Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

ИРКУТСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

|  |
| --- |
| Институт информационных технологий и анализа данных |

наименование института

ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ № 2

по дисциплине:

|  |
| --- |
| **Методы анализа данных** |
| **«Корреляционный и регрессионный анализ»** |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Выполнил | АСУб-20-2 |  |  |  | Арбакова А.В. |
|  | шифр группы |  | подпись |  | Фамилия И.О. |
| Проверил |  |  |  |  | Бучнев О.С. |
|  | должность |  | подпись |  | Фамилия И.О. |

Иркутск 2022 г.

**Содержание**

[Введение 3](#_Toc116638193)

[Ход выполнения работы 4](#_Toc116638194)

[Заключение 11](#_Toc116638195)

[Список литературы 12](#_Toc116638196)

# **Введение**

**Цель лабораторной работы:**

1. Выбрать среду программирования для языка Python. Использовать набор данных согласно номеру варианта.
2. Загрузить набор данных, подготовить данные для анализа. С использованием Python выполнить корреляционный и регрессионный анализ данных. При этом необходимо:

* получить значения основных показателей описательной статистики и построить диаграммы рассеяния для признаков, измеренных в количественной шкале
* построить корреляционную матрицу и график «тепловая карта»
* проверить гипотезы о значимости коэффициентов корреляции
* для заданных пар количественных признаков построить уравнения линейной регрессии, построить диаграммы остатков, оценить дисперсию остатков
* для заданных пар количественных признаков найти коэффициенты уравнения полиномиальной регрессии, построить его график и построить диаграммы остатков, оценить дисперсию остатков, сравнить диаграммы остатков с диаграммами остатков для линейной регрессии
* для заданных пар количественных признаков оценить индекс корреляции, сравнить со значением коэффициента корреляции;
* выбрать переменную отклика и переменные факторы, построить для них модель множественной регрессии, объяснить полученные значения коэффициентов множественной регрессии

1. Выполнить анализ полученных на каждом этапе результатов и оформить отчет по лабораторной работе.

**Язык программирования:** Python

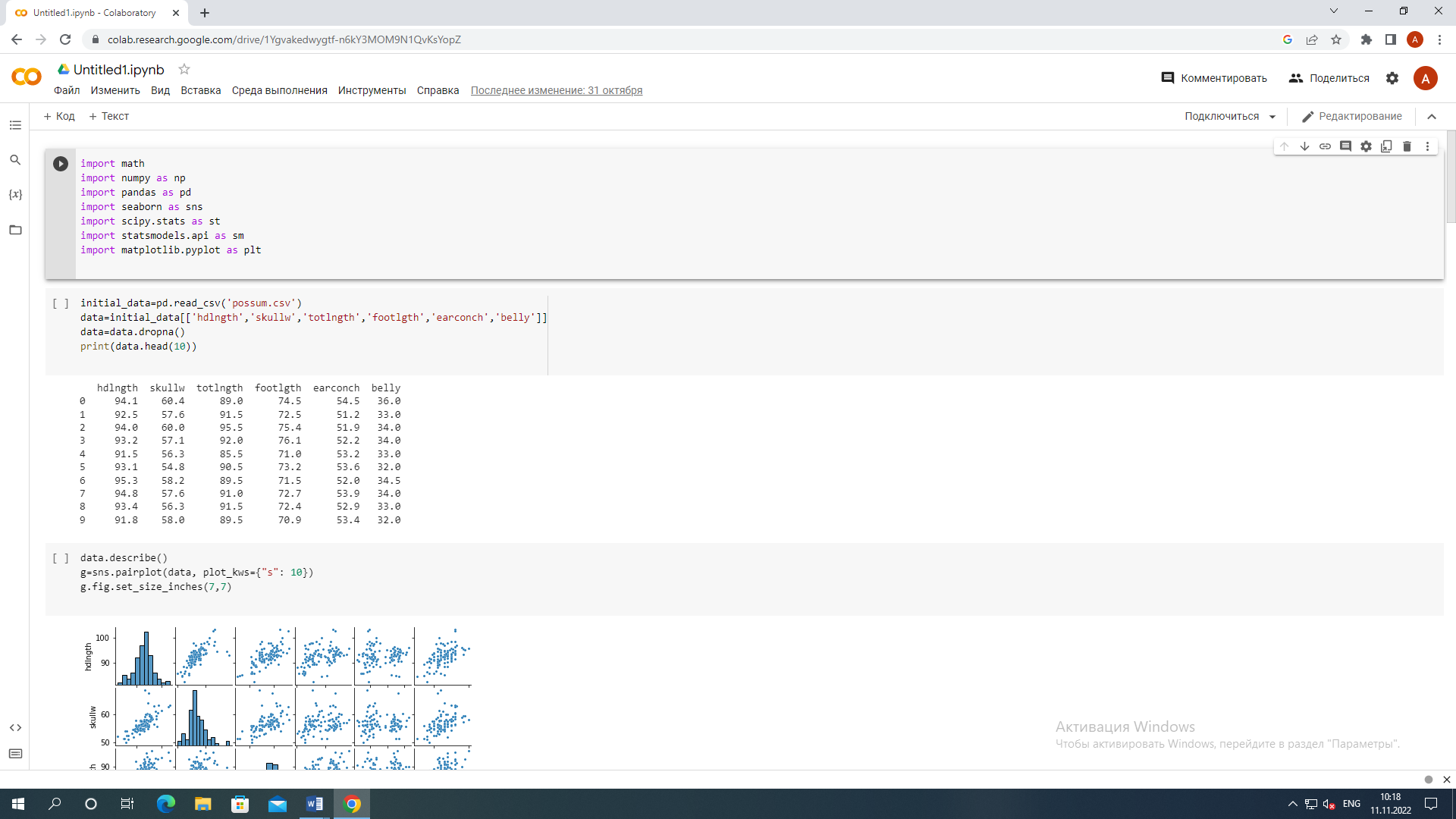
**Номер варианта:** 3

|  |  |  |
| --- | --- | --- |
| 3 | Possum.csv (<https://www.kaggle.com/abrambeyer/openintro-possum>) | Фактор – skullw;  Отклик – belly. |

