**Федеральное государственное образовательное бюджетное учреждение**

**высшего образования**

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ РФ»**

**Департамент анализа данных и машинного обучения**

**Отчет по практике №2**

по дисциплине «системы управления базами данных»

Студента группы ПМ23-1

Факультета информационных технологий и анализа больших данных

Тищенко И.С.

Преподаватель

Михайлова С.С.

Москва 2024

Анализ данных компаний по добыче нефти и сырого газа.

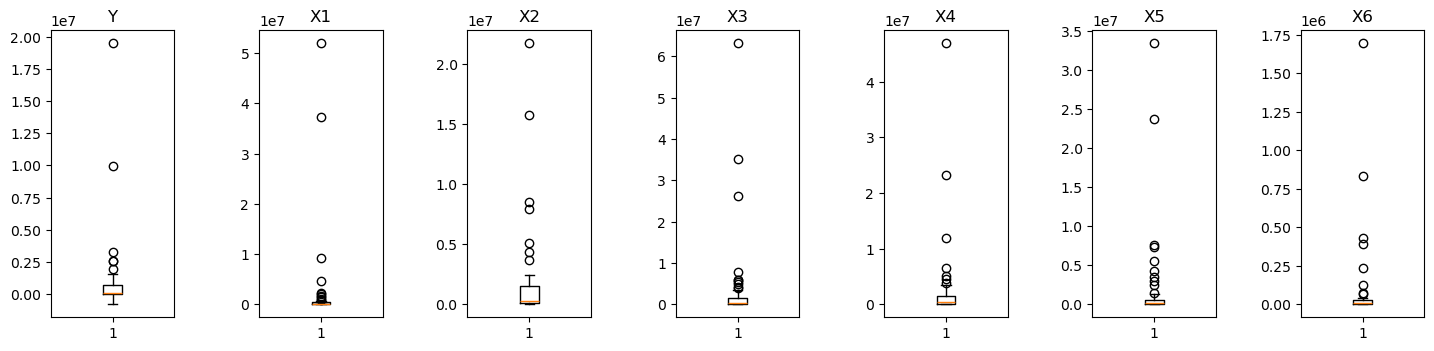
Получим данные из файла **«Домашнее задание №2 ММР.xlsx»** и представим их в виде объекта класса pandas.DataFrame. Для удобства работы с данными выделим столбцы с 1-го по 8-й и переведем их заголовки на английский язык. Создадим объект analysis класса BoxAn, в котором заблаговременно пропишем все нужные функции, предоставим как инициализирующий атрибут датасет из таблицы, загруженный ранее.

Здесь показана не вся таблица, а ее часть для иллюстрации.



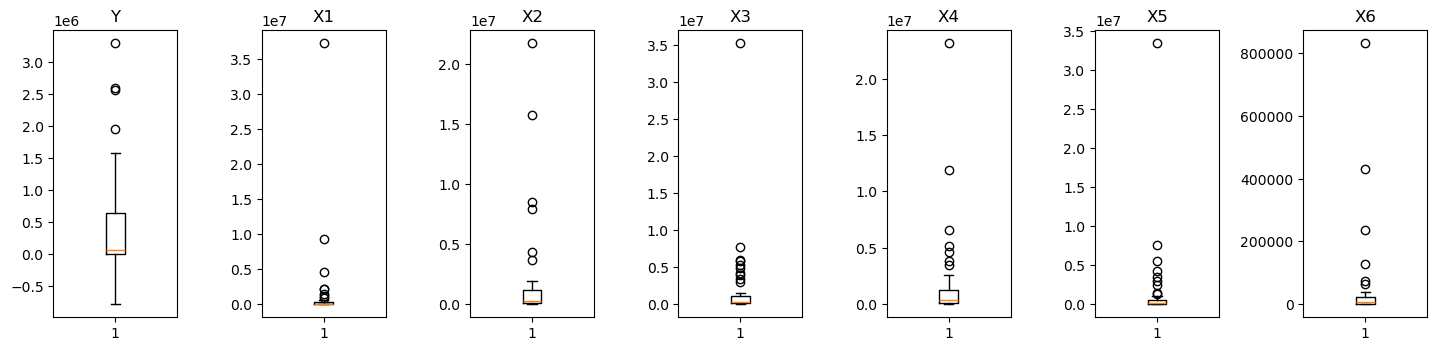
Задание 1: проанализировать ряды данных. Посмотреть, если ли выбросы, если необходимо удалить. Обосновать удаление.

Построим диаграммы вида Ящик с усами для каждого из столбцов с помощью методов библиотеки matplotlib:



# **Модель I**

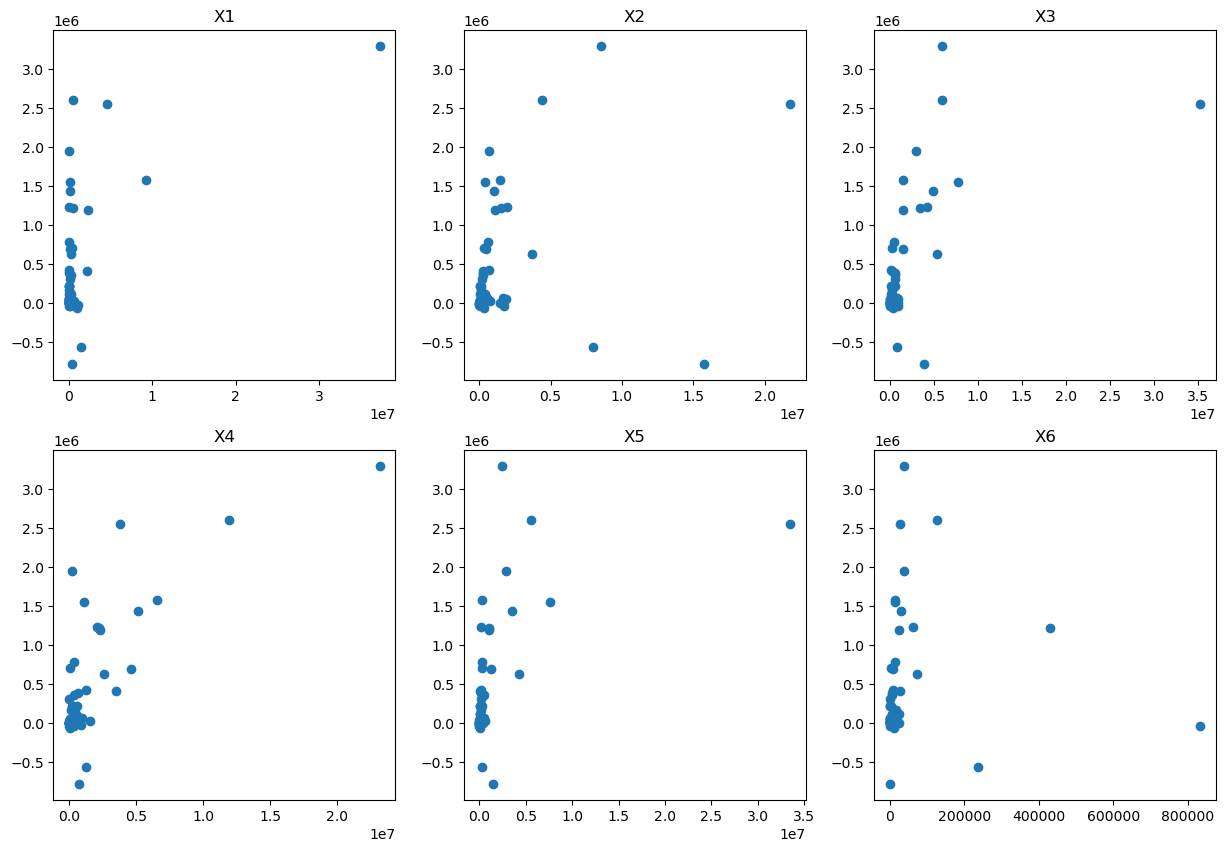
Заметим, что Y есть слишком сильные выбросы сверху. Удалим строки с ними и посмотрим, насколько улучшаться диаграммы:



Посчитаем удаление части выбросов для Y успешным.

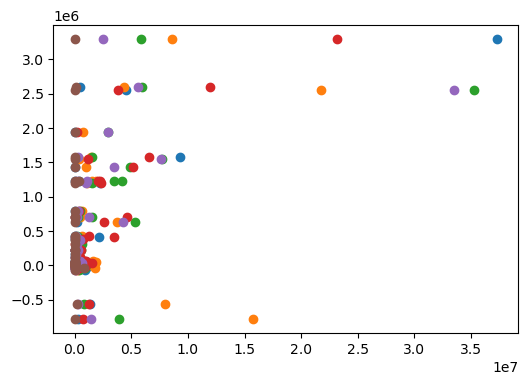
## Задание 2: построить диаграммы рассеяния Прибыли с регрессорами. Провести визуальный анализ.

Построим диаграммы рассеяния прибыли с каждым регрессором по отдельности.



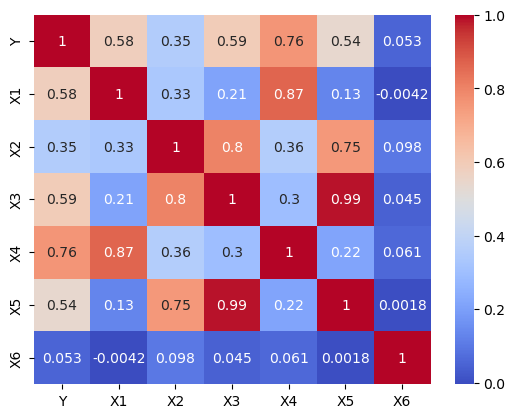
Как видно из каждой диаграммы, в каждой паре регрессор-прибыль наблюдается прямая связь.

Построим общую диаграмму рассеяния прибыли с регрессорами.

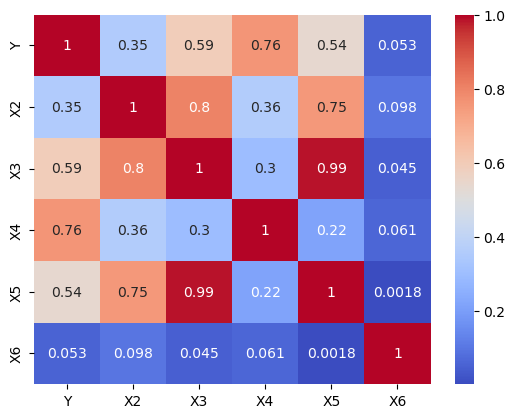


## Задание 3: построить корреляционную матрицу и провести ее анализ.

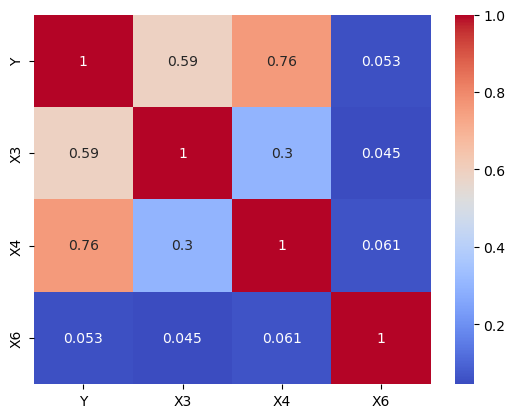
Построим корреляционную матрицу и представим ее в виде тепловой карты:



Заметим, что между атрибутами Х1 и Х4 – сильная связь, а значит они почти коллинеарные, а уже это, в свою очередь, говорит о том, что один из них нужно удалить. Проверим, какой из них больше влияет на Y: корреляция между Х1 и Y равна 0,58, а между Х4 и Y равна 0,76. Из этого сравнения можно сделать вывод, что атрибут Х1 можно убрать из анализируемой выборки:



Аналогичным методом удалим атрибуты Х2 и Х5:

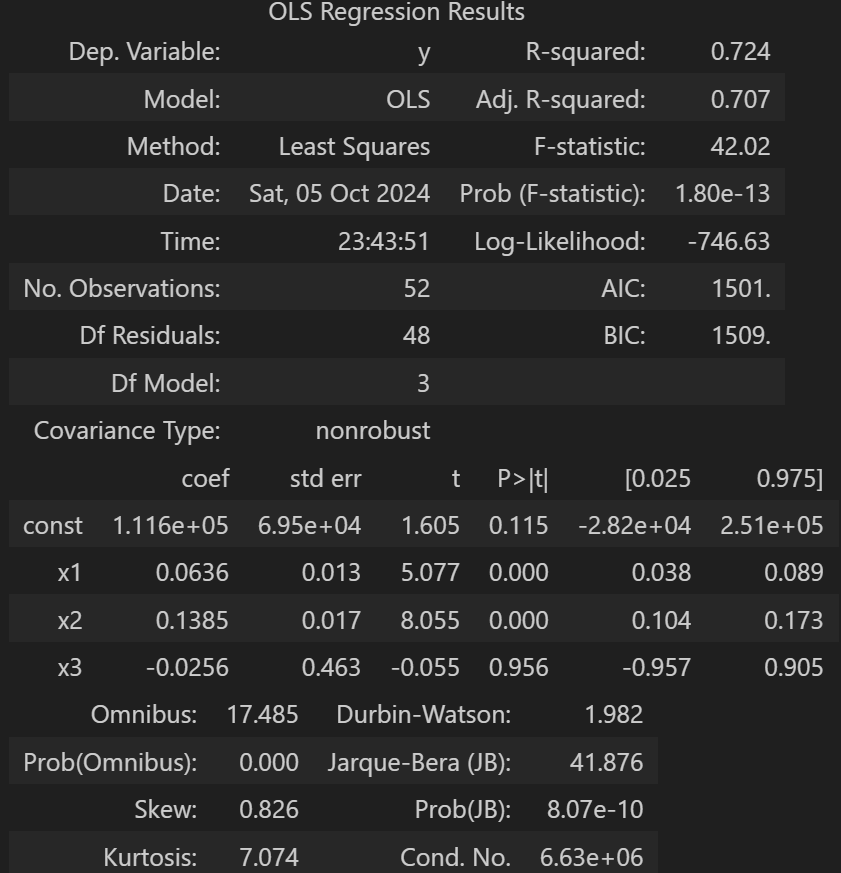


Проанализируем полученные связи между Y и атрибутами:

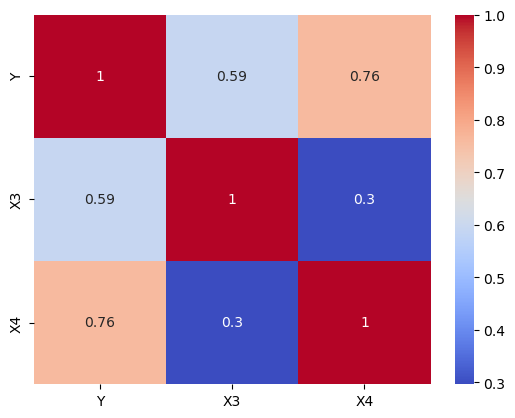
* Связь между Y и X3 - ПРЯМАЯ УМЕРЕННАЯ
* Связь между Y и X4 - ПРЯМАЯ СИЛЬНАЯ
* Связь между Y и X6 - ПРЯМАЯ ОЧЕНЬ СЛАБАЯ

## Задание 4: построить модель со значимыми факторами. Факторы отбирать методом пошагового отбора. Сделать выводы и обосновать.

