|  |
| --- |
| МИНОБРНАУКИ РОССИИ  Федеральное государственное  бюджетное образовательное учреждение  высшего образования  **«МИРЭА – Российский технологический университет»**  **РТУ МИРЭА** |

Институт информационных технологий

Кафедра корпоративных информационных систем

|  |  |  |
| --- | --- | --- |
| **ОТЧЕТ**  **по лабораторной работе №4** | | |
| **по дисциплине** | | |
| **«Структуры и алгоритмы обработки данных»**  **Тема лабораторной работы: «**Корреляционный и регрессионный анализ**»** | | |
| Студент группы | ИКБО-07-18 | Зейналов М.Г. |
| Принял | ассистент кафедры КИС | Габриелян Г.А. |
|  |  |  |
| Выполнено | «\_\_\_» \_\_\_\_\_\_ 201\_\_ г. | *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* |
|  |  | *(подпись студента)* |
| Зачтено | «\_\_\_» \_\_\_\_\_\_ 201\_\_ г. | *\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_* |
|  |  | *(подпись преподавателя)* |

Москва 2020

1. **Часть №1**
   1. **Теоретическое часть**

**1.1.1 Интеллектуальный анализ данных**

**Интеллектуальный анализ данных** (Data Mining) – это раздел информатики, изучающий процессы обработки данных с целью получения полезной информации и принятия решений.

Основная польза разрабатываемых методов анализа данных заключается в некоторой предсказательной способности: проанализировав некоторый набор данных, информационная система анализа данных должна обучиться для дальнейшего распознавания или прогнозирования некоторых участков данных в ситуациях, когда часть данных утеряна или неизвестна.

Кроме этого, системы анализа данных могут решать задачи редуцирования объёма данных с целью устранения избыточности, визуализации данных для их удобного восприятия человеком, моделирования новых данных по имеющимся данным и др.

Впервые понятие Data Mining появилось в 1989 году. Изначально оно было связано с автоматизацией и оптимизацией запросов к крупным базам данных. Между тем понятие анализ данных (Data Analysis) существовало намного раньше и означало обработку и интерпретацию данных, полученных в ходе экспериментов, в основном научных.

С развитием науки и техники эти понятия расширялись и обобщались, стали очень близки друг к другу, и в настоящий момент тесно связаны как с анализом больших данных (Big Data), так и с понятием машинного обучения (Machine Learning).

**1.1.2 Генеральная совокупность**

**Генеральная совокупность** — совокупность всех объектов (единиц), относительно которых предполагается делать выводы при изучении конкретной задачи.

**Выборка** или **выборочная совокупность** — часть генеральной совокупности элементов, которая охватывается экспериментом (наблюдением, опросом).

**Репрезентативность** — соответствие характеристик выборки характеристикам генеральной совокупности в целом. Репрезентативность определяет, насколько возможно обобщать результаты исследования с привлечением определённой выборки на всю генеральную совокупность, из которой она была собрана.



Рисунок 1 – Переменные в статистическом анализе

**1.1.3 Основные меры центральной тенденции**

**Арифметическое среднее** — сумма всех наблюденных значений, делённая на их количество.

**Медиана** — значение, которое делит упорядоченные по возрастанию (убыванию) наблюдения пополам.

**Мода** — наиболее часто встречающееся значение.

**1.1.4 Меры изменчивости**

**Размах** — разность между максимальным и минимальным значением.

**Дисперсия** — средний квадрат отклонений индивидуальных значений признака от их средней величины.

**Среднеквадратическое отклонение** — квадратный корень из дисперсии.

**1.1.5 Нормальное распределение**

Нормальное распределение - распределение вероятностей, которое в одномерном случае задаётся функцией плотности вероятности, совпадающей с функцией Гаусса.

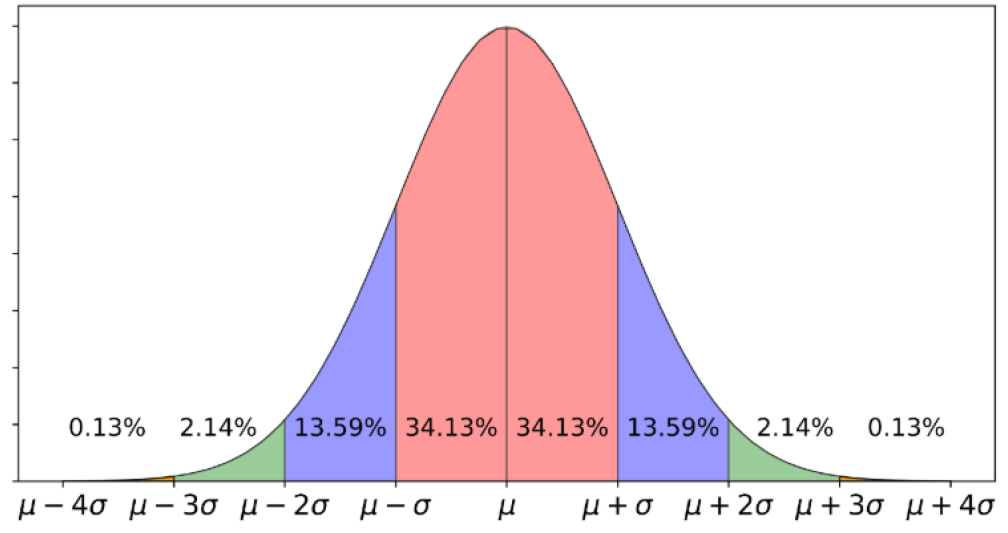


Рисунок 2 – Нормальное распределение

* 1. **Практическая часть**

В соответствии с вариантом входные данные следующие (Таблица 1).

