САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Институт «Компьютерных наук и кибербезопасности»

«Высшая школа технологий искусственного интеллекта»

Направление «Математическое обеспечение и администрирование информационных систем»

**Отчёт по заданию**

«Введение в науки о данных и визуализации»

Преподаватель: Э.Р. Хольгер

Выполнили:

Бай Борис группа: 5130203/40002 семестр: 2

Мартьянов Иван группа: 5130203/40002 семестр: 2

Тимур Гилязов группа: 5130203/40002 семестр: 2

**Введение:**

Для выполнения данного задания был использован набор данных, содержащий информацию о пассажирах Титаника. Набор включает как числовые, так и категориальные переменные. Обработка и анализ этих данных были реализованы с использованием баз данных phpMyAdmin, работающих на локальном сервере (localhost), а также с применением языков программирования PHP и JavaScript с использованием фреймворков Bootstrap и AnyChart.

**Визуализация данных:**

1. **Распределение пассажиров по классам кают (Приложение 1)**

В данном анализе была представлена таблица, которая отображает распределение пассажиров по классам кают (1-й, 2-й и 3-й классы). Эти данные позволяют увидеть, какая категория пассажиров была наиболее представлена на борту, а также узнать, сколько пассажиров из каждого класса выжило или погибло. На основе визуализации можно сделать вывод, что подавляющее большинство пассажиров находилось в 3-м классе, что соответствует историческим данным о социальной структуре на Титанике.

1. **Распределение пассажиров по полу (Приложение 2)**

Визуализация соотношения мужчин и женщин среди пассажиров Титаника, а также расчет общей стоимости билетов для каждого пола. Это важно для дальнейшего анализа демографической структуры и выживаемости, так как известно, что при эвакуации приоритет отдавался женщинам и детям.

1. **Корреляция между числовыми характеристиками (Приложение 3)**

В таблице представлена корреляция между возрастом пассажиров, стоимостью их билетов и количеством родственников, находящихся с ними на борту. Например, выявлена отрицательная корреляция между возрастом и количеством братьев/сестёр — чем старше пассажир, тем меньше у него братьев и сестёр на борту. Также наблюдается положительная корреляция между количеством братьев/сестёр и родителей/детей, что позволяет сделать вывод о взаимосвязи этих данных.

1. **Столбчатая диаграмма для категориальных переменных (Приложение 4)**  
   Столбчатая диаграмма позволяет наглядно продемонстрировать частоту каждого класса и пола в сочетании с их стоимостью билетов. Из диаграммы видно, что стоимость билетов первого класса значительно выше, чем у второго и третьего классов, а для женщин стоимость билетов первого класса превышает мужскую на 35$.
2. **Круговая диаграмма для анализа выживаемости (Приложение 5)**  
   Эта диаграмма показывает, что большинство мужчин не выжило при крушении Титаника, что может быть связано с тем, что мужчины уступали свои места женщинам и детям.
3. **Диаграмма выживаемости женщин и мужчин (Приложение 6)**  
   Как и ожидалось, на круговой диаграмме видно, что процент выживших женщин значительно выше, чем мужчин. Это подтверждает гипотезу о предпочтении в эвакуации женщин и детей.
4. **Диаграмма для числовых данных (Приложение 7)**  
   Диаграмма помогает оценить разброс возрастных данных среди пассажиров. На основе этих данных можно сделать вывод, что пассажиры первого класса были старше, чем пассажиры второго и третьего классов, что подтверждает социальную структуру на борту.
5. **Гистограмма распределения возраста пассажиров (Приложение 8)**   
   Гистограмма отображает возрастное распределение пассажиров Титаника. Преобладающая группа — это люди трудоспособного возраста, но также присутствуют дети и пожилые пассажиры, что подтверждается корреляцией с количеством родителей/детей и братьев/сестёр.
6. **Групповой анализ стоимости билетов по классу и полу (Приложение 9)**  
   На данной диаграмме показана стоимость билетов в зависимости от пола и класса. Видно, что в первом классе женщины платили больше, чем мужчины, что может быть связано с высококачественным обслуживанием и удобствами.

