 64.1 ...
 $ PassengerID      : chr  "983" "676" "822" "586" ...
 $ Age              : int  20 79 37 48 42 26 65 31 69 23 ...
 $ Gender           : chr  "Female" "Female" "Female" "Female" ...
 $ Nationality      : chr  "Indian" "Chinese" "American" "Indian" ...
 $ TicketClass      : chr  "Premium Economy" "Economy" "Economy" "First Class" ...
 $ error2           : num  5.94 7.19 6.78 5.44 6.06 ...
 $ TotalFlightsTaken: num  10.9 26.9 16 17.4 16.6 ...
```



```
str (Customer_details)
```



```
'data.frame':	37 obs. of  13 variables:
 $ PassengerID      : chr  "703" "703" "999" "313" ...
 $ Age              : int  62 62 60 23 51 79 27 73 72 52 ...
 $ Gender           : chr  "Male" "Male" "Male" "Female" ...
 $ Nationality      : chr  "Chinese" "Chinese" "Indian" "Australian" ...
 $ FlightNumber     : chr  "QF400" "QF400" "QF549" "QF151" ...
 $ TicketClass      : chr  "Economy" "Economy" "Economy" "Premium Economy" ...
 $ error2           : num  8.34 8.34 7.44 7.24 7.51 ...
 $ TotalFlightsTaken: num  23.8 23.8 22.4 13 20.3 ...
 $ FeedbackID       : int  37 200 115 187 137 51 84 96 110 23 ...
 $ FeedbackDate     : Date, format: "2020-02-13" "2021-09-09" ...
 $ Category         : chr  "Punctuality" "Food Quality" "Comfort" "In-flight Entertainment" ...
 $ Rating           : num  4.2 4.8 3.6 4.3 4.1 4.4 4.7 6.1 3.6 4.8 ...
 $ Comments         : chr  "Very satisfied" "Neutral" "Very dissatisfied" "Very dissatisfied"
...
```

Data Conversion

based on structure of combined datasets

Flight Details Dataset

Passenger ID

Numeric to be converted Character

Ticket Class

