|  |  |  |
| --- | --- | --- |
| Машинное обучение и анализ данных | | |
| Кафедра САПРиПК  Лабораторная работа №2  «Разведочный анализ данных» | ФИО студента | Козякин А.Д., Орлов Н.Е., Полешко Р.А. |
| Группа | ИВТ-363 |
| Дата выполнения |  |
| Дата отчета |  |
| ФИО преподавателя | Фокин Р.О. |
| Оценка |  |
| Подпись преподавателя |  |

Для выбранного или произвольного набора данных из репозитория

UC Irvine Machine Learning Repository необходимо выполнить следующие

вычисления:

1. найти среднее значение, медиану, экстремумы любого параметра;

2. произвести предобработку данных (удалить шумы, дубликаты)

3. найти дисперсию и стандартное отклонение любого параметра;

4. найти линейную зависимость между двумя любыми параметрами;

5. построить диаграммы рассеяния;

6. построить гистограмму, определить тип распределения одного из

параметров;

7. подвести итог (summary)

**Выбран датасет**: прогнозирование производительности комплекта данных сотрудников швейной промышленности.

**Информация о наборе данных**:

Швейная промышленность является одним из ключевых примеров промышленной глобализации этой современной эпохи. Это очень трудоемкая отрасль с большим количеством ручных процессов. Удовлетворение огромного мирового спроса на швейные изделия в основном зависит от производительности производства и доставки сотрудников в швейных компаниях. Таким образом, крайне желательно, чтобы лица, принимающие решения в швейной промышленности, отслеживали, анализировали и прогнозировали производительность рабочих команд на своих фабриках. Этот набор данных может быть использован для целей регрессии путем прогнозирования диапазона производительности (0-1) или для целей классификации путем преобразования диапазона производительности (0-1) в различные классы.

**Информация об атрибутах**:

1) date: Дата в ММ-ДД-ГГГГ.

2) day: День недели.

3) quarter: Часть месяца. Месяц был разделен на четыре квартала.

4) department: Ассоциированный отдел с экземпляром.

5) team: Связанный номер команды с экземпляром

6) no\_of\_workers: Количество работников в каждой команде

7) no\_of\_style\_change: Количество изменений в стиле конкретного продукта

8) targeted\_productivity: Целевая производительность, установленная Органом для каждой команды на каждый день.

9) smv: Стандартное значение минуты, это выделенное время для задачи

10) wip: Незавершенное производство. Включает количество незавершенных позиций для продуктов

11) over\_time: Представляет собой количество сверхурочной работы каждой команды в минутах

12) incentive: Представляет собой сумму финансового стимула (в БРЭ), который позволяет или мотивирует определенный курс действий.

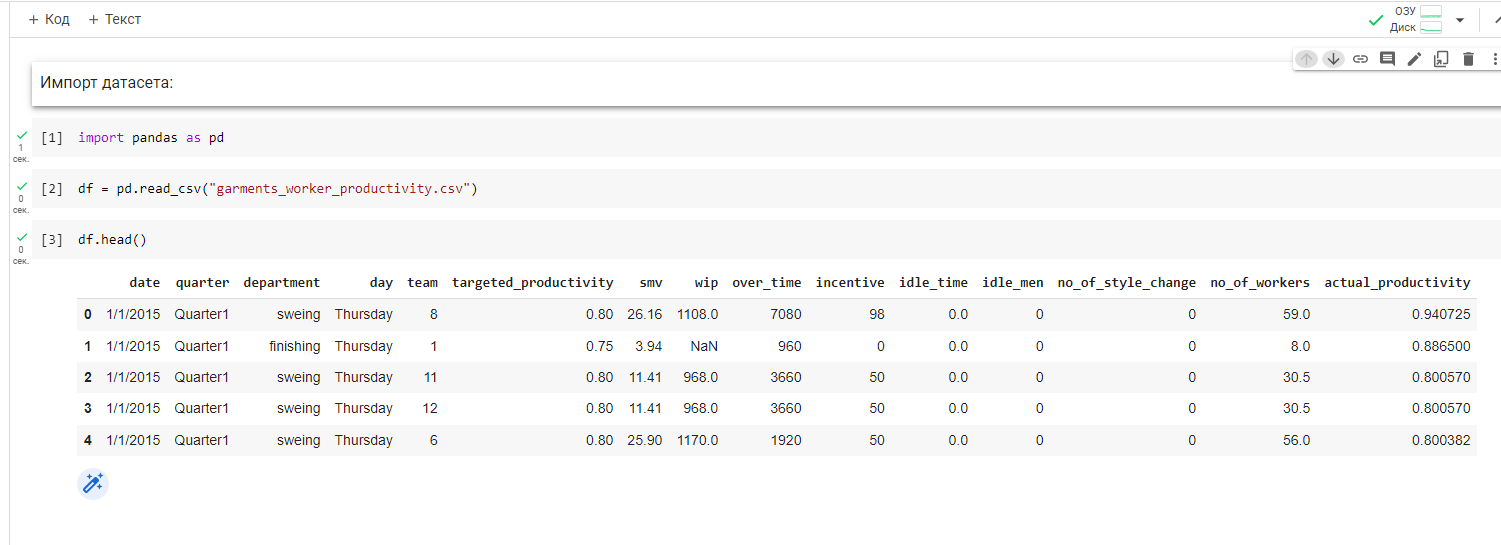
13) idle\_time: Количество времени, когда производство было прервано по нескольким причинам