**Заключение:**

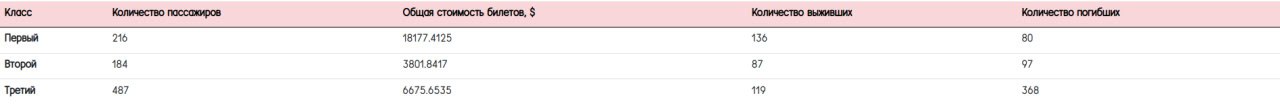
В ходе выполнения работы использовались стандартные инструменты для обработки и визуализации данных. Для работы с базой данных была использована phpMyAdmin и MySQL на локальном сервере, что обеспечивало удобное управление данными. Программирование на PHP и JavaScript позволило обработать данные и создать интерактивные графики, используя фреймворки Bootstrap и AnyChart для визуализации.

Особое внимание уделялось анализу как категориальных, так и числовых данных. Для числовых переменных использовались корреляции, диаграммы с усами и гистограммы, а для категориальных — столбчатые и круговые диаграммы. Такой подход помог выявить закономерности, например, влияние возраста на количество родственников на борту.

Использование этих инструментов позволило эффективно визуализировать данные и выявить значимые тренды, делая анализ доступным и понятным. Работа показала важность правильного выбора технологий для обработки и представления данных.

**Приложение:**

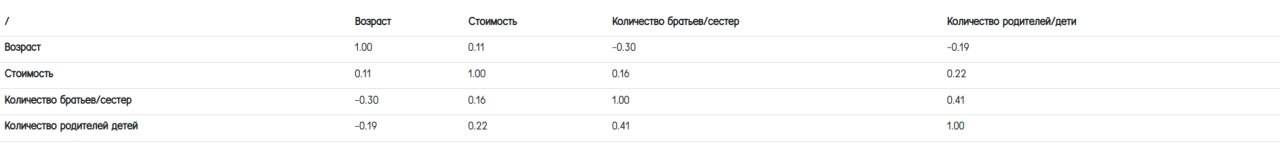
**Приложение 1** (Фото: распределение пассажиров по классам)