Should be ordered factor as it contains categorical values with ranks

Customer Details Dataset

Passenger ID

Numeric to be converted Character

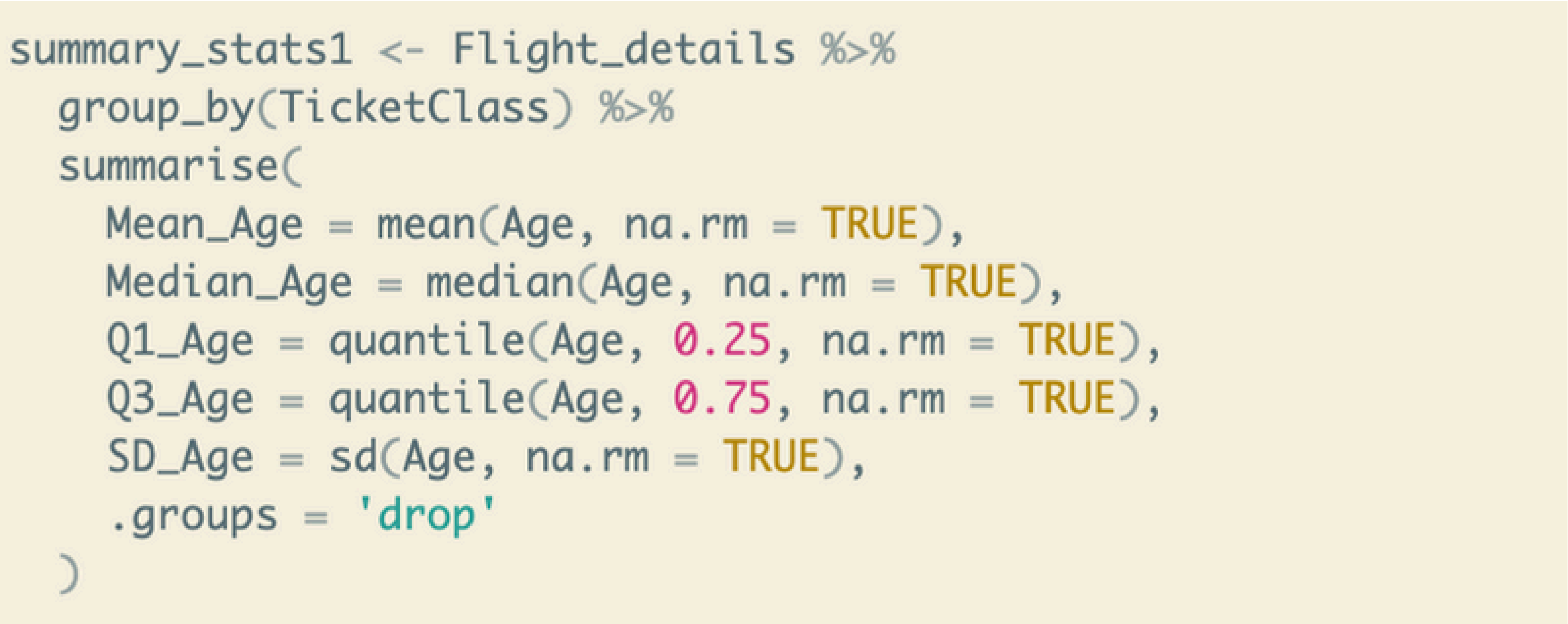
Ticket Class

Should be ordered factor as it contains categorical values with ranks

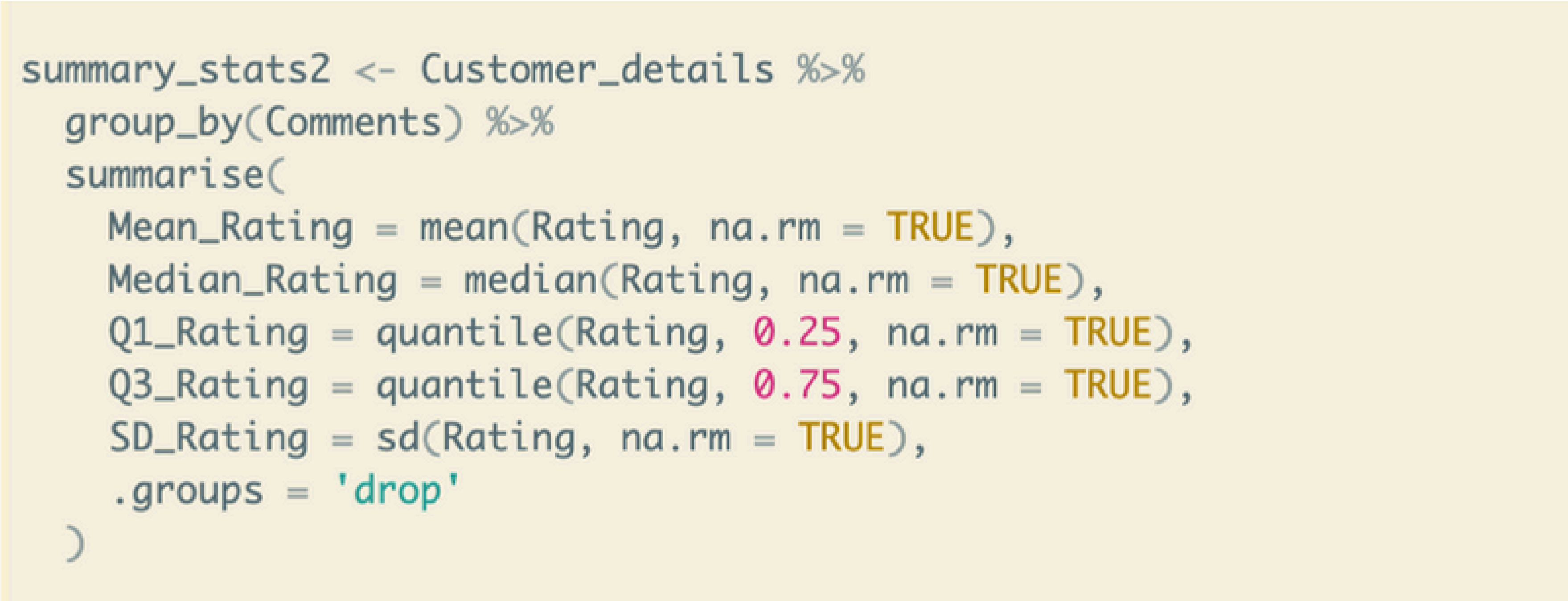
Comments

Should be ordered factor as it contains categorical values with ranks

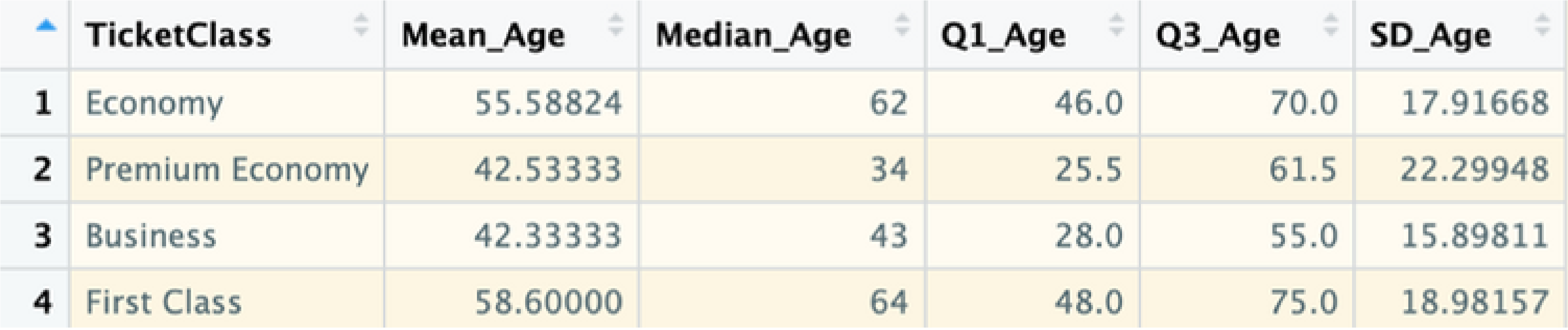
Summary Statistics



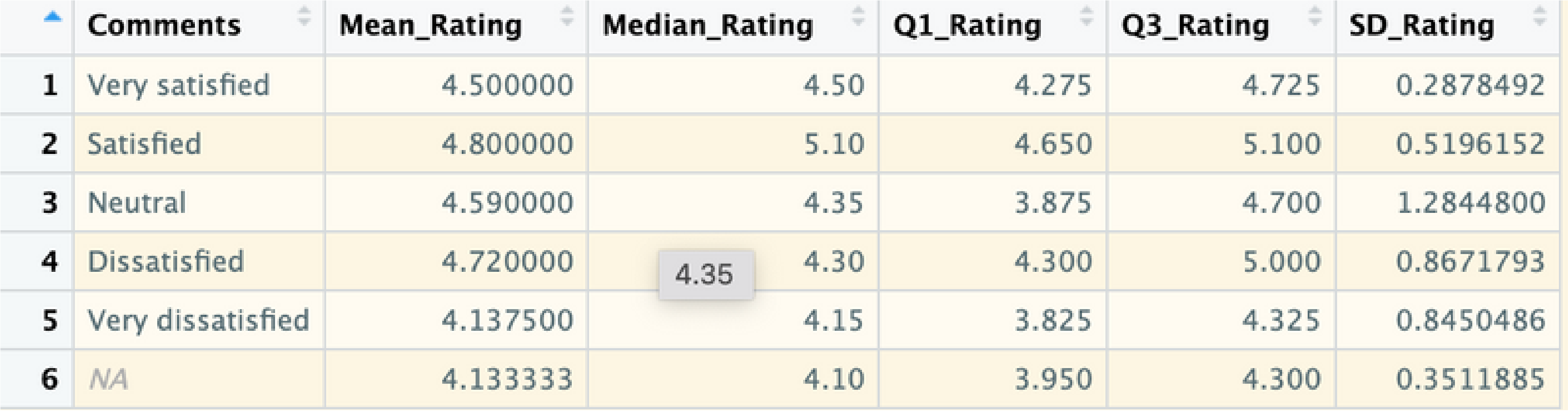
```
summary_stats1 <- Flight_details %>%
  group_by(TicketClass) %>%
  summarise(
    Mean_Age = mean(Age, na.rm = TRUE),
    Median_Age = median(Age, na.rm = TRUE),
    Q1_Age = quantile(Age, 0.25, na.rm = TRUE),
    Q3_Age = quantile(Age, 0.75, na.rm = TRUE),
    SD_Age = sd(Age, na.rm = TRUE),
    .groups = 'drop'
  )
```



