МГТУ им. Н. Э. Баумана, кафедра ИУ5

курс “Технологии машинного обучения”

Лабораторная работа №1

# «[Разведочный анализ данных. Исследование и визуализация данных](https://github.com/ugapanyuk/ml_course/wiki/LAB_EDA_VISUALIZATION)»

ВЫПОЛНИЛ:

Фонканц Р.В.

Группа: ИУ5-61Б

ПРОВЕРИЛ:

Гапанюк Ю.Е.

Москва 2020

**Цель лабораторной работы:** изучение различных методов визуализации данных.

**Задание:**

* Выбрать набор данных (датасет).

Для лабораторных работ не рекомендуется выбирать датасеты большого размера.

* Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

* Сформировать отчет и разместить его в своем репозитории на github.

**Выполненная работа:**

В сфере добычи и продажи драгоценных камней необходимо отслеживать актуальные цены и прочие данные для успешного ведения бизнеса. Давайте проанализируем такие данные.

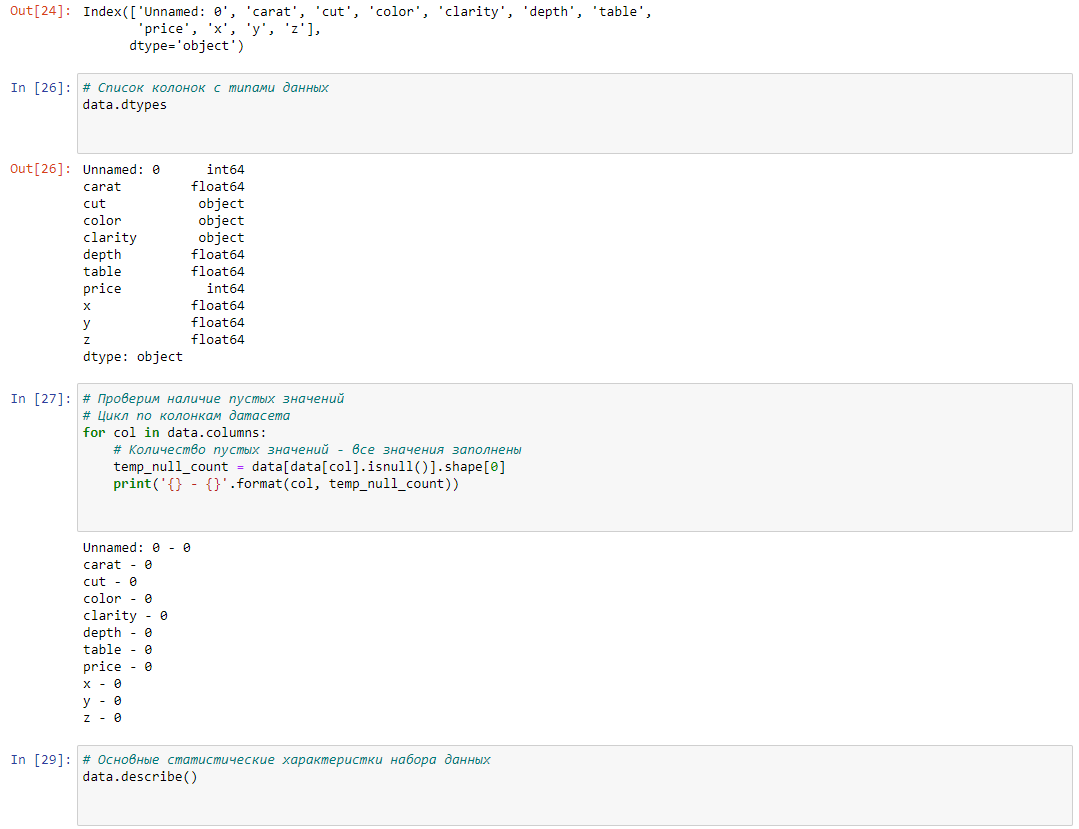
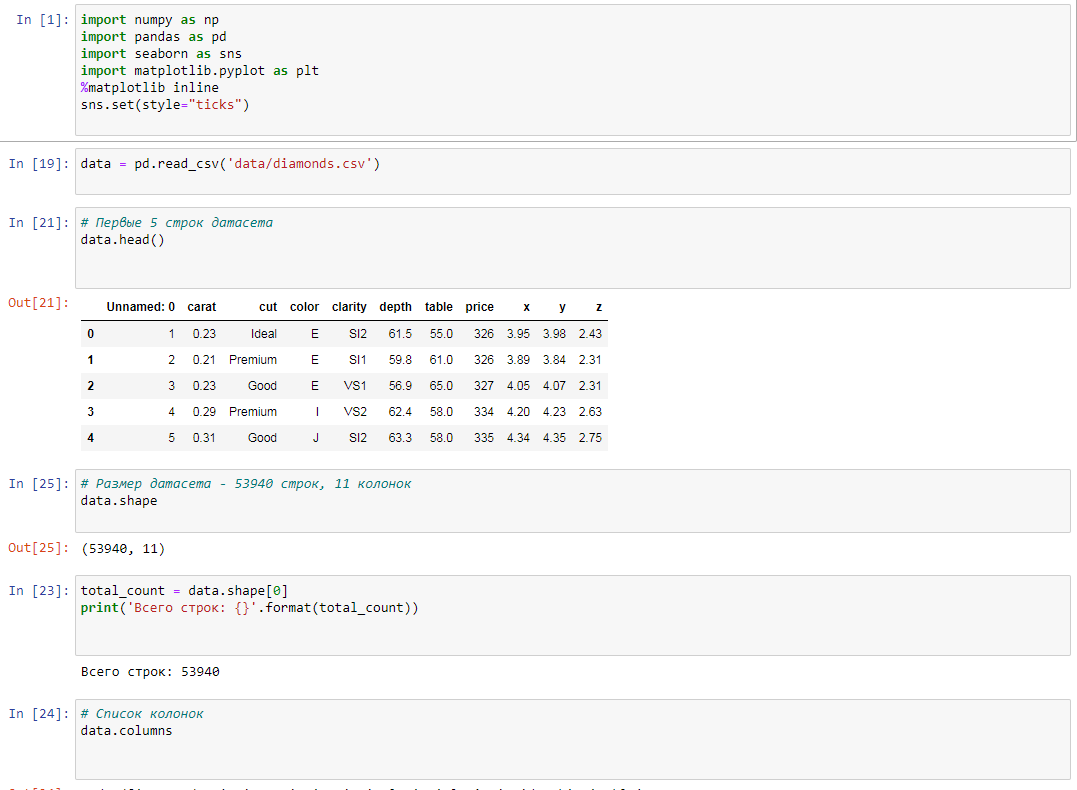
В качестве датасета для данной работы, были выбраны данные о более, чем полусотни тысяч алмазов с различными характеристиками:

* <https://www.kaggle.com/shivam2503/diamonds/data>

Данный набор данных содержит цены и другие атрибуты почти 54 000 алмазов, такие, как:

* **price**: цена в долларах США (\$326--\$18,823).
* **carat**: масса алмаза (0,2-5,01).
* **cut**: качество реза (нормальное, хорошее, очень хорошее, премиумное, идеальное)
* **color**: цвет алмаза, от J (худший) до D (лучший)
* **clarity**: прозрачность алмаза (I1 (худший), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (лучший))
* **x**: длина в мм (0--10.74)
* **y**: ширина в мм (0--58,9)
* **z**: глубина в мм (0--31,8)
* **depth**: общая глубина в процентах = z / среднее значение (x, y) = 2 \* z / (x + y) (43 -- 79)
* **table**: ширина вершины алмаза относительно самой широкой точки (43--95)

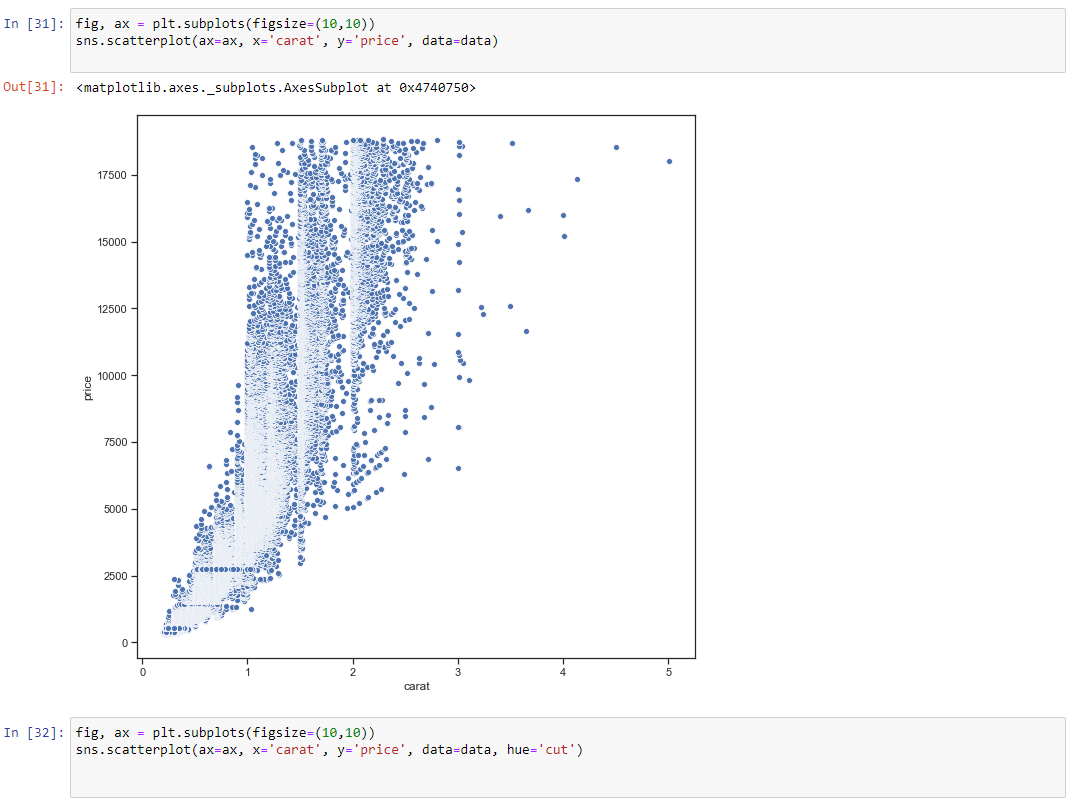
**Текст программы и экранные формы с примерами выполнения программы**:

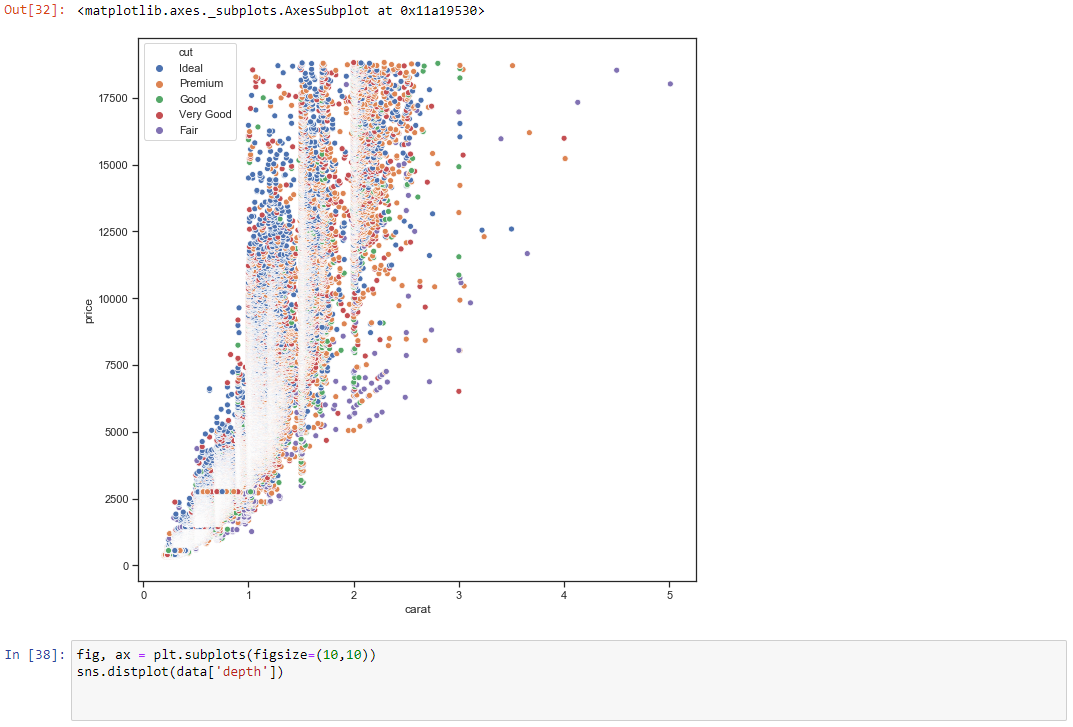
 

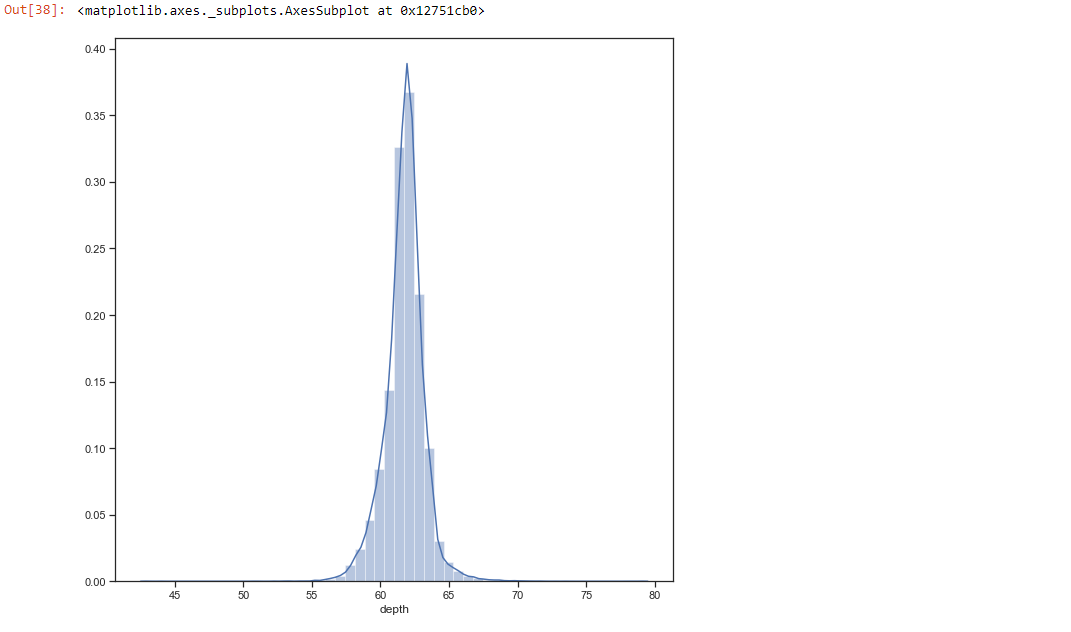
**Визуальное исследование датасета.**

Визуально исследовать наш датасет мы будем при помощи диаграмм рассеивания и гистограмм. С помощью диаграммы рассеивания мы сможем оценить существуют ли отношения или корреляция между этими двумя переменными, например, для максимальных и минимальных значений.

Можно видеть, что между полями присутствует зависимость, которая визуально близка к экспоненциальной.