Таблица 1 – Входные данные

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **data** | **params** | **type** | **n** | **m** | **s** |
| bands.csv | roughness,  press speed | cylinder size | 190 | 1 | 0.75 |

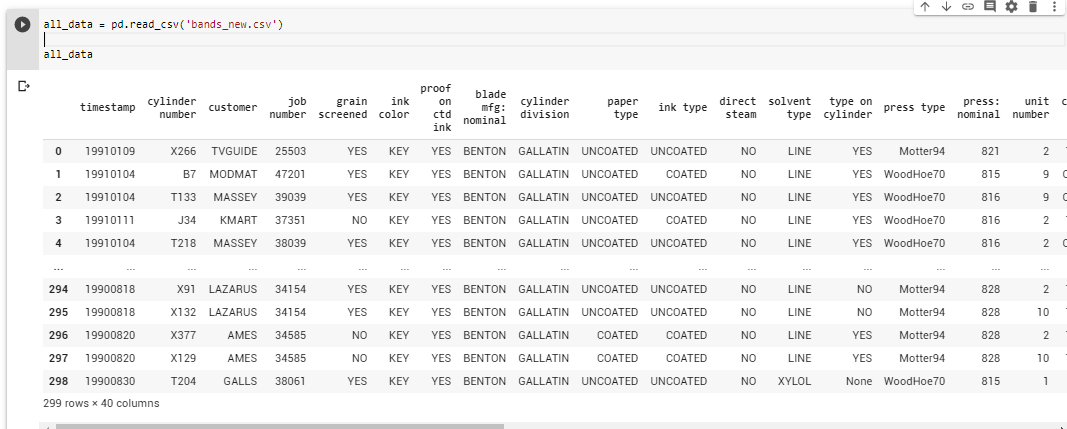


Рисунок 3 – Вывод данных

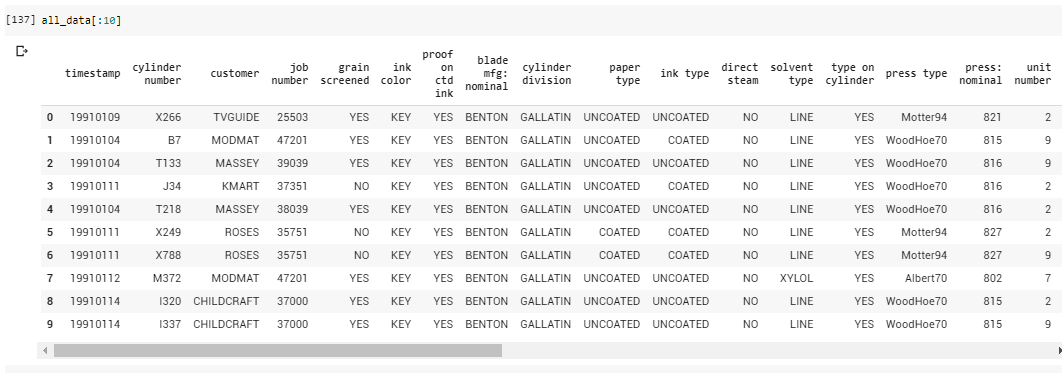


Рисунок 4 – Вывод первых 10-ти строк

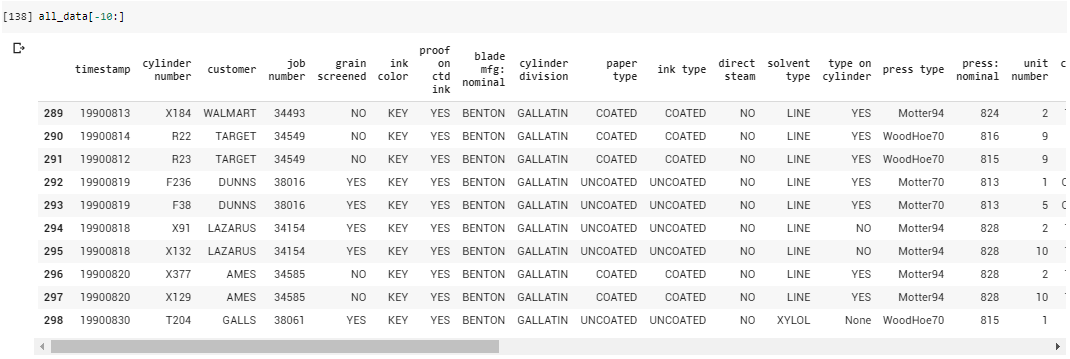


Рисунок 5 – Вывод последних 10-ти строк

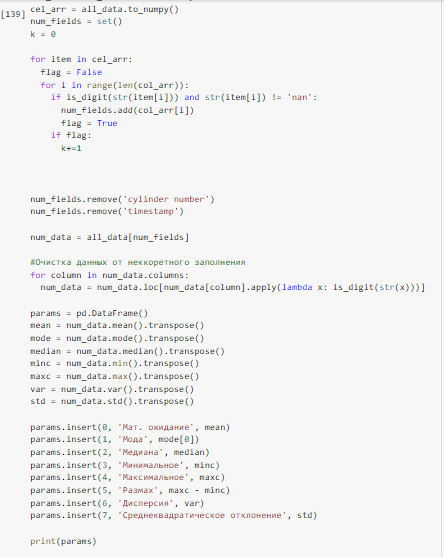


Рисунок 6 – Подсчёт основных характеристик

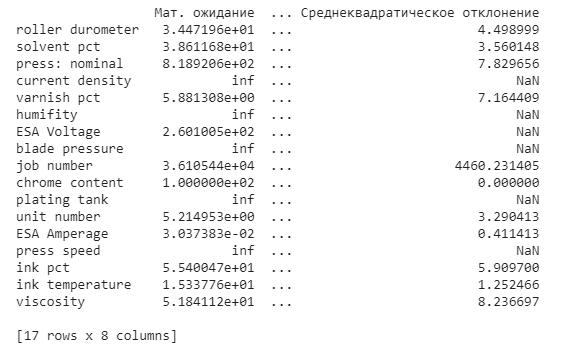


Рисунок 7 – Вывод основных характеристик

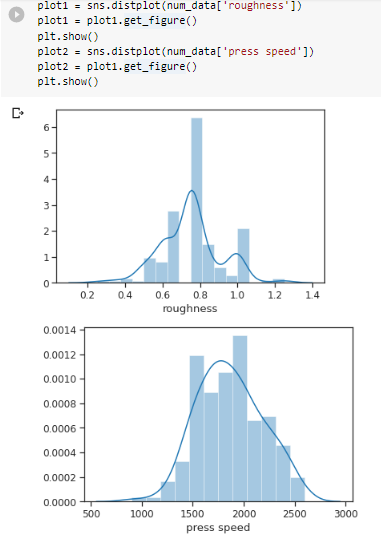


Рисунок 8 – Вывод гистограмм по показателям

