ПРОЕКТНАЯ РАБОТА

**ВЛИЯНИЕ ОСНОВНЫХ ХАРАКТЕРИСТИК БРИЛЛИАНТОВ НА ИХ СТОИМОСТЬ**

Выполнила:

Потапова Софья Андреевна

# 

# ЧАСТЬ 1 (Supervised Learning)

## Начальный анализ данных

Используемый датасет, который был взят из библиотеки R: ggplot2, содержит цены на более чем 50000 бриллиантов круглой огранки. Типы данных – integers, float, strings. Переменные следующие:

* price: цена в долларах US ($326-$18,823);
* carat: вес бриллианта (0.2-5.01);
* cut: качество огранки (Fair, Good, Very Good, Premium, Ideal);
* color: цвет бриллианта (from D – best to J – worst);
* clarity: чистота бриллианта (I1 – worst, SI2, SI1, VS2, VS1, VVS2, VVS1, IF – best);
* 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 (43-79)

;

* table: ширина вершины бриллианта относительно самой широкой точки (43-95).

Изначально выполнили предварительный анализ данных: проверили на наличие пропусков (в итоге выявлено ни одного) и перешли к выбросам. Для этого воспользовались методом Z-score и методом межквартильного диапазона (IQR). Благодаря этому избавились от выбросов и получили конечный набор данных, состоящий из 10 переменных и 46669 наблюдений.

На этом же шаге после подготовительного этапа обработки данных была выведена описательная статистика для каждой используемой в анализе числовой переменной (Таблица 1). В среднем бриллиант стоит $3000 и весит 0.7 каратов. При этом его длина, ширина и глубина относительно небольшие: 5.5, 5.5 и 3.4 соответственно. Невысокие средние значения в данном случае могут свидетельствовать, что в выборке преобладающее большинство – небольшие бриллианты с относительно низкими ценами.

После приступили к разведочному анализу данных (EDA). На данном этапе провели одномерный анализ: построили столбчатые диаграммы для каждой категориальной переменной (cut – качество огранки, color – цвет бриллианта, clarity – чистота бриллианта). Нетрудно заметить, что в выборке преобладающее большинство – чуть больше 40% – это бриллианты идеального качества огранки, а бриллианты премиум и очень хорошего качества содержатся в выборке приблизительно в одном количестве – 25% и 23%. (Приложение 1). В отличие от дифференциации в количестве на каждый тип огранки, у цвета подобного не наблюдается. Категории цвета, относящиеся к среднему (E, F, G, H), составляют примерно 80% всей выборки, в то время как самого лучшего цвета (D) – всего 13%, а самые худшие составляют только около одной десятой всей выборки (Приложение 2).

Таблица 1

Описательная статистика используемых переменных

| index | carat | depth | table | price | x | y | z |
| --- | --- | --- | --- | --- | --- | --- | --- |
| count | 46669 | 46669 | 46669 | 46669 | 46669 | 46669 | 46669 |
| mean | 0,70 | 61,80 | 57,24 | 3023,78 | 5,53 | 5,53 | 3,42 |
| std | 0,36 | 1,08 | 2,01 | 2634,10 | 0,97 | 0,96 | 0,60 |
| min | 0,20 | 59 | 52 | 326 | 3,73 | 3,68 | 1,53 |
| 25% | 0,37 | 61,20 | 56 | 882 | 4,63 | 4,63 | 2,85 |
| 50% | 0,59 | 61,90 | 57 | 2016 | 45021 | 5,41 | 3,33 |
| 75% | 1,01 | 62,50 | 59 | 4548 | 6,39 | 6,39 | 3,97 |
| max | 1,96 | 64,60 | 63,50 | 11228 | 44934 | 8,21 | 45082 |

Похожая тенденция видна и на графике чистоты бриллианта: можно сказать, что по мере улучшения чистоты количество бриллиантов уменьшается, но при этом в выборке содержится всего 459 наблюдений, относящихся к категории самой худшей чистоты – I1 (Приложение 3).

Так как по теоретическим соображениям стоит предполагать корреляцию между переменными, на следующем этапе были построены диаграммы рассеивания, показывающие зависимость между весом бриллианта и его ценой, весом бриллианта и его длиной, ценой бриллианта и его длиной (см. код). Благодаря этому выявили, что корреляция данных присутствует: по мере увеличения одной переменной происходит увеличение другой; а также: имеет смысл рассматривать зависимость между данными переменными.

Далее был проведен одномерный анализ числовых переменных, то есть построены гистограммы, чтобы понять принадлежат ли переменные какому-либо виду распределения данных (Приложение 4). Сделали вывод, что данные распределены довольно хаотично.

В качестве завершающего этапа EDA был проведен многомерный анализ. В качестве примера категориального многомерного анализа в данном случае была приведена числовая переменная цена бриллианта (Приложение 5, Приложение 6, Приложение 7). Смотря на столбчатые диаграммы, нельзя заключить, что существует прямая закономерность: чем выше качество огранки, лучше цвет или чистота, тем выше цена. Это связано с теоретическим соображением о том, что на цену бриллианта влияет несколько факторов в совокупности, то есть нельзя делать вывод на основе только одного из них. Похожая ситуация наблюдается среди других числовых переменных, но так как караты и параметры размера не влияют на огранку, цвет и чистоту, эту зависимость не следует учитывать, что также подкреплено теорией.

По завершению был также построен график пар, чтобы еще раз убедиться в характере влияния одних переменных на другие (см. код). Кроме того, для окончательного вывода по поводу корреляции и возможного влияния переменных была построена корреляционная матрица (Рисунок 1). Благодаря этому была подтверждена высокая положительная корреляция между весом бриллианта и его ценой и параметрами размера, а также ценой и параметрами размера.

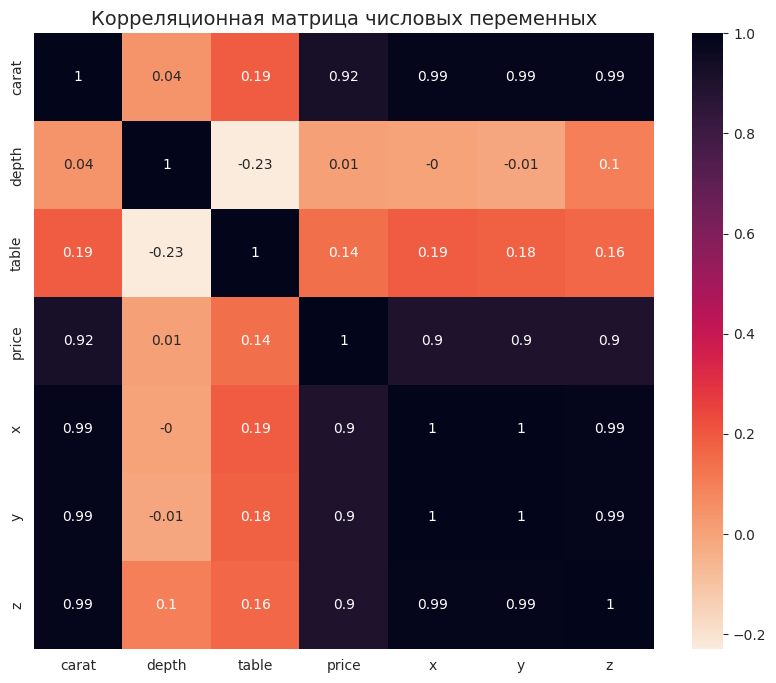


Рис.1. Корреляционная матрица числовых переменных

Таким образом, благодаря проведенному EDA, можем сделать вывод, что благоразумнее всего будет исследование влияния качества огранки бриллианта, его цвета и чистоты на параметры размера, вес и цену бриллианта. Другими словами, можно ли предсказать качество огранки бриллианта, его цвет или чистоту, зная вес и цену бриллианта, а также его длину, ширину и глубину. Это имеет смысл, а также практическую ценность. Если представить ситуацию, в которой индивид хочет купить бриллиант хорошего качества (лучшей категории), но учесть, что он является новичком в данной тематике и прочел только минимум необходимой информации о бриллиантах и их характеристиках. Из этого следует, что он не может на практике точно определить, какой бриллиант хорош, а от какого стоит отказаться. Все, чем он может эффективно оперировать, решая, какой бриллиант приобрести, – это цена и вес бриллианта, а также (в некоторых случаях) – длина, ширина и глубина. При этом, он не может судить о качестве бриллианта по главным признакам его оценки: качество огранки, цвет и чистота, так как еще не обладает достаточной для этого информацией и опытом. По этой причине, задача данного исследования, понять: можно ли приобрести качественный бриллиант, не обладая и не понимая главную информацию о нем. Стоит также отметить, что, так как цель – приобрести лучший бриллиант, анализ фокусируется на самых лучших категориях каждого признака, то есть качество огранки было выбрано Ideal и Premium, цвет – D и E, чистота – IF и VVS1.