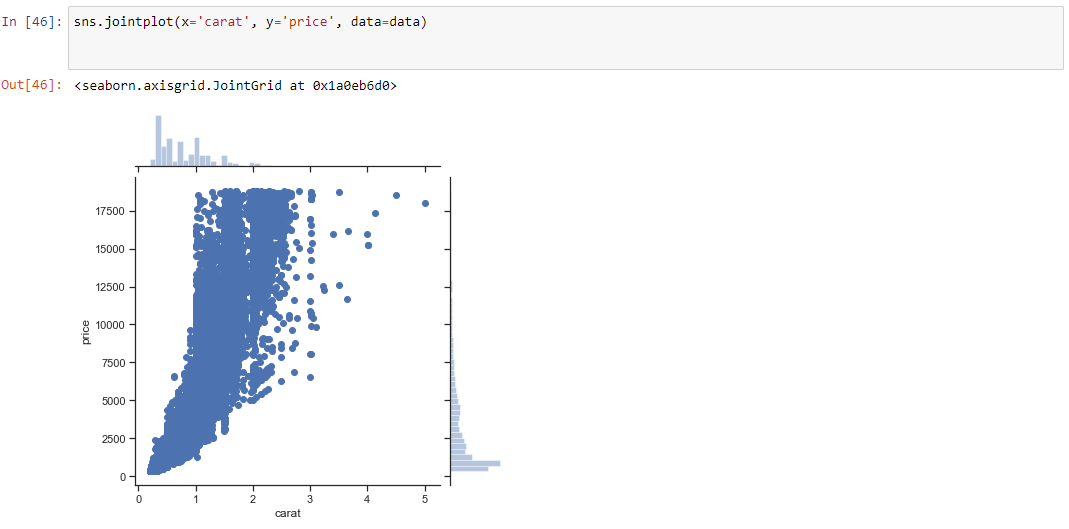






С помощью гистограммы мы можем оценить плотность вероятности распределения данных для поля “price”. Изображение на данной гистограмме соответствует практически идеальному нормальному распределению (распределению Гаусса).

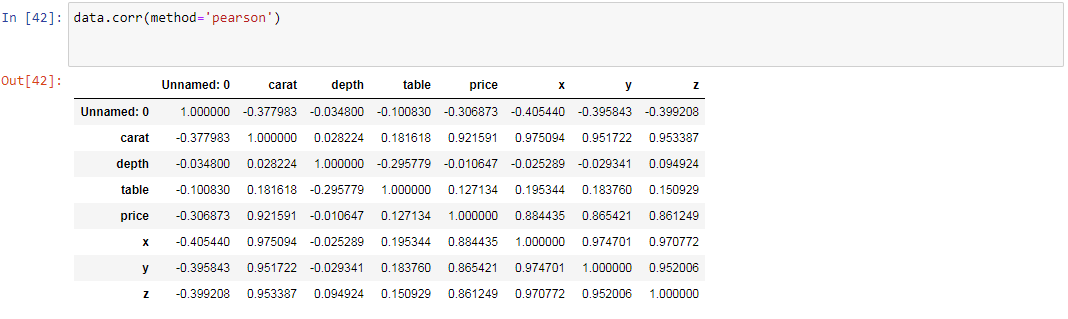
Также для наглядности можно построить jointplot - комбинацию гистограмм и диаграмм рассеивания.