# **Ход выполнения работы**

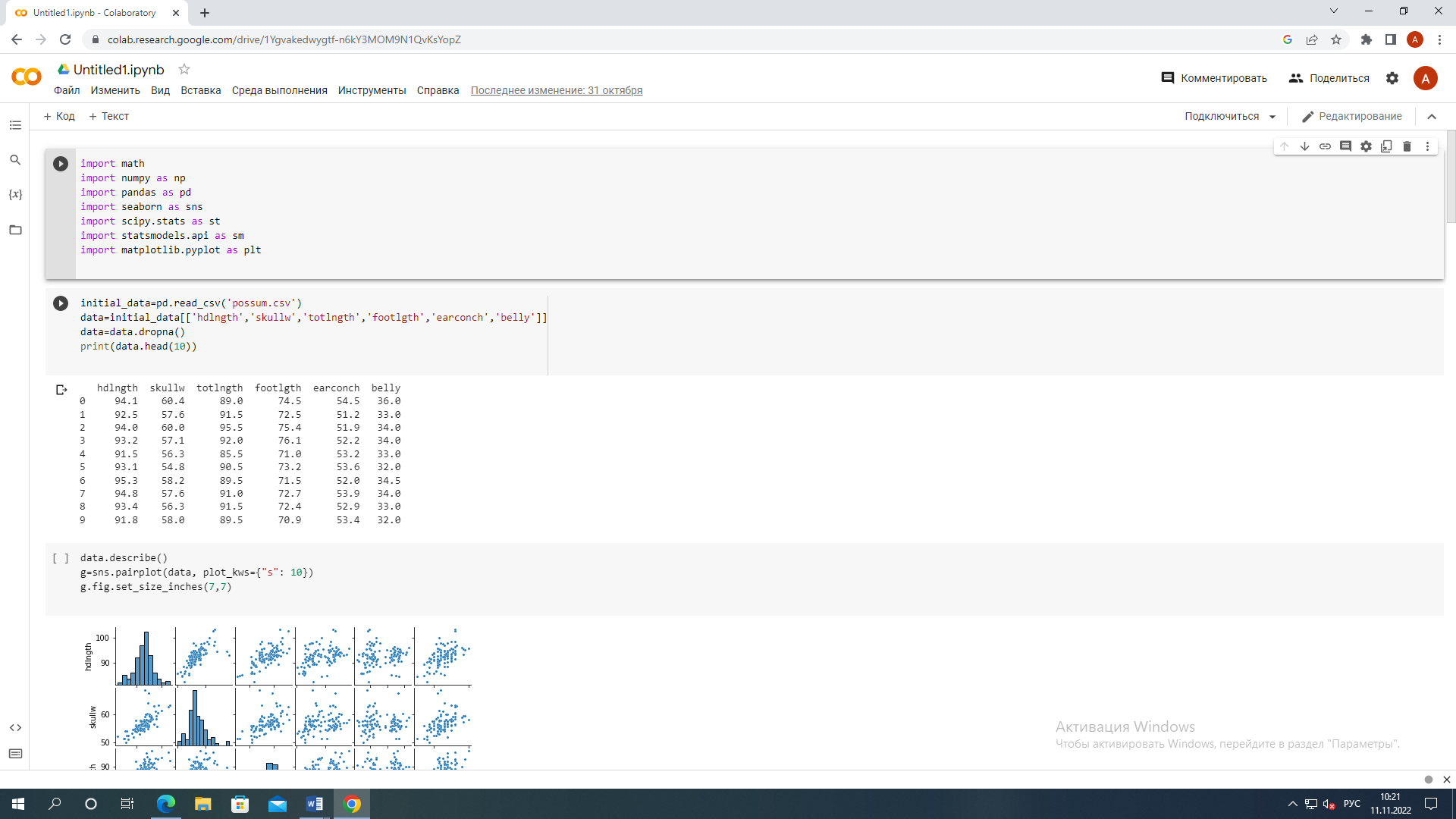
Для выполнения лабораторной работы были использованы данных из Excel-файла Possum.crv. В файле 104 наблюдений о характеристиках и параметрах опоссумов. Для наблюдения заданы значения 14 признаков. Загрузим данные, отберем часть признаков и подготовим их для анализа.

Перед выполнением лабораторной работы подключим необходимые библиотеки.



Изображение 1 – Библиотеки языка Python

Загрузим данные об опоссумах.

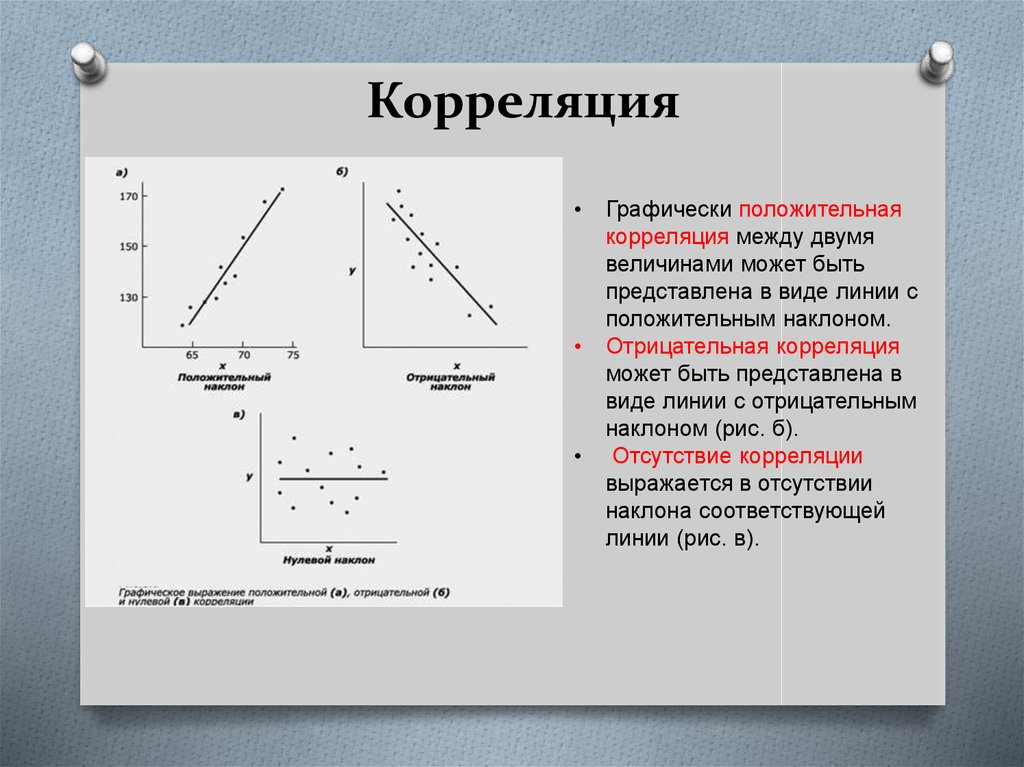


Изображение 2 – Загруженные данные из файла Excel

Корреляция – связь, соотношение между объективно существующими явлениями и процессами. Функциональная и корреляционная связь – два основных типа связи, определяющих соотношение между явлениями и процессами.