## Классификация методом kNN

Первым шагом на данном этапе был выбор переменных. Так как цель – ответить на вопрос о том, можно ли предсказать качество огранки, цвет и чистоту бриллианта, зная его цену, вес и параметры размера, а также основываясь на теоретических соображениях и результатах EDA, в качестве features были выбраны: цена бриллианта, вес бриллианта, длина, ширина и глубина бриллианта, а в качестве target: качество огранки, цвет и чистота бриллианта. Изначально были проведены базовые процедуры: разделение выборки на обучающую и тестовую, нормализация данных (во избежание отклонения результатов), после чего приступили непосредственно к классификации. Таким образом, должно получится три классификации по kNN (по причине трех target variables).

Классификацию начали с выяснения, можно ли предсказать качество огранки бриллианта помощью его цены, веса и параметров размера. Для начала выставили значение гиперпараметра k, равное восьми. Получили следующий прогноз:

Predictions: [0 1]; Actual Values: [0 1],

где 1 – качество огранки Premium, а 0 – Ideal,

причем точность прогноза составляет 75%. Далее воспользовались перекрестной проверкой (Cross Validation), чтобы смоделировать k, так как рандомный подбор гиперпараметра чаще всего нежелателен из-за недо- или переоценивания модели и прогнозов. Благодаря этому выяснили, что оптимальное значение гиперпарметра составляет 28, а точность прогноза не изменилась.

Те же шаги были применены в случае цвета бриллианта в качестве target variables, причем оптимальное значение k в данном случае было равно 25. Был получен прогноз:

Predictions: [0 1]; Actual Values: [0 0],

где 1 – цвет категории E, а 0 – D,

точность которого составила 61% и не изменилась после моделирования значения k.

Наконец, проанализировали, можно ли предсказать чистоту бриллианта. Был получен следующий прогноз, оптимальное значение k которого составило 29:

Predictions: [0 1]; Actual Values: [0 0],

где 1 – чистота категории VVS1, а 0 – IF,

с точностью 68%, также не изменилась после использования лучшего значения гиперпараметра.

## Линейная и Логистическая Регрессии

Решая задачу линейной регрессии, в качестве зависимой переменной была выбрана цена как более интересующая среди числовых переменных, а в качестве объясняющих – вес бриллианта, его длина, ширина и глубина, качество огранки, цвет и чистота.

Благодаря построенной модели было выявлено, что по мере увеличения категории каждого признака увеличивается и степень влияния (Таблица 2). Чистота бриллианта IF и VVS1, качество огранки Ideal и Premium, цвет D и E имеют наибольшее влияние среди своих признаков, то есть чем выше цена бриллианта, тем выше его чистота, цвет и качество огранки. Из этого следует, что вероятнее всего при покупке дорогого бриллианта, его характеристики будут высокими, то есть сам бриллиант будет качественным.

Таблица 2

Линейная регрессия

| Переменная | Коэффициент |
| --- | --- |
| intercept | -530.64 |
| Вес бриллианта | 10845.08 |
| Чистота бриллианта (IF) | 3460.34 |
| Чистота бриллианта (VVS2) | 3387.57 |
| Чистота бриллианта (VVS1) | 3352.43 |
| Чистота бриллианта (VS1) | 3098.71 |
| Чистота бриллианта (VS2) | 2821.31 |
| Чистота бриллианта (SI1) | 2365.95 |
| Чистота бриллианта (SI2) | 1651.81 |
| Ширина бриллианта | 715.95 |
| Качество огранки бриллианта (Ideal) | 425.02 |
| Качество огранки бриллианта (Premium) | 324.89 |
| Качество огранки бриллианта (Very Good) | 282.20 |
| Качество огранки бриллианта (Good) | 198.39 |
| Цвет бриллианта (E) | -153.98 |
| Цвет бриллианта (F) | -190.67 |
| Цвет бриллианта (G) | -324.76 |
| Цвет бриллианта (H) | -684.66 |
| Глубина бриллианта | -959.11 |
| Цвет бриллианта (I) | -1082.80 |
| Длина бриллианта | -1322.00 |
| Цвет бриллианта (J) | -1725.76 |

Необходимо добавить, что получившаяся модель имеет довольно высокую объяснительную способность – 93%, а также относительно неплохую прогнозную способность:

Predictions: [5747.38470686 5423.51230183]; Actual Values: [6036 5004]

Далее перешли к логистической регрессии, зависимые переменные которой соответствовали интересующим признакам (качество огранки, цвет, чистота). В данном случае анализ основывался на том, что были взяты исключительно по две лучших категории каждого признака.

Для первой логистической регрессии с зависимой переменной cut (качество огранки), была рассчитана вероятность того, будет ли приобретенный бриллиант Premium или Ideal качества, если учесть, что нам известны его цена, вес, длина, ширина и глубина. В 68% случаев модель сделала верный прогноз, другими словами: было верно предсказано 5410 значений, и 2581 – неверно. Также была построена ROC-кривая, чтобы понять насколько точно модель предсказывает результаты (Рисунок 2).

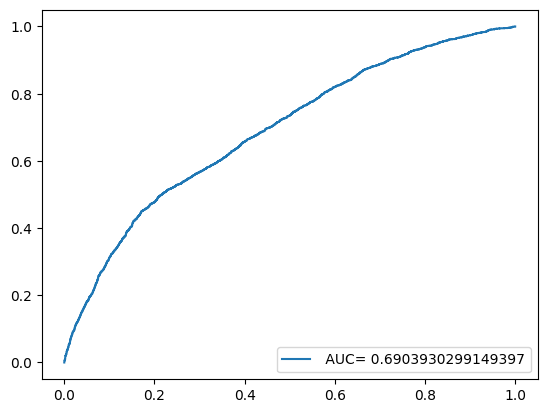


Рис.2. ROC-кривая при зависимой переменной cut

Следующим была рассчитана вероятность того, будет ли приобретенный бриллиант D или E категории цвета, если известны его цена, вес, длина, ширина и глубина. В 59% случаев модель сделала верный прогноз, другими словами: было верно предсказано 2181 значений, и 1530 – неверно. Также была построена ROC-кривая, чтобы понять насколько точно модель предсказывает результаты (Рисунок 3).

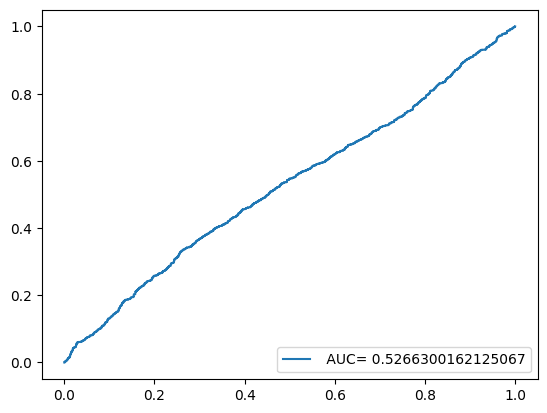


Рис.3. ROC-кривая при зависимой переменной color

Наконец, была рассчитана вероятность того, будет ли приобретенный бриллиант IF или VVS1 категории чистоты при его известных значений цены, веса, длины, ширины и глубины. В 68% случаев модель сделала верный прогноз, другими словами: было верно предсказано 856 значений, и 404 – неверно. Также была построена ROC-кривая, чтобы понять насколько точно модель предсказывает результаты (Рисунок 4).