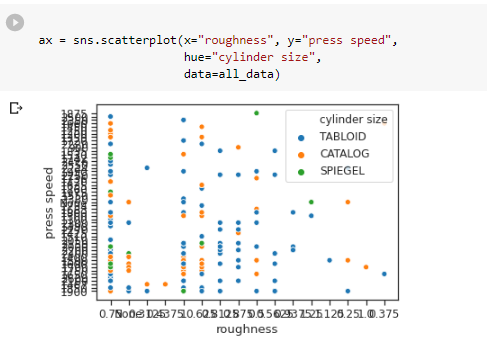


Рисунок 9 – Вывод точечной диаграммы

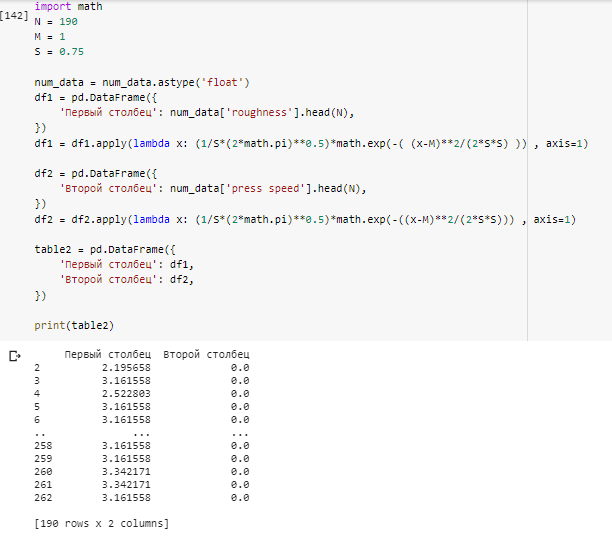


Рисунок 10 – Создание таблицы из двух столбцов

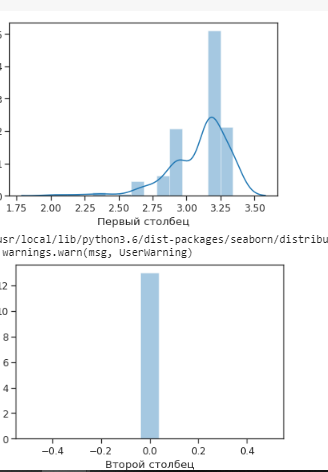


Рисунок 11 – Вывод гистограммы по столбцам

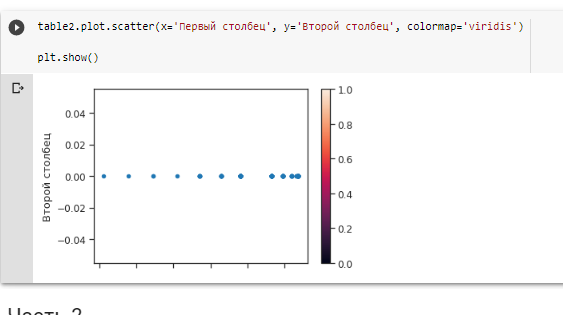


Рисунок 12 – Вывод точечной диаграммы

1. **Часть №2**
   1. **Теоретическое часть**

**2.1.1 Корреляционная зависимость**

**Корреляция** или **корреляционная зависимость** — статистическая взаимосвязь двух или более случайных величин (либо величин, которые можно с некоторой допустимой степенью точности считать таковыми). При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин.

