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

```
- 1 -------------------------
Число наблюдений:  891
Число переменных:  12
Переменые и их типы
'data.frame':   891 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...
 $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
 $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...

----------------------------
```

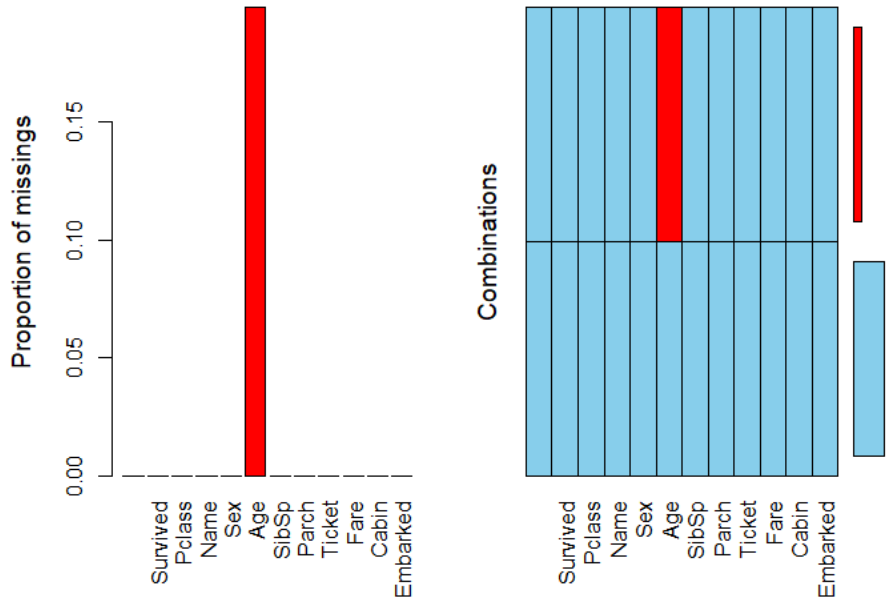
2

```
- 2 -------------------------
Всего полностью заполненных строк:  714
-----------------------------
```

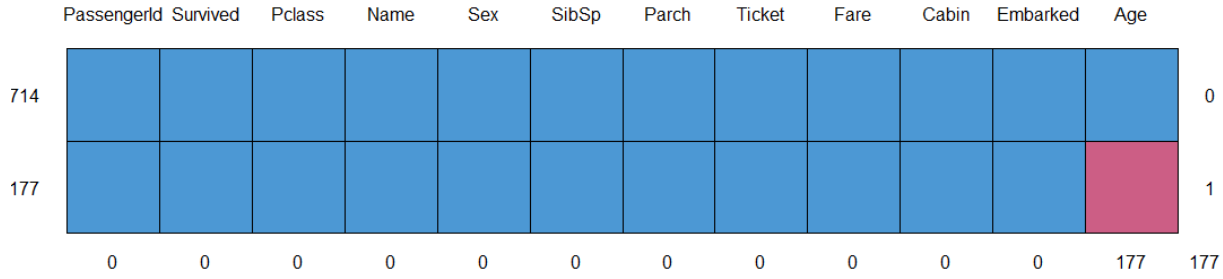
df_na 177 obs. of 12 variables

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
6	6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583		Q
18	18	1	2	Williams, Mr. Charles Eugene	male	NA	0	0	244373	13.0000		S
20	20	1	3	Masselmani, Mrs. Fatima	female	NA	0	0	2649	7.2250		C
27	27	0	3	Emir, Mr. Farred Chehab	male	NA	0	0	2631	7.2250		C
29	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	NA	0	0	330959	7.8792		Q
30	30	0	3	Todoroff, Mr. Lalio	male	NA	0	0	349216	7.8958		S
32	32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	NA	1	0	PC 17569	146.5208	B78	C
33	33	1	3	Glynn, Miss. Mary Agatha	female	NA	0	0	335677	7.7500		Q
37	37	1	3	Mamee, Mr. Hanna	male	NA	0	0	2677	7.2292		C
43	43	0	3	Kraeff, Mr. Theodor	male	NA	0	0	349253	7.8958		C
46	46	0	3	Rogers, Mr. William John	male	NA	0	0	S.C./A.4. 23567	8.0500		S
47	47	0	3	Lennon, Mr. Denis	male	NA	1	0	370371	15.5000		Q

