

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ М. В. ЛОМОНОСОВА
Факультет вычислительной математики и кибернетики

Кафедра исследования операций

ОТЧЕТ ПО ЗАДАНИЮ №4

Выполнили студенты:

Битиев Алексей
Кулакова Мария
Сат Санчай

Преподаватель:

Гусева Юлия

Москва
2020

Постановка задачи:

После оглушительного успеха в освобождении Астапора, Миэрина и Юнка от власти работоторговцев Дейенерис Бурерожденная открыла себе доступ к Летнему морю, а следовательно – путь в Вестерос.

Для ведения войны с Семью Королевствами нужно оружие, а для оружия нужна сталь. Нет никаких сомнений в кузнечном искусстве Безупречных, однако поставщики стали не столь надежны.

Два основных поставщика стали – это Westeros Inc. и Harpy & Co. На протяжении нескольких месяцев мы покупаем сталь у обеих компаний, и каждая из них предлагает ощутимую скидку при заключении эксклюзивного договора на поставку.

Советник королевы Тирион Ланнистер знает о твоём умении принимать взвешенные рациональные решения и просит помощи в объективном решении вопроса о том, с какой из компаний следует заключить эксклюзивный договор на поставку стали.

У Тириона есть записи о производстве мечей каждым из кузнецов-безупречных, а также данные о количестве сломанных мечей в каждый из месяцев ведения боевых действий.

Исходные данные:

Данные о производстве оружия и количестве единиц сломанного оружия за каждый месяц каждым из кузнецов, предоставлены в CSV-файле "production.csv". Работа всех кузнецов одинакова и зависит только от качества поставляемой стали. Файл состоит из 6 столбцов:

	unsullen.id	production.date	report.date	produced	defects	supplier
0	1.0	1	1	103.0	0.0	harpy.co
1	1.0	1	2	0.0	2.0	harpy.co
2	1.0	1	3	0.0	4.0	harpy.co
3	1.0	1	4	0.0	5.0	harpy.co
4	1.0	1	5	0.0	13.0	harpy.co

- 1) **unsullen.id** - номер кузнеца, который ковал мечи;
- 2) **production.date** - дата производства мечей;
- 3) **report.date** - дата отчета о произведенных; мечях;
- 4) **produced** - количество произведенных мечей;
- 5) **defects** - количество дефектов в мечях;
- 6) **supplier** - название поставщика стали;

Цель:

Необходимо провести разведывательный анализ данных с целью ответа на вопрос: *"С каким из поставщиков стали следует заключить договор?"*

Разведывательный анализ данных — анализ основных свойств данных, нахождение в них общих закономерностей, распределений и аномалий, построение начальных моделей, зачастую с использованием инструментов визуализации.

Понятие введено математиком Джоном Тьюки, который сформулировал цели такого анализа следующим образом:

- максимальное «проникновение» в данные,
- выявление основных структур,
- выбор наиболее важных переменных,
- обнаружение отклонений и аномалий,
- выявление основных структур,
- разработка начальных моделей.

Решение:

Нулевой этап:

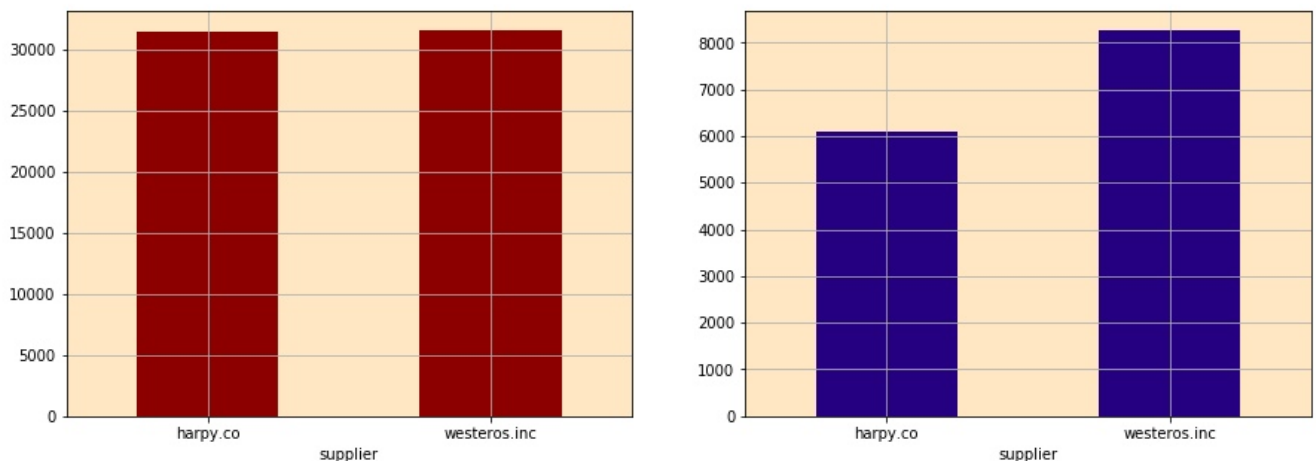
Считываем данные из CSV-файла в удобном для работы формате с помощью функции библиотеки `pandas read_csv('production.csv', sep=';')`.

