**Сборный проект**

Поздравляем! Вы прошли первый блок курсов в тренажёре. Самое время проверить знания и решить аналитический кейс. Выполнять работу вы будете самостоятельно.

Как закончите, отправьте работу на проверку ревьюеру. В течение суток вы получите комментарии. Их нужно учесть: доработать проект и вернуть ревьюеру обновлённый вариант.

Скорее всего, вы будете дорабатывать кейс по комментариям несколько раз. Это нормально.

Проект завершён, когда одобрены все доработки.

**Описание проекта**

Вы работаете в интернет-магазине «Стримчик», который продаёт по всему миру компьютерные игры. Из открытых источников доступны исторические данные о продажах игр, оценки пользователей и экспертов, жанры и платформы (например, *Xbox* или *PlayStation*). Вам нужно выявить определяющие успешность игры закономерности. Это позволит сделать ставку на потенциально популярный продукт и спланировать рекламные кампании.

Перед вами данные до 2016 года. Представим, что сейчас декабрь 2016 г., и вы планируете кампанию на 2017-й. Нужно отработать принцип работы с данными. Неважно, прогнозируете ли вы продажи на 2017 год по данным 2016-го или же 2027-й — по данным 2026 года.

В наборе данных попадается аббревиатура *ESRB (Entertainment Software Rating Board)* — это ассоциация, определяющая возрастной рейтинг компьютерных игр. *ESRB* оценивает игровой контент и присваивает ему подходящую возрастную категорию, например, «Для взрослых», «Для детей младшего возраста» или «Для подростков».

**Инструкция по выполнению проекта**

**Шаг 1. Откройте файл с данными и изучите общую информацию**

Путь к файлу: */datasets/games.csv*.

**Шаг 2. Подготовьте данные**

* Замените названия столбцов (приведите к нижнему регистру);
* Преобразуйте данные в нужные типы. Опишите, в каких столбцах заменили тип данных и почему;
* Обработайте пропуски при необходимости:
  + Объясните, почему заполнили пропуски определённым образом или почему не стали это делать;
  + Опишите причины, которые могли привести к пропускам;
  + Обратите внимание на аббревиатуру 'tbd' в столбце с оценкой пользователей. Отдельно разберите это значение и опишите, как его обработать;
* Посчитайте суммарные продажи во всех регионах и запишите их в отдельный столбец.

**Шаг 3. Проведите исследовательский анализ данных**

* Посмотрите, сколько игр выпускалось в разные годы. Важны ли данные за все периоды?
* Посмотрите, как менялись продажи по платформам. Выберите платформы с наибольшими суммарными продажами и постройте распределение по годам. За какой характерный срок появляются новые и исчезают старые платформы?
* Возьмите данные за соответствующий **актуальный период.** Актуальный период определите самостоятельно в результате исследования предыдущих вопросов. Основной фактор — эти данные помогут построить прогноз на 2017 год.
* Не учитывайте в работе данные за **предыдущие годы**.
* Какие платформы лидируют по продажам, растут или падают? Выберите несколько потенциально прибыльных платформ.
* Постройте график «ящик с усами» по глобальным продажам игр в разбивке по платформам. Опишите результат.
* Посмотрите, как влияют на продажи внутри одной популярной платформы отзывы пользователей и критиков. Постройте диаграмму рассеяния и посчитайте корреляцию между отзывами и продажами. Сформулируйте выводы.
* Соотнесите выводы с продажами игр на других платформах.
* Посмотрите на общее распределение игр по жанрам. Что можно сказать о самых прибыльных жанрах? Выделяются ли жанры с высокими и низкими продажами?

**Шаг 4. Составьте портрет пользователя каждого региона**

Определите для пользователя каждого региона (*NA, EU, JP*):

* Самые популярные платформы (топ-5). Опишите различия в долях продаж.
* Самые популярные жанры (топ-5). Поясните разницу.
* Влияет ли рейтинг ESRB на продажи в отдельном регионе?

**Шаг 5. Проверьте гипотезы**

* Средние пользовательские рейтинги платформ *Xbox One* и *PC* одинаковые;
* Средние пользовательские рейтинги жанров *Action* (англ. «действие», экшен-игры) и *Sports* (англ. «спортивные соревнования») разные.

Задайте самостоятельно пороговое значение *alpha*.

Поясните:

* Как вы сформулировали нулевую и альтернативную гипотезы;
* Какой критерий применили для проверки гипотез и почему.

**Шаг 6. Напишите общий вывод**

**Оформление:** Выполните задание в *Jupyter Notebook*. Заполните программный код в ячейках типа *code,* текстовые пояснения — в ячейках типа *markdown*. Примените форматирование и заголовки.

**Описание данных**

* *Name* — название игры
* *Platform* — платформа
* *Year\_of\_Release* — год выпуска
* *Genre* — жанр игры
* *NA\_sales —* продажи в Северной Америке (миллионы проданных копий)
* *EU\_sales* — продажи в Европе (миллионы проданных копий)
* *JP\_sales* — продажи в Японии (миллионы проданных копий)
* *Other\_sales —* продажи в других странах (миллионы проданных копий)
* *Critic\_Score* — оценка критиков (максимум 100)
* *User\_Score* — оценка пользователей (максимум 10)
* *Rating* — рейтинг от организации *ESRB* (англ. *Entertainment Software Rating Board).* Эта ассоциация определяет рейтинг компьютерных игр и присваивает им подходящую возрастную категорию.

Данные за 2016 год могут быть неполными.

**Как будут проверять мой проект?**

Мы подготовили критерии оценки проекта. Прежде чем решать кейс, внимательно изучите их.

На что обращают внимание при проверке проекта:

* Как вы описываете выявленные в данных проблемы?
* Как готовите датасет к анализу?
* Какие строите графики для распределений и как их объясняете?
* Как рассчитываете стандартное отклонение и дисперсию?
* Формулируете ли альтернативную и нулевую гипотезы?
* Какие методы применяете, чтобы их проверить?
* Объясняете результат проверки гипотезы или нет?
* Соблюдаете ли структуру проекта и поддерживаете аккуратность кода?
* Какие выводы делаете?
* Оставляете ли комментарии к шагам?

Всё необходимое, чтобы выполнить проект, есть в шпаргалках и конспектах прошлых тем.

Успехов!