**Моделирование цена на бензин А-95**

Сущность работы состоит в исследовании зависимости цены бензина от наиболее значительных факторов.

Для построения моделей были выделены наиболее значимые из них. Объясняемой переменной (Y) является цена бензина А-95 (Price)

Объясняющими переменными являются:

1. Курс доллара (Kurs\_dollar, грн)

Курс доллара в значительной степени оказывает влияние на цену бензина на украинских АЗС. Украина не может в полной мере обеспечить свои потребности в бензине, поэтому около 80 % бензина, представленного на украинских АЗС, покупается за рубежом за доллары США.

1. Курс евро (Kurs\_euro, грн)

Очень существенным фактором также является курс евро. Рост закупочной цены на нефтепродукты на 1% приводит к росту цены 1 литра на 10 коп., а снижение курса национальной валюты на 1 грн ведет к удорожанию 1 литра горючего на 70 - 80 коп. Это связано с тем, что в стоимости топлива присутствуют две переменные, отечественные покупатели оплачивают в валюте: цена закупки

и акциз, который оплачивается в евро (с 1 литра бензина А 95 трейдеры платят 0.213 евро импортного акциза)

1. Цена нефти (Oil\_price, $ за баррель)

Стоимость нефти оказывает наибольшее влияние на формирование цены бензина (может достигать 30% окончательной цены).

1. Количество использованного бензина а-95 (Used, тыс. тонн)

Бензин является рыночным товаром, поэтому его цена также определяется взаимосвязью между спросом и предложением.

Все переменные в модели являются количественными. Для моделирования будем использовать язык программирования Python, а в качестве среды для работы jupyter notebook.

Для анализа взяты ежедневные данные по цене бензина А-95

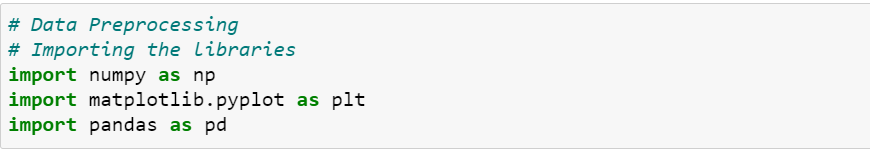
1. **Подготовка данных**

Для получения осмысленных результатов модели машинного обучения необходимо выполнить предварительную подготовку данных. Данные будет возможно использовать для обучения модели только в том случае, если они будут представлены в правильном масштабе, формате, будут выявлены и заменены или исправлены неполные, неправильные, неточные или несуществующие данные. Иначе построенная модель будет неточна и сможет использоваться для построения достоверных прогнозов.

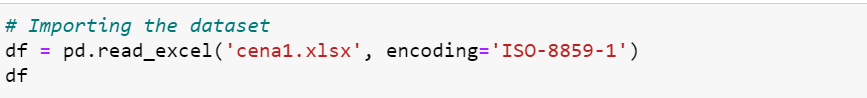
Загрузка данных

Для того, чтобы иметь возможность работать с данными их необходимо загрузить в оперативную память и подключить необходимые библиотеки в jupyter notebook для начала работы с ними.

In (Подключение библиотек):



In (Импортирование датасета и вывод его на экран):



In (Вывод датасета на экран):



Out (Результат выведения датасета на экран):



Результат работы программы показал нам загруженный массив данных, содержащих 882 строчки и 4 столбца.

In (Описание основных статистик):



Для описания статистик выводим все столбцы, кроме Даты(Date).

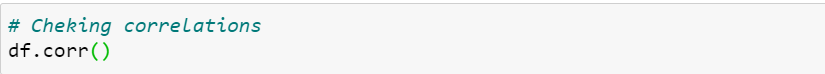
Out (Основные статистики):



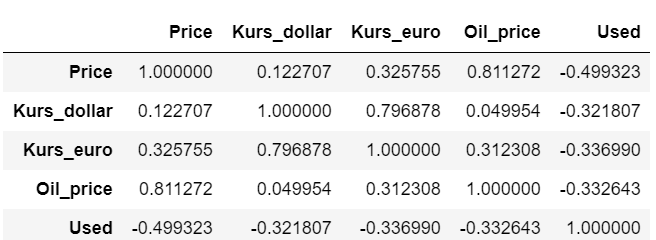
По каждой переменной видим различные параметры: количество записей (count), средние значения (mean), стандартное отклонение (std), минималльные значения (min), количество значений, которые попадают в 25%, 50%, 75% и максимальное значение.

Проведем анализ датасета на наличие выбросов. Девяносто девять процентов данных, распределенных по нормальному закону, находятся всередине интервала, ограниченного утроенным значением стандартного отклонения. Соответственно все значения, которые не попадают в этот интервал будут являться выбросами. Исходя из таблицы, делаем вывод, что не имеет смысла устранять выбросы не в одной из переменных (так как (mean+std\*3) > max) во всех случаях).

In (Построение корреляционной матрицы):



Out:

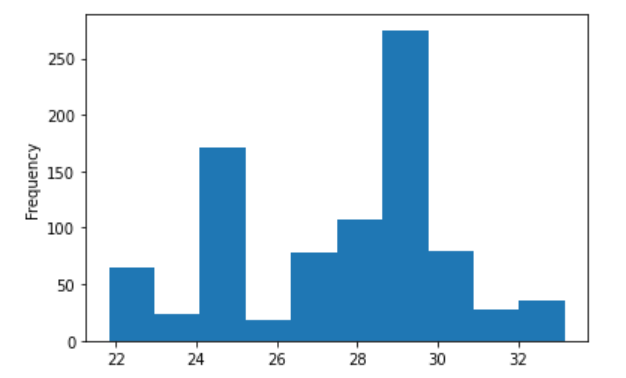


Наибольшая связь наблюдается между ценой нефти и ценой бензина (81,1%).

In (Визуализация переменной «Price»):



Out (Визуализация переменной «Price»):

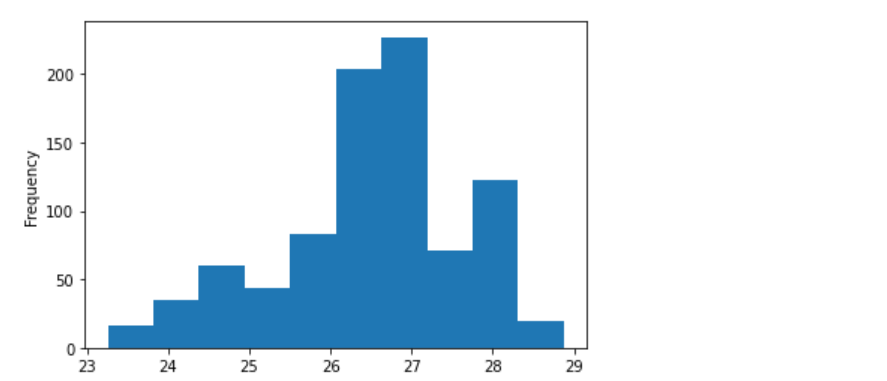


Наиболее частой ценой бензина в рассматриваемом периоде времени является цена в пределах от 29 грн. до 30 грн.

In (Визуализация переменной «Kurs\_dollar»):

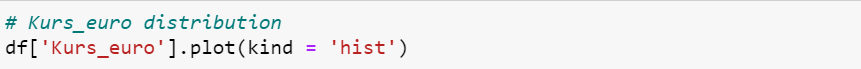


Out (Визуализация переменной «Kurs\_dollar»):

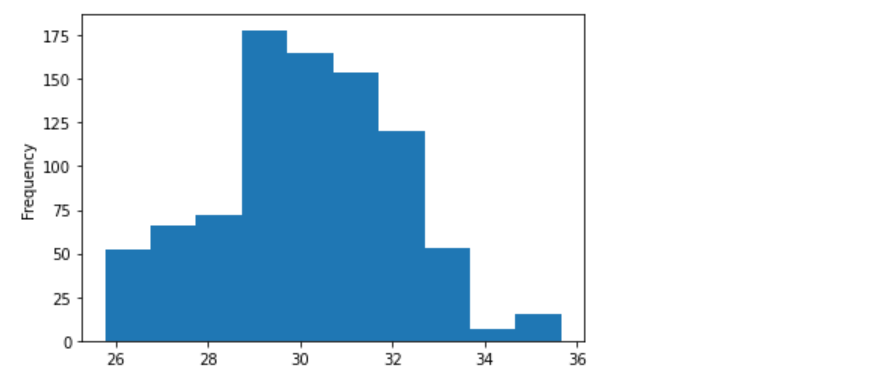


Наиболее частым курсом доллара в рассматриваемом периоде времени является курс в пределах от 26 грн. до 27 грн. за 1 доллар.

In (Визуализация переменной «Kurs\_euro»):

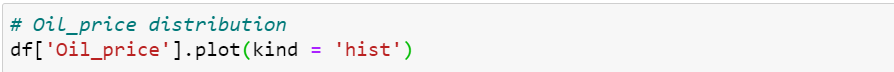


Out (Визуализация переменной «Kurs\_dollar»):

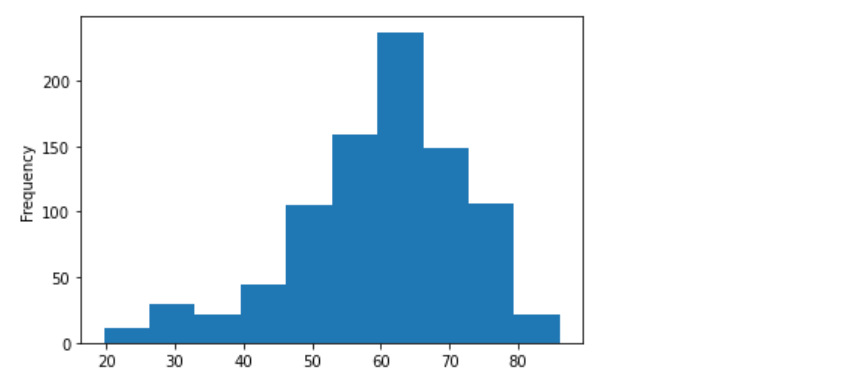


Наиболее частым курсом евро в рассматриваемом периоде времени является курс в пределах от 29 грн. до 30 грн. за 1 доллар.

