

Анализ SQL-запросов для исследования выживаемости на примере датасета судна “Титаник”

by Павлова Мария

1. ДАТАСЕТ

1. Ссылка на датасет: <https://www.kaggle.com/code/alexisbcook/titanic-tutorial>
2. Общее количество строк - 887.
3. Атрибуты:
 - ❑ survived (указаны значения "1" - если пассажир выжил, "0" - если не выжил);
 - ❑ pclass (класс обслуживания пассажира);
 - ❑ name (имя пассажира);
 - ❑ sex (пол пассажира);
 - ❑ age (возраст пассажира);
 - ❑ siblings_spouses_aboard (количество сестер/братьев/супругов на борту);
 - ❑ parents_children_aboard (количество родителей/детей на борту);
 - ❑ fare (стоимость проезда).
4. адрес расположения проекта: <https://github.com/Maria-Pav1987/titanic>

2. ПОСТАНОВКА ЗАДАЧ

1. Выяснить средний процент выживших.
2. Выявить зависимость выживания, в зависимости от пола и класса обслуживания.
3. Выяснить, как возраст пассажиров повлиял на их выживаемость.
4. Выявить закономерности выживания пассажиров, в зависимости от наличия или отсутствия семьи.
 - 4.1. Выявить, как повлиял на выживаемость размер семьи.
 - 4.2. Выявить, какая именно категория членов семьи влияла больше на выживаемость.

3. ПЕРВИЧНАЯ ОБРАБОТКА ДАННЫХ

```
1  #импорт нужных библиотек
2
3  import sqlite3
4  import pandas as pd
5
6  #подключение к бд
7  con = sqlite3.connect(database: 'C:/Users/User/Desktop/project_sql_Pavlova', timeout=10)
8  cur = con.cursor()
9
10 #подготовка таблицы с данными в формате pandas dataframe
11
12 df = pd.read_csv('C:/Users/User/Desktop/titanic.csv')
13
14 #проверка датасета
15 df.info()
16 #вся числовая информация (int, float) представлена в необходимых типах данных.
17
```

4. ПЕРВИЧНАЯ ОБРАБОТКА ДАННЫХ. ЗАГРУЗКА ДАННЫХ

```
19 #поиск пропусков
20 print(df.isnull().mean() * 100)
21 #вся числовая информация (int, float) представлена полностью, пропуски отсутствуют.
22
23 #поиск дубликатов
24 doubles = df[df.duplicated()]
25 print(doubles)
26
27 # naming_convention
28 df.columns = df.columns.str.replace(' ', '_')
29 df.columns = df.columns.str.replace('/', '_')
30 df.columns = df.columns.str.lower()
31
32 #загрузка таблицы в бд
33 df.to_sql(con=con, name='titanic_table', if_exists='replace', index=False)
34
35 #считывание данных из таблицы
36 data_test = cur.execute('select * from titanic_table')
37 con.commit()
38 cur.fetchall()
```

5.1. Исследование данных с помощью SQL

Первая задача: выяснить средний процент выживших (необходим для дальнейших сравнений отклонений выживаемости в зависимости от факторов).

КОД	РЕЗУЛЬТАТ								
<pre>SELECT SUM(survived) AS surv_passenger, COUNT(*) AS total_passenger, SUM(survived) * 100 / COUNT(*) AS surv_percent FROM titanic_table tt</pre> <p>-- Выжившие обозначены в атрибуте "survived" как 1, погибшие как 0. Соответственно, сумма выживших - сумма всех значений атрибута "survived".</p>	<div>Результат 1 ✕</div> <div>SELECT SUM(survived) AS surv_passenger, COUNT(*) AS total_passenger ↕ ↕ Везде SQL</div> <table><tr><th></th><th>123 surv_passenger</th><th>123 total_passenger</th><th>123 surv_percent</th></tr><tr><td>1</td><td>342</td><td>887</td><td>38</td></tr></table>		123 surv_passenger	123 total_passenger	123 surv_percent	1	342	887	38
	123 surv_passenger	123 total_passenger	123 surv_percent						
1	342	887	38						
<p>Вывод: получили базовый средний показатель выживаемости всех пассажиров (38%). Цель: для дальнейшего исследования отклонений от него, в зависимости от факторов.</p>									

5.2. Вторая задача: выявить зависимость выживания, в зависимости от пола и класса обслуживания.