14) idle\_men: Количество рабочих, которые простаивали из-за перерыва

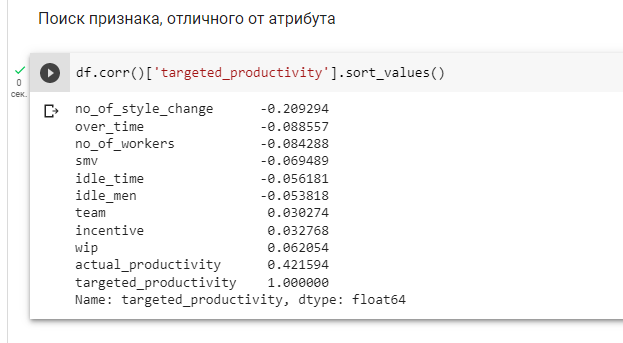
в производстве

15) actual\_productivity: фактический процент производительности, которую обеспечивали рабочие. Он колеблется от 0-1.

Импортируем датасет:



Воспользуемся методом Pandas corr() для поиска признака, отличного от атрибута(например, targeted\_productivity), который наиболее коррелирует с целевой производительностью:



Дальше рассматриваем столбцы actual\_productivity и targeted\_productivity.

**1. найти среднее значение, медиану, экстремумы любого параметра:**

Изображение выглядит как текст

Автоматически созданное описание