С помощью библиотеки statsmodels.api создадим модель на основе факторов, которые есть сейчас и фактора Х0 – столбца нулей – и выведем информацию о модели:



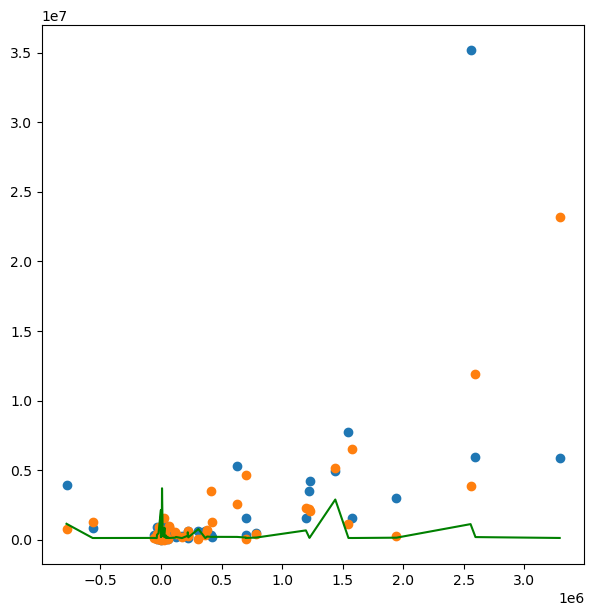
Как видно из результирующей таблицы по модели, факторы Х0 и Х6 не подходят по значению P-величины, поэтому их нужно исключить. Построим конечную тепловую карту матрицы корреляций.



Создадим новую модель на основе полученных данных и выведем ее результирующую таблицу:



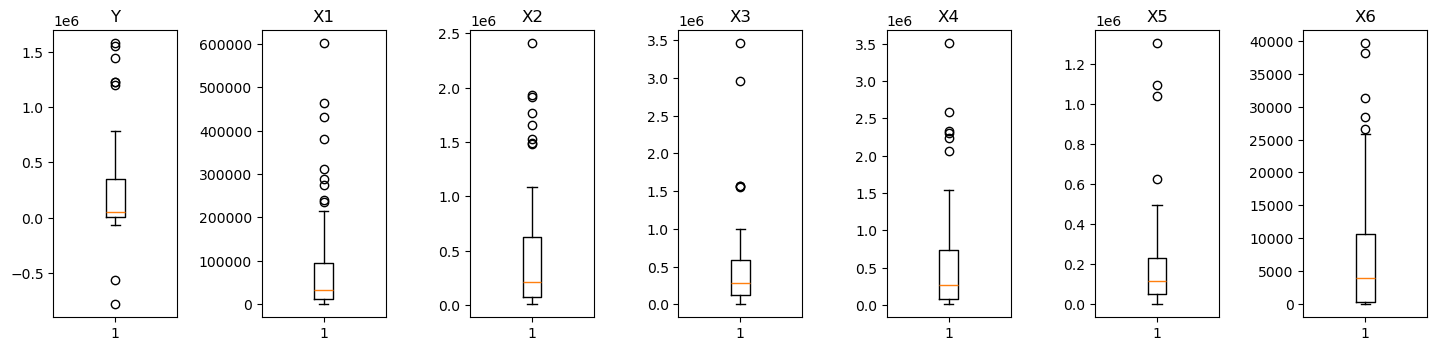
На основе обработанных данных создадим регрессионную модель из библиотеки scikit learn и построим конечную кривую и диаграмму рассеяния:



Как видно из коэффициента детерминации, модель точна на 78%. Да, это много, но можно ли сделать такую модель на основе этих данных, чтобы она имела больший коэффициент детерминации?

# **Модель II**

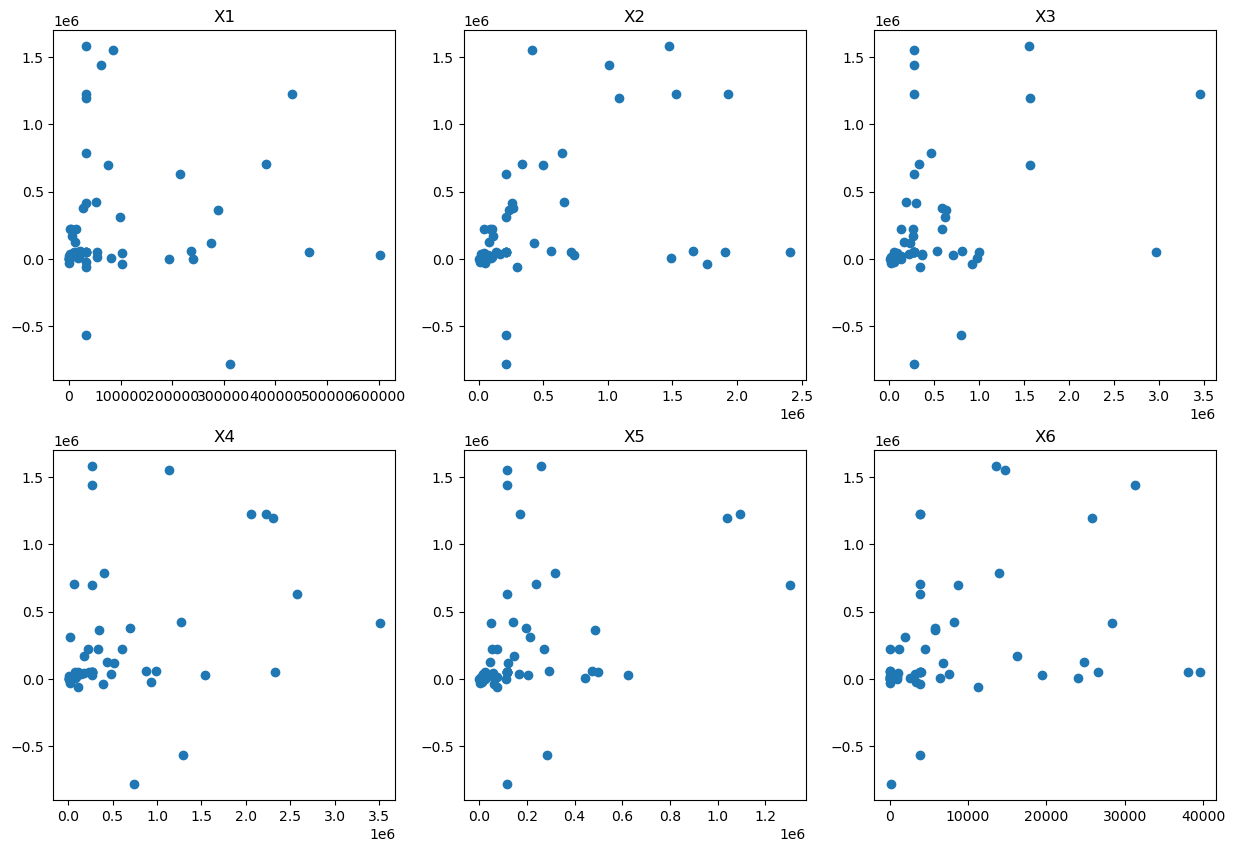
Предполагается, что в изначальных данных было больше выбросов, поэтому все выбросы, не входящие в межквартильный размах стоит заменить на медиану и провести все те же действия, кроме удаления строк из-за выбросов Y.