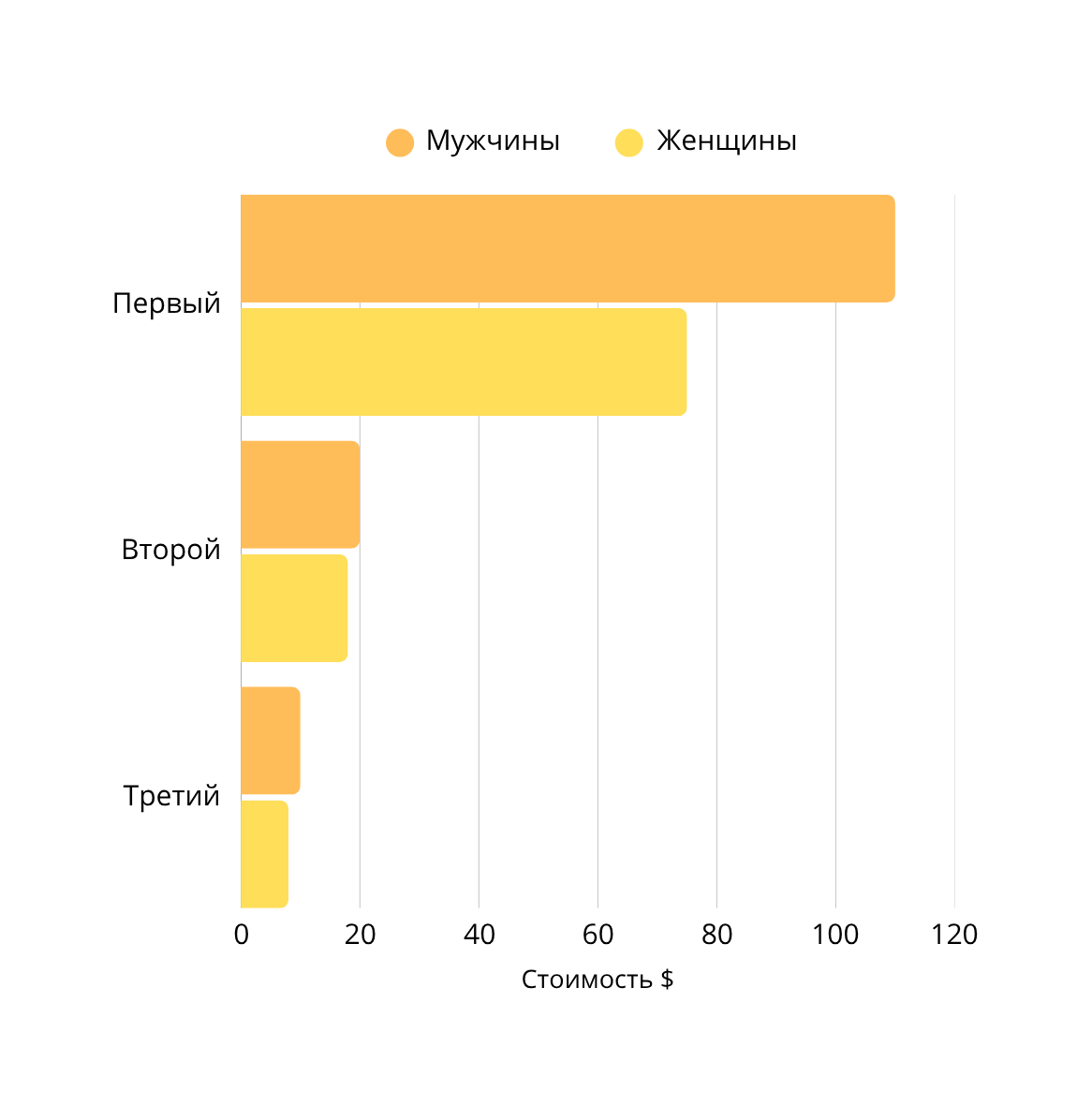
**Приложение 2** (Фото: соотношение мужчин и женщин)