In (Визуализация переменной «Oil\_price»):



Out (Визуализация переменной «Oil\_price»):

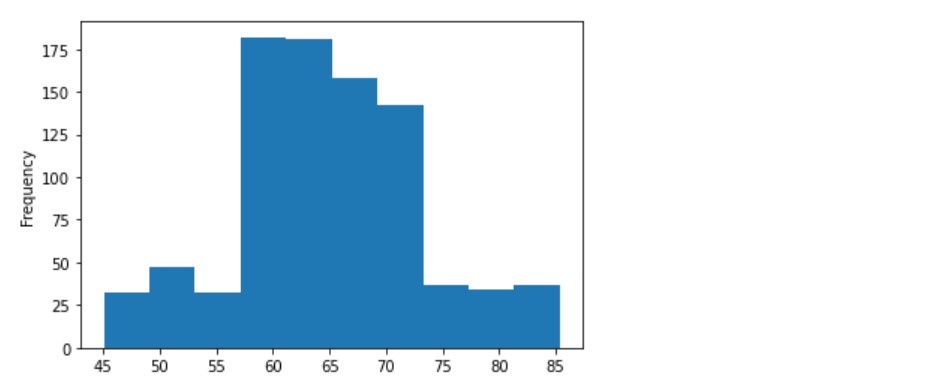


Наиболее частым значением цены нефти в рассматриваемом периоде времени является цена в пределах от 60 $ до 70 $ за 1 баррель.

In (Визуализация переменной «Used»):



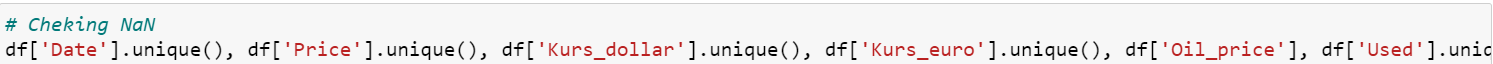
Out (Визуализация переменной «Used»):



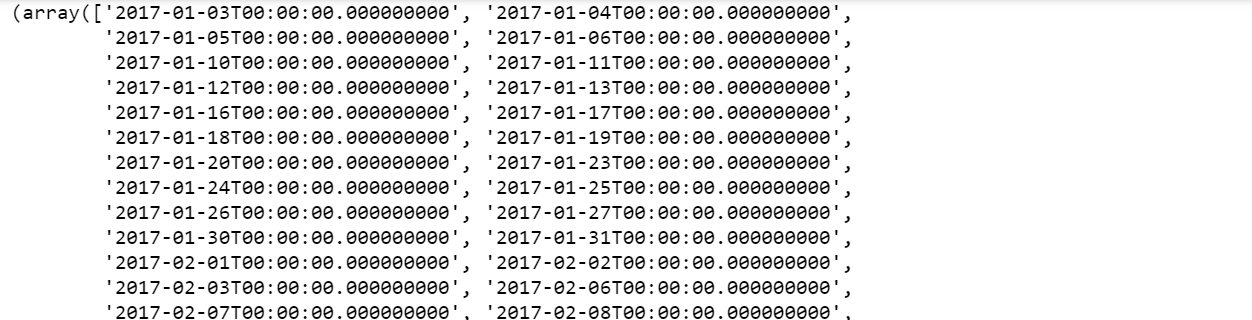
Наиболее частым значением используемого бензина в рассматриваемом периоде времени является количество в пределах от 57 $ до 63 $ тысяч тонн.

Анализируя графики распределения делаем вывод, что графики достаточно симметричны и близки к нормальному распределению.

In (Проверка датасета на наличие пропущенных значений):

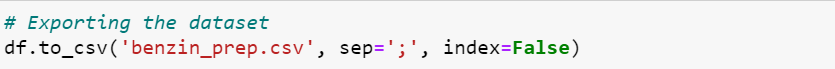


Out (Уникальные значения по каждой из переменных):



В результате проверки делаем вывод, что пропущенные значения отсутствуют.

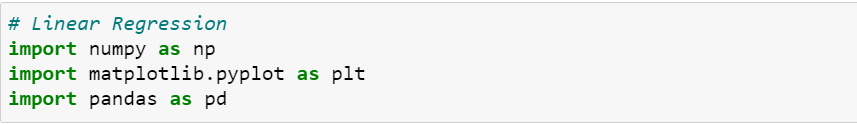
In (Экспорт данных в файл benzin.csv)



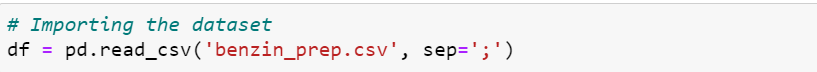
1. **Линейная регрессия**

**2.1 Подготовка модели**

In (Подключение библиотек, необходимых для линейной регрессии):



In (Импорт подготовленного на первом этапе датасета):



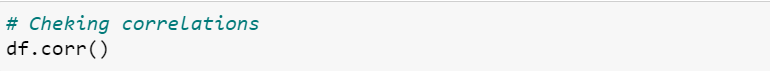
In (Вывод датасета на экран):



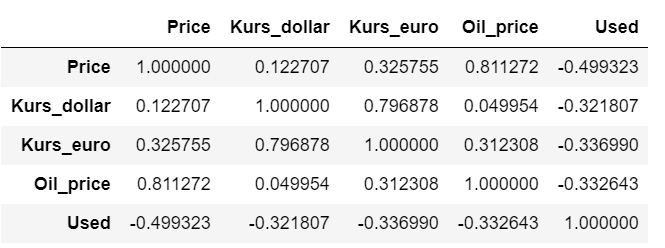
Out (Результат вывода датасета на экран):



In (Построение корреляционной матрицы):

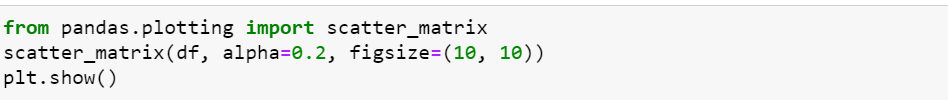


Out (Корреляционная матрица):

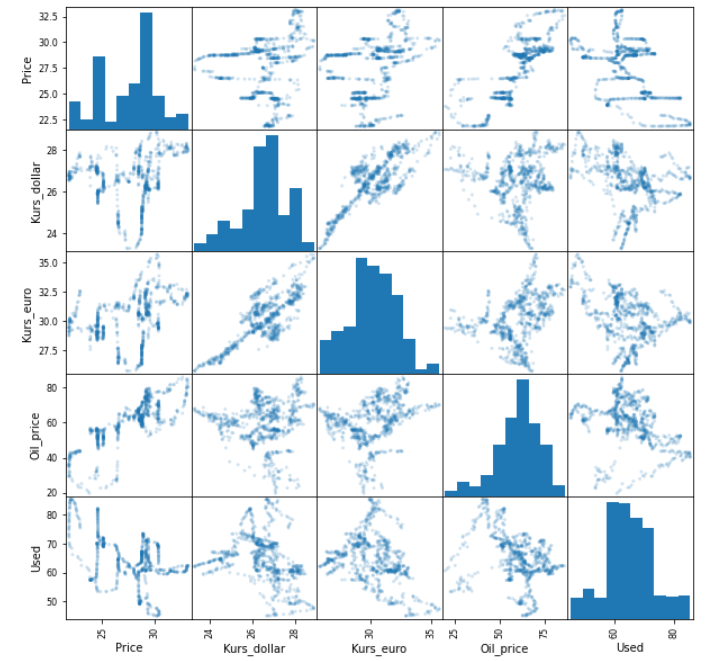


На объясняемую переменную больше всего оказывает влияние цена нефти.

In (Построение корреляционной матрицы в виде матрицы рассеивания):

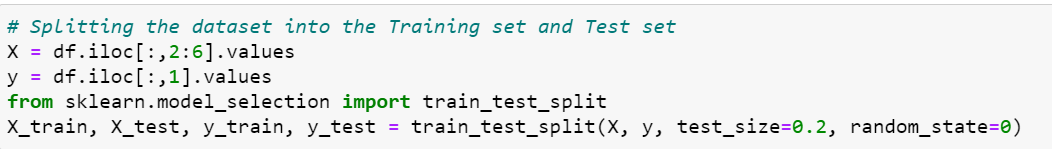


Out (Матрица рассеивания):



На объясняемую переменную больше всего оказывает влияние цена нефти.

In (Разделение датасета на обучающую и тестовую выборки):



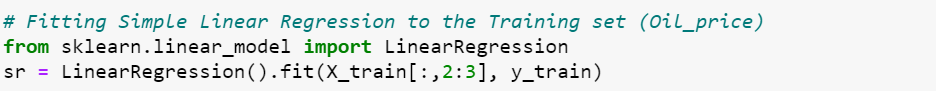
В качестве объясняющих переменных берём все строки и следующие столбцы: Kurs\_dollar, Kurs\_euro, Oil\_price, Used.

В качестве объясняемой переменной берём все строки и столбец Price.

Для проверки прогнозного качества модели разделим все наблюдения на обучающую (80%) и тестовую выборки (20%).

