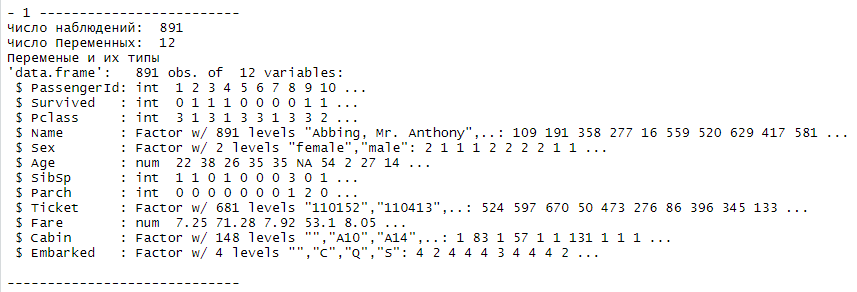
# **Задания**

## **Задача 1:**

1. Сколько в базе данных наблюдений? Сколько переменных? Какие это переменные? Какого типа?
2. Сколько в базе данных строк, которые не содержат пропущенных значений? Сохраните строки, содержащие пропущенные значения, в отдельную базу данных df\_na.
3. Постройте график, который показывал бы частоту, с которой встречаются пропущенные значения в каждой из переменных базы данных. В какой переменной больше всего пропущенных значений?
4. Постройте график, который позволит определить паттерны пропущенных значений. Можно ли по полученным результатам сделать вывод о том, что значения в базе пропущены “системно” (часто нет ответов на определенный вопрос или вопросы)? Может ли это быть связано со спецификой самих вопросов?
5. Удалите в базе данных пропущенные значения.