```
summary_stats2 <- Customer_details %>%
  group_by(Comments) %>%
  summarise(
    Mean_Rating = mean(Rating, na.rm = TRUE),
    Median_Rating = median(Rating, na.rm = TRUE),
    Q1_Rating = quantile(Rating, 0.25, na.rm = TRUE),
    Q3_Rating = quantile(Rating, 0.75, na.rm = TRUE),
    SD_Rating = sd(Rating, na.rm = TRUE),
    .groups = 'drop'
  )
```



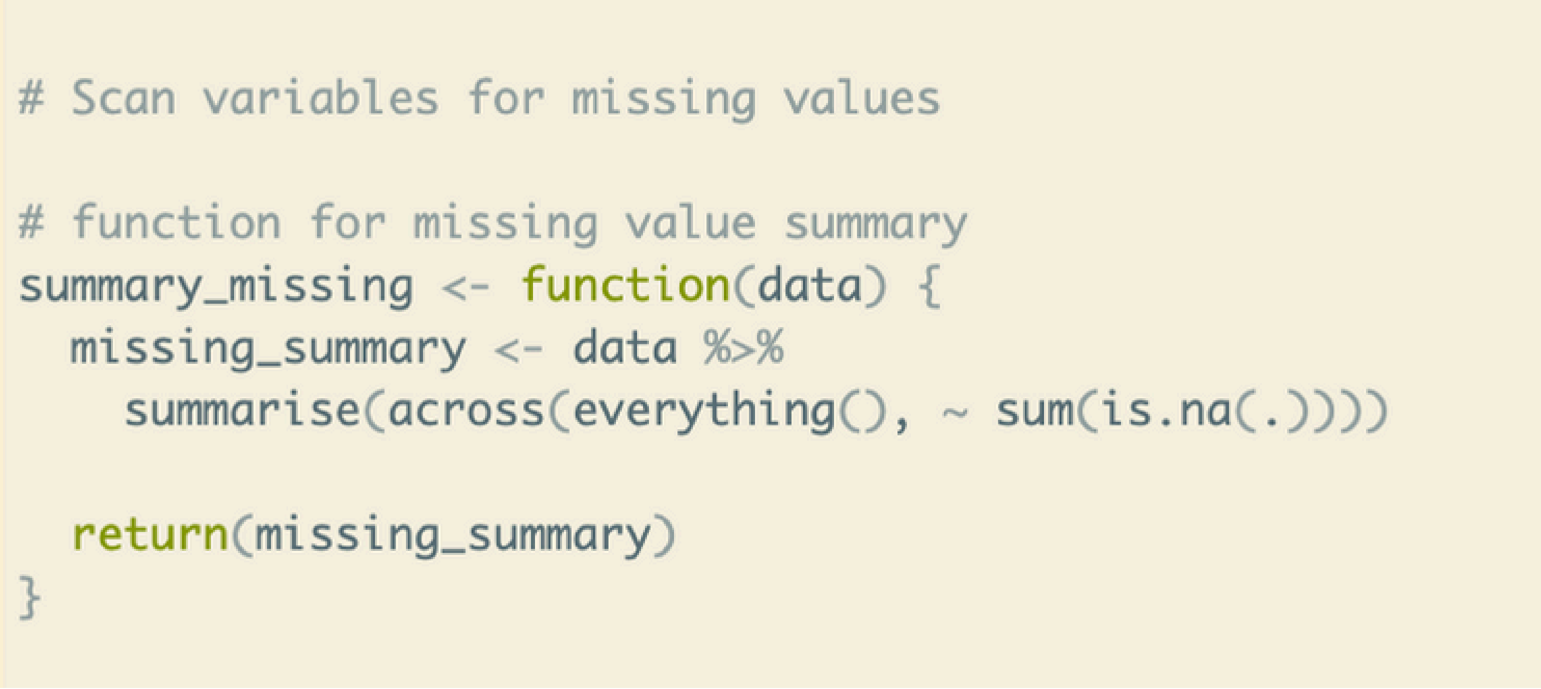
	TicketClass	Mean_Age	Median_Age	Q1_Age	Q3_Age	SD_Age
1	Economy	55.58824	62	46.0	70.0	17.91668
2	Premium Economy	42.53333	34	25.5	61.5	22.29948
3	Business	42.33333	43	28.0	55.0	15.89811
