МДЗ2:

Давайте в качестве упражнения проанализируем вероятно самый популярный dataset для иллюстрации всевозможной статистики и алгоритмов машинного обучения – «Titanic». Исходный dataset можно найти по ссылке: https://www.kaggle.com/c/titanic/ (полный dataset поделен на train и test для верификации алгоритмов, Вы можете взять любой или оба).
В архиве приложен чуть обработанный (в плане форматирования) датасет «Titanic_train_edit.csv» (для импорта на сервер и написания запросов) и этот же датасет в формате xls для написания формул в SpreadSheet (Excel, Open Office и тд). Вы можете пользоваться этими датасетами, а можете скачать исходные с сайта kaggle (особенно, если мой csv не с теми разделителями, которые требует Ваша ОС).

Описание данных:
Частичная расшифровка полей (полная расшифровка доступна по ссылке):
Survival – 0 – No, 1 – Yes
pClass – класс пассажира (Первый (Бизнес, в самолетных терминах), Второй (Комфорт), Третий (Эконом))
Embarked – порт отправления: (C = Cherbourg; Q = Queenstown; S = Southampton)
Данные находятся в csv файле, который Вам нужно проимпортировать на сервер.

Описание импорта на сервер:
Кратко как это сделать в **MS SQL**:
- правой мышкой в дереве объектов на имя базы, в которую Вы хотите проимпортировать данные;
- выберите пункт: Tasks далее ImportData;
- в качестве источника данных выбираете Flat File Source (для csv, можно выбрать и xls файл, тогда и тип источника - Excel);
- далее выбираете, где Ваш файл находится;
Параметры:
    Format – Delimited;
    Header row delimiter – {CR}/{LF}
    Locale – Russian, Code Page – 1251 (ANSI-Cyrillic).
    Галка «Column names in the first data row» - должна быть проставлена (это значит, что первая строка используется для заголовков столбцов).
В разделе Columns или Preview можно посмотреть, как данные будут храниться в таблице. В Columns указывается, что используется в качестве разделителя между столбцами. В данном случае – “;”.
В разделе «Advanced» можно сразу изменить типы полей (иначе они будут импортироваться как строки):
    PassengerID, Survived, Pclass, SibSp, Parch-> int (Выберите подходящий)
    Name длину замените на 200 (в принципе достаточно 100).
    Age, Fare -> float

- далее отобразится на какой сервер и в какую базу Ваши данные попадут.
- далее отобразится, в какую таблицу данные попадут (таблицы может еще не существовать). Назовите таблицу «Titanic». По кнопке «Edit Mappings» можно посмотреть, в какие типы Ваши данные будут переведены. При необходимости можно поправить.

- далее «Run immediately» и Finish. Далее идет импорт.

Если по каким-то причинам все равно никак не получается проимпортировать, то можно скопировать данные из csv или xls (открываться должны в Excel каждая колонка в отдельной колонке) и вставить их в созданную Вами таблицу с соответствующими полями (метод ctrl-c ctrl-v).
В любом случае при неудаче с импортом – обязательно внимательно прочитайте, какую ошибку выдает сервер (по каким причинам не может проимпортировать данные).

Есть и иные способы: посмотрите, например, Bulk import.
Какие проблемы могут возникнуть при импорте: в csv файле разделитель может отличаться от Вашего в региональных настройках. Вы можете временно поменять региональные настройки

(поменять стандартный разделитель, а после импорта вернуть обратно). Либо скачать с сайта kaggle csv в другом формате.

Кратко как это сделать в **Access**:
Файл -> Внешние Данные -> Импорт (аналогично, можно выбрать csv, а можно xls)
Далее просто следуете инструкциям, разделитель «;».

В **PostGre**: есть отдельные команды для импорта файлов.

Запросы на SQL (каждый пункт – один запрос):

1. Выведите суммарное колво пассажиров и колво выживших. Вычислите долю выживших.
2. Посчитайте по каждому классу билета суммарное колво пассажиров и колво выживших. Вычислите долю выживших по каждому классу билета.
3. По каждому классу билета и полу пассажира посчитайте: суммарное колво пассажиров, колво выживших и долю выживших.
4. По каждому порту отправления посчитайте колво пассажиров, колво выживших и долю выживших.
5. Выведите порт отправления с наибольшим колвом пассажиров.
6. Посчитайте средний возраст пассажиров и средний возраст выживших в группировке по классу билета и полу. При подсчете среднего возврата посмотрите, а как у Вас проимпортировались данные, где возраст указан не был. Если как NULL, то средний возраст посчитается верно при использовании AVG (позже изучим обработку NULL значений), иначе, если неизвестные возраста заменились на 0, подумайте, как правильно посчитать средний возраст только по ненулевым age? (хинт: CASE WHEN)
7. Выведите первые 10 строк по убыванию стоимости билета. Как Вы считаете, стоимость билета указана на человека?
8. Проверьте, есть ли билеты, для которых цена в разных строках отличается? Аналогично для порта отправления (можно в два запроса).
9. Для каждого номера билета, класса, цены и порта отправления посчитайте колво строк (колво пассажиров), колво выживших пассажиров.
10. Выведите билеты, для которых колво пассажиров более 1 и все пассажиры выжили.
11. Напишите запрос, который посчитает вероятность выжить, если Вас зовут Elizabeth, если Вас зовут Mary (достаточно посчитать, что такая подстрока должна входить в имя пассажира)

Мы не ограничиваем Вас этими запросами. Вы можете посчитать статистику в любом разрезе, каком Вам хочется. На kaggle (ссылка выше) было открыто соревнование по анализу данных с Титаника (для лучшего прогноза выживаемости для пассажира). Вы можете поучаствовать.

Упражнения в Excel (или любом другом удобном Вам spreadsheet: Open Office и тд):
В xls файле на странице «Exercise» предлагается выполнить упражнения в xls. Знания xls никак не будут проверяться в контрольной, поэтому выполнять эти задания или нет полностью на Ваше усмотрение.
Упражнения – так же посчитать статистику по выживаемости в различных разрезах. Для примера колво пассажиров в делении на пол и класс билета уже сделано.
Эти результаты можно сравнить с тем, что у Вас получилось в SQL (аналогично обратите внимание на случаи, когда возраст не указан).
Для Excel:
Кратко: роль $ - знак фиксатора, который показывает, что либо столбец, либо колонка (смотря где указан этот знак) остаются фиксированными при перетаскивании, растягивании или копировании формулы.
Такую формулу =SUM(IF(($A4=Titanic!$E$2:$E$892)*(B$3=Titanic!$C$2:$C$892),1,0)) недостаточно просто прописать, так как это «массивная» формула. Ее нужно «включить». Для включения достаточно войти в режим редактирования (F2) и нажать Ctrl-shift-Enter одновременно.
Если Вы работаете в других SpreadSheet, то вместо такой формулы можно использовать SUMIF (как и в Excel). Почитайте про их использование, там так же нет ничего сложного.