Первый этап:

Анализ количества произведенных и сломанных мечей за все время. Чтобы подсчитать и проанализировать количество произведенных и сломанных мечей за все время, разделим данные по поставщикам стали `'harpy.co'` и `'westeros.inc'` с помощью метода `loc`.

Далее просуммируем все произведенные и сломанные мечи с помощью методов `sum` и `groupby` для каждого поставщика в отдельные переменные и найдем процент дефектных мечей от общего числа.

По полученным данным строим гистограммы с помощью `matplotlib.pyplot`:

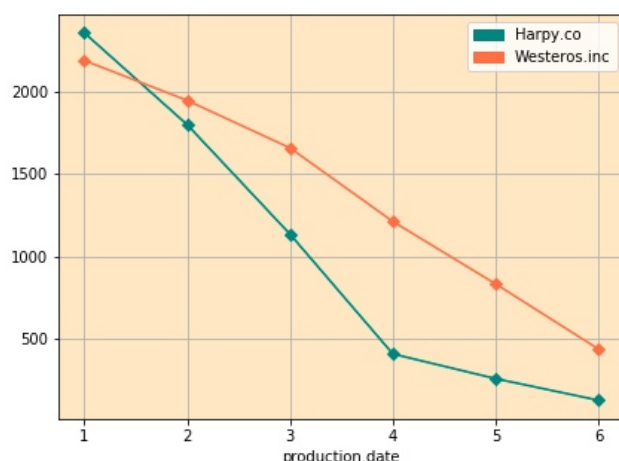
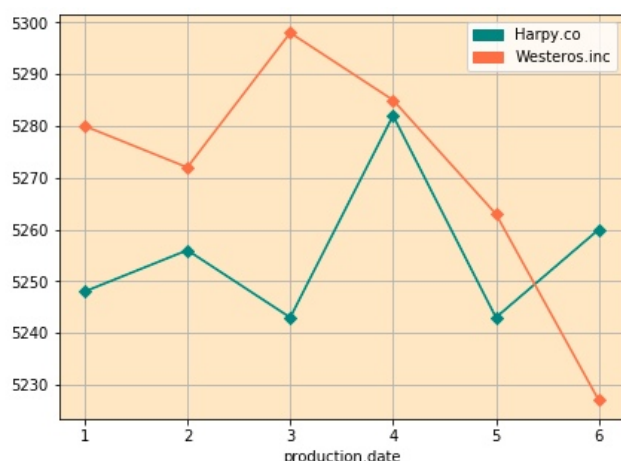


Второй этап:

Анализ количества произведенных и сломанных мечей за месяц.

Для подсчета и анализа количества произведенных и сломанных мечей за месяц сгруппируем произведенные мечи по кузнецам и датам и просуммируем с помощью методов `groupby` и `sum`.