Корреляция между RXY вещественными случайными величинами X и Y определяется как

где M – математическое ожидание случайной велечины.

Корреляция характеризует уровень взаимосвязи между случайными величинами X и Y, то есть, чем сильнее корреляция отличается от нуля, тем больше при изменении одной из величин меняется другая.

С учётом свойства линейности математического ожидания легко убедиться, что формула корреляции может быть переписана в виде

Значительная корреляция между случайными величинами всегда означает, что присутствует некая взаимосвязь между значениями конкретной выборки, но при другой выборке связь вполне может отсутствовать. Поэтому при нахождении взаимосвязи не нужно делать поспешных выводов о причинно-следственном характере величин, а следует рассмотреть наиболее полную выборку, чтобы делать какие-либо выводы. Коэффициенты корреляции устанавливают лишь статистические взаимосвязи, но не более того.

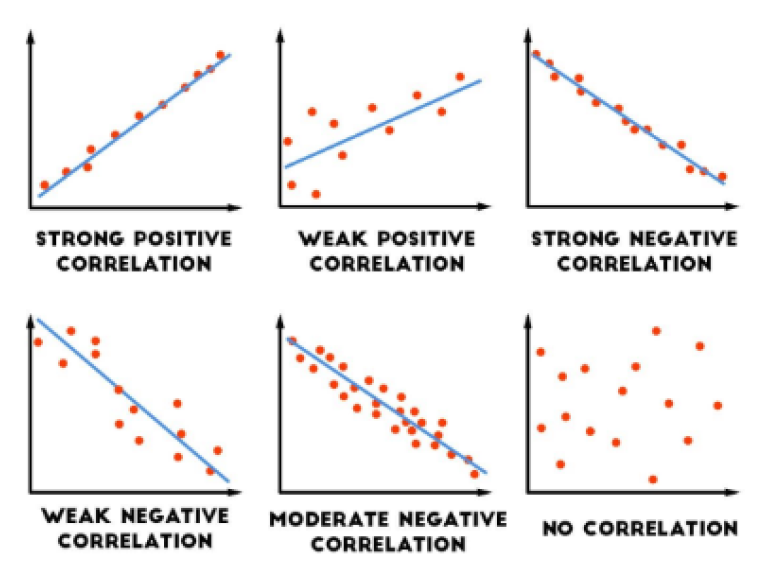


Рисунок 13 – Виды графиков корреляции

**2.1.2 Коэффициент корреляции**

Если произвести корреляцию на среднее геометрическое дисперсий, то полученная характеристика не будет зависеть от масштабов самих случайных величин. Нормированная таким образом корреляция называется коэффициентом корреляции Пирсона:

Свойства коэффициента корреляции:

1. Коэффициент корреляции между вещественными случайными величинами коммутативен: rXY = rYX;
2. Коэффициент корреляции между независимыми случайными величинами равен нулю;
3. Коэффициент корреляции по модулю не превосходит единицы: .

Пусть имеется выборка из значений независимых копий некоторой случайной величины X.

Известно, что состоятельной, несмещённой и эффективной статистической оценкой математического ожидания случайной величины X является выборочное среднее .

Таким образом, состоятельной оценкой коэффициента корреляции является.

где , ,

Чем ближе оценка к нулю, тем меньше связаны признаки и . С другой стороны, если эта оценка по модулю близка к единице, то это означает сильную взаимосвязь между признаками. Например, они могут вообще линейно зависеть друг от друга. В этом случае, как правило, нет смысла использовать для анализа оба признака, а достаточно оставить только один из них. Такие соображения позволяют редуцировать данные, уменьшить объём выборки, устранить избыточность информации, которая может негативно влиять на качество многих других методов анализа. Кроме того, иногда наличие взаимосвязи между некоторыми признаками может быть интересно само по себе.