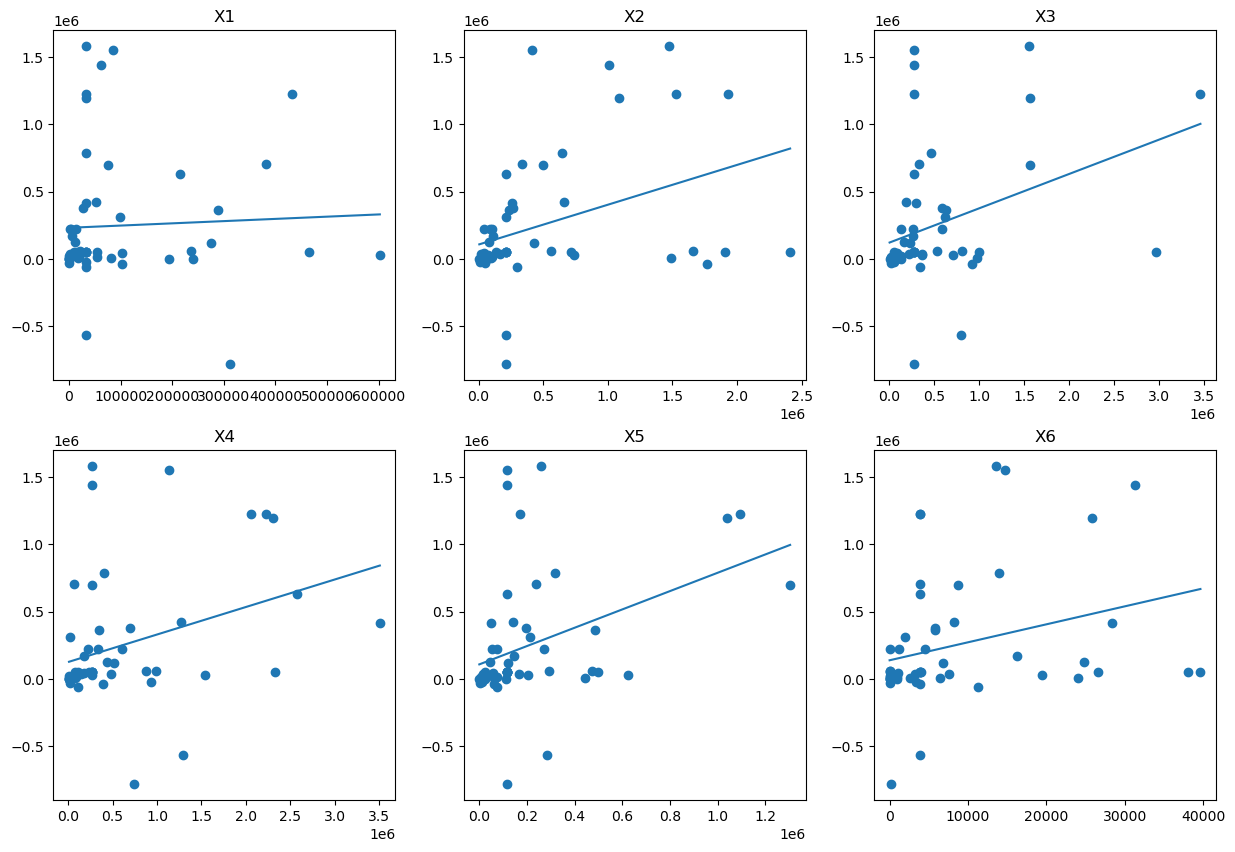
Как видно из диаграмм ящик с усами, выбросов стало гораздо меньше.

## Задание 2: построить диаграммы рассеяния Прибыли с регрессорами. Провести визуальный анализ.

Построим диаграммы рассеяния прибыли с каждым регрессором по отдельности.