Виды корреляции: относительно характера корреляции – положительная корреляция и отрицательная корреляция; относительно числа переменных - простая (парная), множественная и частная; относительно формы связи – линейная и нелинейная; относительно типа соединения – непосредственная, косвенная и ложная.

Положительная корреляция – значения переменных увеличиваются вместе. Отрицательная корреляция – одна переменная увеличивается, пока другая уменьшается.



Изображение 3 – Положительная и отрицательная корреляции

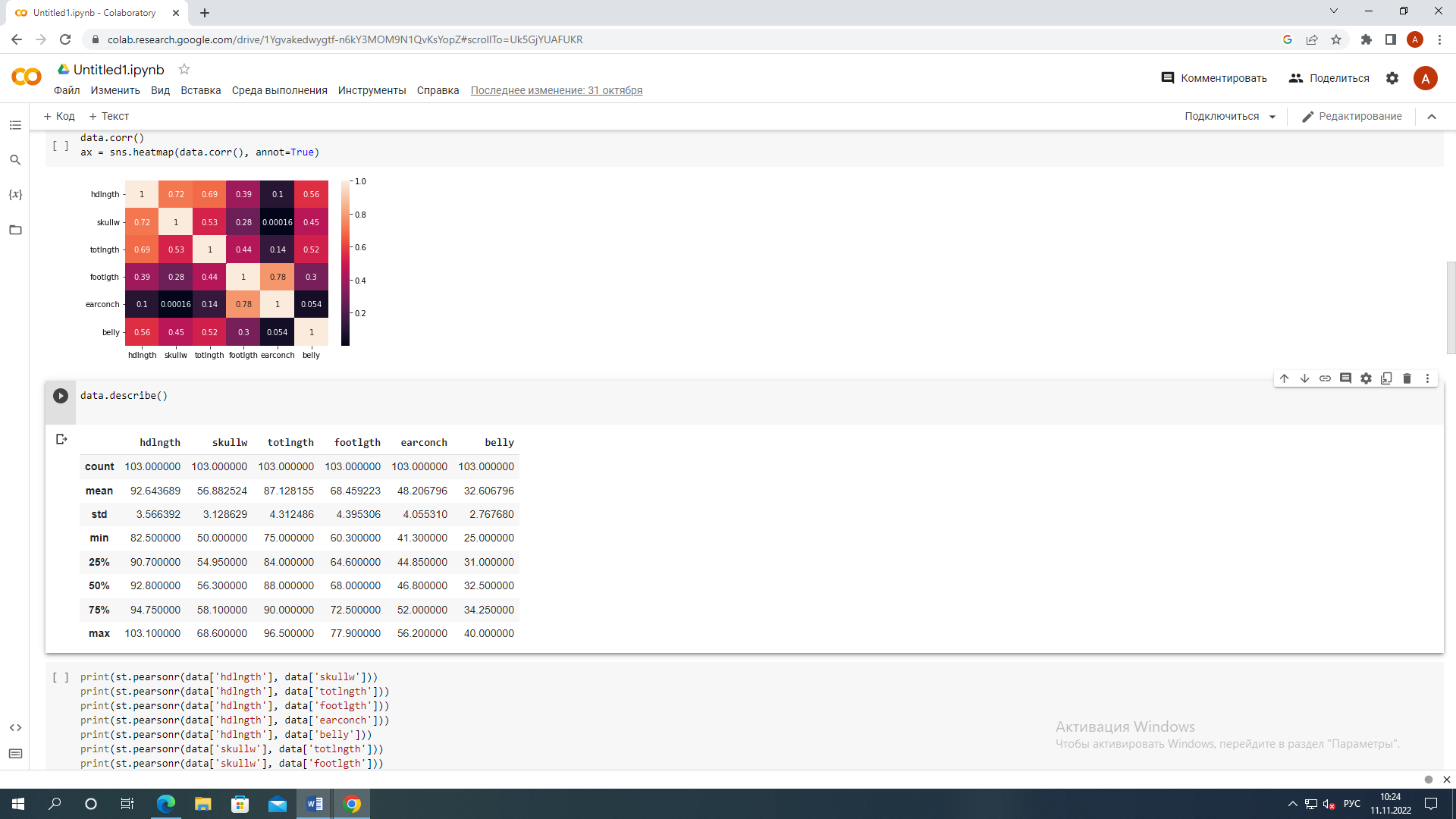
Парная корреляция ‑ связь между двумя признаками (результативным и факторным, или двумя факторными). Частная корреляция ‑ зависимость между результативным и одним факторным признаками при фиксированном значении других факторных признаков. Множественная корреляция ‑ зависимость результативного и двух или более факторных признаков, включенных в исследование. Результативными называются признаки, изменяющиеся под воздействием факторных признаков. Факторные признаки – те признаки, которые характеризуют причину.

Линейная корреляция – если с увеличением или уменьшением одной переменной Х, вторая переменная Y в среднем либо также растет, либо убывает. Нелинейная корреляция – при увеличении одной величины, характер изменения второй не линеен, а описывается другими законами.