КОД	РЕЗУЛЬТАТ	
<pre>SELECT sex, pclass, SUM(survived) * 100 / COUNT(*) AS surv_percent FROM titanic_table GROUP BY sex, pclass</pre>	ПО ПОЛУ:	ПО КЛАССУ:

ABC sex	123 surv_percent
female	74
male	19

123 pclass	123 surv_percent
1	62
2	47
3	24

Вывод: В основном выжили женщины, а не мужчины (74% против 19 %).

Кроме того, выжило больше пассажиров 1-го (более высокого класса обслуживания):

62% - 1 класс,

47% - 2 класс,

24% - 3 класс.

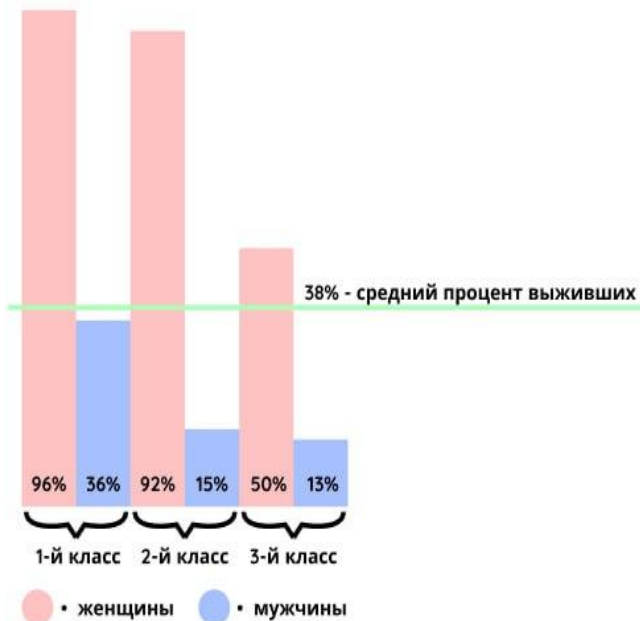
5.2. Вторая задача: выявить зависимость выживания, в зависимости от пола и класса обслуживания.

Визуализация (агрегированные данные по полу и классу):

```
' SELECT sex, pclass, SUM(survived) * 100 / COUNT(*) AS surv_percent
```

	ABC sex ▼	123 pclass ▼	123 surv_percent ▼
1	female	1	96
2	female	2	92
3	female	3	50
4	male	1	36
5	male	2	15
6	male	3	13

График количества выживших,
в зависимости от пола и класса обслуживания (в процентах)



5.3. Третья задача: выясним, как возраст пассажиров повлиял на их выживаемость.

КОД	РЕЗУЛЬТАТ																																																
<pre>SELECT sex, pclass, SUM(survived) * 100 / COUNT(*) AS surv_percent, CASE WHEN age BETWEEN 0 AND 12 THEN 'minor(0-12)' WHEN age BETWEEN 13 AND 19 THEN 'teenager(13-19)' WHEN age BETWEEN 20 AND 30 THEN 'young_man(20-30)' WHEN age BETWEEN 31 AND 50 THEN 'adult(31-50)' ELSE 'aged' END AS generation FROM titanic_table tt GROUP BY pclass, generation</pre>	<table><tr><th>123 pclass</th><th>123 surv_percent</th><th>ABC generation</th></tr><tr><td>1</td><td>64</td><td>adult(31-50)</td></tr><tr><td>1</td><td>42</td><td>aged</td></tr><tr><td>1</td><td>75</td><td>minor(0-12)</td></tr><tr><td>1</td><td>82</td><td>teenager(13-19)</td></tr><tr><td>1</td><td>68</td><td>young_man(20-30)</td></tr><tr><td>2</td><td>46</td><td>adult(31-50)</td></tr><tr><td>2</td><td>20</td><td>aged</td></tr><tr><td>2</td><td>100</td><td>minor(0-12)</td></tr><tr><td>2</td><td>47</td><td>teenager(13-19)</td></tr><tr><td>2</td><td>40</td><td>young_man(20-30)</td></tr><tr><td>3</td><td>19</td><td>adult(31-50)</td></tr><tr><td>3</td><td>5</td><td>aged</td></tr><tr><td>3</td><td>39</td><td>minor(0-12)</td></tr><tr><td>3</td><td>27</td><td>teenager(13-19)</td></tr><tr><td>3</td><td>23</td><td>young_man(20-30)</td></tr></table>	123 pclass	123 surv_percent	ABC generation	1	64	adult(31-50)	1	42	aged	1	75	minor(0-12)	1	82	teenager(13-19)	1	68	young_man(20-30)	2	46	adult(31-50)	2	20	aged	2	100	minor(0-12)	2	47	teenager(13-19)	2	40	young_man(20-30)	3	19	adult(31-50)	3	5	aged	3	39	minor(0-12)	3	27	teenager(13-19)	3	23	young_man(20-30)
123 pclass	123 surv_percent	ABC generation																																															
1	64	adult(31-50)																																															
1	42	aged																																															
1	75	minor(0-12)																																															
1	82	teenager(13-19)																																															
1	68	young_man(20-30)																																															
2	46	adult(31-50)																																															
2	20	aged																																															
2	100	minor(0-12)																																															
2	47	teenager(13-19)																																															
2	40	young_man(20-30)																																															
3	19	adult(31-50)																																															
3	5	aged																																															
3	39	minor(0-12)																																															
3	27	teenager(13-19)																																															
3	23	young_man(20-30)																																															

