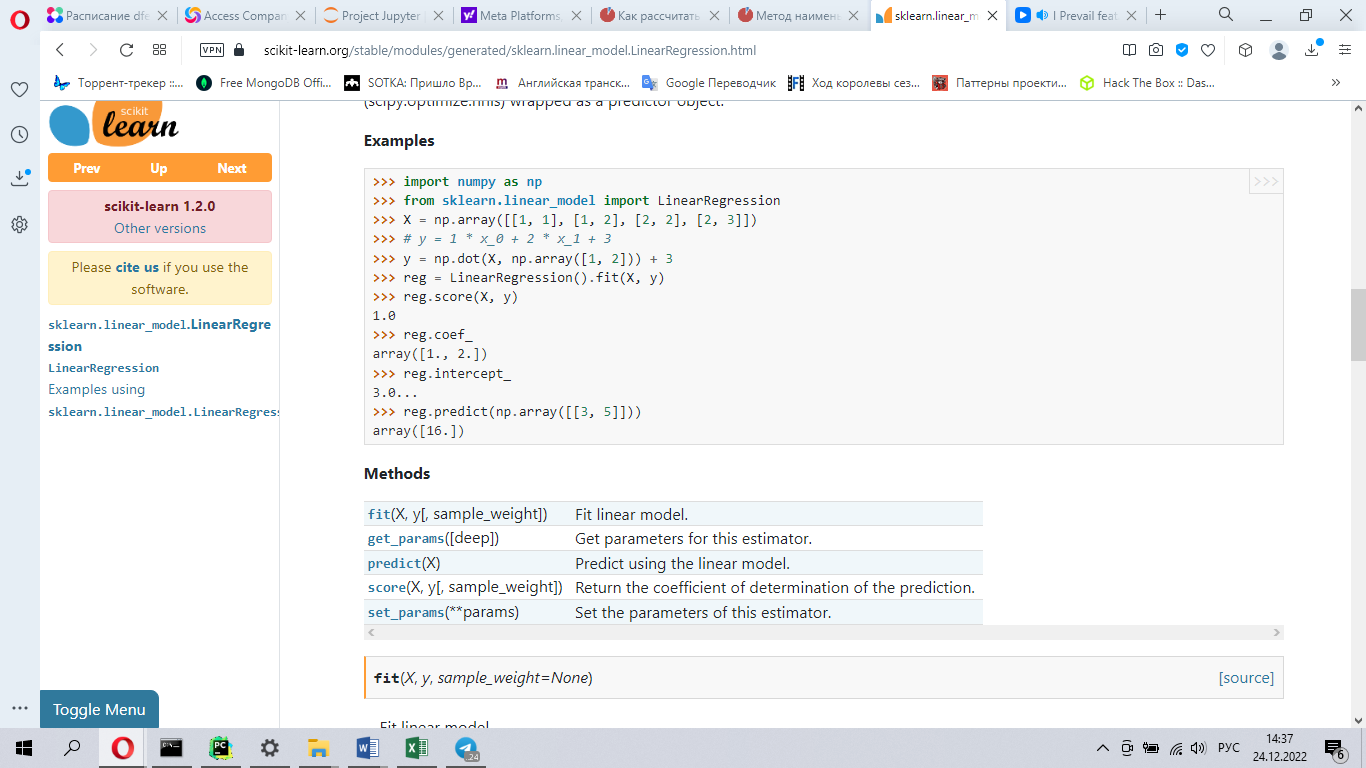


**Каптуров Александр Васильевич (учащийся Школы Приора)**

Выполнение тестового задания на позицию

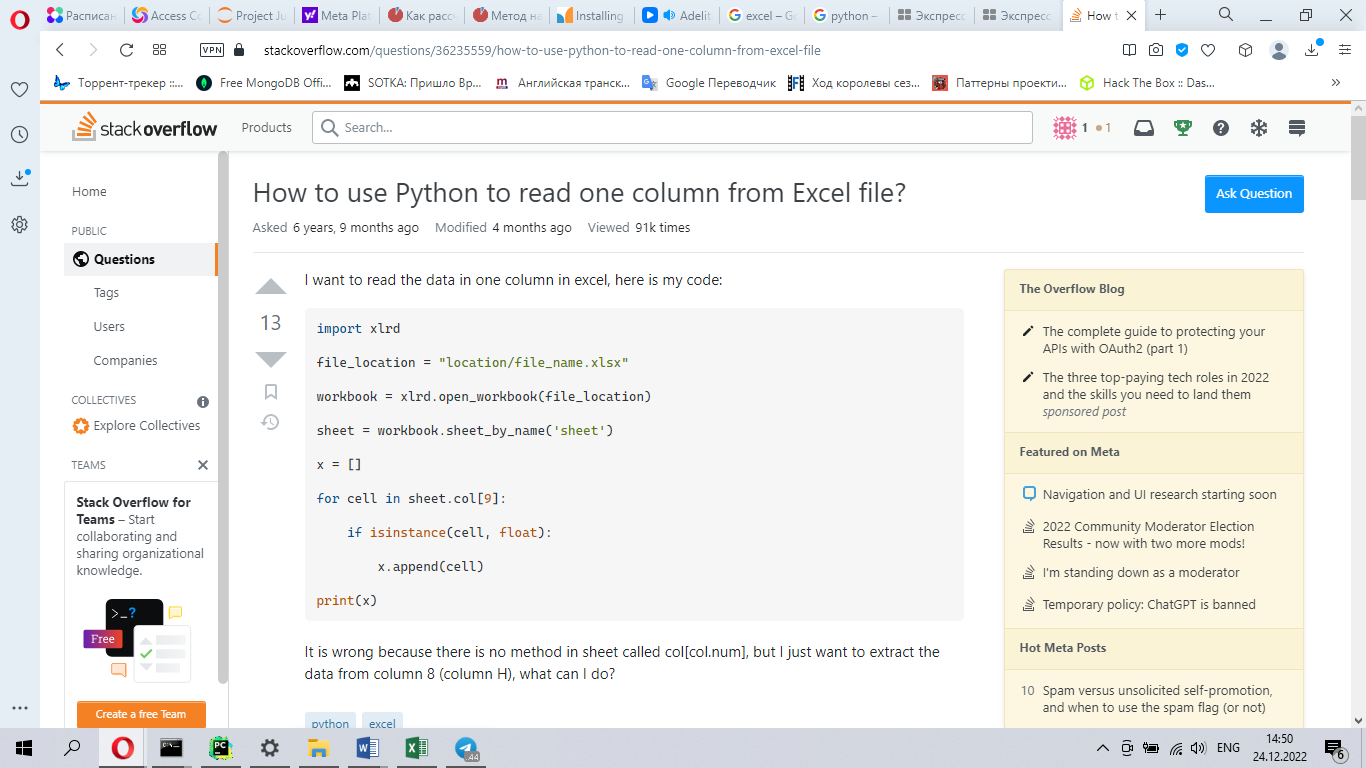
математика-экономиста в департамент кредитных рисков