**2.2 Однофакторная линейная регрессия**

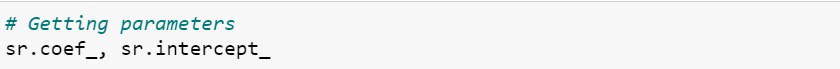
In (Обучение простой однофакторной линейной регрессии):



Для обучения простой линейной регрессии берём наиболее значимый фактор – цену нефти (Oil\_price).

Обучение модели производим на множестве X\_train (переменная Oil\_price) и Y\_train (переменная Price)

In (Получение параметров полученной модели):

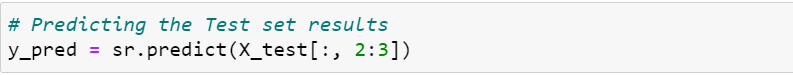


Out (Параметры модели):

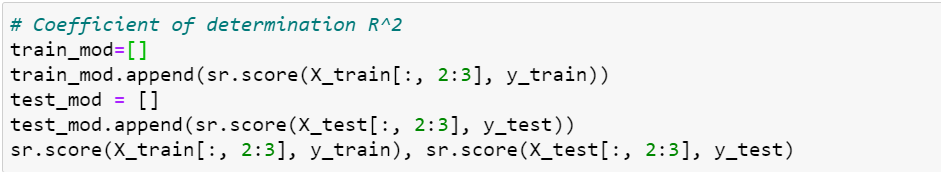


Коэффициент при переменной - 0,18693878

Свободный член – 16,25356

In (Построение прогноза на тестовом множестве):Для предсказания используем значение наиболее значимого фактора (Oil\_price) тестового множества.

In (Коэффициент детерминации для тестовой и обучающей выборки):

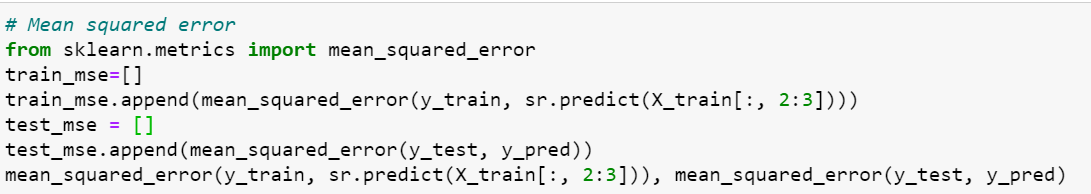


Out (Коэффициент детерминации для тестовой и обучающей выборки):



Коэффициент детерминации на обучающей выборке (0,673) больше, чем на тестовой (0,579).

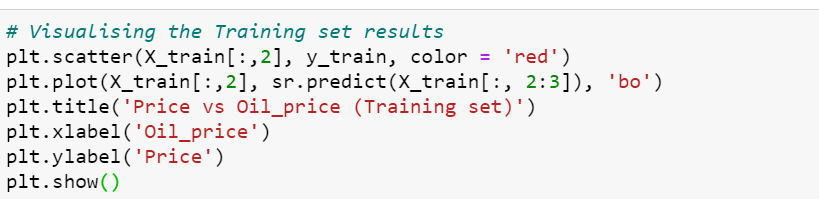
In (Получение среднеквадратической ошибки):



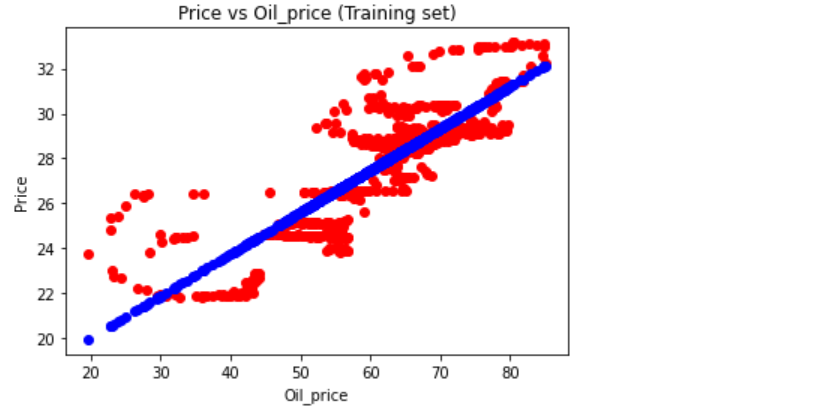
Out (Получение среднеквадратической ошибки):



In (Визуализация обучающей выборки линейной модели):

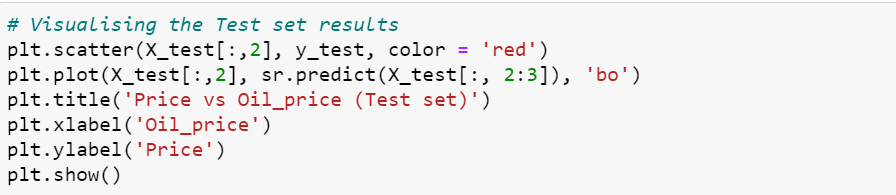


Out (Визуализация обучающей выборки линейной модели):

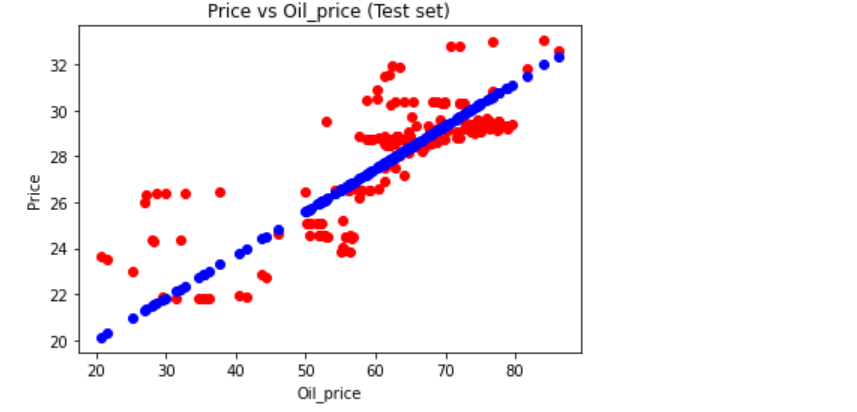


На графике отображена зависимость цены бензина от цены нефти на обучающей выборке. Красными кругами отображены реальные данные, синими – прогнозные.

In (Визуализация тестовой выборки линейной модели):



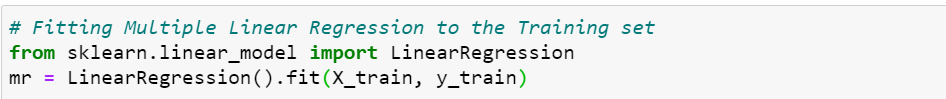
Out (Визуализация тестовой выборки линейной модели):



На графике отображена зависимость цены бензина от цены нефти на тестовой выборке. Красными кругами отображены реальные данные, синими – прогнозные. Разброс регрессии здесь больше, чем на обучающей выборке.

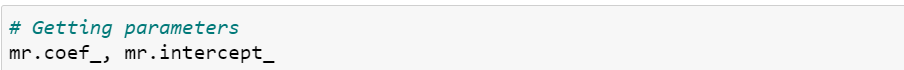
**2.3 Построение множественной регрессии**

In (Обучение модели множественной регрессии):

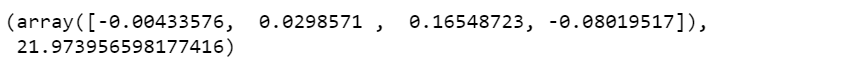


Для построения множественной регрессии используем все объясняющие переменные тестового множества (X\_train).

In (Получение параметров полученной модели):



Out (Параметры модели):



Коэффициент при переменной x1 – (-0.00433576)

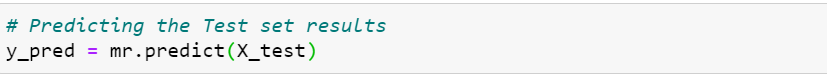
Коэффициент при переменной x2 – (0.0298571)

Коэффициент при переменной x3 – (0.16548723)

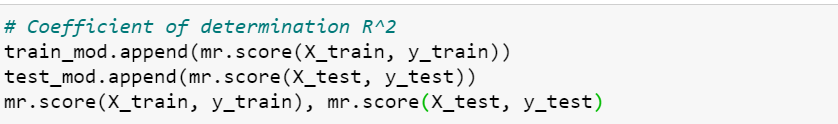
Коэффициент при переменной x4 – (-0.08019517)

Свободный член – 21.973956598177416

In (Построение прогноза на тестовом множестве):



In (Коэффициент детерминации для тестовой и обучающей выборки):



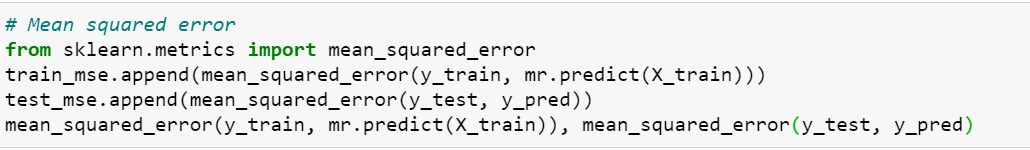
Out (Коэффициент детерминации для тестовой и обучающей выборке):



Коэффициент детерминации на обучающей выборке (0,725) больше, чем на тестовой (0,677).

По сравнению с моделью с одним фактором (цена нефти) качество модели возросло на обучающей выборке с 0,673 до 0,725, а на тестовой с 0,579 до 0,677.