5.3. Третья задача: выясним, как возраст пассажиров повлиял на их выживаемость.

Выводы:

1. Пожилых людей (> 50 лет) выжило меньше всех во всех классах -

самая уязвимая категория.

2. Самая выживаемая группа - дети до 18.

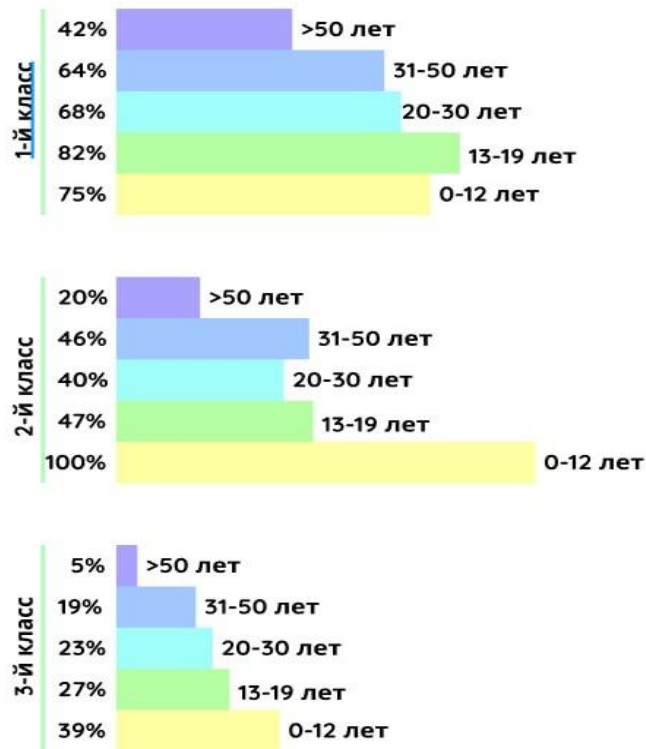
3. Подкатегории людей в возрасте от 19 до 50 лет выживают

по-разному, в зависимости от класса, но в целом тенденция

выживаемости более молодой подкатегории людей

(20-30 лет) и далее по убыванию (31-50).

Графики количества выживших
с разбивкой на группы по возрастам и классам



5.4.1. Четвертая задача (первая подзадача): посчитаем количество выживших, в зависимости от наличия/отсутствия членов семьи на борту

