ыСАНКТ-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:

Потеряев Родион группа: 5130203/40002 семестр: 2

Сикорский Александр группа: 5130203/40002 семестр: 2

Слинкина Василина группа: 5130203/40002 семестр: 2

Чигринский Захар группа: 5130203/40002 семестр: 2

В качестве базы данных для выполнения данной работы мы взяли набор данных с табличной структурой данных, содержащий информацию о пассажирах Титаника, с числовыми и категориальными видами данных. Выглядит он таким образом:



Наш исходный код работает с базой данных phpMyAdmin, лежащей на localhost.

Реализован на языке PHP, JS в VSC с использованием фреймворков Bootstrap и AnyChart.

Ниже представлена визуализация данных этой таблицы в различных формах.

1) Визуализация категориальных переменных (по классу пассажира)

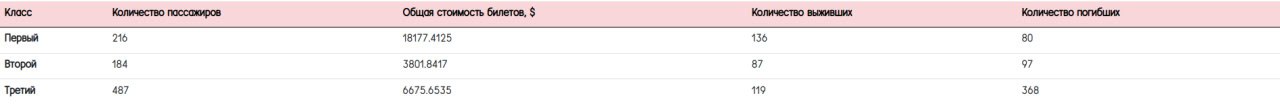


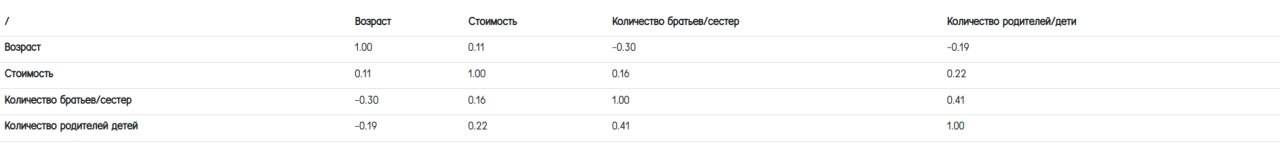
Таблица показывает распределение пассажиров Титаника по классам (1-й, 2-й, 3-й). Это позволяет оценить, какая категория пассажиров преобладала на борту и том числе узнать общую стоимость билетов каждого класса, а также количество выживших и погибших. Можно заметить, что большинство пассажиров относилось к 3-му классу, что соответствует историческим данным о социальной структуре пассажиров.

2) Визуализация категориальных переменных (по полу пассажира)



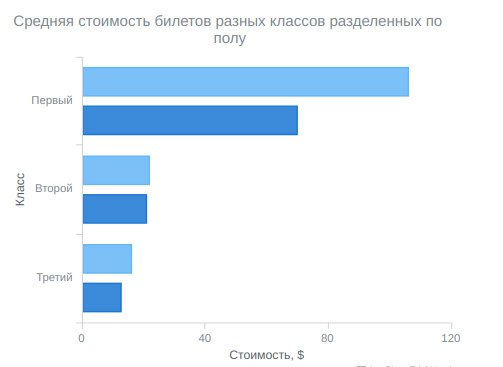
Здесь представлено соотношение мужчин и женщин среди пассажиров, а также общая стоимость билетов для каждого пола. Это важно для анализа демографического состава и, например, выживаемости, так как известно, что при эвакуации приоритет отдавался женщинам и детям.

3) Таблица корреляции числовых характеристик



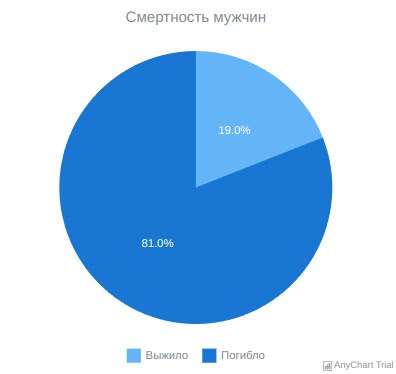
В этой таблице отражены коэффициенты корреляции между числовыми параметрами по возрасту, стоимости билета и количеству родственников на борту. Корреляция помогает выявить взаимосвязи. Из наиболее влиятельных можно выделить, например, отрицательную корреляцию между возрастом и количеством братьев и сестёр, что говорит нам о том, что чем старше пассажир, тем в среднем меньше у него братьев/сестёр на борту, и наоборот. А также положительная корреляция между количеством братьев/сестёр и количеством родителей/детей можно натолкнуть на вывод, что пассажиры, у которых было больше братьев/сестёр на борту, чаще имели и больше родителей/детей на борту, и наоборот. Остальные имеют более случайную закономерность.