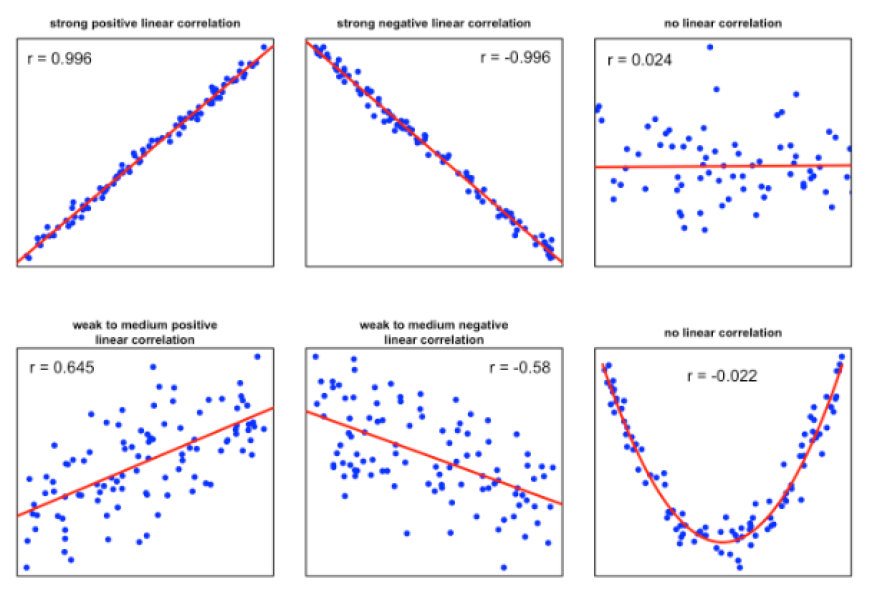


Рисунок 14 – виды графиков корреляции

**2.1.3 Гипотеза о некоррелированности**

Требуется оценить коэффициент корреляции генеральной совокупности (совокупность всех мысленно возможных объектов данного вида, над которыми проводятся наблюдения) на основе полученного значения коэффициент корреляции нашей выборки. Гипотезы формулируются следующим образом:

1. Основная гипотеза: (не существует корреляции между признаками и в генеральной совокупности)
2. Алтернативная гипотеза: (корреляции между признаками и в генеральной совокупности значима)

Когда основная гипотеза отвергается на опредленном уровне значимости, это значит, что существует значимое различие между p и 0. Когда основная гипотеза принимается, это значит, что значение p не сильно отличиается 0 и является случайным.

Для проверки статистической гипотезы о некоррелированности признаков и используется статистика

Которая имеет **t-распределение Стьюдента** с количеством степеней свободы . Здесь – это оценка коэффициента корреляции Пирсона между признаками и , а – объём выборки.

Пусть обозначает квантиль (значение, которое заданная случайная величина не превышает с фиксированной вероятностью) на уровне t-распределения Стьюдента с степенями свободы.

Вероятность того, что статистика будет отличаться от нуля больше, чем на величину при условии, что в действительности признаки и являются независимыми, составляет . Это значит, что можно считать признаки и зависимыми на уровне значимости , если выполняется неравенство

Чем с меньшим уровнем значимости отвергается гипотеза о некоррелированности, тем меньше вероятность того, что это сделано ошибочно. Таким образом, уровень значимости статистической гипотезы – это вероятность отвергнуть эту гипотезу, при условии, что на самом деле она верна. Если решить неравенство относительно , то можно найти граничный уровень значимости , для которого гипотеза о некоррелированности всё ещё отвергается, при том, что при чуть меньшем уровне значимости она уже была бы принята:

где – это функция распределения t-распределения Стьюдента с степенями свободы.

Это значение носит название **p-value**.

**p-value (p-критерий)** — вероятность получить для данной вероятностной модели распределения значений случайной величины такое же или более экстремальное значение статистики, по сравнению с ранее наблюдаемым, при условии, что нулевая гипотеза верна.

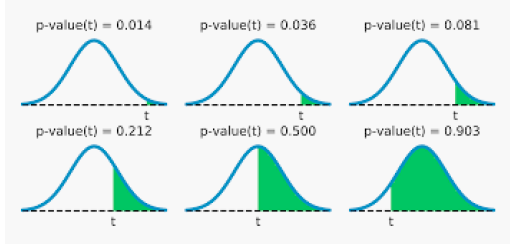


Рисунок 15 – Значение p-value в зависимости от t

Низкое значение для значимости соответствует низкой вероятности того, что экспериментальные результаты получились случайно, и наоборот. Уровни значимости записываются в виде десятичных дробей (таких как 0.01), что соответствует вероятности того, что экспериментальные результаты мы получили случайно (в данном случае вероятность этого 1%).

По соглашению, ученые обычно устанавливают уровень значимости своих экспериментов равным 0.05, или 5%. Это означает, что экспериментальные результаты, которые соответствуют такому критерию значимости, только с вероятностью 5% могли получиться чисто случайно. Другими словами, существует 95% вероятность, что результаты были вызваны тем, как ученый манипулировал экспериментальными переменными, а не случайно. Для большинства экспериментов 95% уверенности наличия связи между двумя переменными достаточно, чтобы считать, что они «действительно» связаны друг с другом.

Допустим мы видим корреляцию между признаками и в выборке и , следовательно, на 5% уровне значимости есть основания отвергнуть нулевую гипотезу о равенстве коэффициента корреляции нулю. Раз эту гипотезу отвергаем, считаем, что коэффициент корреляции в генеральной совокупности не равен 0, а следовательно, связь между признаками и действительно есть.

* 1. **Практическая часть**

В соответствии с вариантом входные данные следующие (Таблица 2).