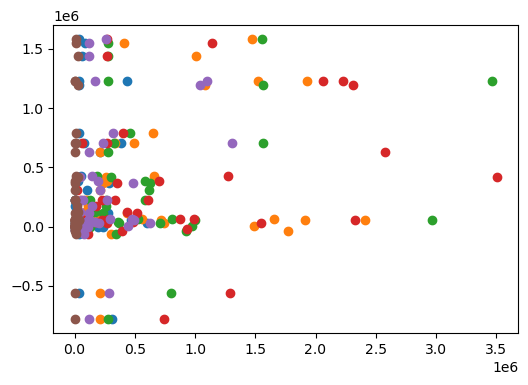


В этом случае мы не видим ничего определенного из диаграмм рассеяния, но, может быть, картина будет яснее с использованием регрессионных прямых.



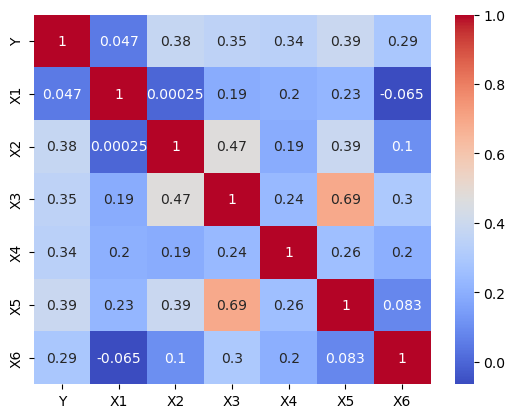
Такое угол прямых говорит о прямой связи между каждым атрибутом и прибылью.

Построим общую диаграмму рассеяния прибыли с регрессорами.

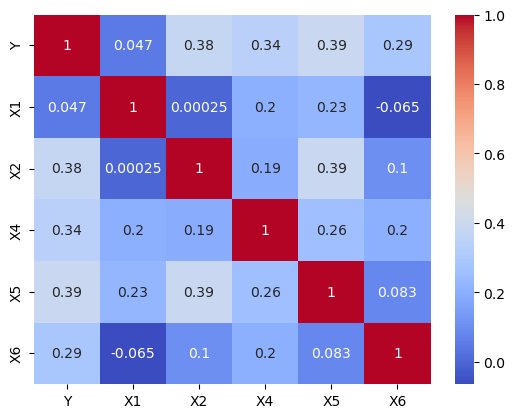


## Задание 3: построить корреляционную матрицу и провести ее анализ.

Построим корреляционную матрицу и представим ее в виде тепловой карты:



Заметим, что между атрибутами Х5 и Х3 – почти сильная, т.е. умеренная связь. Предположим, что удаление одного из факторов повлияет на модель. Проверим, какой из них больше влияет на Y: корреляция между Х5 и Y равна 0,39, а между Х3 и Y равна 0,35. Из этого сравнения можно сделать вывод, что атрибут Х3 можно убрать из анализируемой выборки:

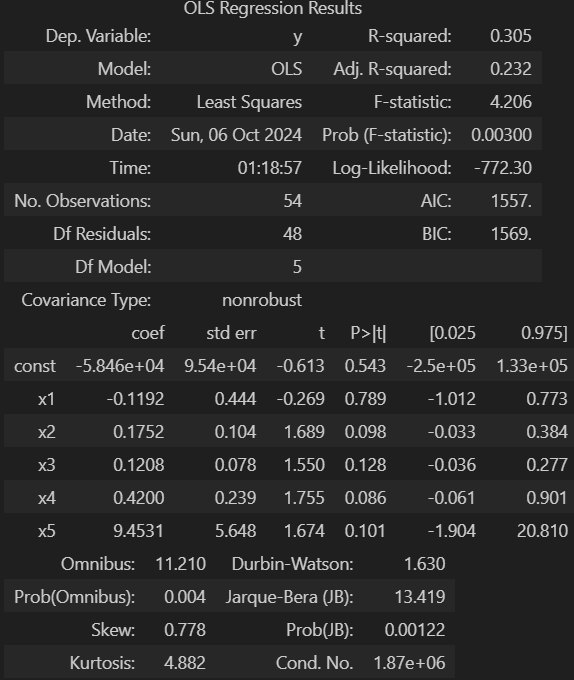


Проанализируем полученные связи между Y и атрибутами:

* Связь между Y и X1 - ПРЯМАЯ ОЧЕНЬ СЛАБАЯ
* Связь между Y и X2 - ПРЯМАЯ СЛАБАЯ
* Связь между Y и X4 - ПРЯМАЯ СЛАБАЯ
* Связь между Y и X5 - ПРЯМАЯ СЛАБАЯ
* Связь между Y и X6 - ПРЯМАЯ ОЧЕНЬ СЛАБАЯ

## Задание 4: построить модель со значимыми факторами. Факторы отбирать методом пошагового отбора. Сделать выводы и обосновать.

С помощью библиотеки statsmodels.api создадим модель на основе факторов, которые есть сейчас и фактора Х0 – столбца нулей – и выведем информацию о модели:

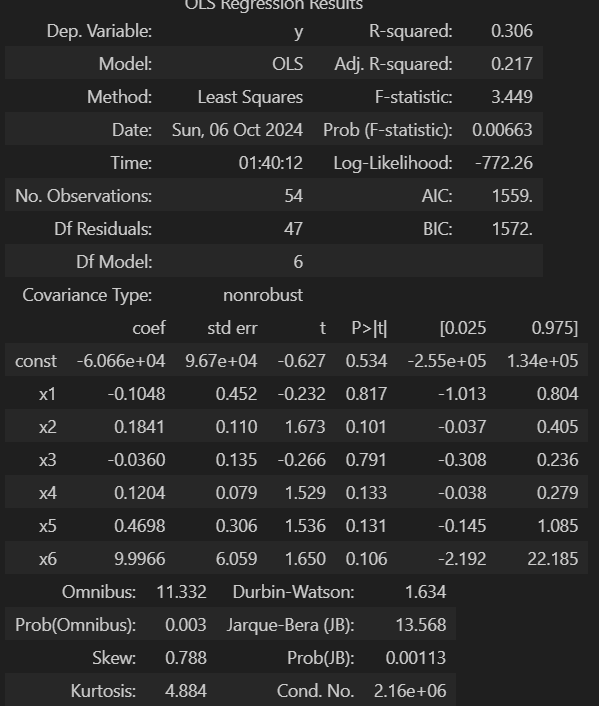


Как видно из результирующей таблицы, коэффициент детерминации равен 30,5%, так что предыдущая модель была лучше.