4	First Class	58.60000	64	48.0	75.0	18.98157



	Comments	Mean_Rating	Median_Rating	Q1_Rating	Q3_Rating	SD_Rating
1	Very satisfied	4.500000	4.50	4.275	4.725	0.2878492
2	Satisfied	4.800000	5.10	4.650	5.100	0.5196152
3	Neutral	4.590000	4.35	3.875	4.700	1.2844800
4	Dissatisfied	4.720000	4.30	4.300	5.000	0.8671793
5	Very dissatisfied	4.137500	4.15	3.825	4.325	0.8450486
6	NA	4.133333	4.10	3.950	4.300	0.3511885

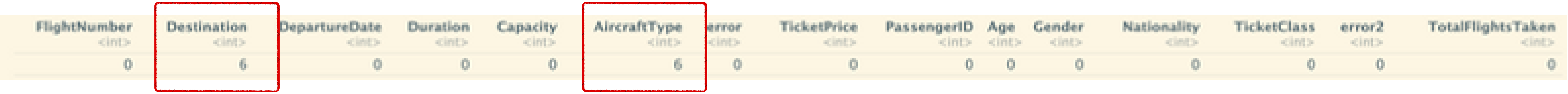
Scanning missing values

Function for scanning total missing values in dataframe



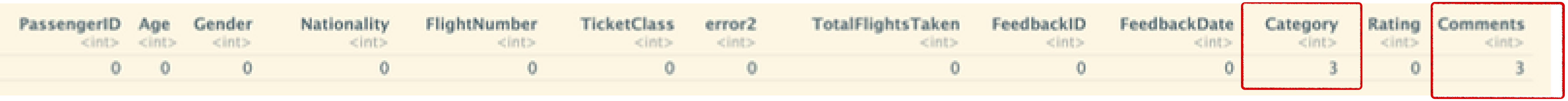
```
# Scan variables for missing values