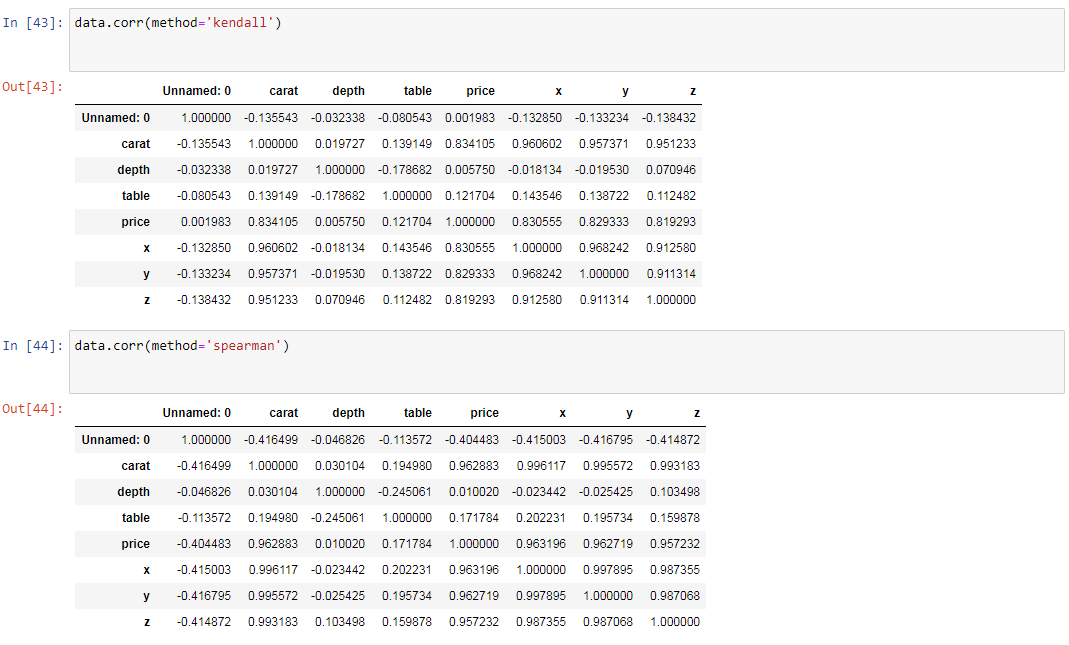


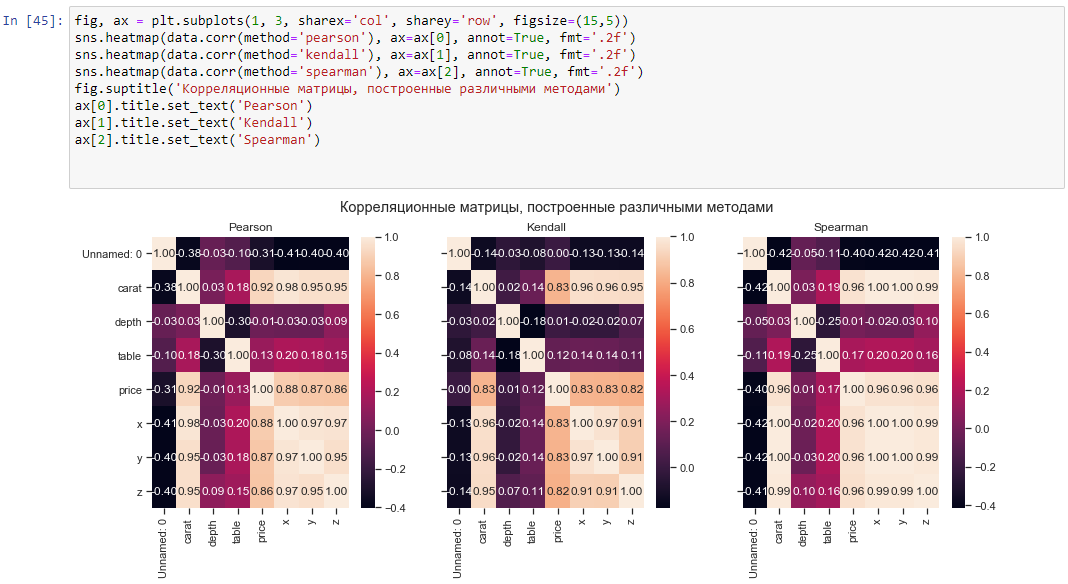
**Информация о корреляции признаков**

Проверка корреляции признаков позволяет решить две задачи:

1. Понять какие признаки (колонки набора данных) наиболее сильно коррелируют с целевым признаком. Именно эти признаки будут наиболее информативными для моделей машинного обучения. Признаки, которые слабо коррелируют с целевым признаком, можно попробовать исключить из построения модели, иногда это повышает качество модели. Нужно отметить, что некоторые алгоритмы машинного обучения автоматически определяют ценность того или иного признака для построения модели.
2. Понять какие нецелевые признаки линейно зависимы между собой. Линейно зависимые признаки, как правило, очень плохо влияют на качество моделей. Поэтому если несколько признаков линейно зависимы, то для построения модели из них выбирают какой-то один признак.







Корреляционная матрица содержит коэффициенты корреляции между всеми парами признаков, она симметрична относительно главной диагонали. На главной диагонали расположены единицы (корреляция признака самого с собой). На основе корреляционной матрицы можно сделать выводы, которые помогут с решениями задач или для определения ненужных в выборке значений.

Данный отчет размещен в репозитории: <https://github.com/Rauger474/TMO>