In (Получение среднеквадратической ошибки):

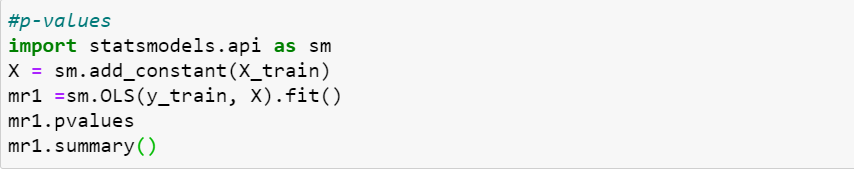


Out (Получение среднеквадратической ошибки):

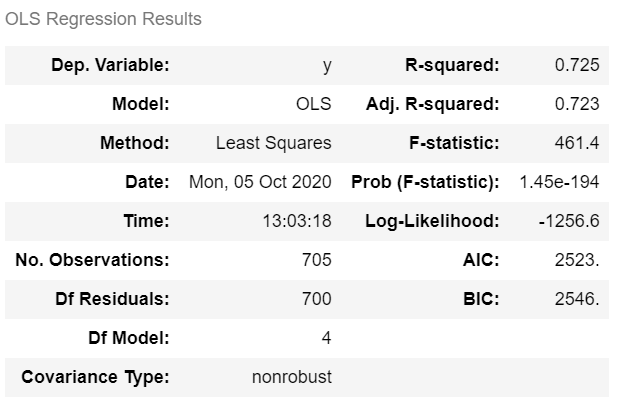


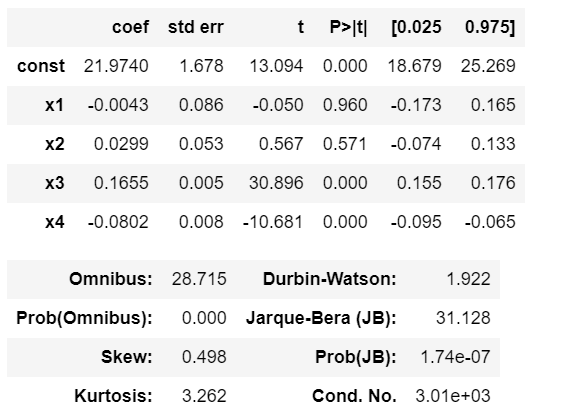
По сравнению с моделью с одним фактором (цена нефти) среднеквадратическая ошибка модели снизилась на обучающей выборке с 2,4566 до 2,0689, а на тестовой с 2,8245 до 2,1625.

In (Построение линейной модели с помощью библиотеки statsmodels.api):



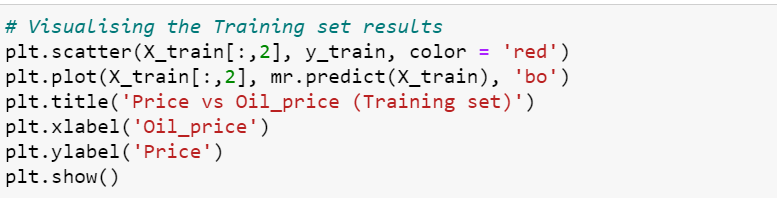
In (Результаты регрессии):



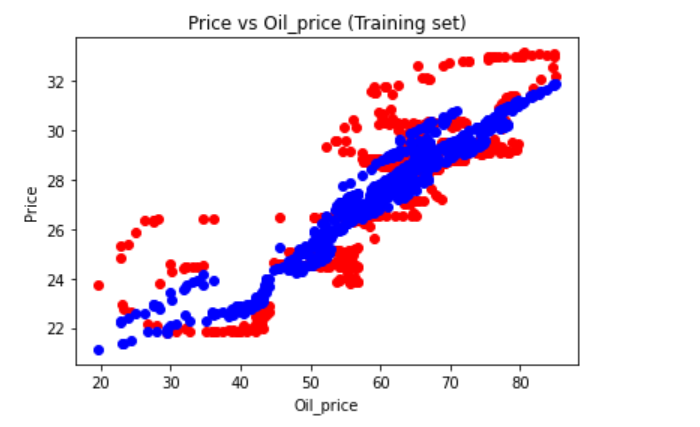


Делаем вывод, что переменные х1(Kurs\_dollar) и х2 (Kurs\_euro) не значимы.

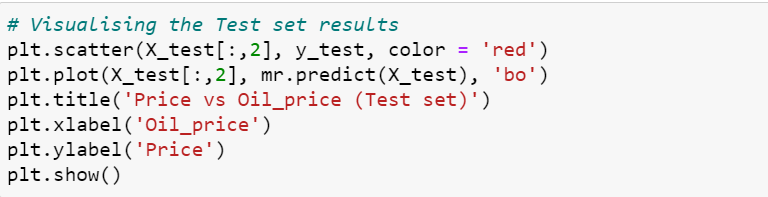
In (Визуализация обучающей выборки линейной модели):



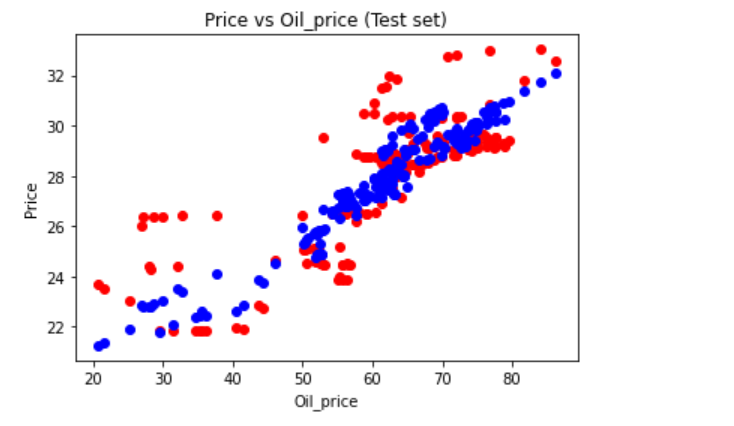
Out (Визуализация обучающей выборки линейной модели):



In (Визуализация тестовой выборки линейной модели):

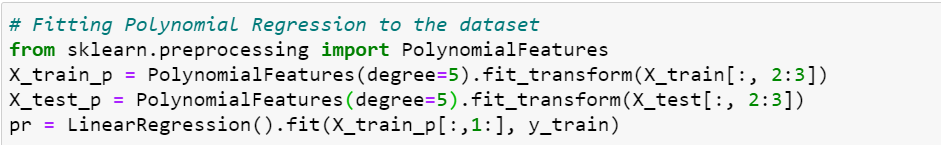


Out (Визуализация тестовой выборки линейной модели):



* 1. **Построение полиномиальной регрессии**

In (Построение модели полиномиальной регрессии):

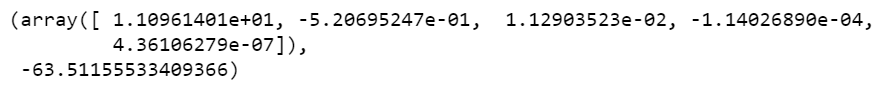


Степень полинома возьмем равной 5 (degree = 5)

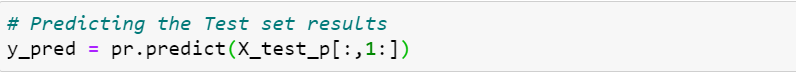
In (Получение параметров полученной модели):



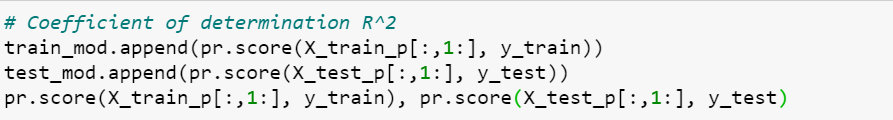
Out (Параметры модели):



In (Построение прогноза на тестовом множестве):



In (Коэффициент детерминации для тестовой и обучающей выборки):



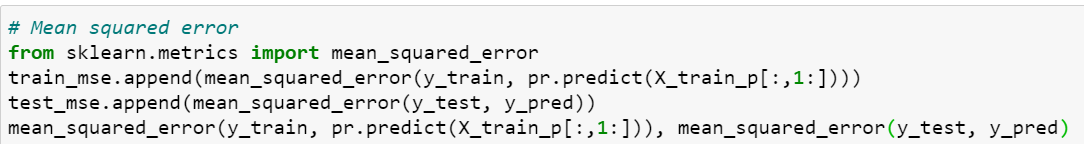
Out (Коэффициент детерминации для тестовой и обучающей выборке):



Коэффициент детерминации на обучающей выборке (0,77) больше, чем на тестовой (0,73). Но в полиномиальной модели эта разница меньше, чем в линейных.

Также повысилось качество модели (по сравнению с множественной линейной регрессией) на обучающей выборке с 0,725 до 0,77, а на тестовой с 0,677 до 0,7386 .