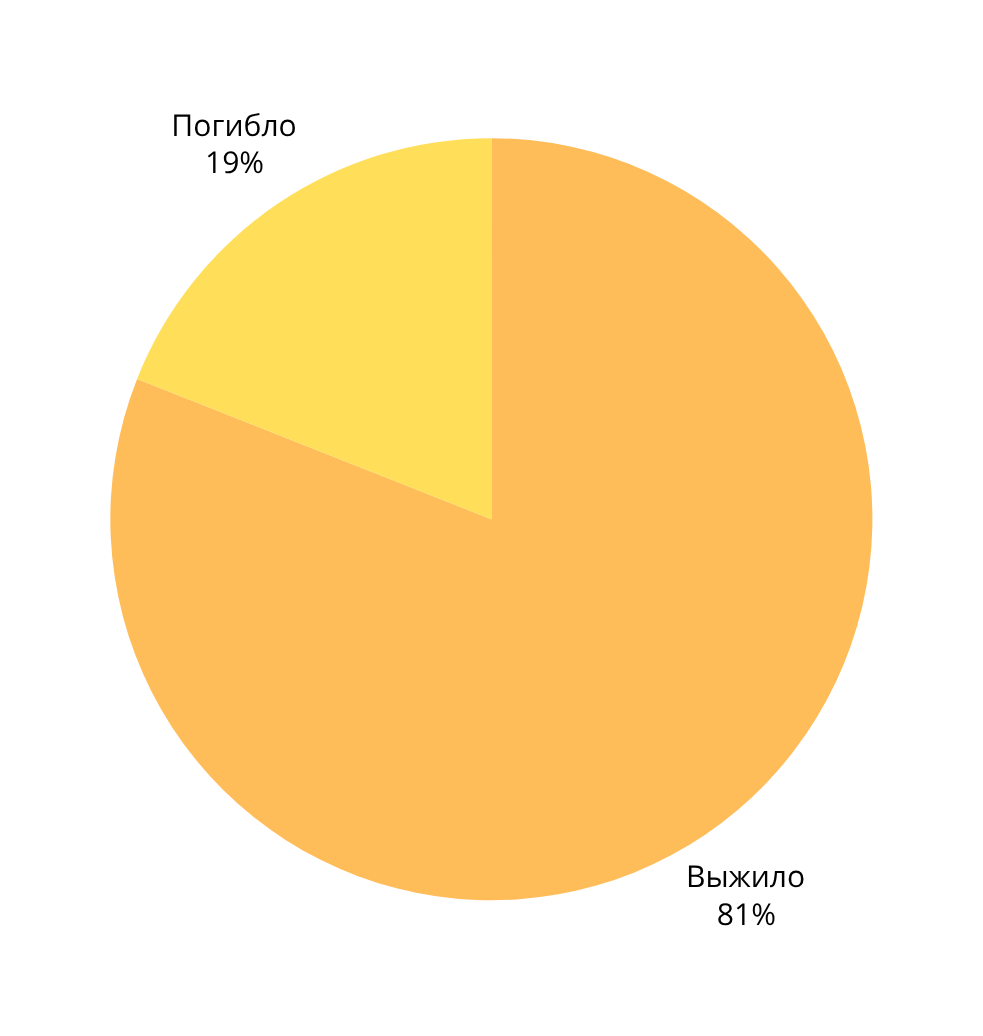


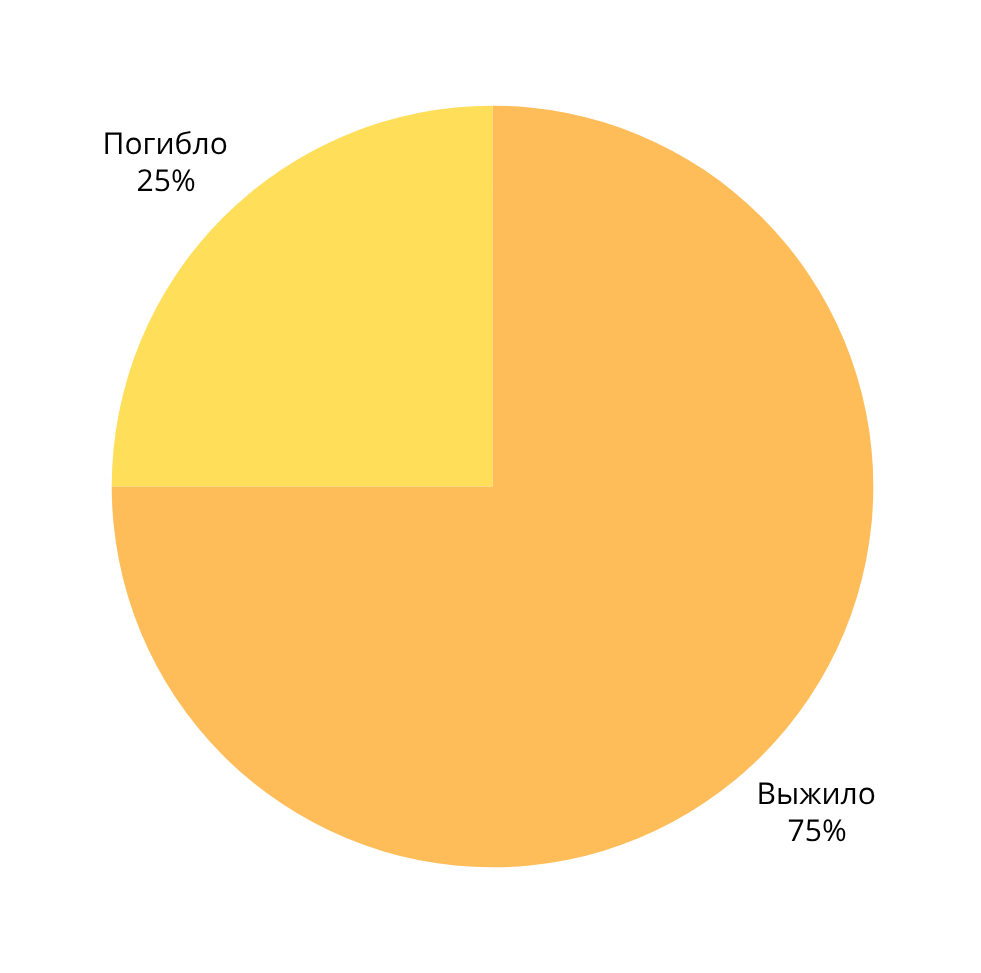
**Приложение 3** (Фото: таблица корреляции)