4) Столбчатая диаграмма для абсолютных частот для каждого экземпляра с использованием всех категориальных переменных



Из данной диаграммы можно сделать несколько выводов: во-первых, средняя стоимость билетов второго и третьего классов отличается незначительно (не более 5$), в то же время средняя стоимость билетов первого класса превышает стоимость остальных билетов в 3-7 раз; во-вторых, разница между билетами для женщин и для мужчин для второго и третьего классов незначительна, однако для первого класса стоимость билетов для женщин превышает стоимость билетов для мужчин примерно на треть своей стоимости (~35$).

5) Круговая диаграмма для относительных частот для каждого экземпляра с использованием категориальных переменных



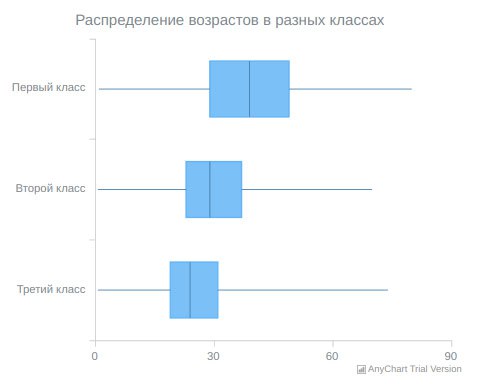
По данной диаграмме можно видеть, что абсолютное большинство мужчин погибло в результате крушения Титаника. Принимая во внимание культурный контекст того времени, можно предположить, что мужчины уступали свои места на спасательных шлюпках женщинам и детям. Если это предположение верно, то мы увидим подтверждение этому на следующем слайде.

6) Круговая диаграмма для относительных частот для каждого экземпляра с использованием категориальных переменных



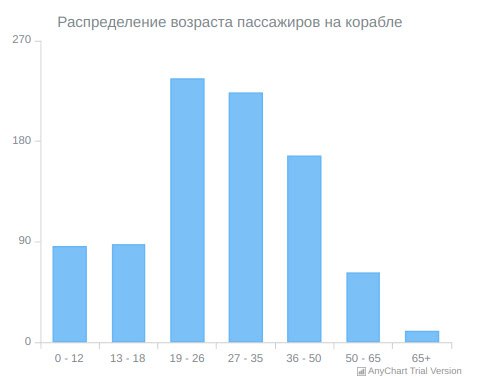
Как и предполагалось на предыдущем пункте, женщин выжило больше в процентном соотношении, чем мужчин. Объяснение этого факта были приведены в 5 пункте.

7) Диаграмма с усами для визуализации числовых переменных



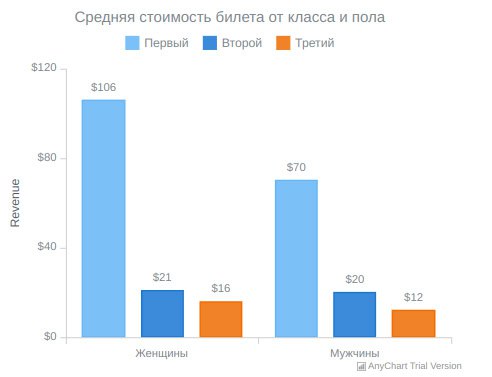
Эта диаграмма отображает медиану, квартили и выбросы в числовых данных. Она помогает оценить разброс значений. Пассажиры первого класса, в среднем, были старше, чем во втором и третьем. Можно сделать вывод, что более взрослые и состоятельные люди предпочитали первый класс, в то время как более молодые и, вероятно, представители рабочего класса ехали в более дешёвых каютах.

8) Гистограмма частот для визуализации числовых переменных



Гистограмма показывает распределение числового признака, в нашем случае возраста, по интервалам. По ней можно понять, все возрастные группы пассажиров на борту, а также какая возрастная группа преобладала среди них. На Титанике преобладали люди трудоспособного возраста, но в то же время было и не мало детей, что подтверждается и статистикой по количеству родителей/детей и братьев/сестёр из таблицы корреляции.

9) Групповой анализ абсолютной частоты с помощью столбчатой диаграммы



Здесь сравниваются несколько категориальных признаков – стоимость билета пассажира в зависимости от класса и пола. Видно, что стоимость билетов женщин, в особенности у первого класса, выше, чем у мужчин. Рост цен на билеты в зависимости от класса не линейный.