Таблица 2 – Входные данные

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **data** | **params** | **n** | **mean** | **cov** |
| bands.csv | roughness, press  speed | 100 | (-1, 0) |  |

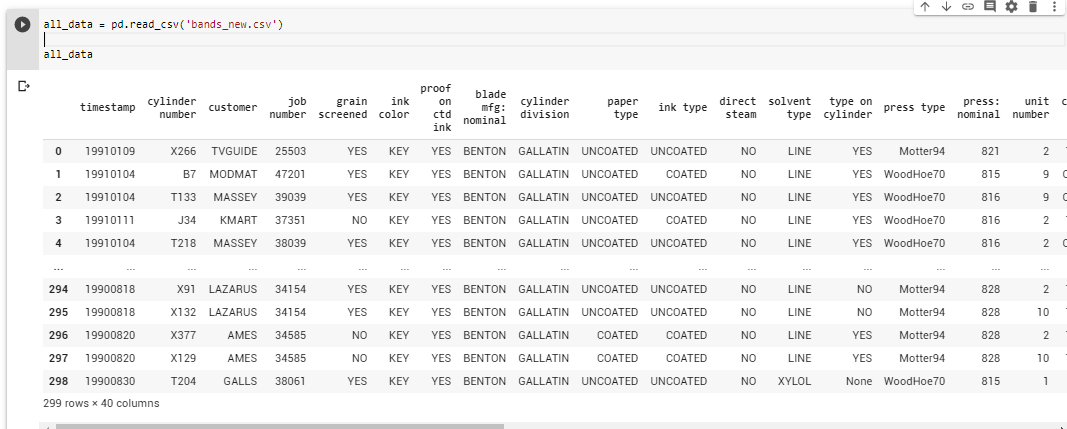


Рисунок 16 – Вывод данных

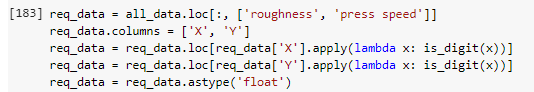


Рисунок 17 – Создание нового набор данных

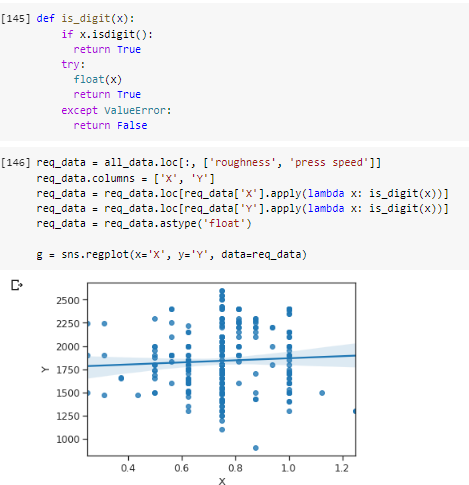


Рисунок 18 – Вывод точечного графика с регрессионной прямой



Рисунок 19 – Расчёт коэффициента Пирсона

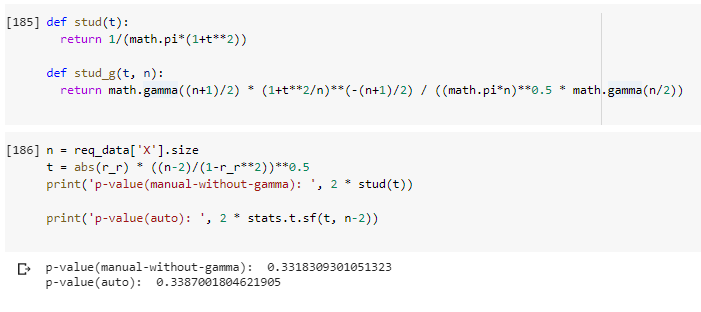


Рисунок 20 – Расчёт p-value

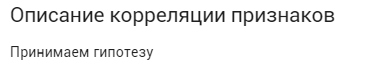


Рисунок 21 – Проверка гипотезы

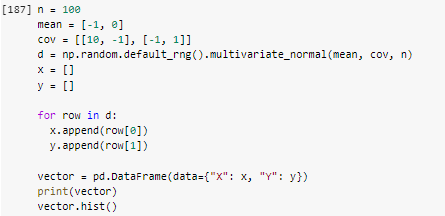


Рисунок 22 – Моделирование вектора

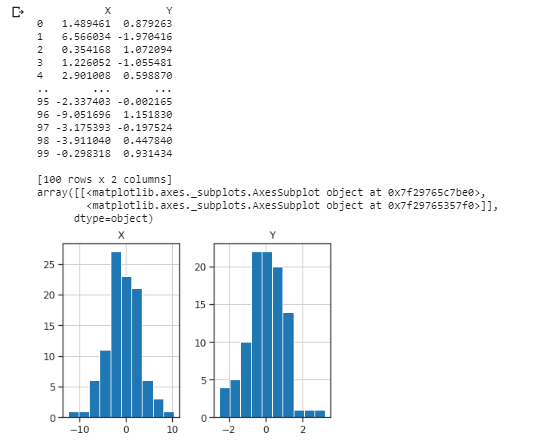


Рисунок 23 – Вывод вектора

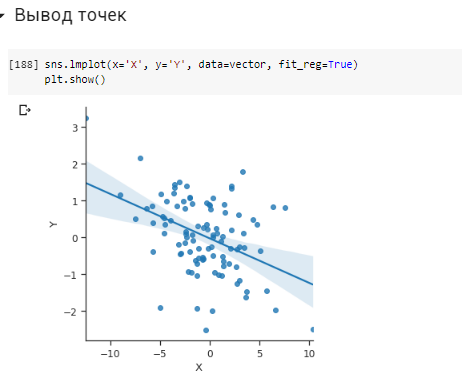


Рисунок 24 – Вывод точечного графика

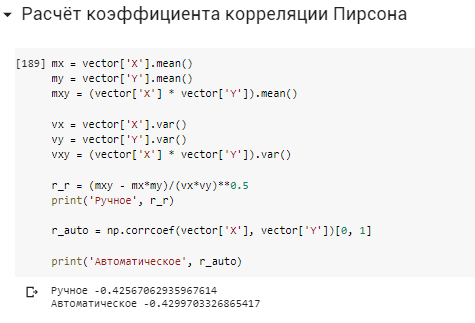


Рисунок 25 – Расчёт коэффициента Пирсона

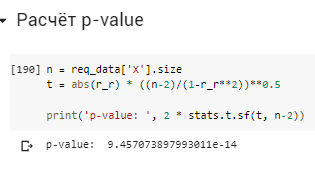


Рисунок 26 – Расчёт p-value

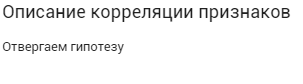


Рисунок 27 – Проверка гипотезы

1. **Часть №3**
   1. **Теоретическое часть**

3.1.1 Регрессионная модель

Пусть объекты наблюдения, характеризуются m признаками , но интерес представляет другой признак этих объектов Y. Требуется научиться по набору наблюдаемых признаков , научиться определять значение признака Y. Предполагается, что признак каким-то неизвестным образом зависит от признаков .

Если считать все эти признаки случайными величинами, то можно определить **регрессию** (regression) Y на X как условное математическое ожидание.

График функции называется **линией регрессии**, признаки – **независимыми переменными** (**предикторами**), а признак – **зависимой** (**выходной**) переменной.