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

df 714 obs. of 12 variables

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
6	6	0	3	Moran, Mr. James	male	NA	0	0	330877	8.4583		Q
18	18	1	2	Williams, Mr. Charles Eugene	male	NA	0	0	244373	13.0000		S
20	20	1	3	Masselmani, Mrs. Fatima	female	NA	0	0	2649	7.2250		C
27	27	0	3	Emir, Mr. Farred Chehab	male	NA	0	0	2631	7.2250		C
29	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	female	NA	0	0	330959	7.8792		Q
30	30	0	3	Todoroff, Mr. Lalio	male	NA	0	0	349216	7.8958		S
32	32	1	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female	NA	1	0	PC 17569	146.5208	B78	C
33	33	1	3	Glynn, Miss. Mary Agatha	female	NA	0	0	335677	7.7500		Q
37	37	1	3	Mamee, Mr. Hanna	male	NA	0	0	2677	7.2292		C
43	43	0	3	Kraeff, Mr. Theodor	male	NA	0	0	349253	7.8958		C

Все в одном

```
> task_1()

- 1 -------------------------
Число наблюдений:  891
Число Переменных:  12
Переменые и их типы
'data.frame':   891 obs. of  12 variables:
 $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
 $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
 $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
 $ Name       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 581 ...
 $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
 $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
 $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
 $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
 $ Ticket     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133 ...
 $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
 $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
 $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...

-----------------------------
- 2 -------------------------
Всего полностью заполненных строк:  714
-----------------------------
- 3 -------------------------
-----------------------------
- 4 -------------------------
-----------------------------
- 5 -------------------------
-----------------------------
>
```

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

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	female
1	1	0	3	Braund, Mr. Owen Harris	male	22.00	1	0	A/5 21171	7.2500		S	0
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.00	1	0	PC 17599	71.2833	C85	C	1
3	3	1	3	Heikkinen, Miss. Laina	female	26.00	0	0	STON/O2. 3101282	7.9250		S	1
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.00	1	0	113803	53.1000	C123	S	1
5	5	0	3	Allen, Mr. William Henry	male	35.00	0	0	373450	8.0500		S	0
7	7	0	1	McCarthy, Mr. Timothy J	male	54.00	0	0	17463	51.8625	E46	S	0
8	8	0	3	Palsson, Master. Gosta Leonard	male	2.00	3	1	349909	21.0750		S	0
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.00	0	2	347742	11.1333		S	1
10	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.00	1	0	237736	30.0708		C	1
11	11	1	3	Sandstrom, Miss. Marguerite Rut	female	4.00	1	1	PP 9549	16.7000	G6	S	1

2

df2 221 obs. of 13 variables

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	female
3	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250		S	1
5	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500		S	0
9	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333		S	1
14	14	0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.2750		S	0
19	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31.0	1	0	345763	18.0000		S	1
21	21	0	2	Fynney, Mr. Joseph J	male	35.0	0	0	239865	26.0000		S	0
22	22	1	2	Beesley, Mr. Lawrence	male	34.0	0	0	248698	13.0000	D56	S	0
26	26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)	female	38.0	1	5	347077	31.3875		S	1
41	41	0	3	Ahlin, Mrs. Johan (Johanna Persdotter Larsson)	female	40.0	1	0	7546	9.4750		S	1

3

```
- 3 -------------------------
>   cat("\nВсего пассажиров женского пола: ", nrow(subset(df, Sex == "female")))

Всего пассажиров женского пола:  261
>   cat("\nВсего пассажиров мужского пола: ", nrow(subset(df, Sex == "male")))

Всего пассажиров мужского пола:  453
>   cat("\n-----------------------------")

-----------------------------
```

4

```
- 4 -------------------------
Самый молодой выживший:  0.42
Самый старый выживший:  80
Средний возраст пассажиров:  35.3682
-----------------------------
```

Все в одном

```
> task_2()

- 1 -------------------------
-----------------------------
- 2 -------------------------
-----------------------------
- 3 -------------------------
Всего пассажиров мужского пола:  261
Всего пассажиров женского пола:  453
-----------------------------
- 4 -------------------------
Самый молодой выживший:  0.42
Самый старый выживший:  80
Средний возраст пассажиров:  35.3682
-----------------------------
>
```

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-----------------------------")
```