**2. произвести предобработку данных (удалить шумы, дубликаты):**

Изображение выглядит как текст, компьютер, внутренний, снимок экрана

Автоматически созданное описание

**3. найти дисперсию и стандартное отклонение любого параметра:**

Изображение выглядит как текст

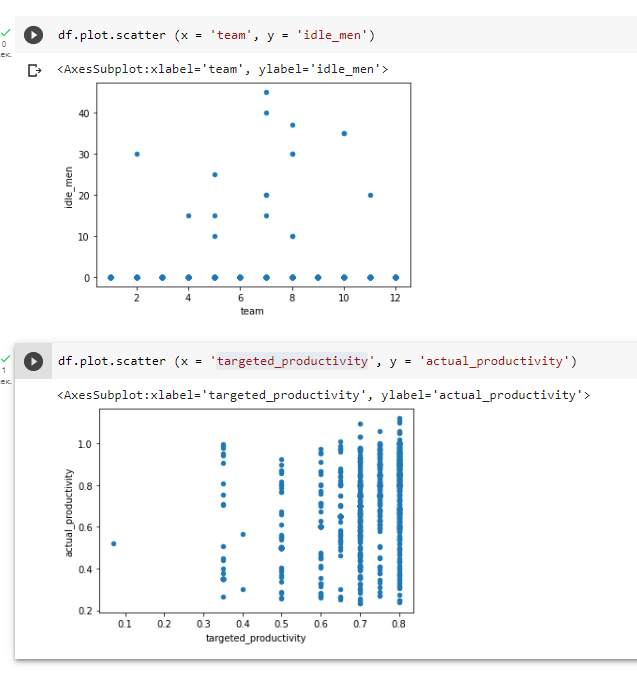
Автоматически созданное описание

**4. найти линейную зависимость между двумя любыми параметрами:**

**Изображение выглядит как текст

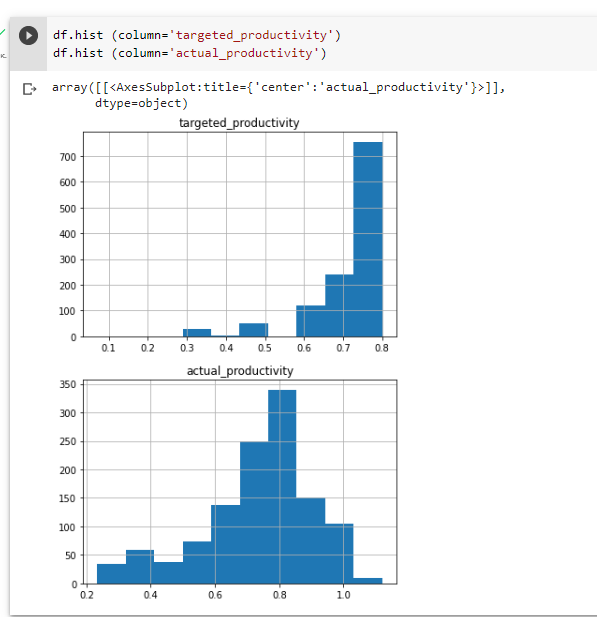
Автоматически созданное описание**

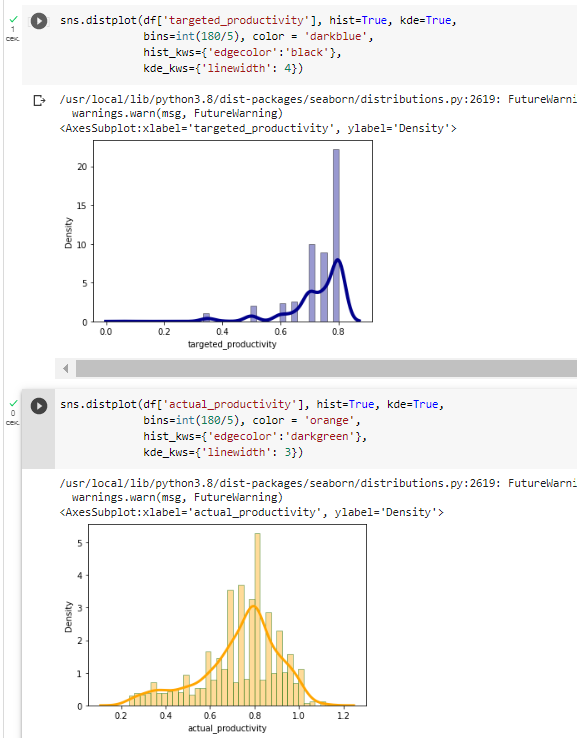
**5. построить диаграммы рассеяния:**

****

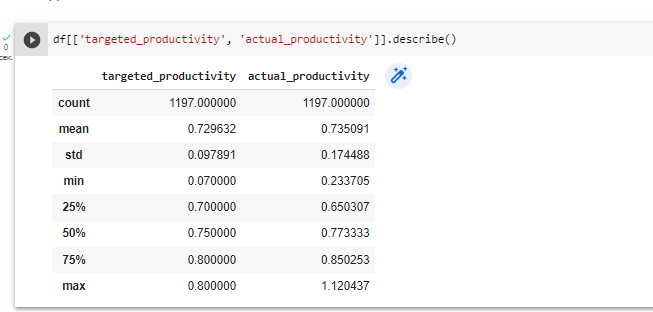
**6. построить гистограмму, определить тип распределения одного из**

**параметров:**

****

****

**7. подвести итог (summary).**

****