Задача восстановления линии регрессии заключается в построении некоторой функции , которая наилучшим образом описывает данные из некоторой обучающей выборки , в которой каждому вектору предикторов ставится в соответствие зависимая переменная .

Понятно, что эту функцию следует искать в некотором конкретном классе функций. Наиболее популярным подходом является предположение, что данные отвечают линейной регрессионной модели вида

где и – это неизвестные коэффициенты в уравнении регрессии, – Гауссовский белый шум с нулевым средним (математическим ожиданием) и некоторой дисперсией .

Итак, будем задавать функцию регрессии линейной функцией от аргумента. Будем считать, нулевой признак равен для всех объектов равен единице (для возможности задания свободного ). Функций будет иметь вид.

Для заданной выборки , состоящей из пар , оптимальные коэффициенты обычно определяются из условия **минимума суммы квадратов отклонений**:

После вычисления производной функции и приравнения к нулю найдем решение в явном виде.

где строки матрицы – это признаковые описания наблюдаемых объектов.

Конечно в одномерном случае – это просто один числовой признак, и модель сильно упростится.

Для оценки качества приближения данных с помощью заданной регрессионной модели используется **коэффициент детерминации**.

где - сумма квадратов регрессионных остатков,

– обща дисперсия,

– соответственно, фактические и расчётные значения объясняемой переменной,

– выборочное среднее.

* 1. **Практическая часть**

В соответствии с вариантом входные данные следующие (Таблица 3).

Таблица 3 – Входные данные

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **data** | **params** | **n** | **a** | **b** | **sigma2** |
| bands.csv | roughness, press  speed | 130 | -2 | -10 | 0.05 |