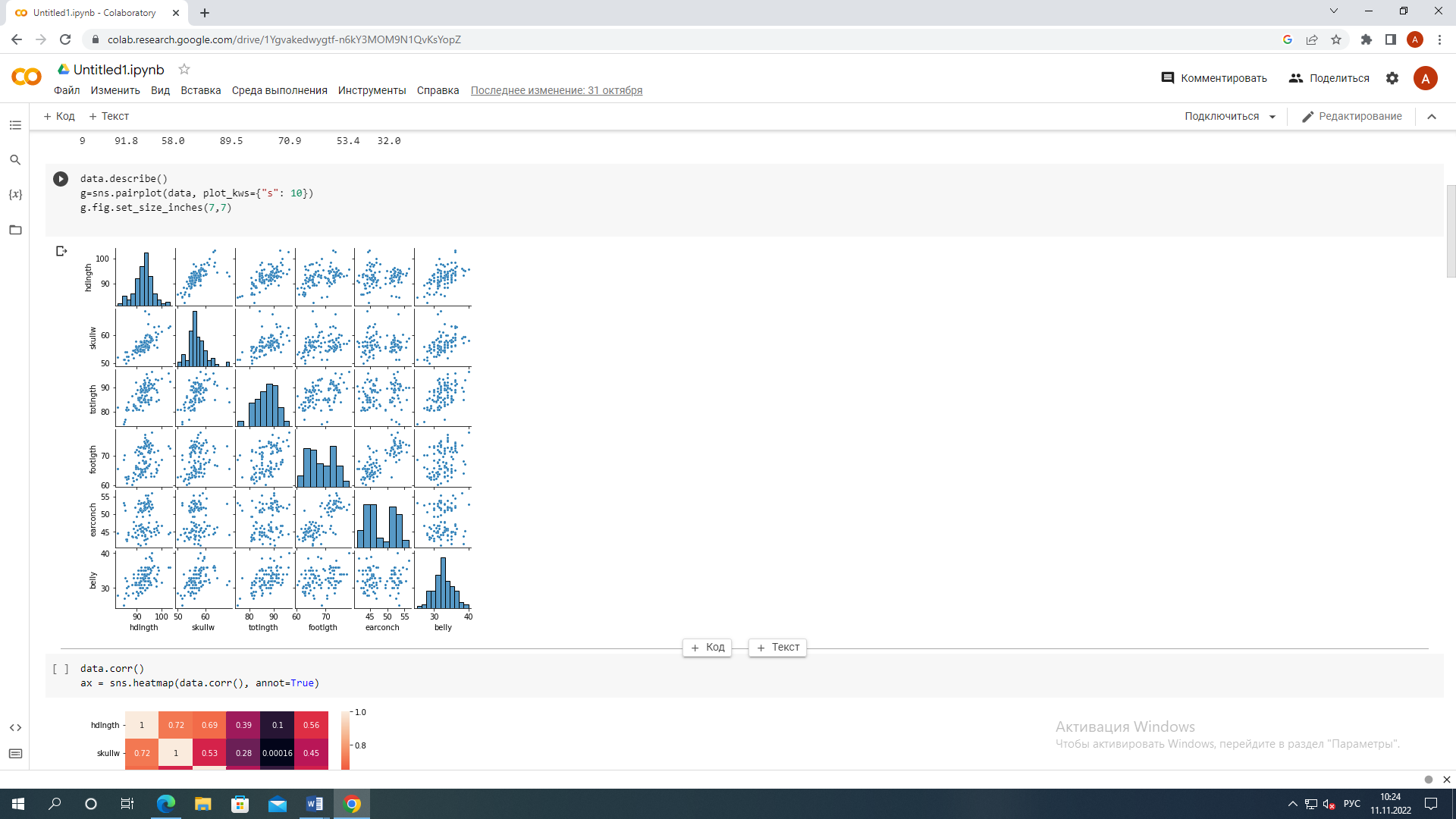
Непосредственная корреляция – исследуемые явления соединены между собой непосредственно. Косвенная корреляция – переменные не имеют непосредственной причинно-следственной связи, а детерминируются общей для них причиной. Ложная корреляция – формальная связь между явлениями, не находящая никакого логического объяснения и основанная лишь на количественном отношении между ними.

Коэффициент корреляции – среднее произведение двух нормированных отклонений.

Получим значения основных показателей описательной статистики и построим диаграммы рассеяния для признаков, измеренных в количественной шкале.