In (Получение среднеквадратической ошибки):

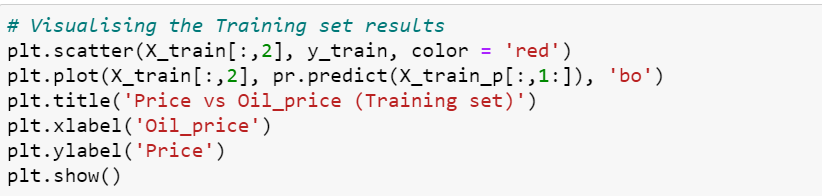


Out (Получение среднеквадратической ошибки):

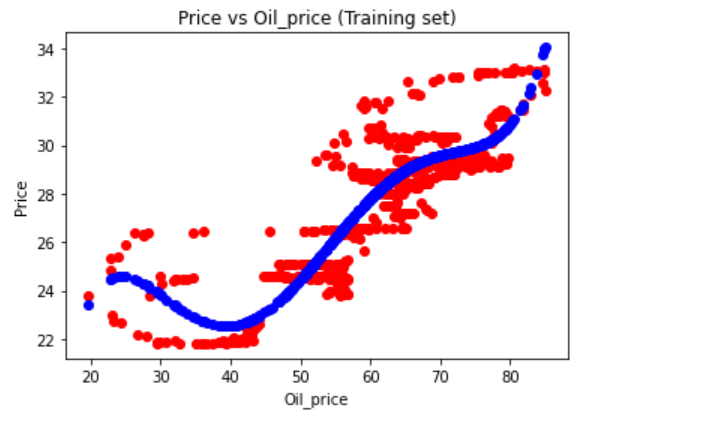


По сравнению с моделью с множественной регрессии среднеквадратическая ошибка модели снизилась на обучающей выборке с 2,0689 до 1,7301, а на тестовой с 2,1625до 1,7521.

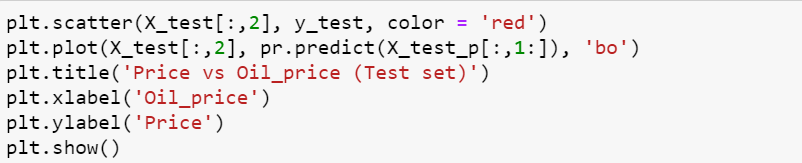
In (Визуализация обучающей выборки линейной модели):



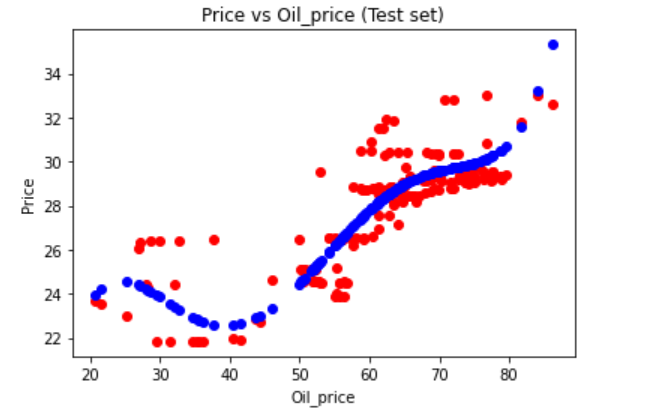
Out (Визуализация обучающей выборки линейной модели):



In (Визуализация тестовой выборки линейной модели):



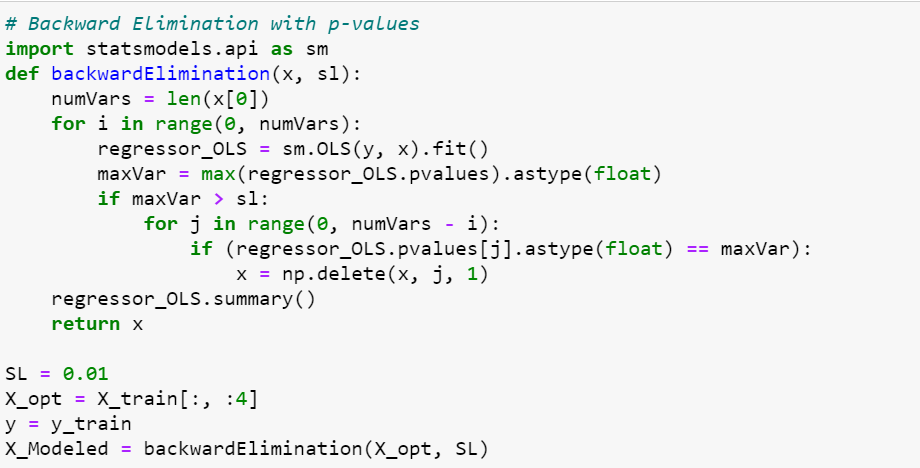
Out (Визуализация тестовой выборки линейной модели):



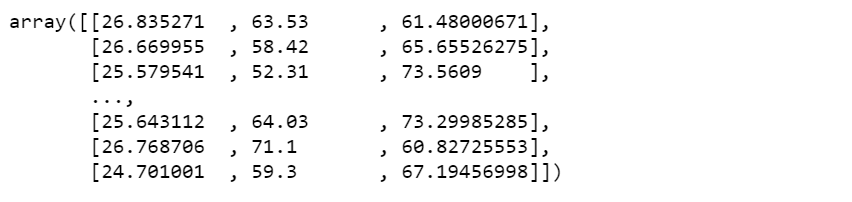
* 1. **Построение линейной модели на значимых факторах**

Построим модель на основании значимых факторов (с уровнем значимости 1%) для чего используем метод обратного исключения (Backward Elimination)

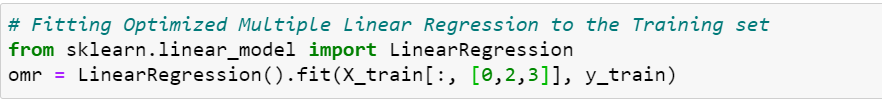
In (Модель на множестве значимых переменных):



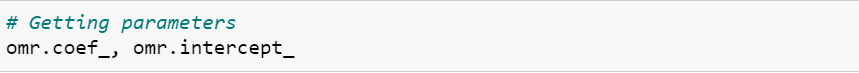




In (Построение модели оптимальной множественной регрессии):



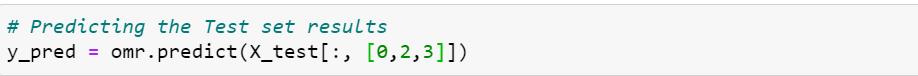
In (Получение параметров полученной модели):



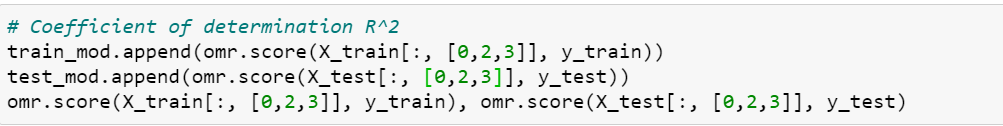
Out (Параметры модели):



In (Построение прогноза на тестовом множестве):



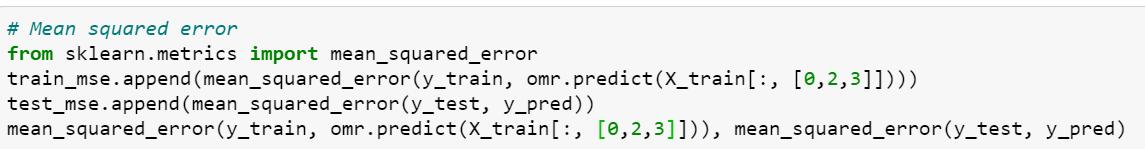
In (Коэффициент детерминации для тестовой и обучающей выборки):



Out (Коэффициент детерминации для тестовой и обучающей выборке):



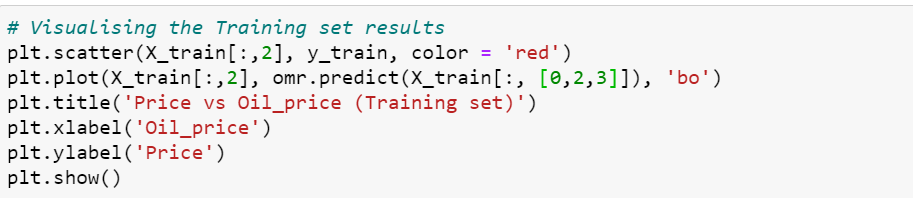
In (Получение среднеквадратической ошибки):



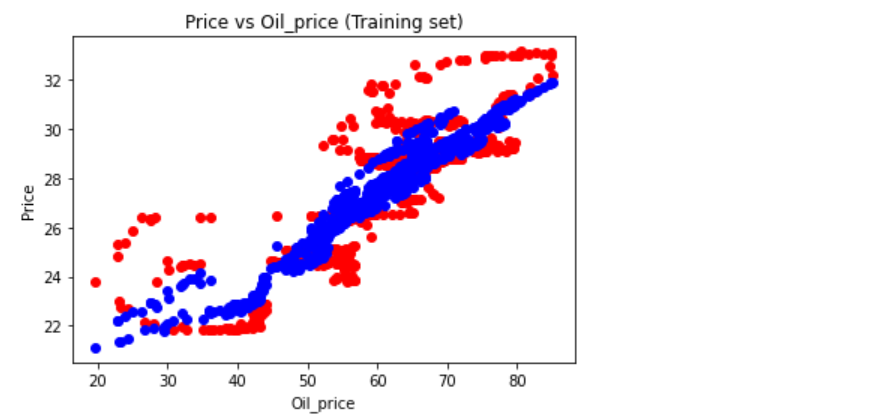
Out (Получение среднеквадратической ошибки):



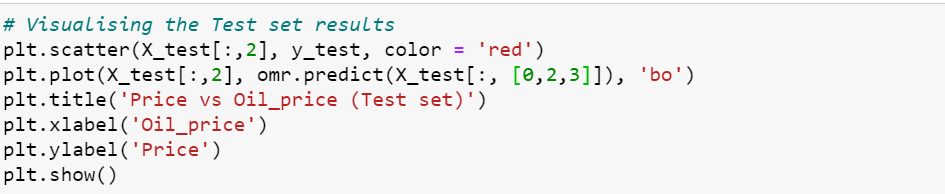
In (Визуализация обучающей выборки линейной модели):