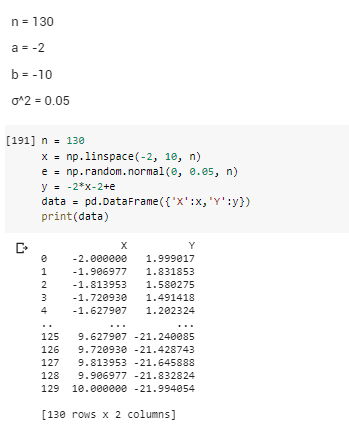


Рисунок 28 – Моделирование вектора



Рисунок 29 – Вывод точечной диаграммы

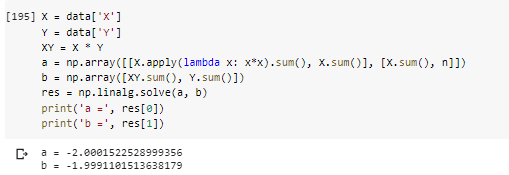


Рисунок 30 – Построение модели линейной регрессии

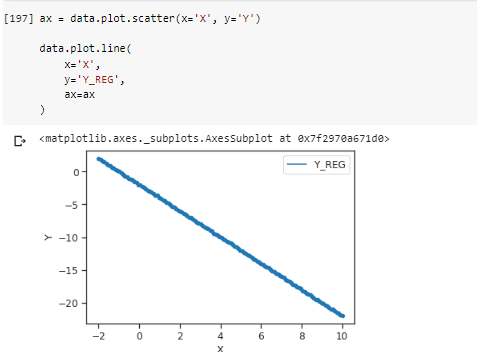


Рисунок 31 – Построение графика

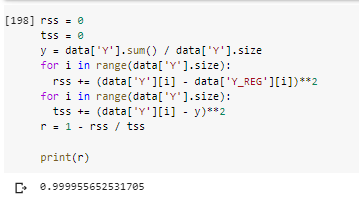


Рисунок 32 – Вычисление коэффициента детерминации

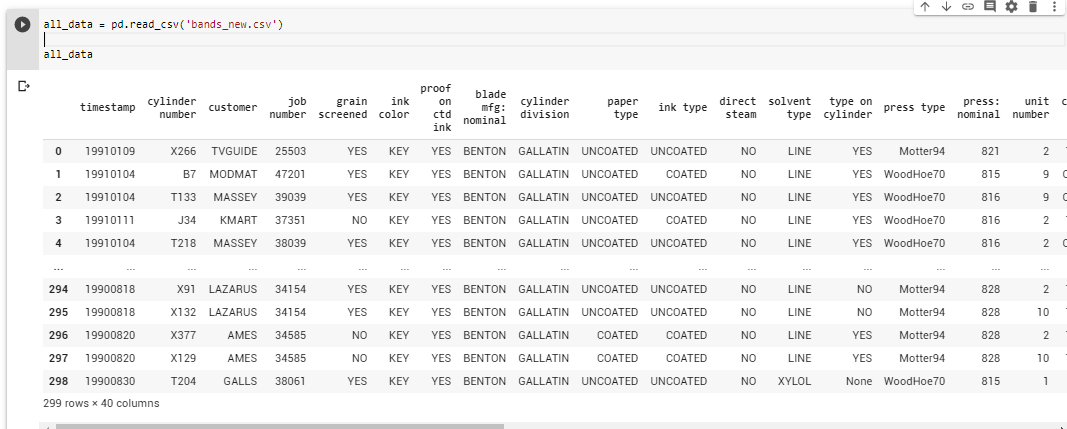


Рисунок 33 – Вывод данных

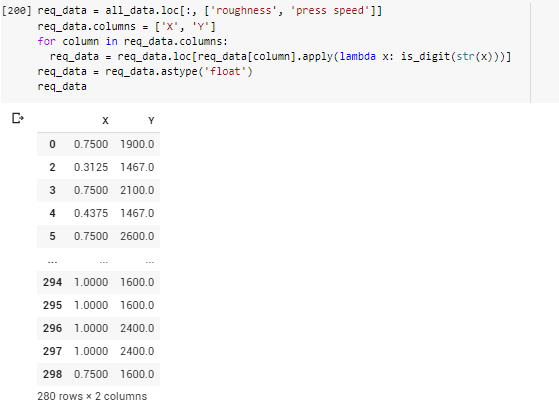


Рисунок 34 – Создание нового набора данных

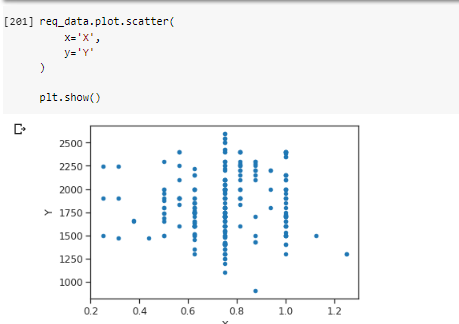


Рисунок 35 – Вывод точечного графика

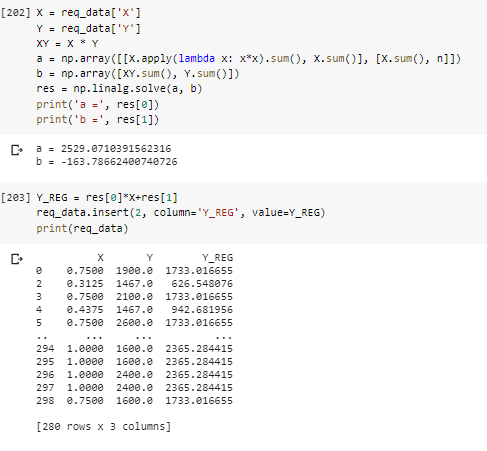


Рисунок 36 – Построение модели линейной регрессии

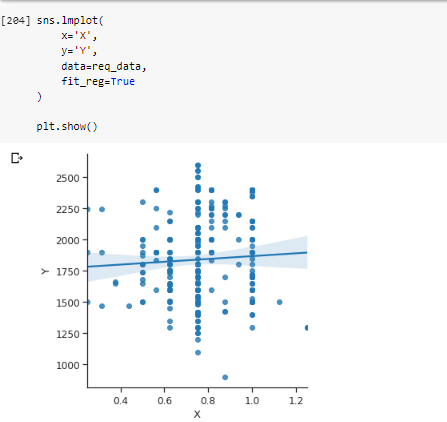


Рисунок 37 – Построение точечного графика

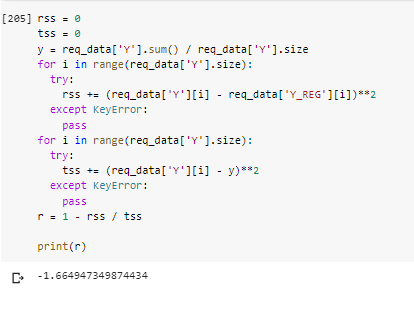


Рисунок 38 – Расчёт коэффициента детерминации