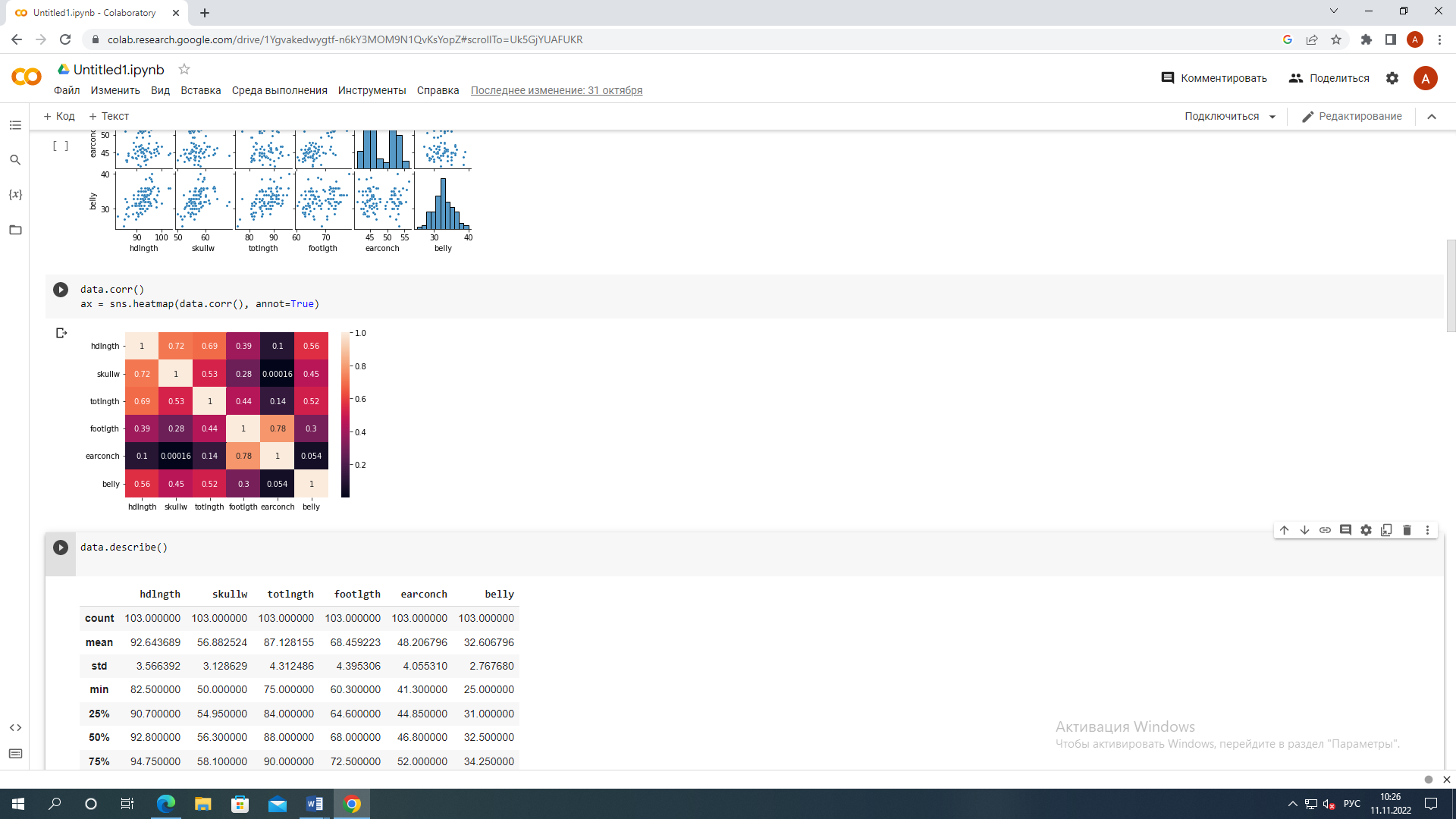


Изображение 4 – Основные показатели описательной статистики



Изображение 5 – Диаграммы рассеяния для признаков

Построим график «тепловая карта».



Изображение 6 – График «Тепловая карта»

Проверим значимость коэффициента корреляции. Нулевая гипотеза состоит в том, что коэффициент корреляции равен нулю, альтернативная - не равен нулю:

http://statistica.ru/upload/medialibrary/ff2/image001.png

http://statistica.ru/upload/medialibrary/252/image003.png

Достаточно большое по абсолютной величине значение величины r будет стремиться опровергнуть нулевую гипотезу.

Для того чтобы проверить гипотезу, мы должны знать распределение величины r. Собственное распределение величины r довольно сложное, поэтому мы применим преобразование:

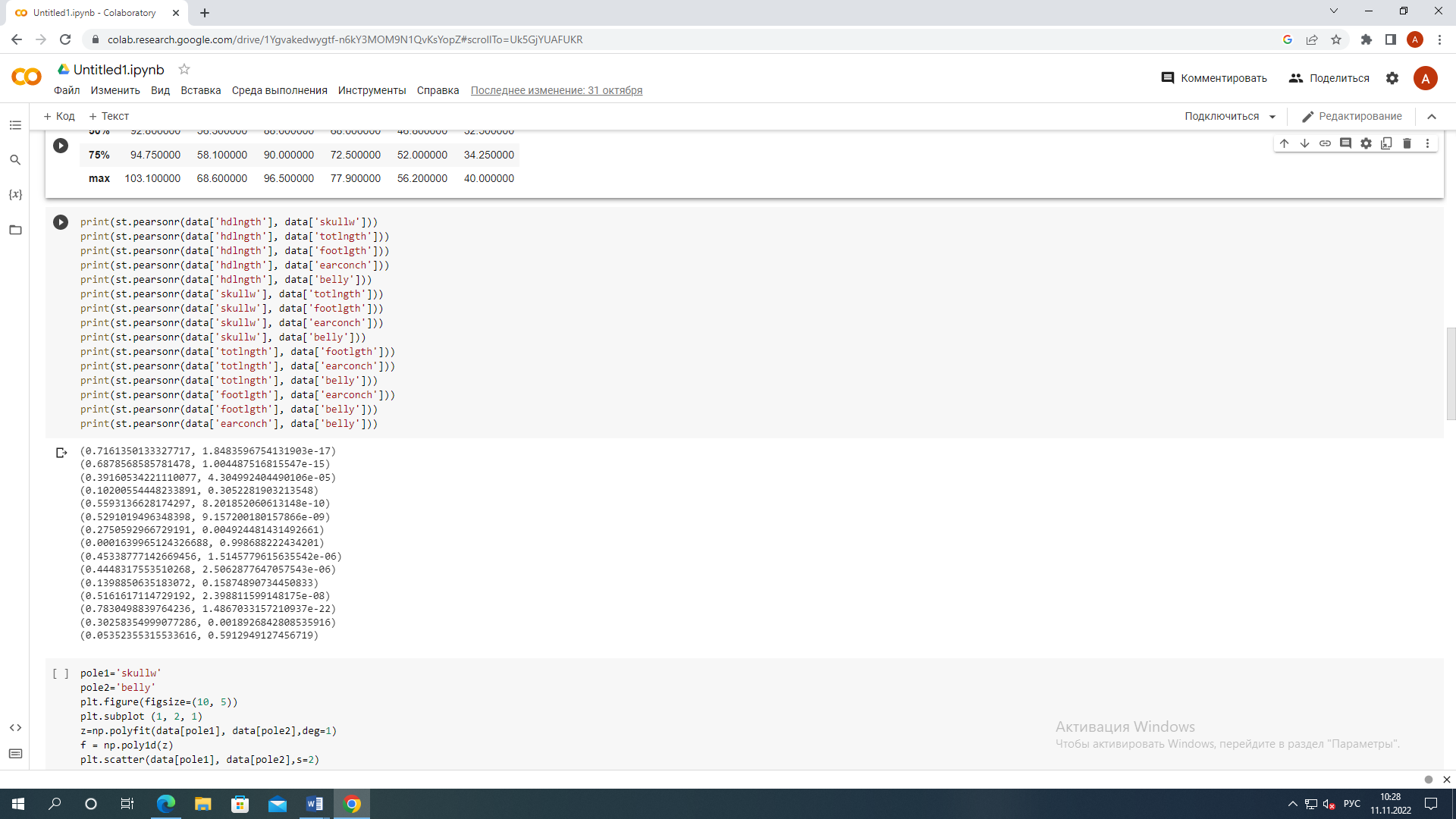
http://statistica.ru/upload/medialibrary/9b0/image005.png

Выборочное распределение этой статистики есть распределение Стьюдента с n-2 степенями свободы. При заданном уровне значимости (α) определяем критическое значение tкр. Принимаем решение об отклонении или не отклонении нулевой гипотезы:

http://statistica.ru/upload/medialibrary/039/image009.png - отклоняем H0

http://statistica.ru/upload/medialibrary/bd5/image007.png - не отклоняем H0

Проверим гипотезы о значимости коэффициентов корреляции.



Изображение 7 – Гипотезы о значимости коэффициентов корреляции

Корреляция говорит о том, насколько отклонения от средней одной переменной совпадают с отклонениями другой переменной, а регрессия говорит о приближении к среднему показателю для популяции в целом.

Регрессия – это метод, используемый для моделирования и анализа отношений между переменными, а также для того, чтобы увидеть, как эти переменные вместе влияют на получение определенного результата. Регрессия используется для анализа воздействия на отдельную зависимую переменную значений одной или более независимых переменных.