Далее строим графики зависимости по месяцам с помощью *matplotlib.pyplot* и *matplotlib.patches*

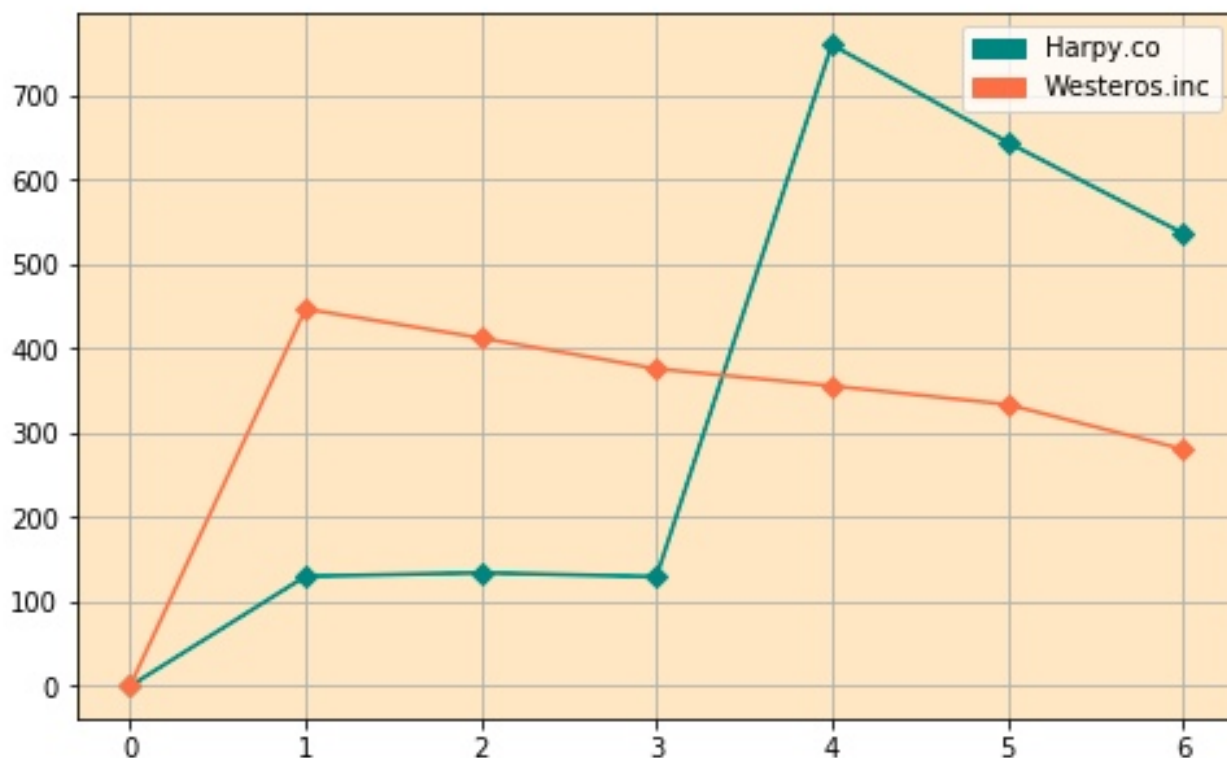


Третий этап:

Среднее количество сломанной продукции после каждого месяца.

Удаляем столбцы с номером кузнеца и количеством произведенных мечей (*unsullen.id* и *produced*) с помощью метода *production.date*.

Далее получим срок службы меча: $report.date = report.date - production.date$, и сгруппируем по этим столбцам методом **groupby** и просуммируем **sum**. Дальше уберем *productin.date* методом **drop**, сгруппируем с помощью **groupby** по столбцу *report.date* и возьмем среднее методом **mean**. После чего для красоты уберем столбец *report.date* (для красоты вывода).



Четвертый этап:

Оценка работы кузнецов каждой из компаний.

Удалим столбцы с датами производства и отчета (*report.date* и *production.date*) с помощью метода ***drop***) и переименуем столбец с номером кузнеца, идентифицировав его названием компании поставщика, применив метод ***rename***. Добавим столбец производительности для каждого кузнеца (*work accuracy*), подсчитанной по формуле:

$$work_accuracy = 100 - \frac{defects}{produced} * 100$$

Теперь мы можем сравнить лучших работников каждой из компаний, для этого к столбцу *work accuracy* применим метод ***max***.

Инструкции по запуску:

Библиотеки:

Язык программирования: python

Используемые библиотеки:

- pandas - библиотека, предоставляющая инструменты для обработки и анализа данных.
- csv - позволяет читать и записывать csv-файлы.
- matplotlib.pyplot - для отрисовки графиков.

Программы

Для написания программы был использован дистрибутив Anaconda.

Реализация и запуск кода осуществлялись в Jupyter Notebook.

Вклад участников

Битиев Алексей, Кулакова Мария, Сат Санчай - совместная работа над кодом, презентацией и readme.