### **Работа программы:**

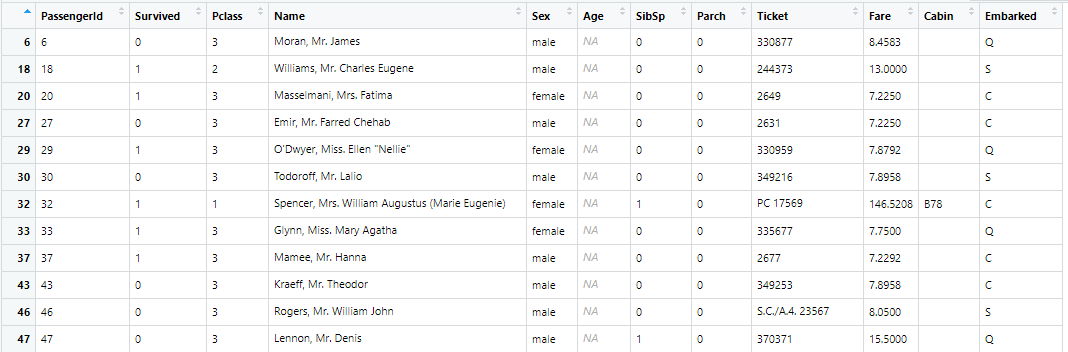
1



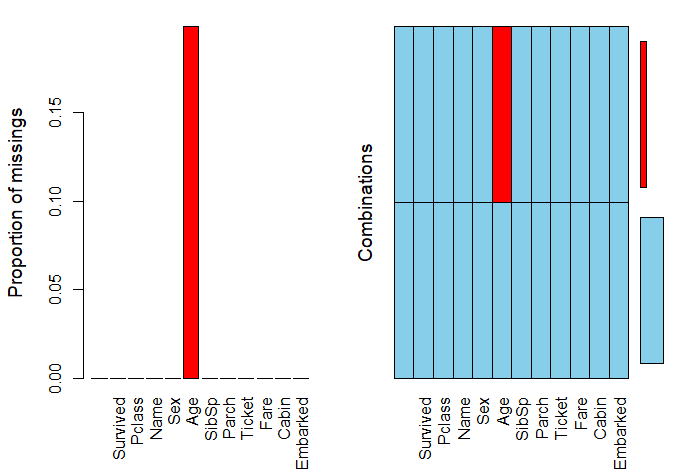
2



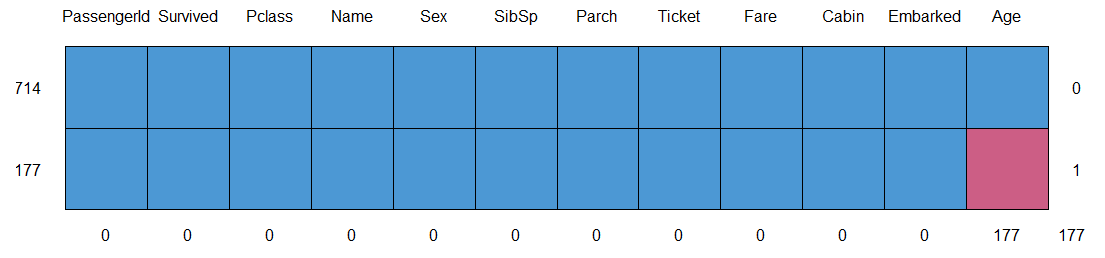




3 Больше всего пропущенно в переменной age

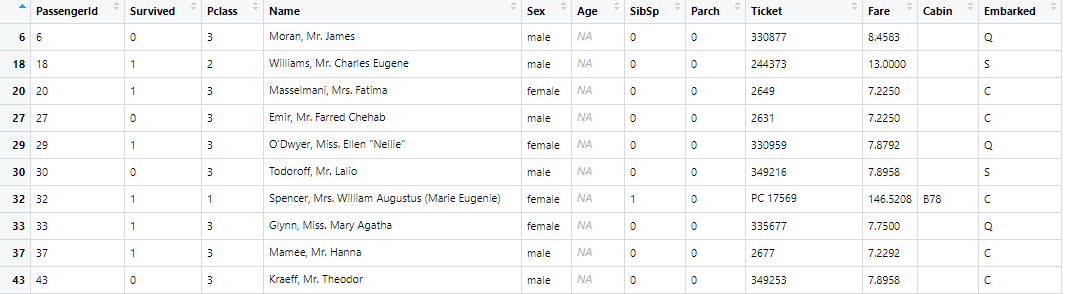


4 Если данная таблица действительно ссылается на реальную статистику то возможно билеты с пропущенным возрастом покупались у перекупщика или разыгрывались из-за чего не указан возраст

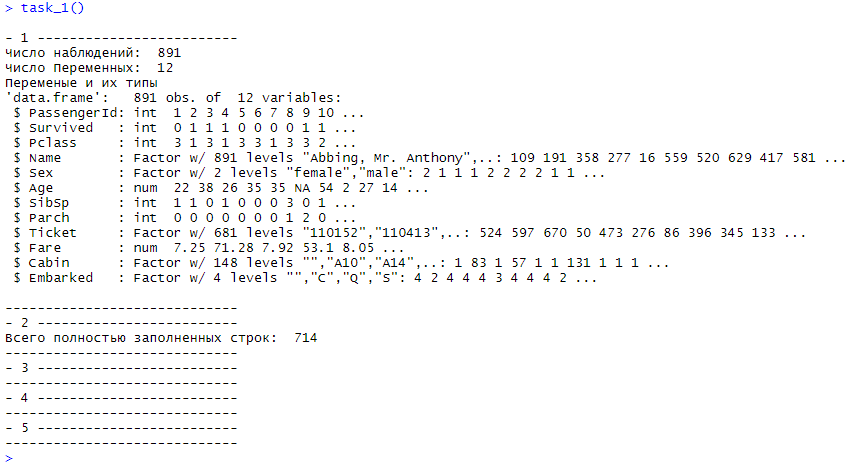


5





Все в одном



### **Листинг:**

library(mice)

library(VIM)

cat("\n- 1 -------------------------")

df <- read.csv("https://raw.githubusercontent.com/agconti/kaggle-titanic/master/data/train.csv")

cat("\nЧисло наблюдений: ", nrow(df))

cat("\nЧисло Переменных: ", ncol(df))

cat("\nПеременые и их типы\n")

str(df)

cat("\n-----------------------------")

cat("\n- 2 -------------------------")

cat("\nВсего полностью заполненных строк: ", sum(complete.cases(df)))

df\_na <- df[!complete.cases(df),]

cat("\n-----------------------------")

cat("\n- 3 -------------------------")

aggr(df)

cat("\n-----------------------------")

cat("\n- 4 -------------------------")

md.pattern(df)

cat("\n-----------------------------")

cat("\n- 5 -------------------------")

df <- na.omit(df)