Виды регрессии: относительно числа явлений – простая (парная) и множественная или частная; относительно формы зависимости – линейная и нелинейная; относительно характера – положительная и отрицательная; относительно типа соединения явлений – непосредственная, косвенная и ложная.

Простая (парная) регрессия представляет собой модель, где среднее значение зависимой переменной Y рассматривается как функция одной независимой переменной X. Множественная регрессия представляет собой модель, где среднее значение зависимой переменной Y рассматривается как функция нескольких независимых переменных X1, X2…



Изображение 8 – Линейная и нелинейная регрессия

В линейной регрессии переменные имеют прямое направление, в отличие от нелинейной.



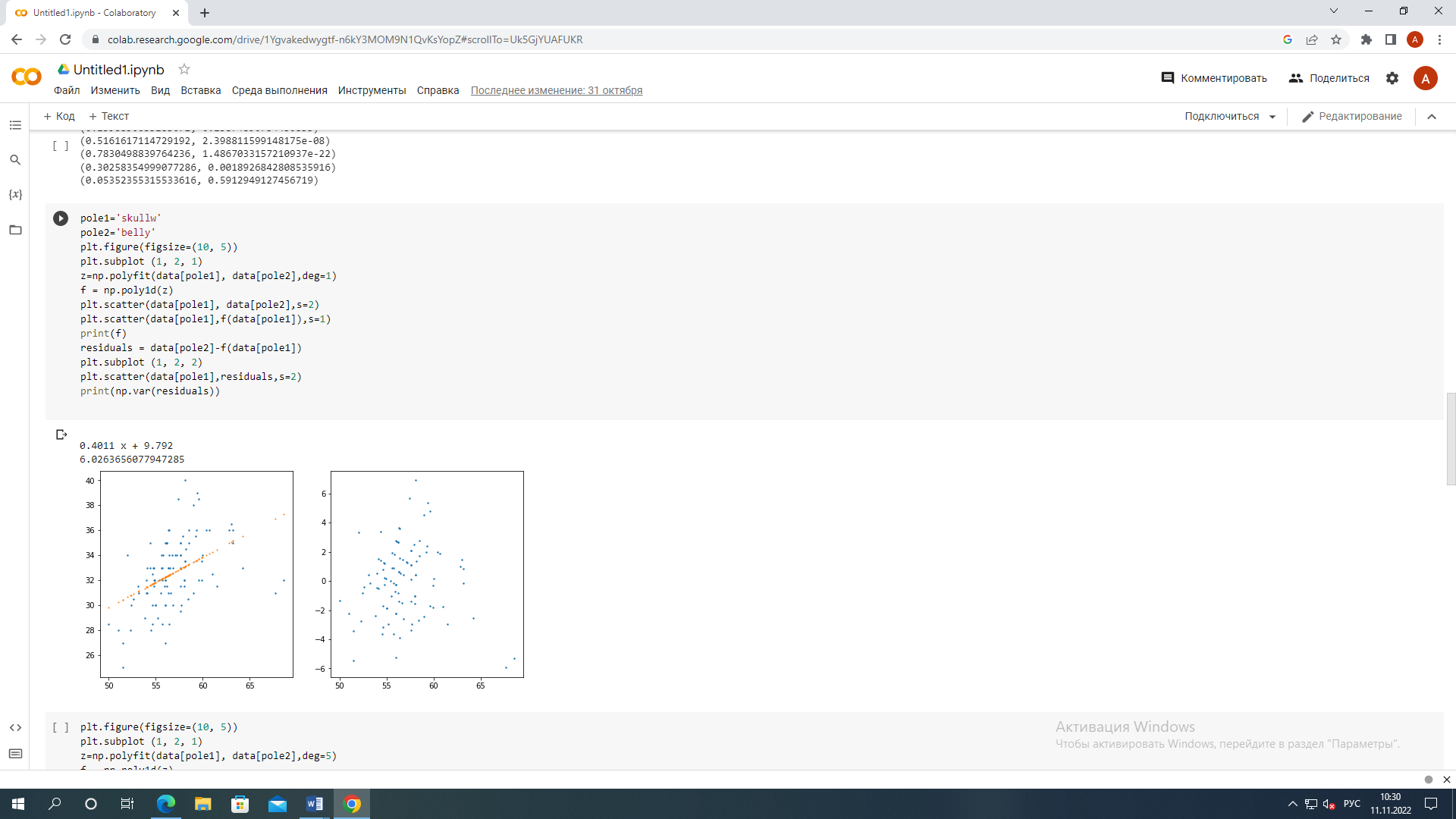
Изображение 9 – Положительная и отрицательная регрессия

К задачам регрессионного анализа относят: установление формы зависимости и определение функции регрессии.

Метод наименьших квадратов – математический метод, основанный на минимизации суммы квадратов отклонений некоторых функций от экспериментальных входных данных.

Полиномиальная регрессия – связь между независимой переменной X и зависимой переменной Y моделируется как полином n-й степени в X.

Для пары количественных признаков «skullw» - «belly» построим уравнение линейной регрессии и диаграмму остатков регрессии, оценим дисперсию остатков.