# function for missing value summary
summary_missing <- function(data) {
  missing_summary <- data %>%
    summarise(across(everything(), ~ sum(is.na(.))))

  return(missing_summary)
}
```

Missing values in Flight Details



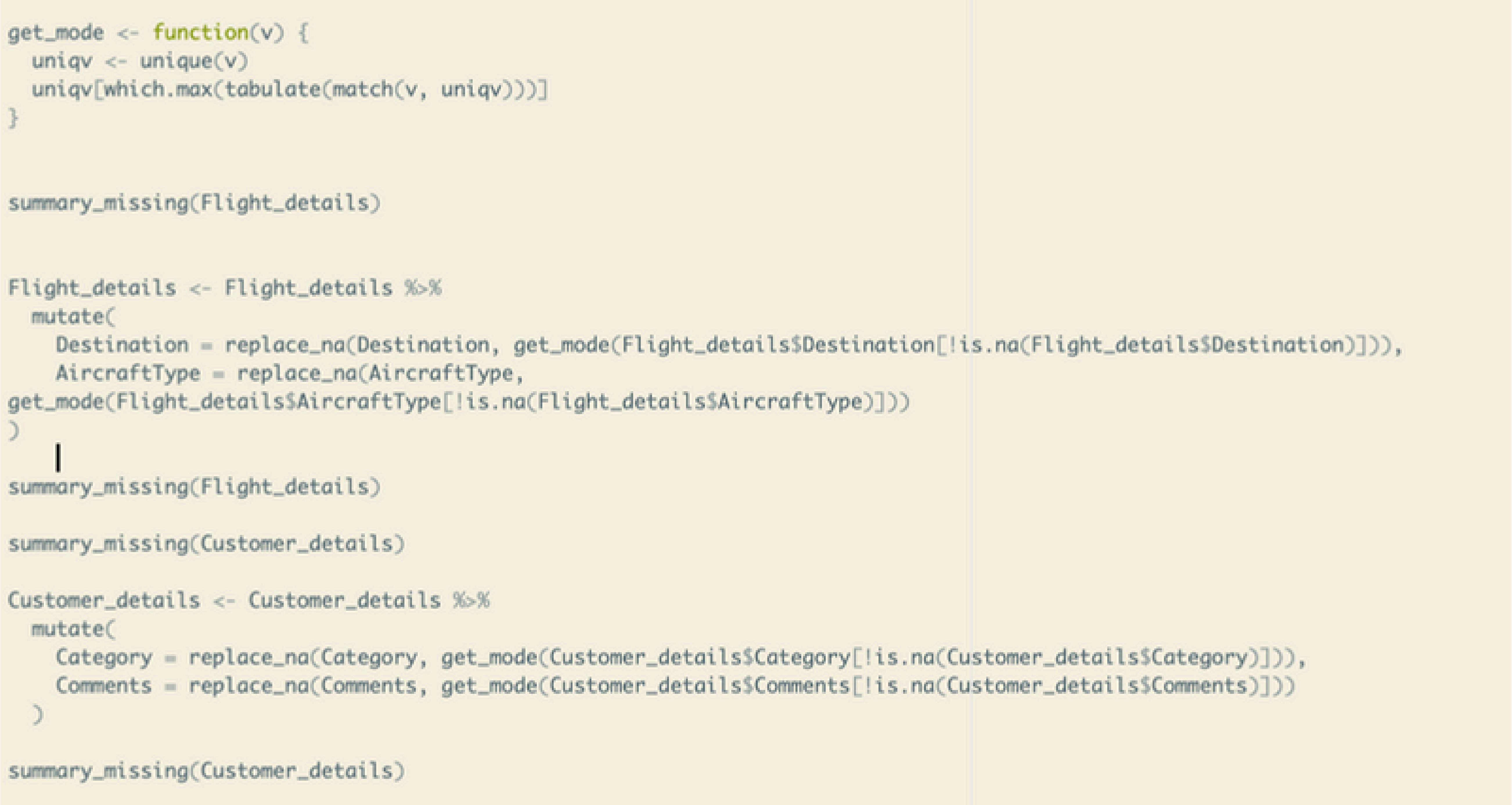
FlightNumber	Destination	DepartureDate	Duration	Capacity	AircraftType	error	TicketPrice	PassengerID	Age	Gender	Nationality	TicketClass	error2	TotalFlightsTaken
<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
0	6	0	0	0	6	0	0	0	0	0	0	0	0	0

Missing values in Customer Details



PassengerID	Age	Gender	Nationality	FlightNumber	TicketClass	error2	TotalFlightsTaken	FeedbackID	FeedbackDate	Category	Rating	Comments
<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
0	0	0	0	0	0	0	0	0	0	3	0	3

Imputing missing values



```
get_mode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}


summary_missing(Flight_details)


Flight_details <- Flight_details %>%
  mutate(
    Destination = replace_na(Destination, get_mode(Flight_details$Destination[!is.na(Flight_details$Destination)])),
    AircraftType = replace_na(AircraftType,
get_mode(Flight_details$AircraftType[!is.na(Flight_details$AircraftType)]))
)

summary_missing(Flight_details)

summary_missing(Customer_details)

Customer_details <- Customer_details %>%
  mutate(
    Category = replace_na(Category, get_mode(Customer_details$Category[!is.na(Customer_details$Category)])),
    Comments = replace_na(Comments, get_mode(Customer_details$Comments[!is.na(Customer_details$Comments)]))
  )

summary_missing(Customer_details)
```

Since out dataset only consisted of missing categorical values, we impute the missing data using the **Mode method**. The Mode method is basically replacing the missing values by the **most frequently occuring value** of the dataset.