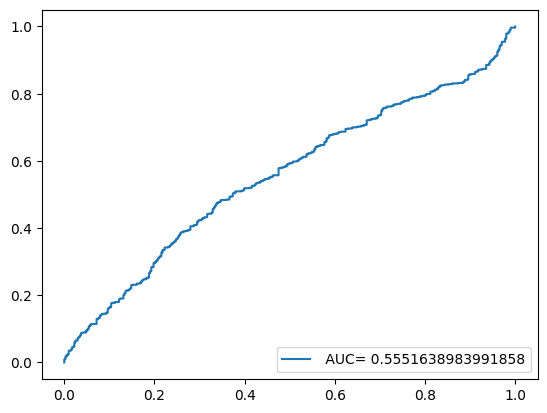


Рис.4. ROC-кривая при зависимой переменной clarity

## Random Forest

Заключительным этапом приступили к методу случайного леса. Так же как и в логистической регрессии, зависимые переменные были определены как качество огранки бриллианта, цвет и чистота. Сначала были также предсказана вероятность того, будет ли приобретенный бриллиант Premium или Ideal качества, зная его цену, вес, длину, ширину и глубину. Модель смогла объяснить только 12% данных, но было выяснено, какие переменные являются наиболее важными в предсказании качества огранки (Рисунок 5).

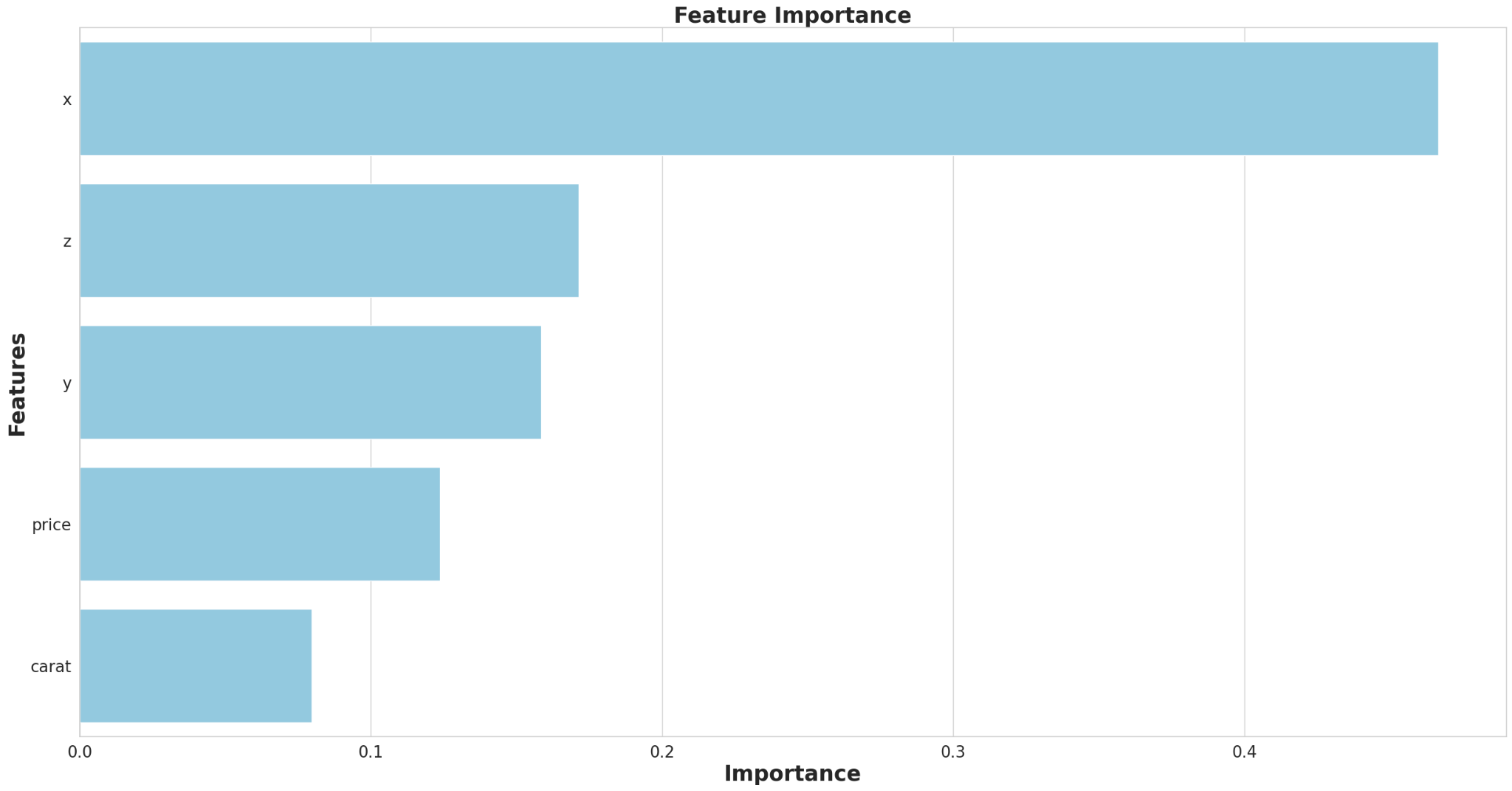


Рис.5. Визуализация важности признаков при предсказании переменной cut

То есть наиболее весомой переменной в определении качества огранки является длина бриллианта, а остальные переменные относительно длины в данном случае приблизительно одинаково значимы (Таблица 3).

Далее предсказали вероятность того, будет ли приобретенный бриллиант D или E категории цвета, зная его цену, вес, длину, ширину и глубину. Модель смогла объяснить только 5% данных, но было выяснено, какие переменные являются наиболее важными в предсказании цвета (Рисунок 6). То есть наиболее весомой переменной в определении цвета является цена бриллианта, а остальные переменные относительно длины в данном случае приблизительно одинаково значимы (Таблица 3).

Заключительным шагом была предсказана вероятность того, будет ли приобретенный бриллиант IF или VVS1 категории чистоты, если известны его цена, вес, длина, ширина и глубина. Модель смогла объяснить только 10% данных, но было выяснено, какие переменные являются наиболее важными в предсказании чистоты (Рисунок 7). То есть наиболее весомой переменной в определении чистоты является цена бриллианта, а остальные переменные относительно длины в данном случае приблизительно одинаково значимы (Таблица 3).