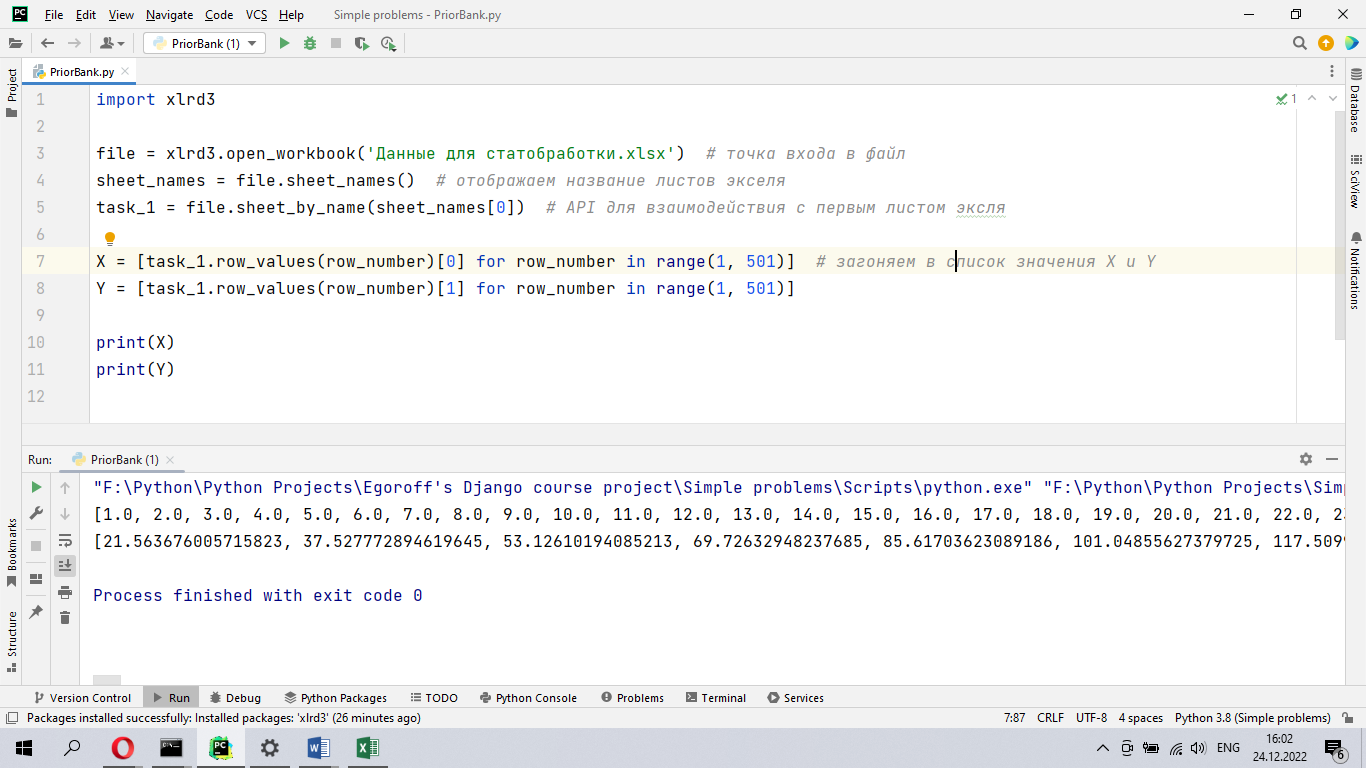


*Задача 1. Для начала попробуем построить модель парной линейной регрессии используя фреймворк машинного обучения scikit-learn.*

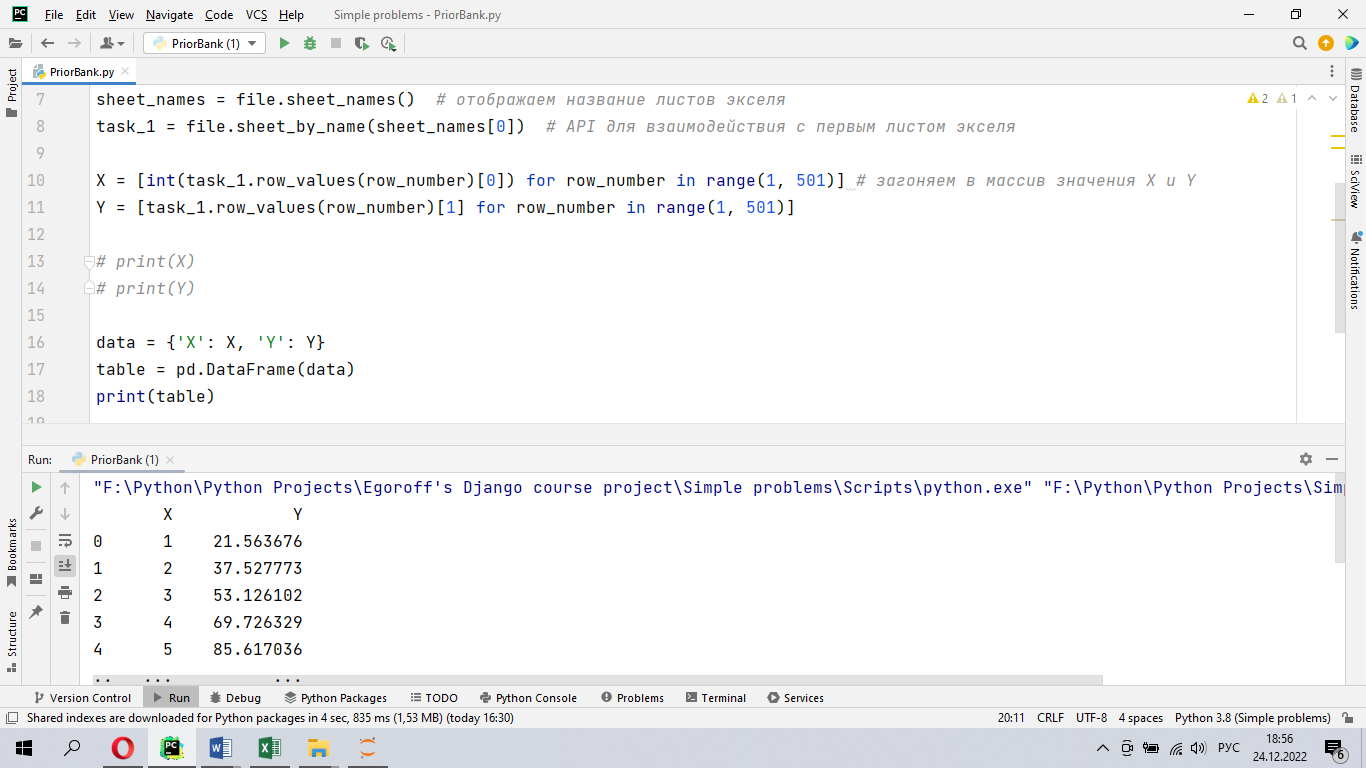
[*https://github.com/scikit-learn/scikit-learn/blob/main/sklearn/linear\_model/\_base.py*](https://github.com/scikit-learn/scikit-learn/blob/main/sklearn/linear_model/_base.py) *По данной ссылке можно перейти к репозиторию, где хранится* class LinearRegression (MultiOutputMixin, RegressorMixin, LinearModel):