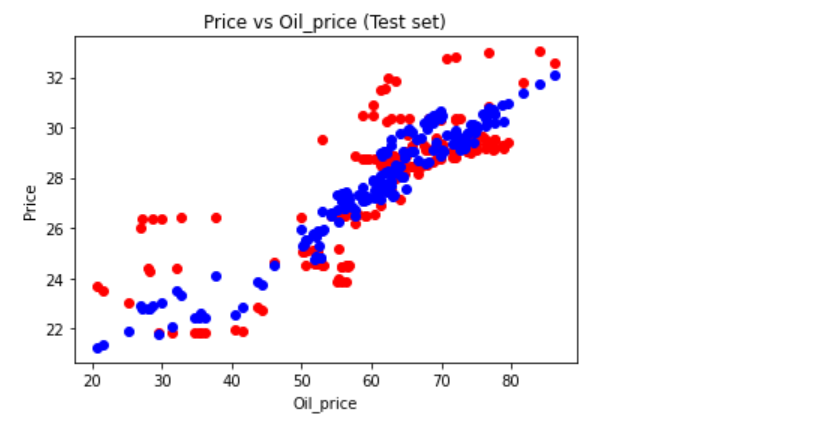
Out (Визуализация обучающей выборки линейной модели):



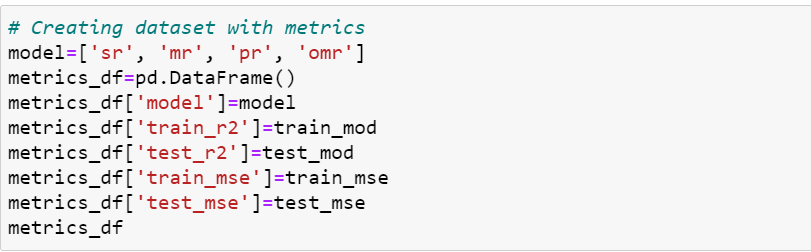
In (Визуализация тестовой выборки линейной модели):



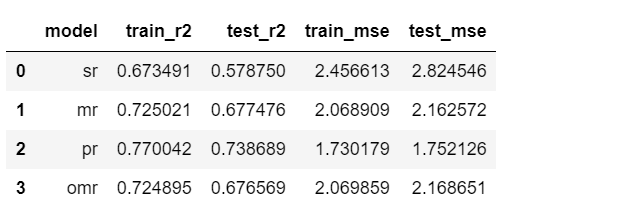
Out (Визуализация тестовой выборки линейной модели):



In (Создание сравнительной таблицы R^2 & mse по различным моделям):

****

Out (Сравнительная таблица R^2 & mse по различным моделям):

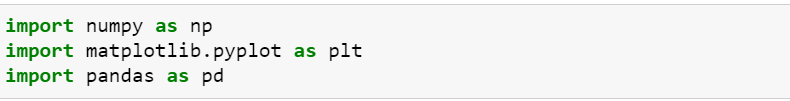
****

**Вывод.** Лучшей моделью с наибольшим коэффициентом детерминации на обучающей и тестовой выборке, а также наименьшей среднеквадратической ошибкой на обучающей и тестовой выборке является полиномиальная модель пятой степени.

1. **Дерево решений и случайный лес**

**3.1. Построение дерева решений на наиболее значимом факторе**

In (Подключение библиотек, необходимых для построения модели):



In (Импорт подготовленного датасета):



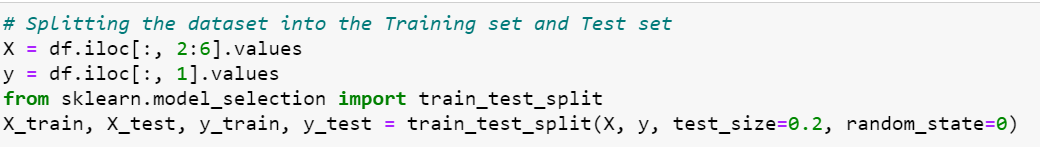
In (Вывод датасета на экран):



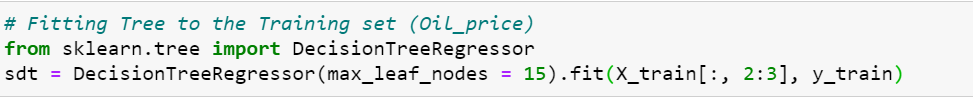
Out (Вывод датасета на экран):



In (Разделение датасета на обучающую и тестовую выборки):

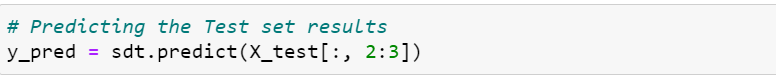


In (Построение дерева решений от фактора Oil\_price):

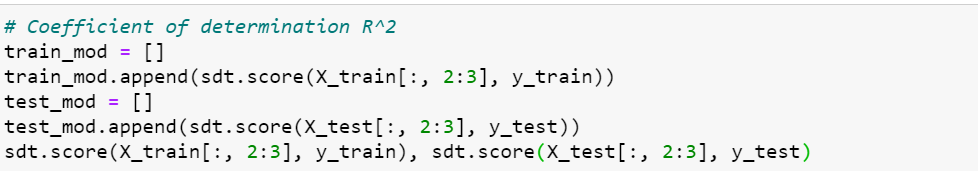


Зададим максимальное количество ветвлений дерево равное 15.

In (Построение прогноза на тестовом множестве):



In (Коэффициент детерминации для тестовой и обучающей выборки):

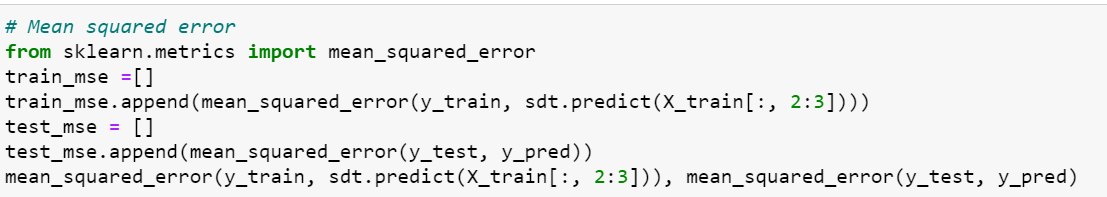


Out (Коэффициент детерминации для тестовой и обучающей выборке):



Расхождение между качеством модели на тестовой и обучающей выборке невелико.

In (Получение среднеквадратической ошибки):

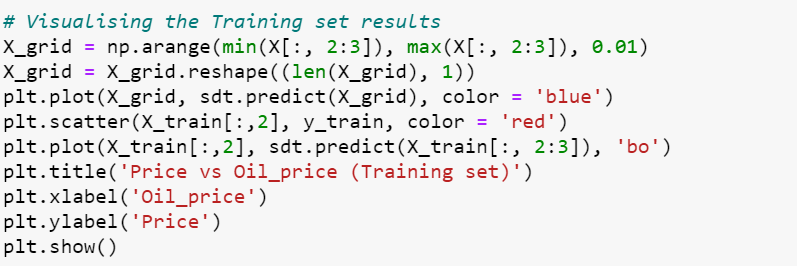


Out (Получение среднеквадратической ошибки):

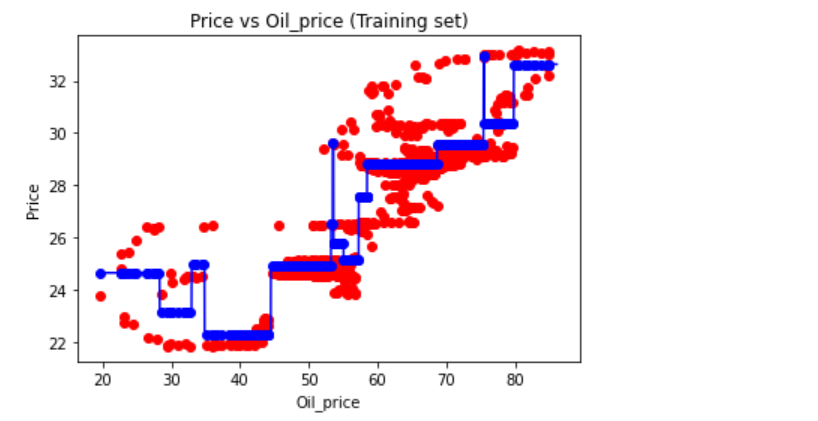


Ошибка прогноза меньше, чем в линейных моделях (в лучшей полиномиальной модели ошибка обучающего множества равнялась 1,73).