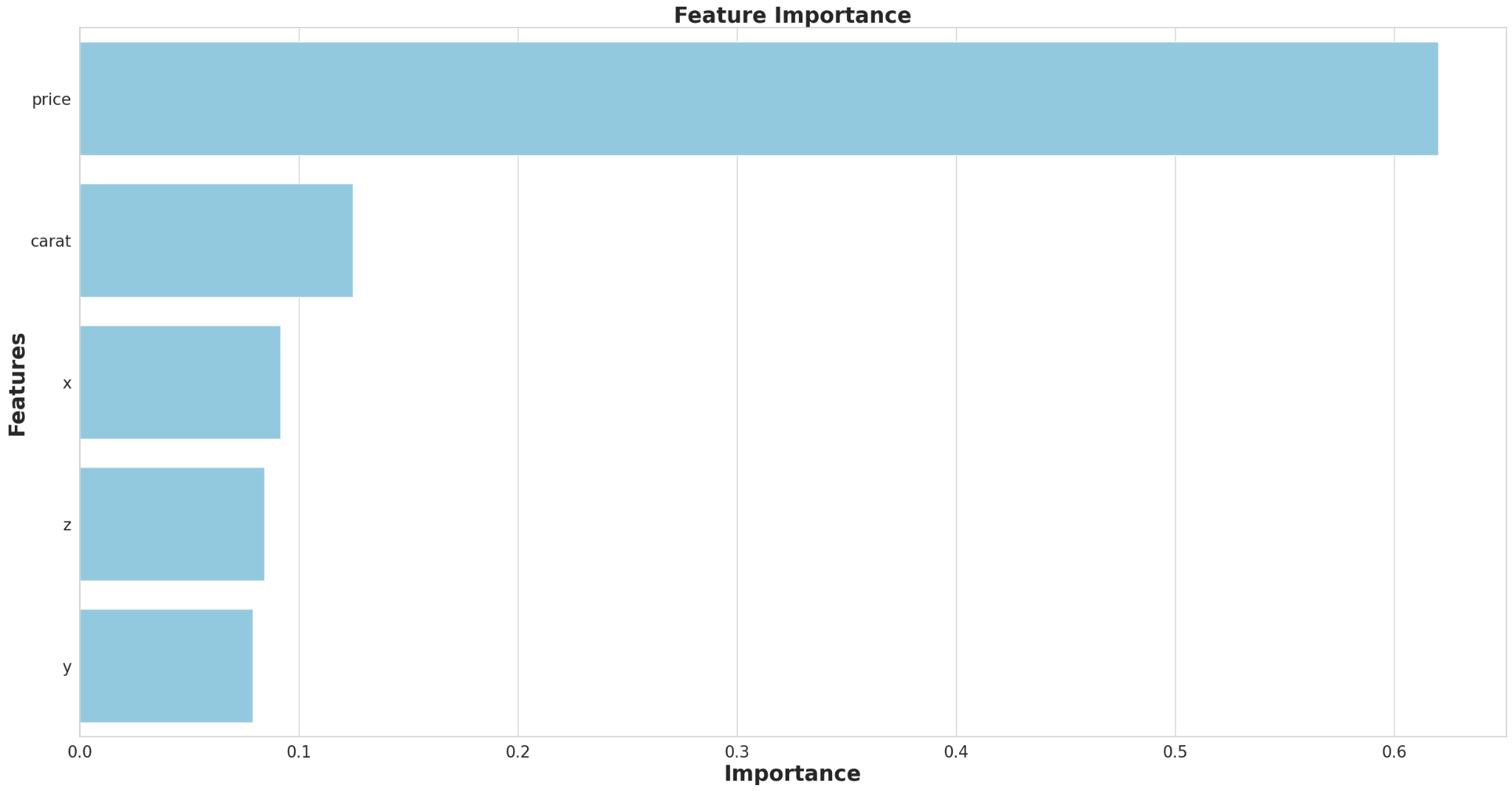


Рис.6. Визуализация важности признаков при предсказании переменной color

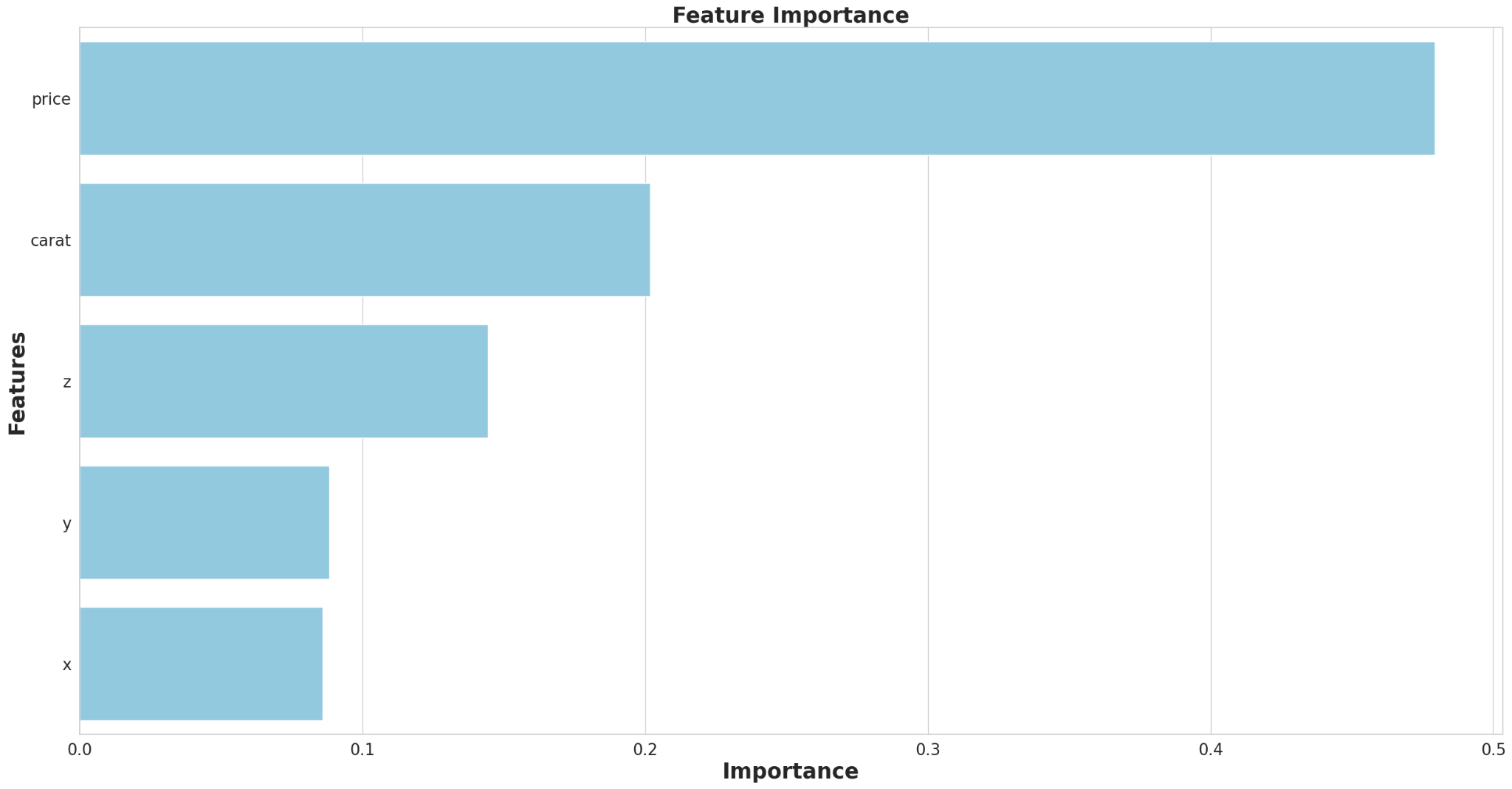


Рис.7. Визуализация важности признаков при предсказании переменной clarity

Таблица 3

Важность признаков

| Предсказание огранки | | Предсказание цвета | | Предсказание чистоты | |
| --- | --- | --- | --- | --- | --- |
| Переменные | Важность | Переменные | Важность | Переменные | Важность |
| x | 0.47 | price | 0.62 | price | 0.48 |
| z | 0.17 | carat | 0.13 | carat | 0.20 |
| y | 0.16 | x | 0.09 | z | 0.15 |
| price | 0.12 | z | 0.08 | y | 0.09 |
| carat | 0.08 | y | 0.08 | x | 0.09 |

## Вывод

Сравнивая все получившиеся модели при помощи значения среднеквадратической ошибки, можно сказать, что лучшую лучшее соответствие данным показал именно метод случайного леса, с RMSE, равными 0.45, 0.48 и 0.44 для качества огранки, цвета и чистоты, соответственно. При этом данный метод имеет самую низкую точность среди всех моделей: 12%, 5% и 10%. Напротив, у линейной регрессии самая высокая точность прогноза, что указывает на ярко выраженную зависимость между ценой бриллианта и выбранными объясняющими переменными. Учитывая содержательную часть, то есть цель данного исследования, линейная регрессия не может в полной мере ответить на заданный вопрос, но может стать промежуточным этапом, который дает понимание, какие переменные имеют большее влияние на цену. А так как цена является более понятной в плане определения качества бриллианта (для заданной гипотетической ситуации), можно сделать вывод насчет огранки, цвета и чистоты. В этом случае появляется понимание, стоит ли подразумевать, что высокая цена гарантирует и высокое качество огранки, цвета и чистоты. С этой логикой целесообразно перейти к методу kNN, модели логистической регрессии и методу случайного леса, и сравнивать данные модели. По этой причине, при выборе между моделями стоит остановиться либо на методе k-ближайших соседей, либо на логистической регрессии, так как оба они имеют лучшую предсказательную силу; а далее выбирая между ними, смотреть на детали и учитывать, что предпочтительнее в той или иной ситуации. В данном исследовании может быть хорош как метод k-ближайших соседей, так и логистическая регрессия.

# ЧАСТЬ 2 (Unsupervised Learning)

## Иерархическая кластеризация

## Перед тем как приступить к решению задачи иерархической кластеризации выведем таблицу по нашим с пятью случайными строками исходной таблицы (Таблица 4).

Таблица 4

Таблица с переменными

| Переменные |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Вес бриллианта | 1.03 | 0.31 | 0.85 | 0.9 | 0.52 |
| Качество огранки | Very Good | Ideal | Ideal | Premium | Very Good |
| Цвет | F | D | E | E | I |
| Чистота | VVS2 | VS2 | SI2 | SI1 | VS1 |
| Общая глубина | 61.5 | 62.2 | 60.1 | 60.8 | 61.5 |
| Ширина вершины | 55.0 | 55.0 | 56.0 | 59.0 | 56.3 |
| Цена | 8998 | 734 | 3787 | 4463 | 1262 |
| Длина | 6.48 | 4.31 | 6.17 | 6.19 | 5.16 |
| Ширина | 6.57 | 4.34 | 6.18 | 6.24 | 5.19 |
| Глубина | 4.01 | 2.69 | 3.71 | 3.78 | 3.18 |

В таблице присутствовал неизвестный столбец (Unnamed), который было решено удалить. Далее необходимо было понять тип данных, в основном это значения в формате float64, int64,object. Однако, нам пришлось изменить формат данных на более легкий( float32, int32), а также вынужденно уменьшить выборку.