КОД	РЕЗУЛЬТАТ												
<pre>SELECT family_member, SUM(survived) * 100 / COUNT(*) AS surv_rate, CASE WHEN family_member = 0 THEN 'without_family' ELSE 'with_family' END AS family_status FROM (SELECT (siblings_spouses_aborboard + parents_children_aborboard) AS family_member, survived FROM titanic_table tt) GROUP BY family_status</pre>	<div>SQL SELECT family_member, SUM(survived) * 100 / COUNT(*) AS surv_rate, CASE WHEN family_member = 0 THEN 'without_family' ELSE 'with_family' END AS family_status FROM (SELECT (siblings_spouses_aborboard + parents_children_aborboard) AS family_member, survived FROM titanic_table tt) GROUP BY family_status</div> <table><tr><th></th><th>123 family_member</th><th>123 surv_rate</th><th>ABC family_status</th></tr><tr><td>1</td><td>1</td><td>50</td><td>with_family</td></tr><tr><td>2</td><td>0</td><td>30</td><td>without_family</td></tr></table>		123 family_member	123 surv_rate	ABC family_status	1	1	50	with_family	2	0	30	without_family
	123 family_member	123 surv_rate	ABC family_status										
1	1	50	with_family										
2	0	30	without_family										
<p>Вывод: наличие семьи на борту влияло на выживаемость пассажиров (30% выживших пассажиров плыли одни и 50% плыли с семьей (т.е. шансы увеличивались на $\frac{2}{3}$).</p>													

5.4.2. Четвертая задача (вторая подзадача): посчитаем количество выживших, в зависимости от количества членов семьи на борту

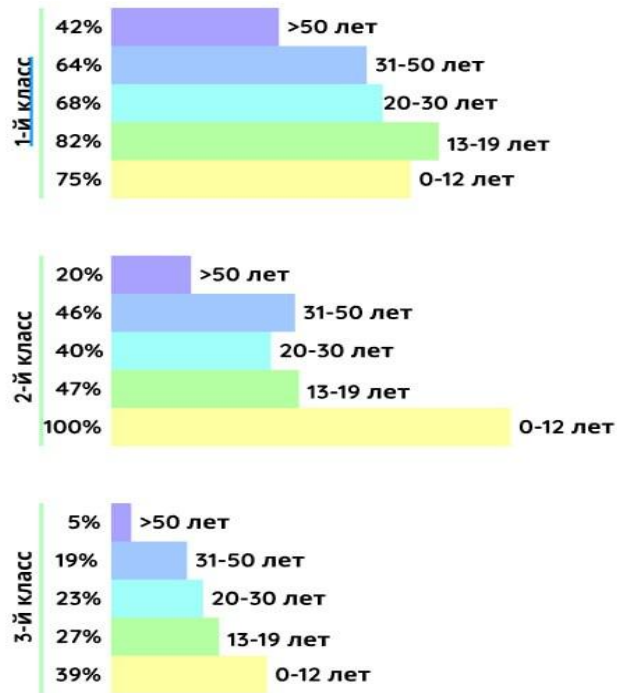
КОД	РЕЗУЛЬТАТ																																																		
<pre>SELECT family_member, SUM(survived) * 100 / COUNT(*) AS surv_rate, CASE WHEN family_member = 0 THEN 'without_family' ELSE 'with_family' END AS family_status, SUM(survived) AS absol_number_surv, count(*) AS total_number_pass FROM (SELECT (siblings_spouses_aborad + parents_children_aborad) AS family_member, survived FROM titanic_table tt) GROUP BY family_member</pre>	<p>с указанием справа абсолютных значений выживших пассажиров (absolute_number_surv) и их общего числа (включая погибших) - total_number_pass:</p> <table><tr><th>123 family_me ▼</th><th>123 surv_rate ▼</th><th>ABC family_status ▼</th><th>123▼</th><th>123 total_ ▼</th></tr><tr><td>0</td><td>30</td><td>without_family</td><td>163</td><td>533</td></tr><tr><td>1</td><td>55</td><td>with_family</td><td>89</td><td>161</td></tr><tr><td>2</td><td>57</td><td>with_family</td><td>59</td><td>102</td></tr><tr><td>3</td><td>72</td><td>with_family</td><td>21</td><td>29</td></tr><tr><td>4</td><td>20</td><td>with_family</td><td>3</td><td>15</td></tr><tr><td>5</td><td>13</td><td>with_family</td><td>3</td><td>22</td></tr><tr><td>6</td><td>33</td><td>with_family</td><td>4</td><td>12</td></tr><tr><td>7</td><td>0</td><td>with_family</td><td>0</td><td>6</td></tr><tr><td>10</td><td>0</td><td>with_family</td><td>0</td><td>7</td></tr></table>	123 family_me ▼	123 surv_rate ▼	ABC family_status ▼	123▼	123 total_ ▼	0	30	without_family	163	533	1	55	with_family	89	161	2	57	with_family	59	102	3	72	with_family	21	29	4	20	with_family	3	15	5	13	with_family	3	22	6	33	with_family	4	12	7	0	with_family	0	6	10	0	with_family	0	7
123 family_me ▼	123 surv_rate ▼	ABC family_status ▼	123▼	123 total_ ▼																																															
0	30	without_family	163	533																																															
1	55	with_family	89	161																																															
2	57	with_family	59	102																																															
3	72	with_family	21	29																																															
4	20	with_family	3	15																																															
5	13	with_family	3	22																																															
6	33	with_family	4	12																																															
7	0	with_family	0	6																																															
10	0	with_family	0	7																																															