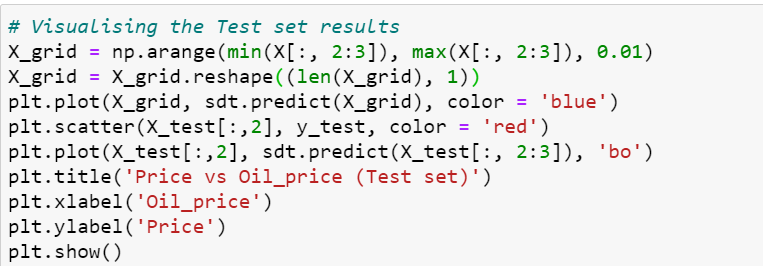
In (Визуализация обучающей выборки линейной модели):



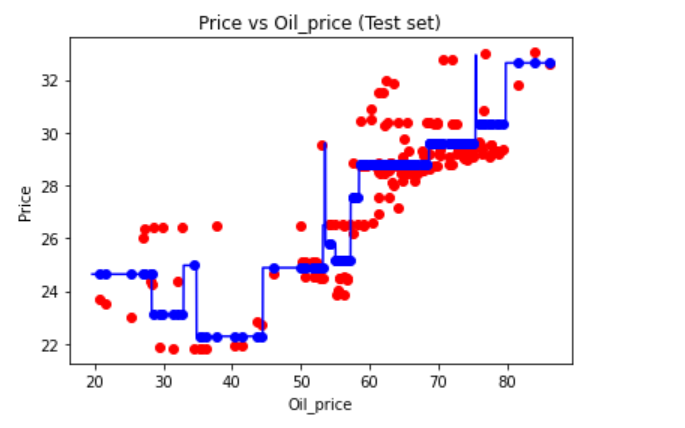
Out (Визуализация обучающей выборки):



In (Визуализация тестовой выборки):

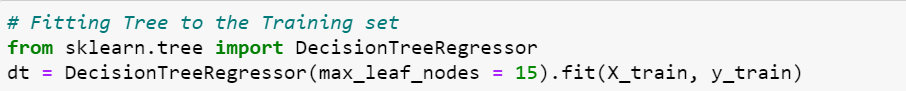


Out (Визуализация тестовой выборки линейной модели):



**3.2. Построение дерева решений на всех объясняющих переменных**

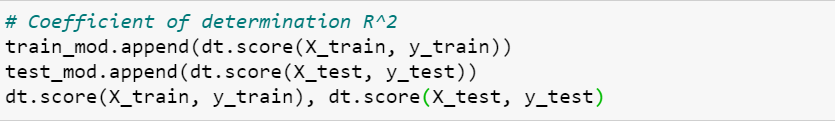
In (Построение дерева решений на всех объясняющих переменных):



In (Построение прогноза на всех объясняющих переменных):



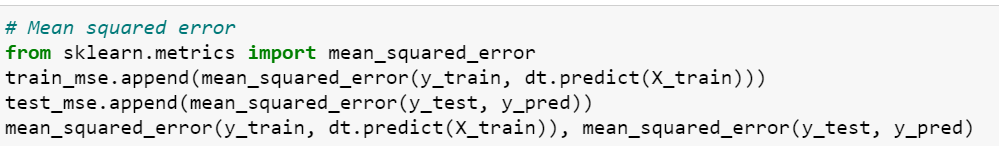
In (Коэффициент детерминации для тестовой и обучающей выборки):



Out (Коэффициент детерминации для тестовой и обучающей выборке):



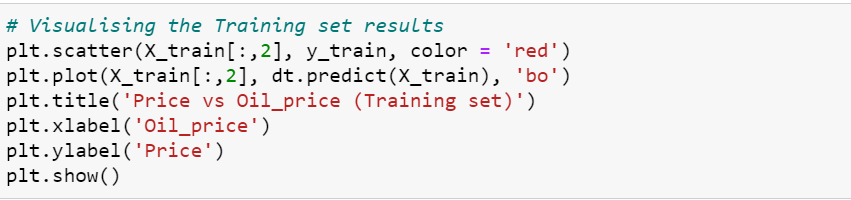
In (Получение среднеквадратической ошибки):



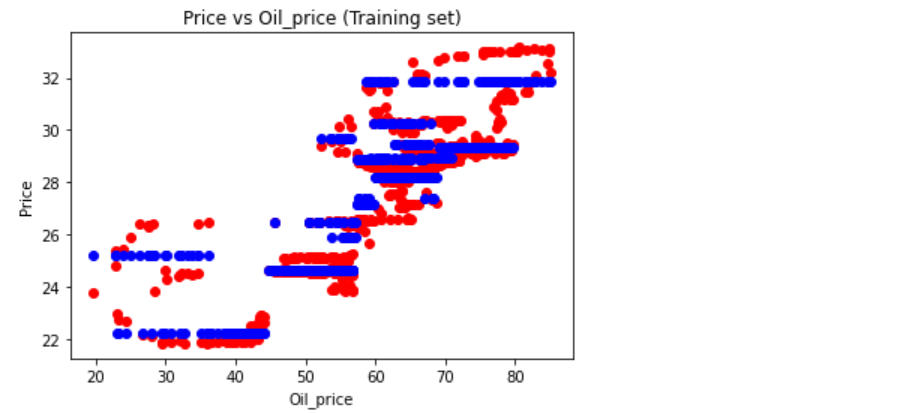
Out (Получение среднеквадратической ошибки):



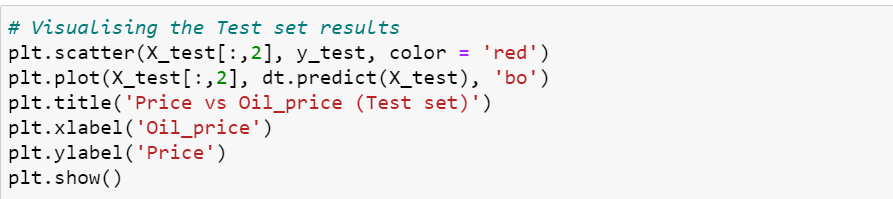
In (Визуализация обучающей выборки):



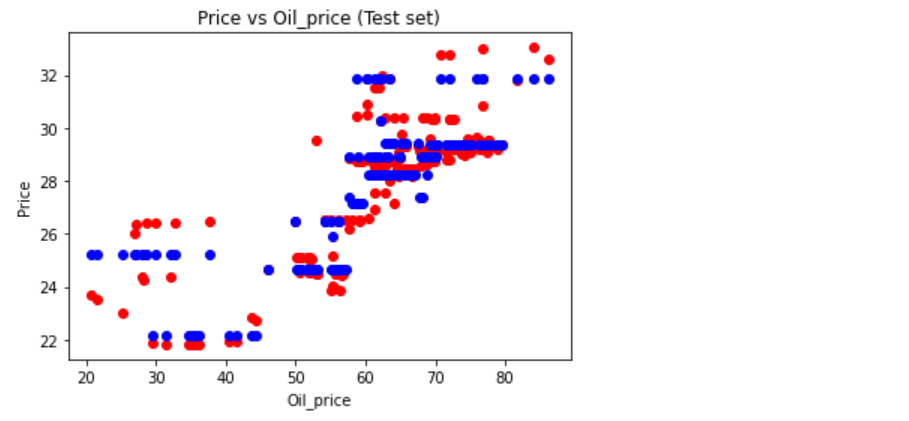
Out (Визуализация обучающей выборки):



In (Визуализация тестовой выборки):



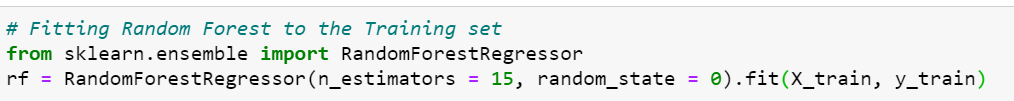
Out (Визуализация тестовой выборки):



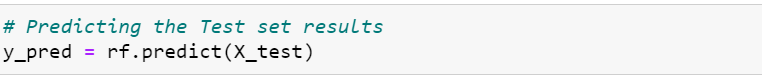
Некоторые прогнозные значения не попадают на реальные, однако степень расслоения не высокая.

**3.3. Построение случайного леса**

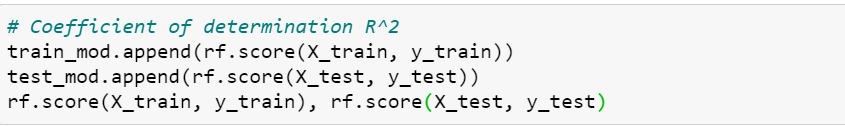
In (Построение случайного леса на всех объясняющих переменных):



In (Построение прогноза на тестовом множестве):



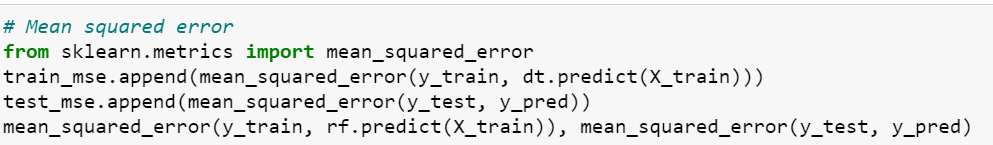
In (Коэффициент детерминации для тестовой и обучающей выборки):



Out (Коэффициент детерминации для тестовой и обучающей выборке):



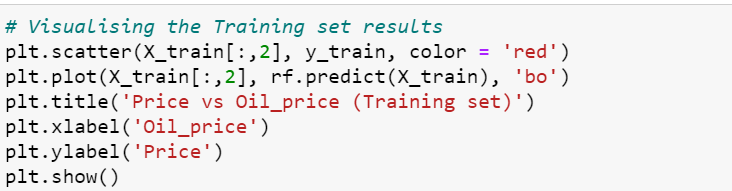
In (Cреднеквадратическая ошибка для тестовой и обучающей выборки):



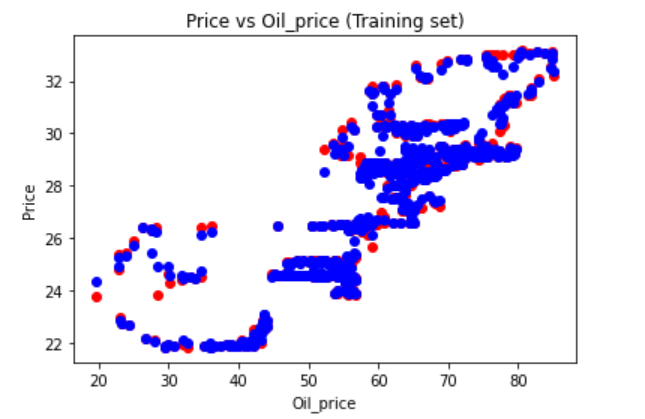
Out (Cреднеквадратическая ошибка для тестовой и обучающей выборки):



In (Визуализация обучающей выборки):