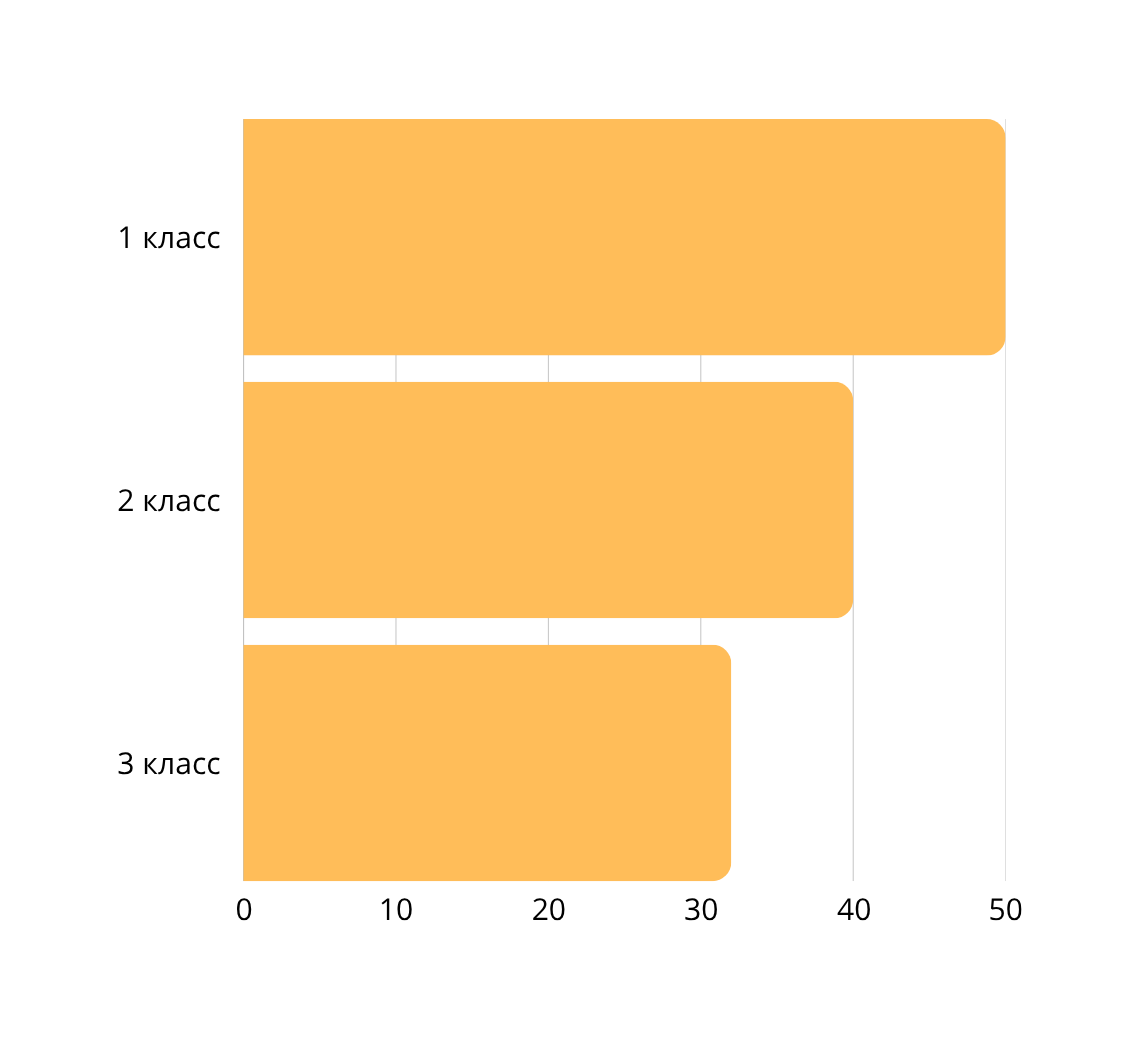


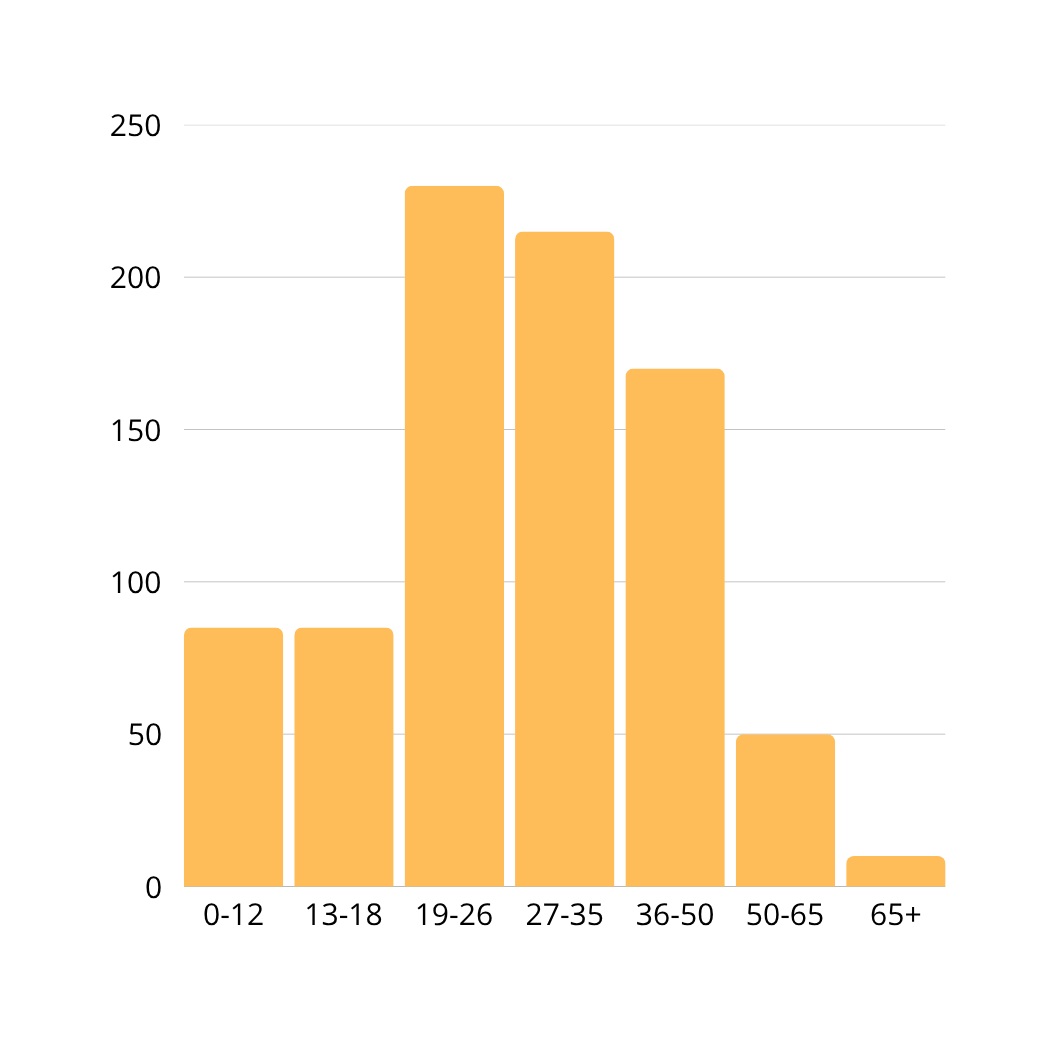
**Приложение 4** (Фото: Средняя стоимость билетов разных классов)