Следующим нашим шагом был выбор столбцов, которые мы будем использовать в дальнейшем анализе. Были выбраны количественные (int, float) столбцы, по которым выполним кластеризацию.  
  
 Для того, чтобы понять специфику данных, построили график рассеивания и гистограммы. По данным графикам рассеивания можно сделать вывод, что присутствует линейная зависимость между переменными. Гистограмма carat указывает на тот факт, что бриллиантов с меньшим весом намного больше, чем с максимально возможным, гистограмма price - более дешевые бриллианты преобладают над дорогими, гистограммы x,y, z, отвечающие за длину, ширину, глубину соответственно, получились достаточно похожими.

Также была построена матрица корреляций, которая показывает, что присутствует положительная взаимосвязь между переменными.

Таблица 5

Матрица корреляций

|  | carat | price | x | y | z |
| --- | --- | --- | --- | --- | --- |
| carat | 1 | 0,92 | 0,98 | 0,98 | 0,98 |
| price | 0,92 | 1 | 0,89 | 0,89 | 0,88 |
| x | 0,98 | 0,89 | 1 | 1,00 | 0,99 |
| y | 0,98 | 0,89 | 1,00 | 1 | 0,99 |
| z | 0,98 | 0,88 | 0,99 | 0,99 | 1 |

Далее проводим нормирование значений, чтобы привести значения к диапазону от 0 до 1. Это необходимо для того, чтобы центры кластеров не смещались вместе с большими значениями.

Для выбора количества кластеров для построения дендрограммы мы воспользовались методом локтя. График показывает как уменьшается расстояние между кластерами. Голубая линия показывает абсолютное изменение расстояния, а желтая - динамику изменений.

По графику видно, что снижение происходит примерно на значении 5, значит определяем количество кластеров как 5. Несмотря на то, что рекомендуемое количество кластеров это 2, мы приняли решение ориентироваться больше на график (Рисунок 8).

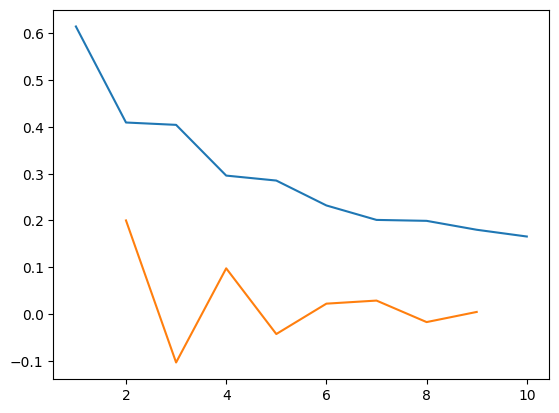
График “Метод локтя”  


Рис.8. Визуализация метода локтя

Строим дендрограмму.

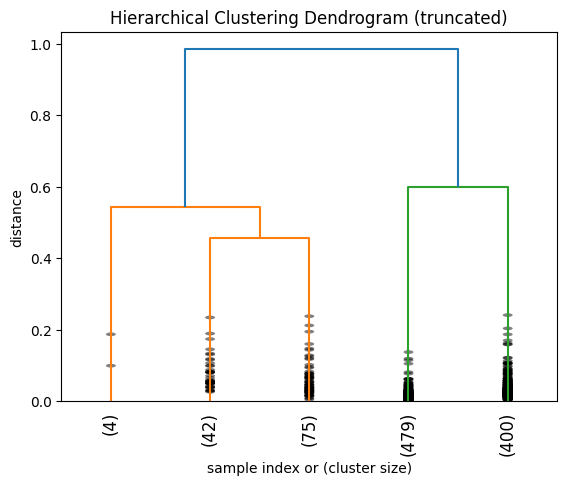


Рис.9. Визуализация дендрограммы

Далее визуализируем наше распределение по кластерам с помощью точечного графика. К примеру посмотрим как выглядит распределение на кластеры при выборе переменных цены и веса бриллианта (Рисунок 10).

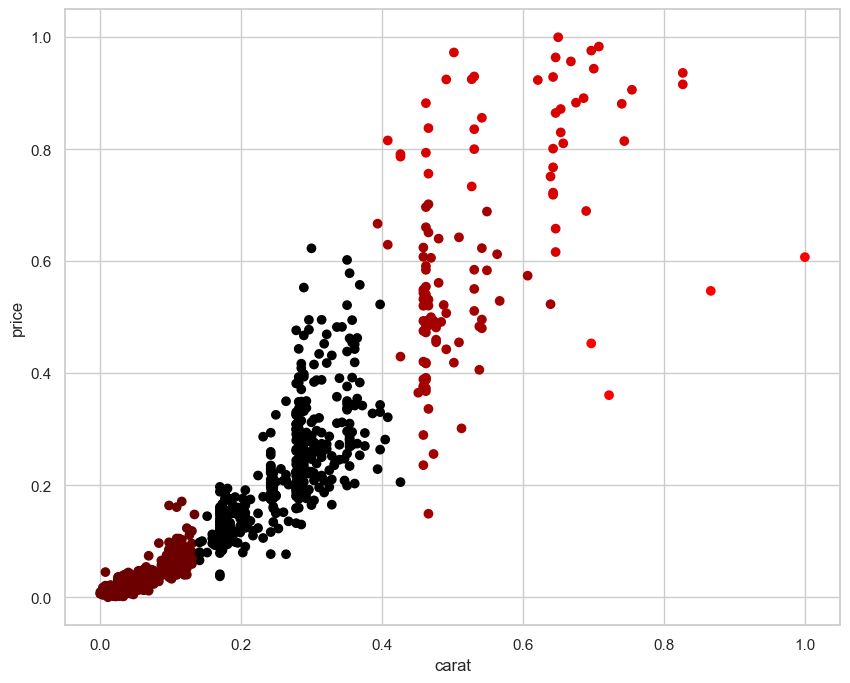


Рис.10 Визуализация иерархической кластеризации

Наблюдаем большое скопление точек в левом нижнем углу, что может означать, что большинство бриллиантов в выборке с меньшим весом и соответственно имеют небольшую цену. Также можно заметить, что график имеет вогнутую форму. Похожая ситуация происходит и при выборе переменных цены и длины, цены и ширины, цены и глубины, однако графики получаются более выгнутыми. Анализируя графики с переменными веса и длины, веса и ширины, веса и глубины можно предположить, что присутствует линейная зависимость. Соответствующие графики находятся в Приложении.

Также посмотрим средние значения по кластерам с помощью таблицы (Таблица 6). По таблице можно сказать, что бриллианты с большим весом, длиной, шириной, глубиной имеют и большую цену, к тому же таких бриллиантов всего 4. Соответственно и более дешевые имеют меньший вес, длину, ширину, глубину, однако в выборке таких бриллиантов наибольшее количество.

Таблица 6

Средние значения по кластерам

| I | carat | price | x | y | z | Количество |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2,51 | 9425,00 | 8,60 | 8,56 | 5,41 | 4 |
| 2 | 1,92 | 15996,07 | 7,97 | 7,95 | 4,90 | 42 |
| 3 | 1,57 | 9448,77 | 7,43 | 7,40 | 4,58 | 75 |
| 4 | 0,40 | 1050,11 | 4,70 | 4,71 | 2,90 | 479 |
| 5 | 0,95 | 4600,49 | 6,29 | 6,29 | 3,88 | 400 |

## Итерационная кластеризация методом k-средних

Следующий способ кластеризации данных это метод k-средних. Мы выбрали 5 кластеров, как и в случае с дендрограммой, так как при выборе количества кластеров, равному 4, возникают трудности при разделении кластеров, в то время как при n\_cluster = 5 кластеры легко идентифицировать и график получается более презентабельным (Рисунок 11).