# **Модель III**

Если же вернуться на несколько шагов назад и не предполагать, что нужно удалить Х3, то модель становится не ощутимо лучше.

## Задание 4: построить модель со значимыми факторами. Факторы отбирать методом пошагового отбора. Сделать выводы и обосновать.



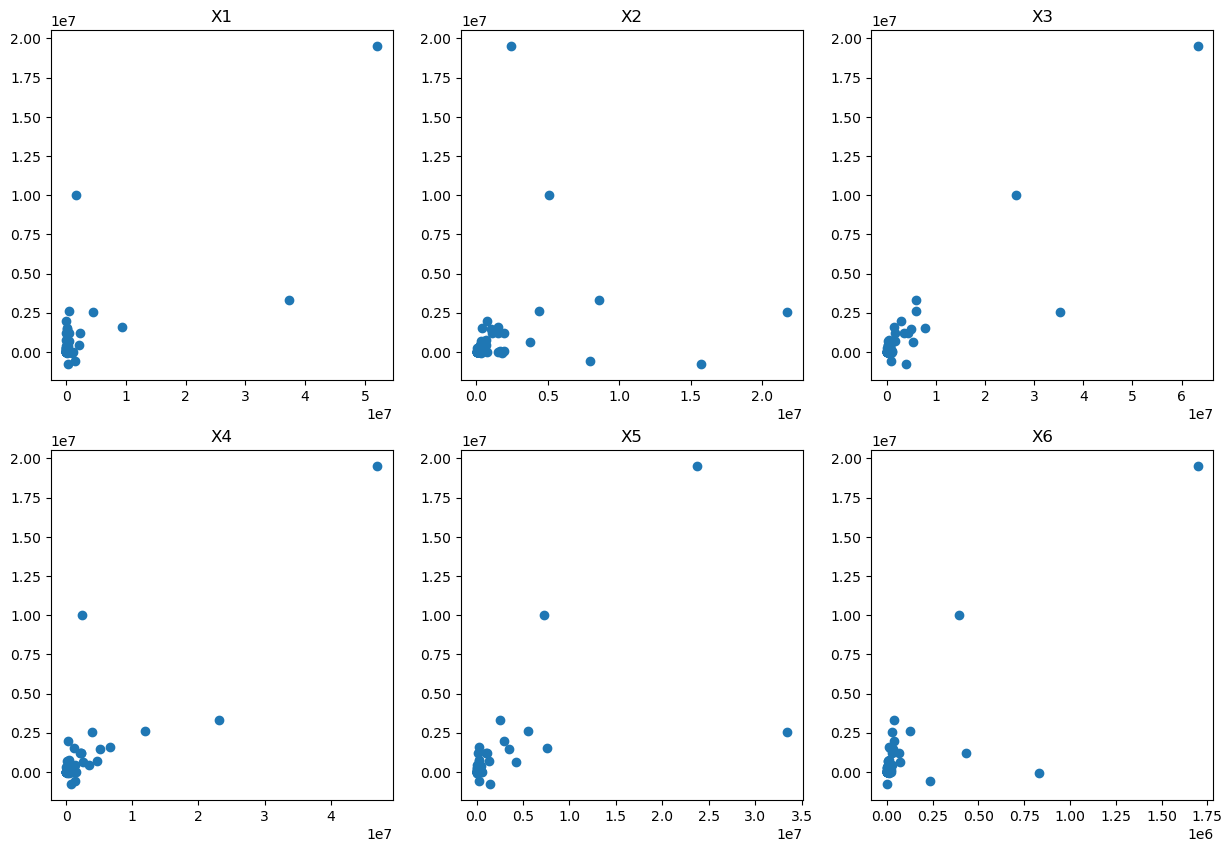
Коэффициент Детерминации равен 30,6%, что все равно меньше, чем этот же коэффициент, но полученный из модели с другими изменениями в изначальных данных.

# **Модель IV**

Предположим, что выбросы, показываемые на диаграммах ящик с усами не имеют смысла в исключении и/или замене, и выполним задания относительно такого набора данных.

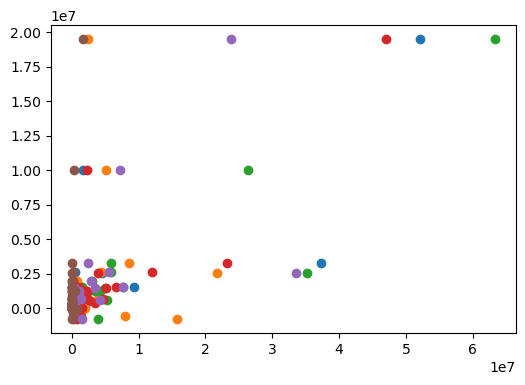
## Задание 2: построить диаграммы рассеяния Прибыли с регрессорами. Провести визуальный анализ.