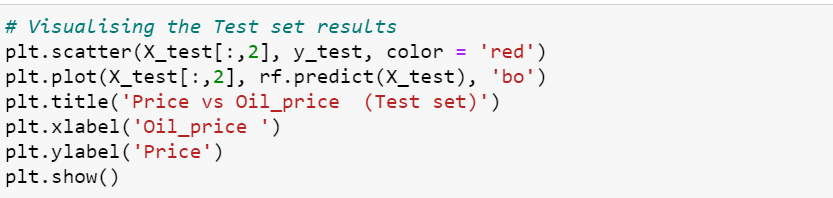


Out (Визуализация обучающей выборки):

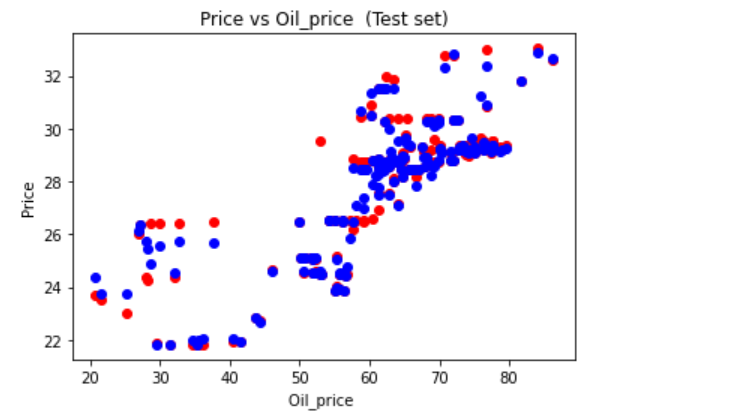


Прогнозные значения практически полностью попадают на реальные в обучающей выборке.

In (Визуализация тестовой выборки линейной модели):



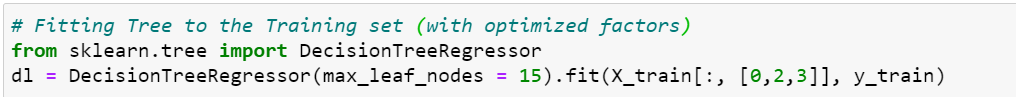
Out (Визуализация тестовой выборки линейной модели):



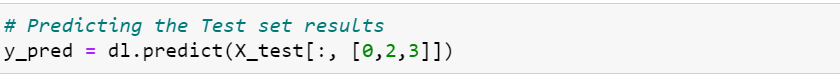
На тестовой выборке расхождение между прогнозными и реальными данными чуть сильнее, чем на обучающей выборке, но всё равно незначительно.

**3.4. Построение дерева решений на наиболее значимых факторах**

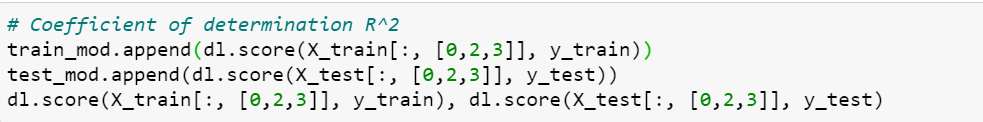
In (Построение дерева решений на наиболее значимых факторах):



In (Построение прогноза на всех объясняющих переменных):



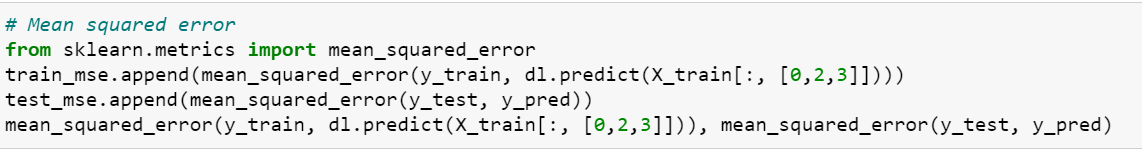
In (Коэффициент детерминации для тестовой и обучающей выборки):



Out (Коэффициент детерминации для тестовой и обучающей выборке):



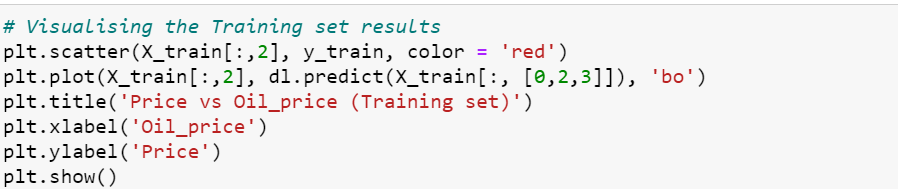
In (Получение среднеквадратической ошибки):



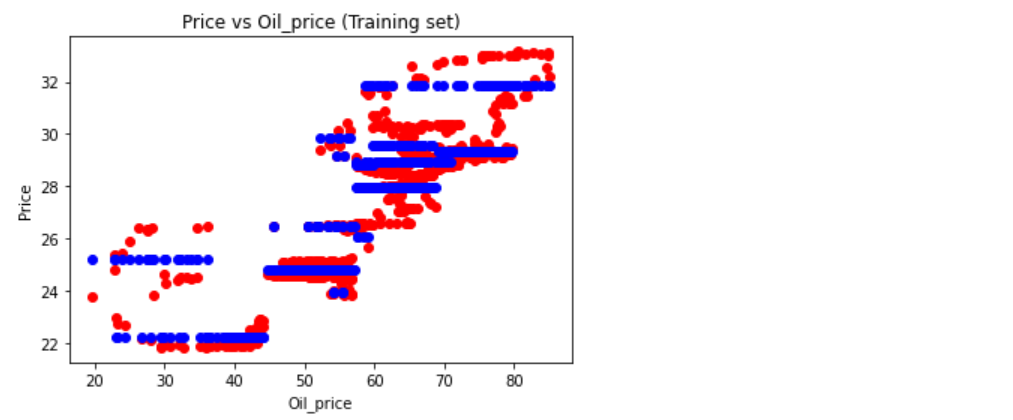
Out (Получение среднеквадратической ошибки):



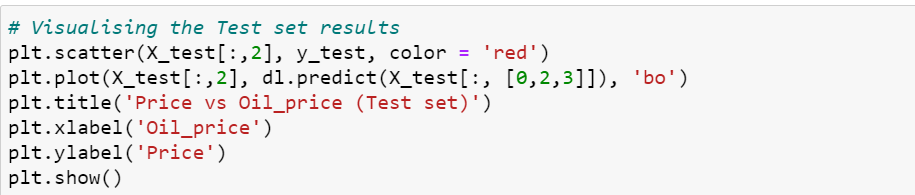
In (Визуализация обучающей выборки):



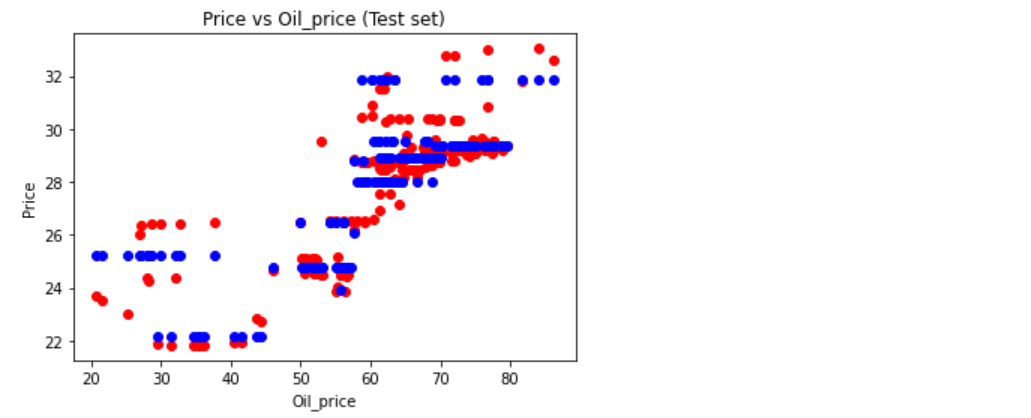
Out (Визуализация обучающей выборки):



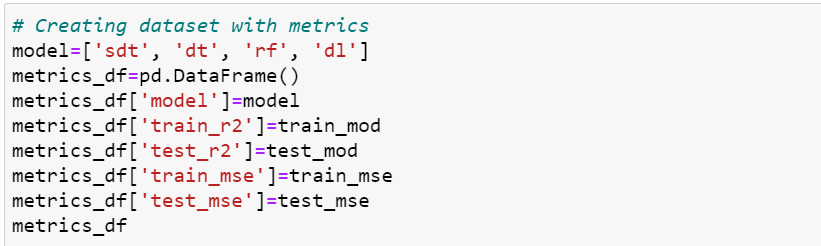
In (Визуализация тестовой выборки):



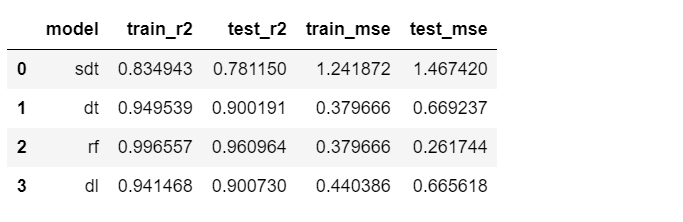
Out (Визуализация тестовой выборки):



In (Создание сравнительной таблицы R^2 & mse по различным моделям):

****

Out (Сравнительная таблица R^2 & mse по различным моделям):

****

**Вывод.** Лучшей моделью с наибольшим коэффициентом детерминации на обучающей и тестовой выборке, а также наименьшей среднеквадратической ошибкой на обучающей и тестовой выборке является модель случайного леса