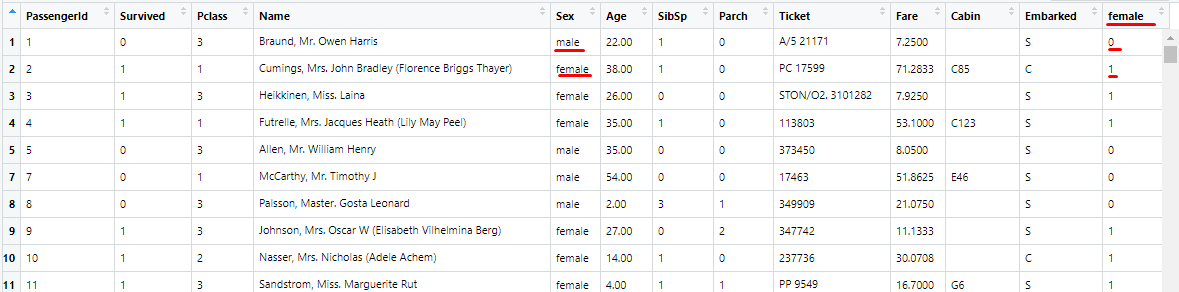
cat("\n-----------------------------")

## **Задача 2:**

1. Добавьте в базу данных бинарную переменную female, где значение 0 соответствует пассажирам мужского пола, а 1 - пассажирам женского пола. Не забудьте: бинарная переменная - всегда числовая (целочисленная). Готовую переменную SexCode использовать нельзя.
2. Представьте, что в исследовании нас интересуют пассажиры старше 25 лет и не старше 45 лет, которые путешествовали вторым или третьим классом. Сохраните соответствующие строки в базу данных df2.
3. Сколько на “Титанике” (согласно базе данных df) было пассажиров мужского пола? Женского пола?
4. Сколько лет было самому молодому пассажиру среди выживших? А самому старому? Каков средний возраст пассажиров первого класса, которые выжили в катастрофе?

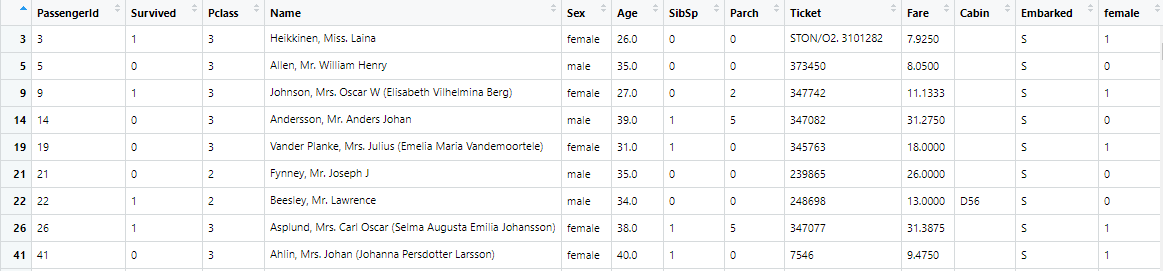
### **Работа программы:**

1

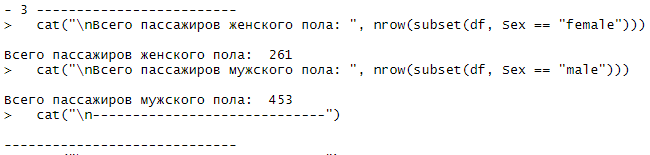


2

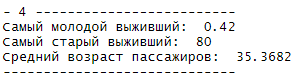




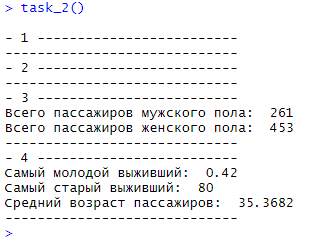
3



4



Все в одном



### **Листинг:**

cat("\n- 1 -------------------------")

df$female <- as.integer(ifelse(df$Sex == "female", 1, 0) )

cat("\n-----------------------------")

cat("\n- 2 -------------------------")

df2 <- subset(df, Age > 25 & Age < 45 & (Pclass == 2 | Pclass == 3))

cat("\n-----------------------------")

cat("\n- 3 -------------------------")

cat("\nВсего пассажиров мужского пола: ", nrow(subset(df, Sex == "female")))

cat("\nВсего пассажиров женского пола: ", nrow(subset(df, Sex == "male")))

cat("\n-----------------------------")

cat("\n- 4 -------------------------")

cat("\nСамый молодой выживший: ", min(subset(df, Survived == 1)$Age))

cat("\nСамый старый выживший: ", max(subset(df, Survived == 1)$Age))

cat("\nСредний возраст пассажиров: ", mean(subset(df, Survived == 1 & Pclass == 1)$Age))

cat("\n-----------------------------")