Построим диаграммы рассеяния прибыли с каждым регрессором по отдельности.



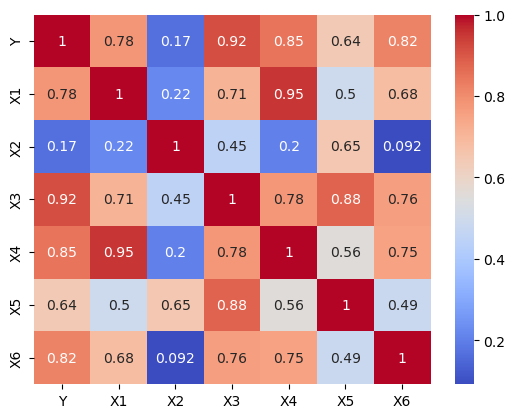
Из диаграмм рассеяния можно извлечь информацию о том, что все парные коэффициенты корреляции между регрессорами и Y будут положительными.

Построим общую диаграмму рассеяния прибыли с регрессорами.

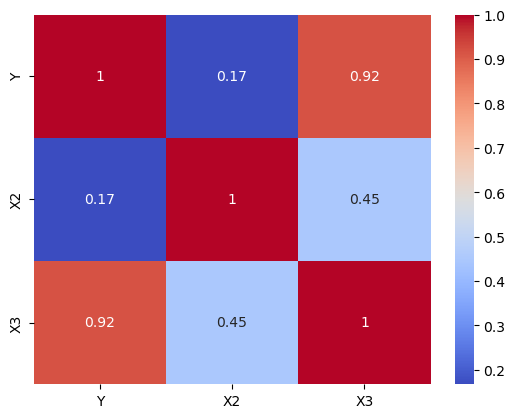


## Задание 3: построить корреляционную матрицу и провести ее анализ.

Построим корреляционную матрицу и представим ее в виде тепловой карты:



Методом, описанным для первой модели, извлечем из рассмотрения все атрибуты, кроме X2 и X3.

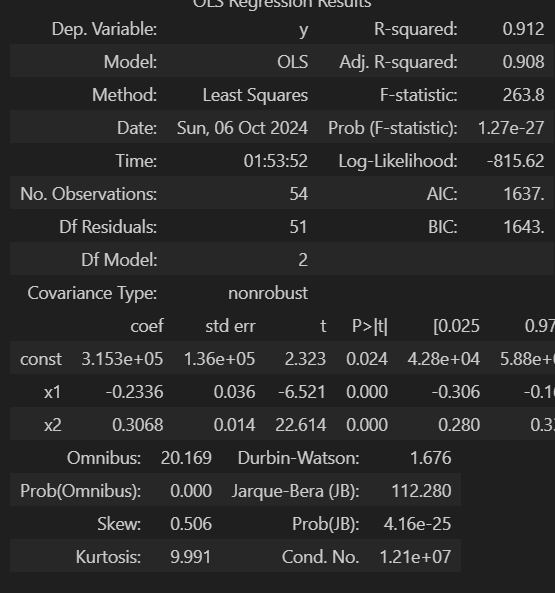


Проанализируем полученные связи между Y и атрибутами:

* Связь между Y и X2 - ПРЯМАЯ ОЧЕНЬ СЛАБАЯ
* Связь между Y и X3 - ПРЯМАЯ СИЛЬНАЯ

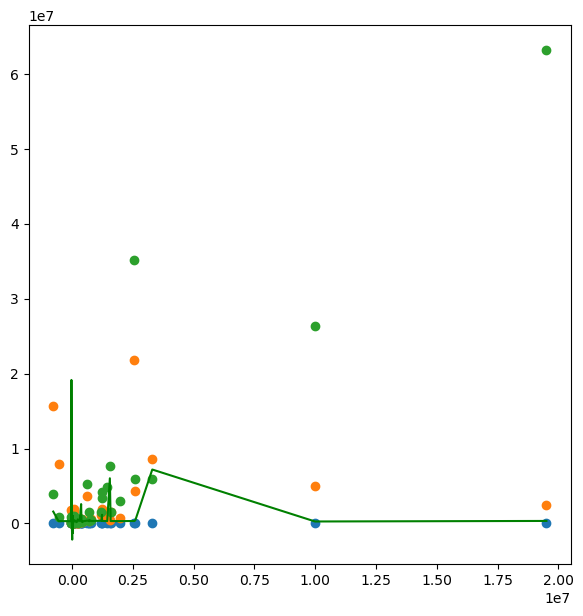
## Задание 4: построить модель со значимыми факторами. Факторы отбирать методом пошагового отбора. Сделать выводы и обосновать.

С помощью библиотеки statsmodels.api создадим модель на основе факторов, которые есть сейчас и фактора Х0 – столбца нулей – и выведем информацию о модели:



Все атрибуты прошли проверку по P- значению, значит они являются значимыми.

На основе обработанных данных создадим регрессионную модель из библиотеки scikit learn и построим конечную кривую и диаграмму рассеяния:



Как видно из результирующей таблицы, коэффициент детерминации для этой модели равен 91,2%, что является наибольшим результатом из всех рассмотренных моделей.

# **Вывод.**

Не всегда предобработка данных улучшает показатели модели. Даже если делать предобработку данных, то нужно понимать, что разные данные могут по-разному влиять в купе при изменениях.