5.4.2. Четвертая задача (вторая подзадача): посчитаем количество выживших, в зависимости от количества членов семьи на борту

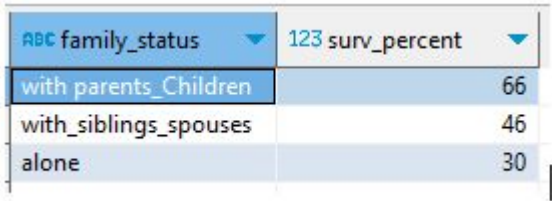
Вывод:

Хотя наличие семьи в целом повышало выживаемость, но только до определенного порога (до 3 членов семьи). Если семья в целом более 4-х человек, выживаемость резко (критически) падала.

Графики количества выживших с разбивкой на группы по возрастам и классам



5.4.2. Четвертая задача (третья подзадача): посчитаем количество выживших, в зависимости от категории членов семьи на борту (*братья/сестры/супруги либо дети/родители*)

КОД	РЕЗУЛЬТАТ								
<pre>SELECT CASE WHEN siblings_spouses_aborad > 0 THEN 'with_siblings_spouses' WHEN parents_children_aborad > 0 THEN 'with parents_Children' ELSE 'alone' END AS family_status, SUM(survived) * 100 / COUNT(*) AS surv_percent FROM titanic_table tt GROUP BY family_status ORDER BY surv_percent DESC</pre>	 <table><thead><tr><th>family_status</th><th>surv_percent</th></tr></thead><tbody><tr><td>with parents_Children</td><td>66</td></tr><tr><td>with_siblings_spouses</td><td>46</td></tr><tr><td>alone</td><td>30</td></tr></tbody></table>	family_status	surv_percent	with parents_Children	66	with_siblings_spouses	46	alone	30
family_status	surv_percent								
with parents_Children	66								
with_siblings_spouses	46								
alone	30								
<p>Вывод: выживало больше пассажиров с детьми, чем путешествовавших с супругами/братьями/сестрами. Однако даже вторые получали преимущество перед одинокими пассажирами (выживаемость была выше практически в полтора раза).</p>									

Выводы (резюме):

1. Выжило гораздо больше женщин, чем мужчин (74% против 19 %). Кроме того, выжило больше пассажиров 1-го (более высокого класса обслуживания): 62% - 1 класс, 47% - 2 класс, 24% - 3 класс.
2. По возрасту выявились следующие зависимости:
Самая выживаемая группа - дети до 18.
Подкатегории людей в возрасте от 19 до 50 лет выживают по-разному в зависимости от класса, но в целом тенденция выживаемости более молодой подкатегории людей (20-30 лет) и далее по убыванию (31-50, >50).
На последнем месте по выживаемости категория пожилых людей (> 50 лет) - самая уязвимая категория.
3. Наличие семьи на борту влияло на выживаемость пассажиров (30% выживших пассажиров плыли одни и 50% плыли с семьей (т.е. шансы увеличивались на $\frac{2}{3}$). Имело значение даже наличие взрослых членов семьи, т.е. не детей (но с детьми выживаемость была более чем в 2 раза выше). Вместе с тем, если количество членов в семье превышало 4, шансы выжить критически уменьшались.

СПАСИБО ЗА ВНИМАНИЕ!