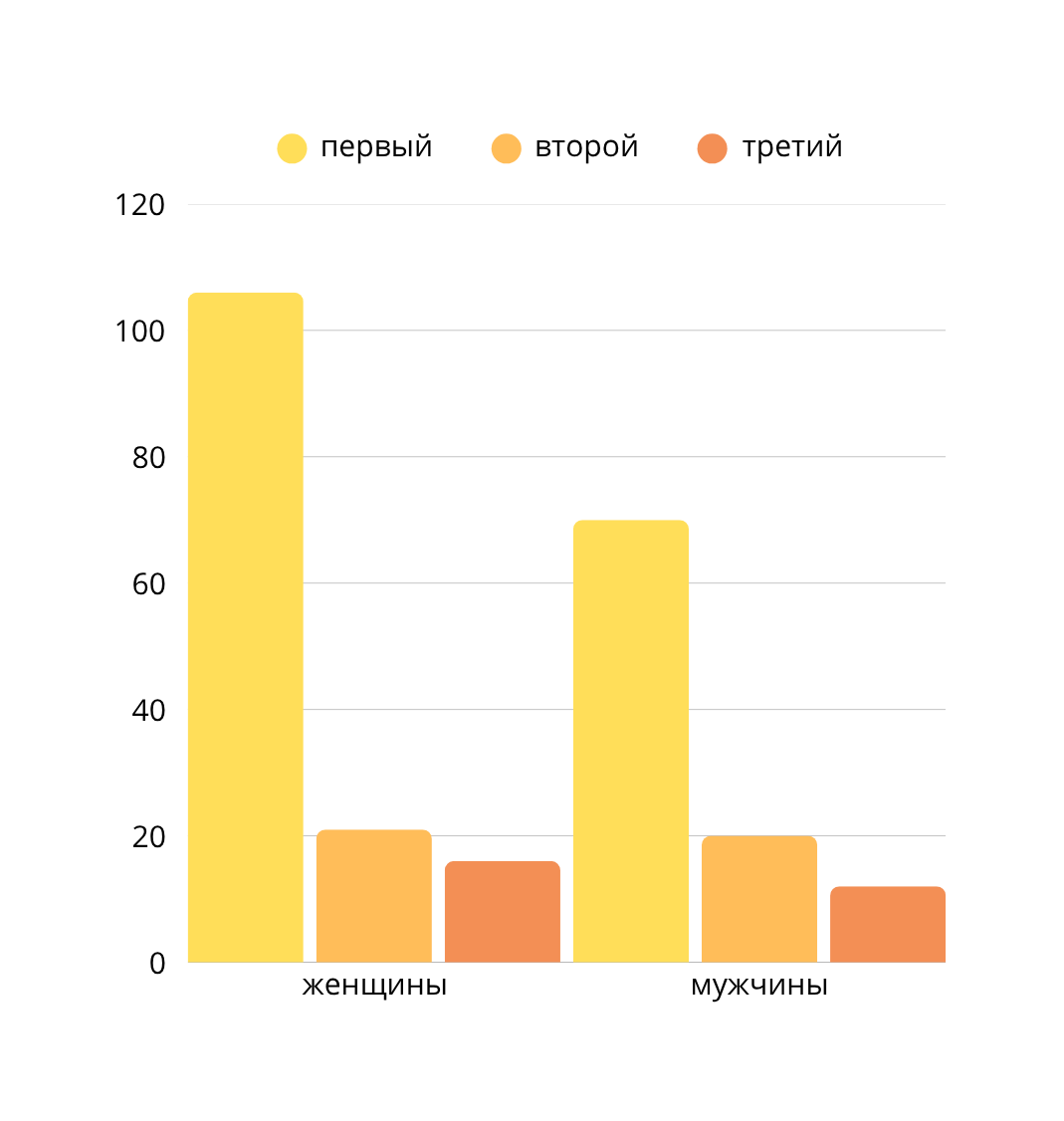


**Приложение 5** (Фото: смертность мужчин)

**Приложение 6** (Фото: смертность женщин)

**Приложение 7** (Фото: распределение возрастов)

**Приложение 8** (Фото: гистограмма возраста пассажиров)

**Приложение 9** (Фото: средняя стоимость билета от класса и пола)