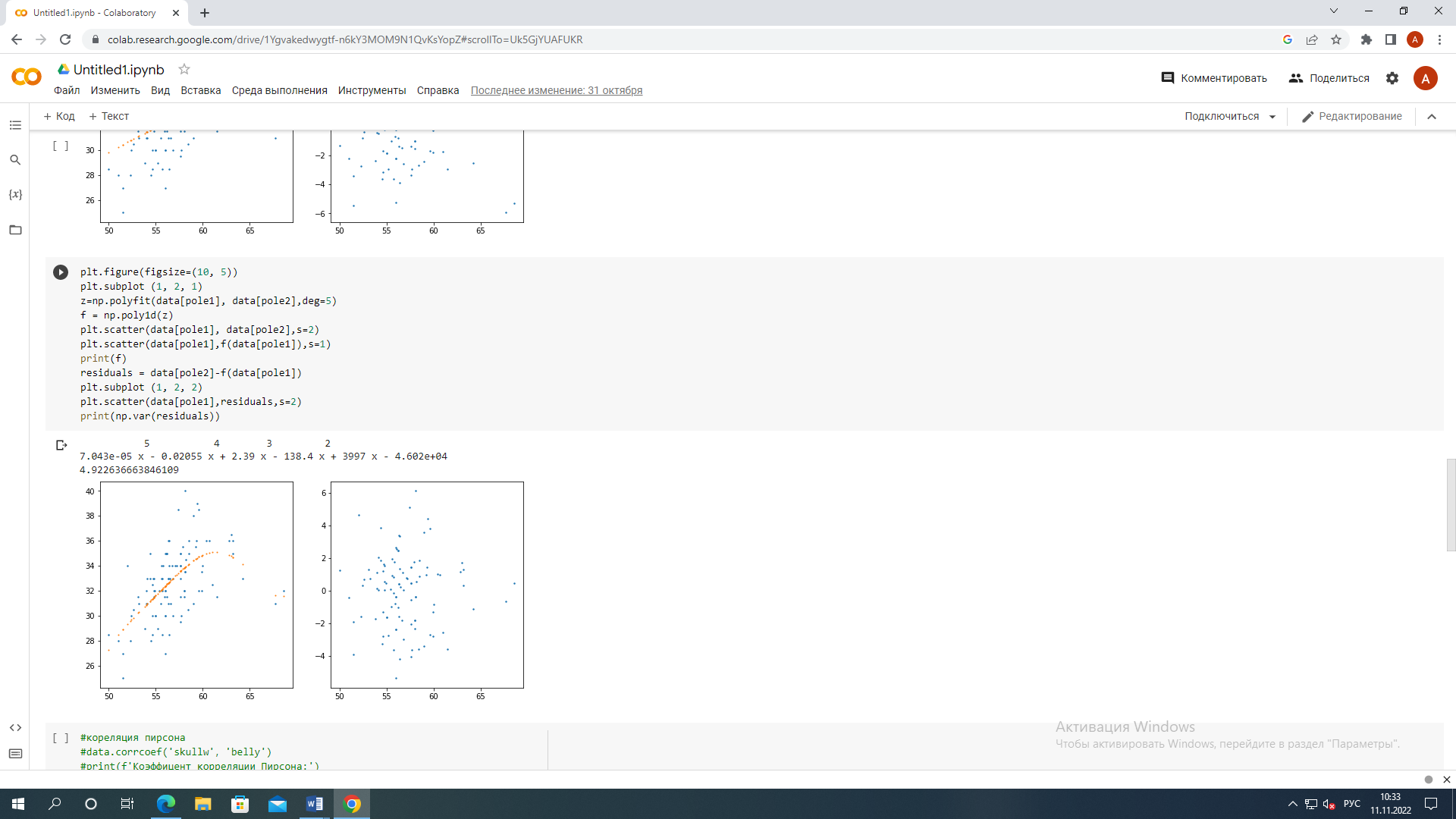


Изображение 10 – Уравнение линейной регрессии и диаграмма остатков регрессии

Для оценки качества регрессионной функции используют коэффициент детерминации R. По значению коэффициента детерминации можно судить о том, какая часть вариации зависимого признака может быть объяснена регрессией. Значение R = 1 соответствует функциональной зависимости между признаками, если R = 0, то корреляционная зависимость между признаками отсутствует.

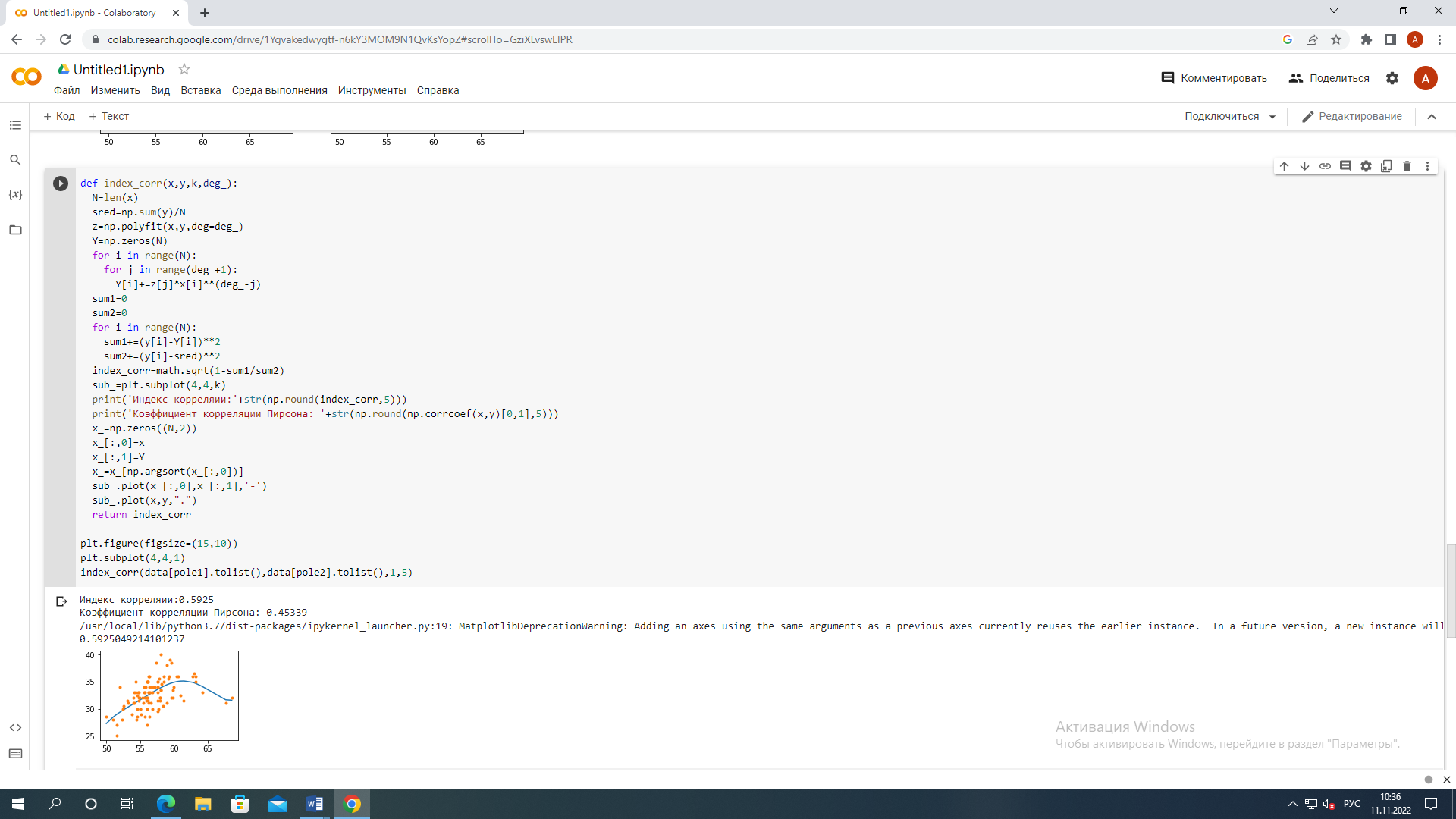
Остаток представляет собой разность между наблюдаемым и предсказанным значениями зависимой переменной при заданном значении. Диаграмма остатков позволяет отфильтровать или удалить линейную зависимость между переменными X и Y и выявить недостаточную точность модели простой линейной регрессии.