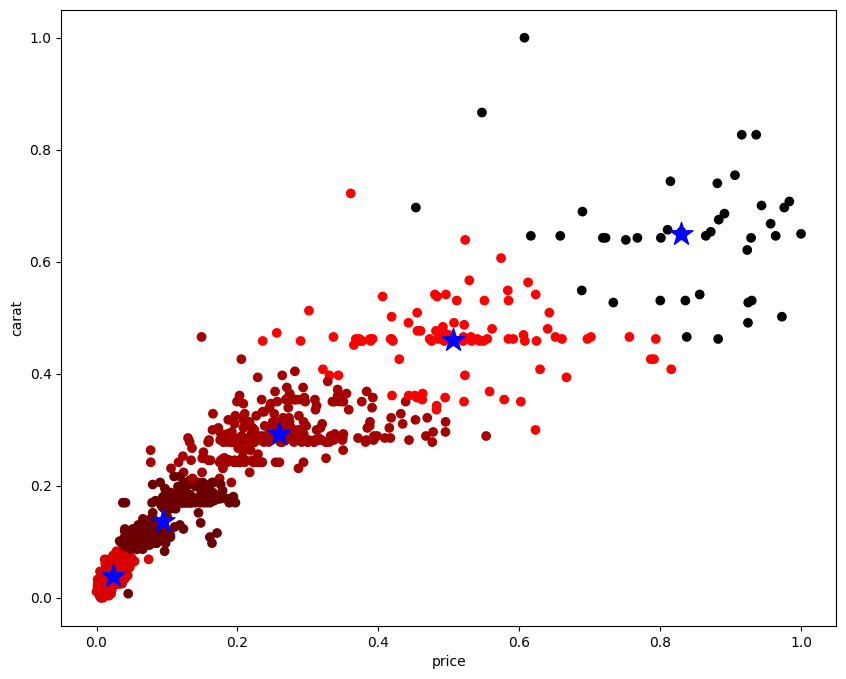


Рис.11 Визуализация метода KMeans

Благодаря методу KMeans мы можем наблюдать центры кластеров. График достаточно похож на тот, который получился в предыдущей кластеризации.

По таблице средних значений по кластерам (Таблица 7) можно сказать, что бриллианты с большим весом, длиной, шириной, глубиной имеют и большую цену, к тому же таких бриллиантов всего 41. Соответственно и более дешевые имеют меньший вес, длину, ширину, глубину, однако в выборке таких бриллиантов наибольшее количество.

Таблица 7

Средние значения по кластерам (KMeans)

| KMeans | carat | price | x | y | z | Количество |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 2,03 | 15637,22 | 8,11 | 8,09 | 4,99 | 41 |
| 2 | 0,61 | 2103,57 | 5,43 | 5,44 | 3,35 | 270 |
| 3 | 1,04 | 5131,77 | 6,49 | 6,48 | 4,00 | 265 |
| 4 | 1,50 | 9670,59 | 7,32 | 7,31 | 4,52 | 97 |
| 5 | 0,34 | 777,79 | 4,47 | 4,48 | 2,76 | 327 |

## 

## 

## Метод главных компонент

Поскольку PCA дает подпространство признаков, которое максимизирует дисперсию вдоль осей, имеет смысл стандартизировать данные, особенно если они были измерены в разных масштабах. Продолжим преобразование данных в единичный масштаб (среднее значение = 0 и дисперсия = 1), что является требованием для оптимальной производительности многих алгоритмов машинного обучения.

Далее визуализируем метод PCA (Рисунок 12). Масштаб был выбран таким для лучшего представления полученных классов.

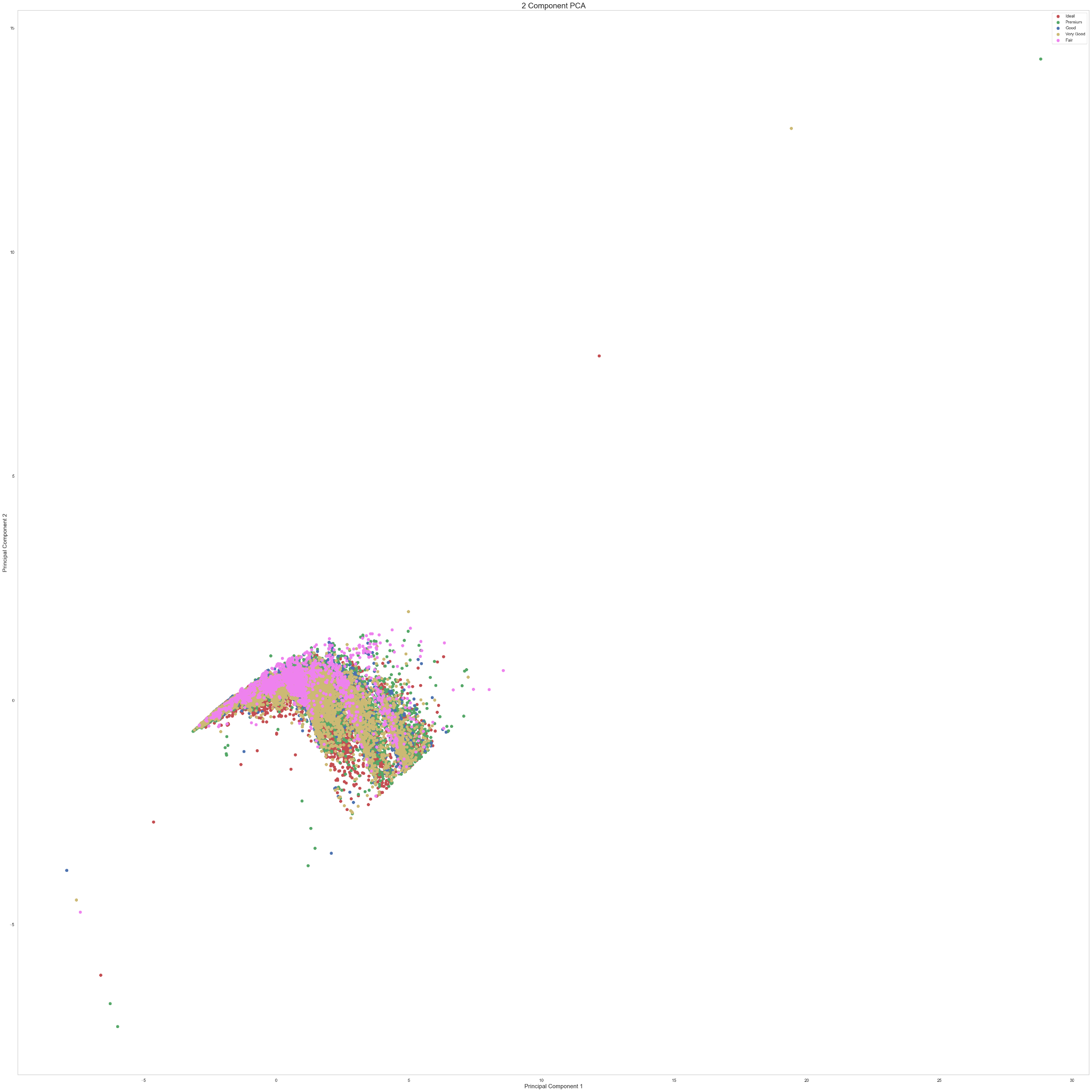


Рис.12. Визуализация PCA

Очевидно, что классы находятся очень близко из-за чего происходят наложения на графике. Можно сделать выводы, что выбранные данные имеют плотное распределение.

Далее рассчитали объясненную дисперсию. Объясненная дисперсия сообщает нам, сколько информации (дисперсии) может быть отнесено к каждому из основных компонентов. Вместе первые два основных компонента содержат 98,30% информации. Первый основной компонент содержит 93,9% дисперсии, а второй основной компонент содержит 4,4% дисперсии. Третий и четвертый основные компоненты содержали остальную часть дисперсии набора данных.

## Неотрицательное матричное разложение

В случае NMF, где мы разлагаем матрицу на несколько компонент, графики представляют собой исходные данные, компоненту 1, компоненту 2 (Рисунок 13).

Смотря на график исходных данных можно увидеть закономерную линейность данных, что уже было выяснено ранее с помощью других способов анализа. По графикам компонент 1 и 2 можно предположить, что компоненты являются идентичными и не особо интерпретируемые. Предполагаем, что данный тип разложения не подходит для наших данных.