Для пары количественных признаков «skullw» - «belly» найдем коэффициенты уравнения полиномиальной регрессии, построим его график и диаграмму остатков регрессии, оценим дисперсию остатков.



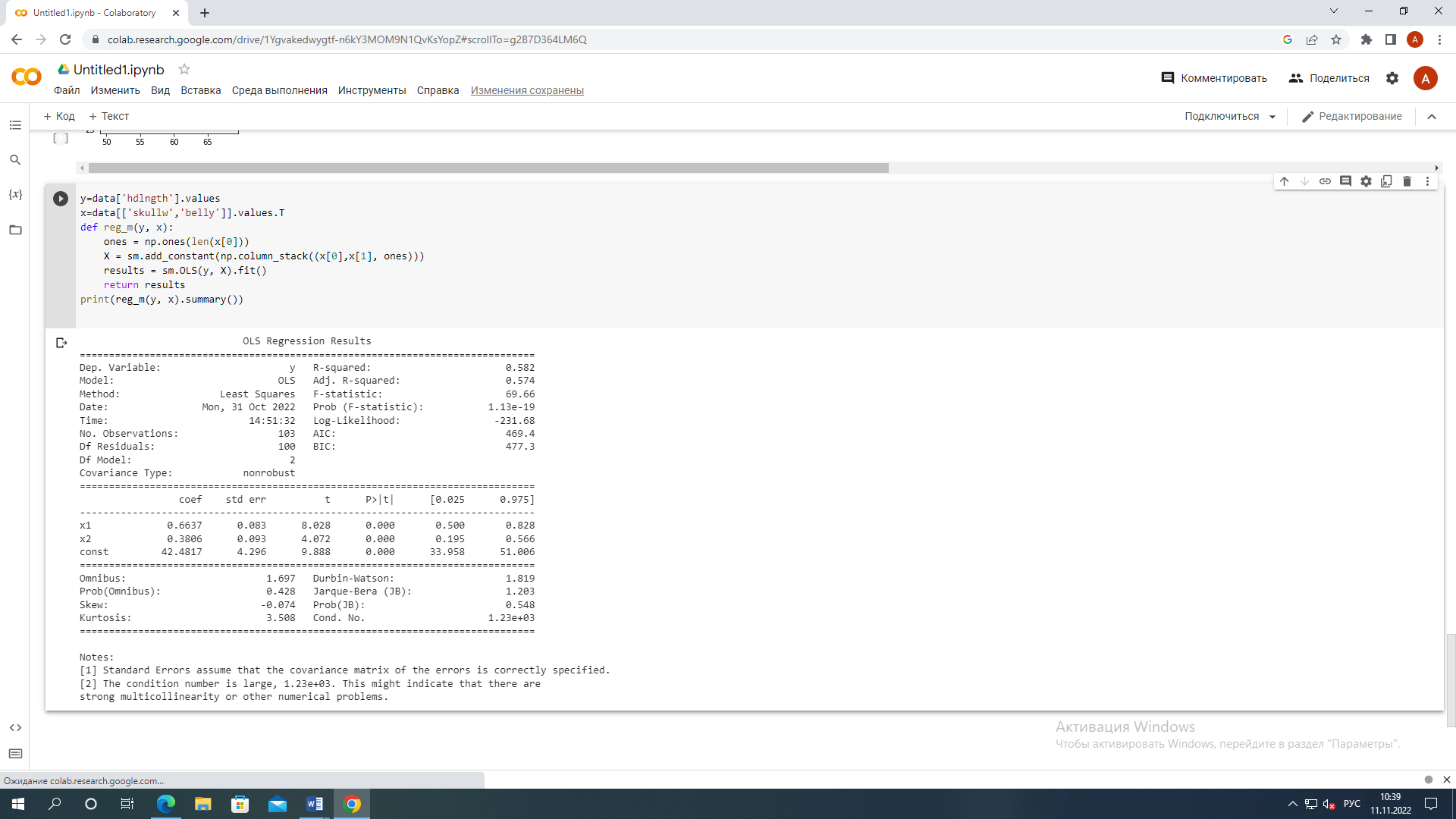
Изображение 11 – Коэффициенты уравнения полиномиальной регрессии, график и диаграмма остатков регрессии

Для пары количественных признаков «skullw» - «belly» найдем значение коэффициента корреляции Пирсона и оценим индекс корреляции.



Изображение 12 – Значение коэффициента корреляции Пирсона и индекс корреляции

Исследуем зависимость «hdlngth» от «skullw» и «belly». Для этого построим модель множественной регрессии.



Изображение 13 – Модель множественной регрессии

# **Заключение**

В ходе выполнения лабораторной работы по теме «Корреляционный и регрессионный анализ» была выполнена цель работы по изучению корреляционного и регрессионного анализа данных.

Было изучены такие темы как: понятие корреляции, виды корреляции, парная линейная корреляция, анализ частных и множественных связей, проверка гипотезы о значимости коэффициента корреляции, регрессия, виды регрессии, задачи регрессионного анализа, парная линейная регрессия, метод наименьших квадратов, множественная линейная регрессия, полиномиальная регрессия, проверка гипотезы о значимости коэффициентов регрессионного уравнения, коэффициент детерминации и диаграмма остатков.

Знания были закреплены во время использования среды разработки Google Colab и использования языка программирования Python, в котором было выполнено задание лабораторной работы.

# **Список литературы**

1. Бучнев О. С. Методы анализа данных: лабораторный практикум – Иркутск, 2022. – 115 с
2. <https://colab.research.google.com>
3. С.Э. Мастицкий, В.К. Шитиков Статистический анализ и визуализация данных с помощью языка R, Хайдельберг – Лондон – Тольятти, 2014 г. – 401 с
4. https://www.kaggle.com/datasets