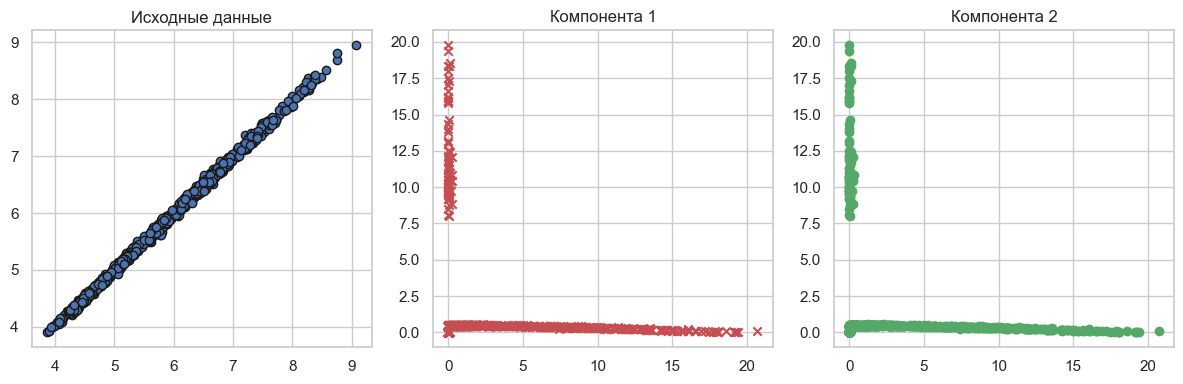
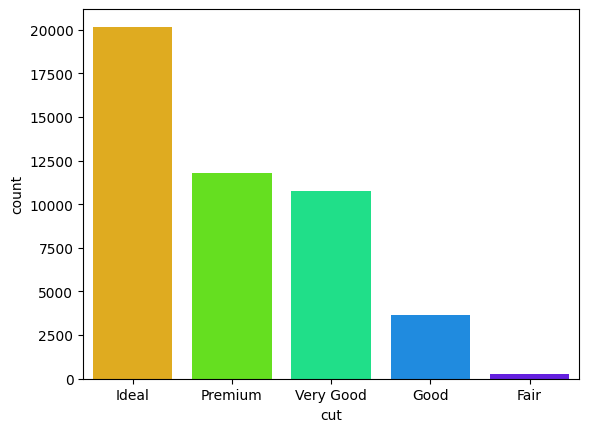


Рис.13. Визуализация NMF

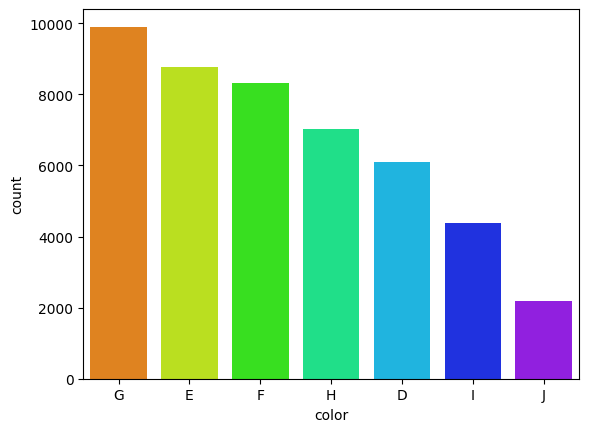
Если сравнивать все вышеприведенные методы кластеризации, то можно сделать выводы, что для нашего датасета более репрезентативными являются метод KMeans и дендрограммы.

# Приложения

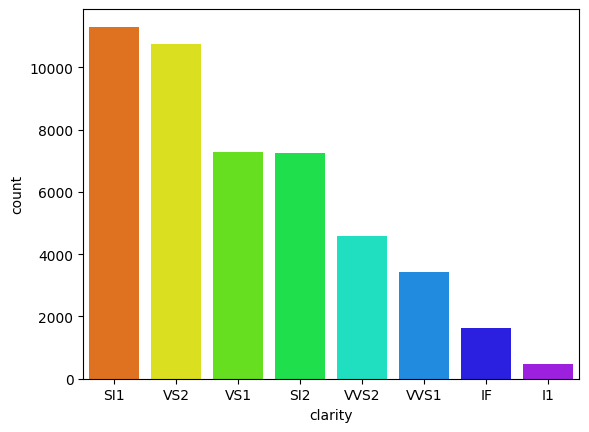
Приложение 1



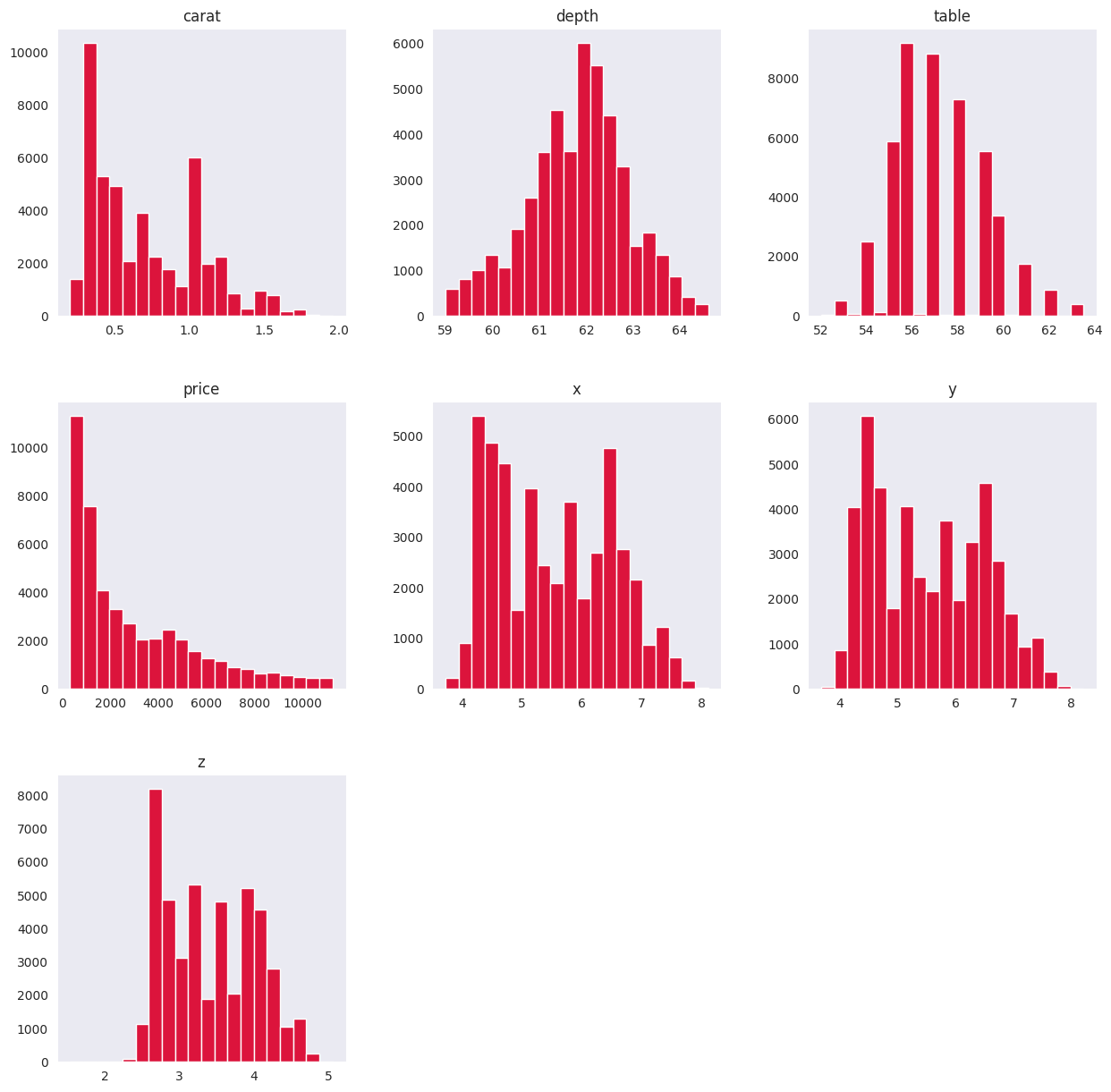
Приложение 2



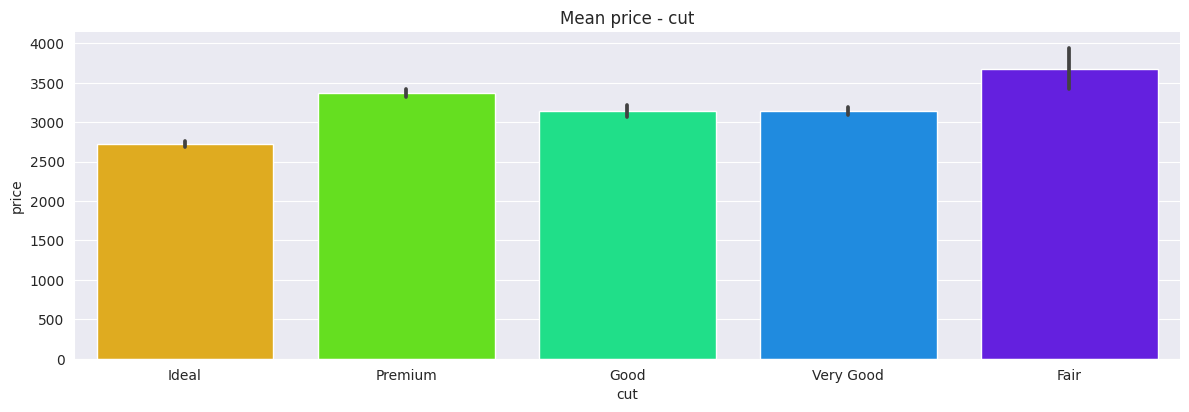
Приложение 3



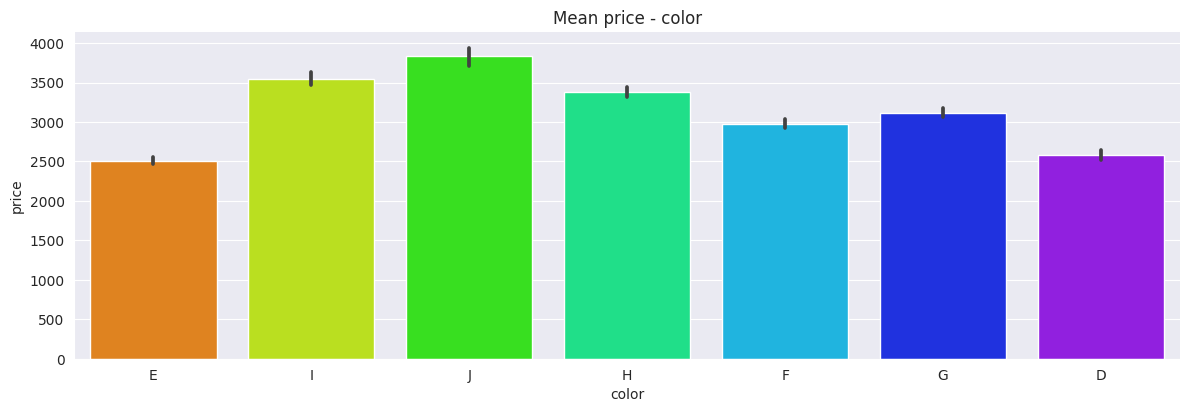
Приложение 4



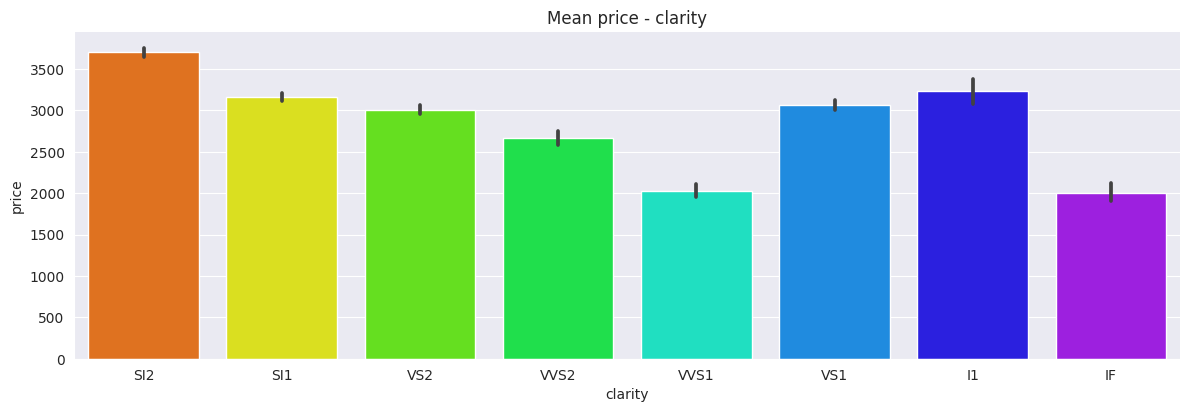
Приложение 5



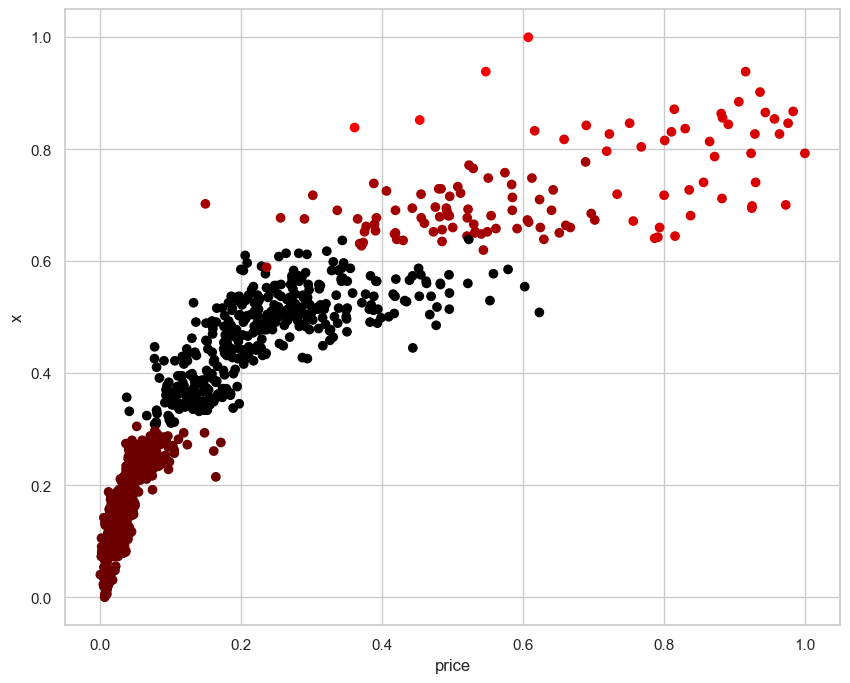
Приложение 6



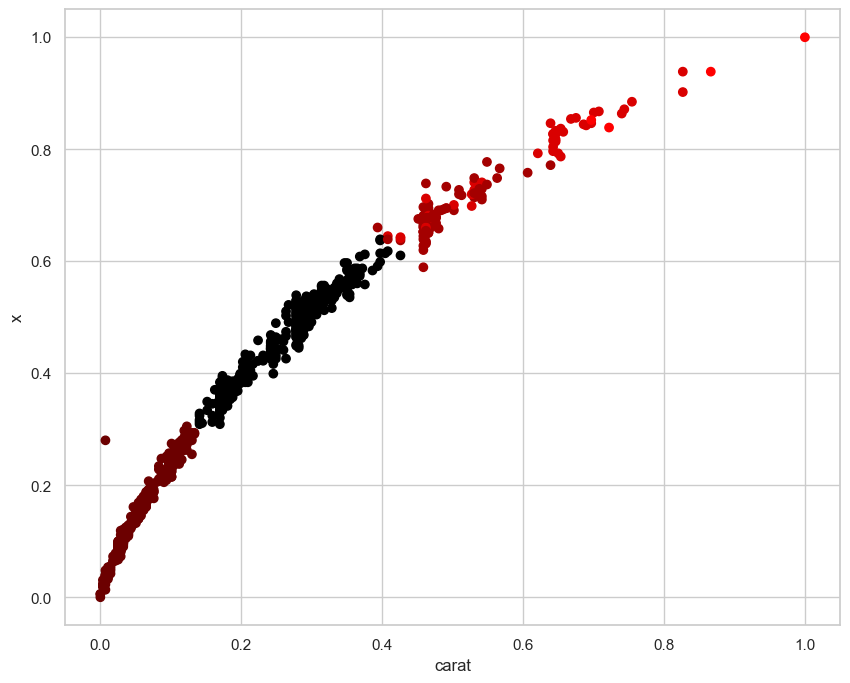
Приложение 7



Приложение 8



Приложение 9



Приложение 10