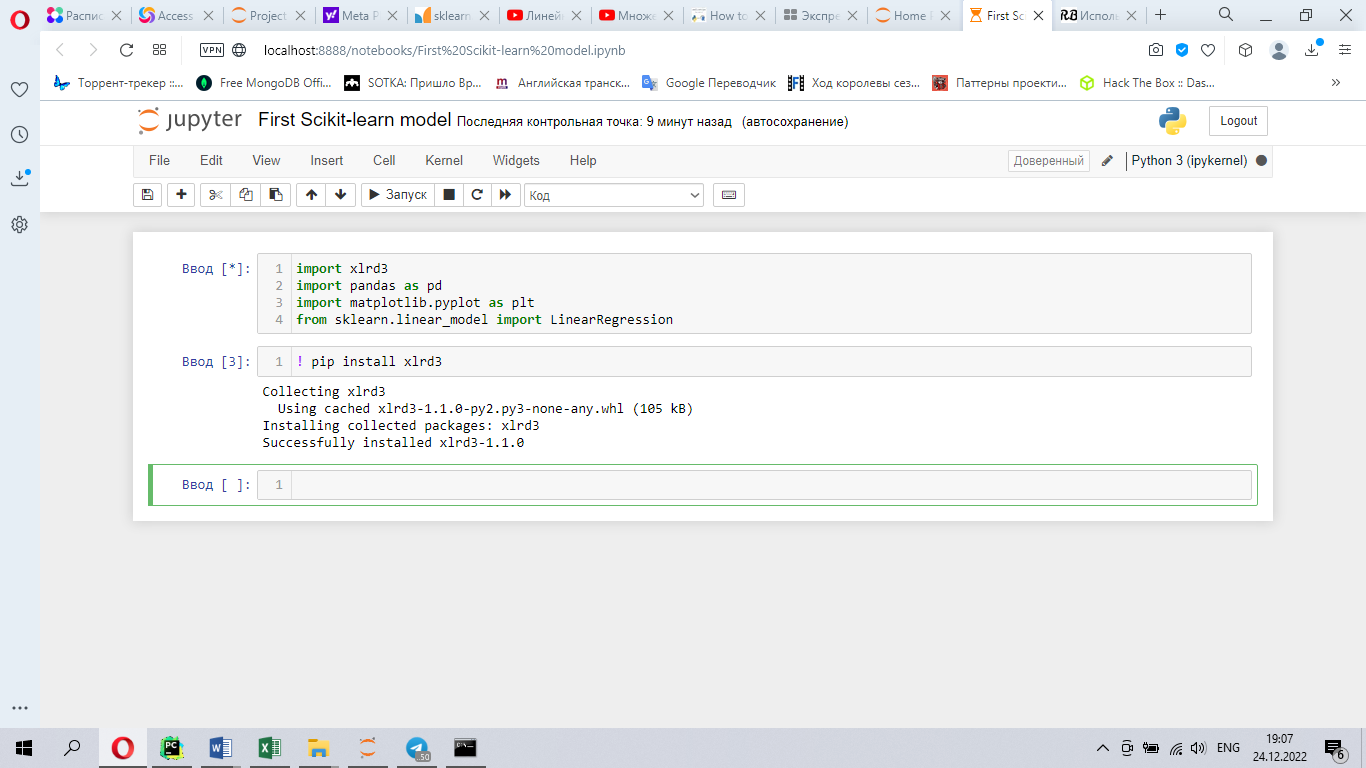
*Загуглим как именно извлечь столбцы X и Y из Excel файла. Данный источник не подошёл, но потратив ещё время на поиски, находим библиотеку xlrd3, которая считывает информацию с эксель файла.*



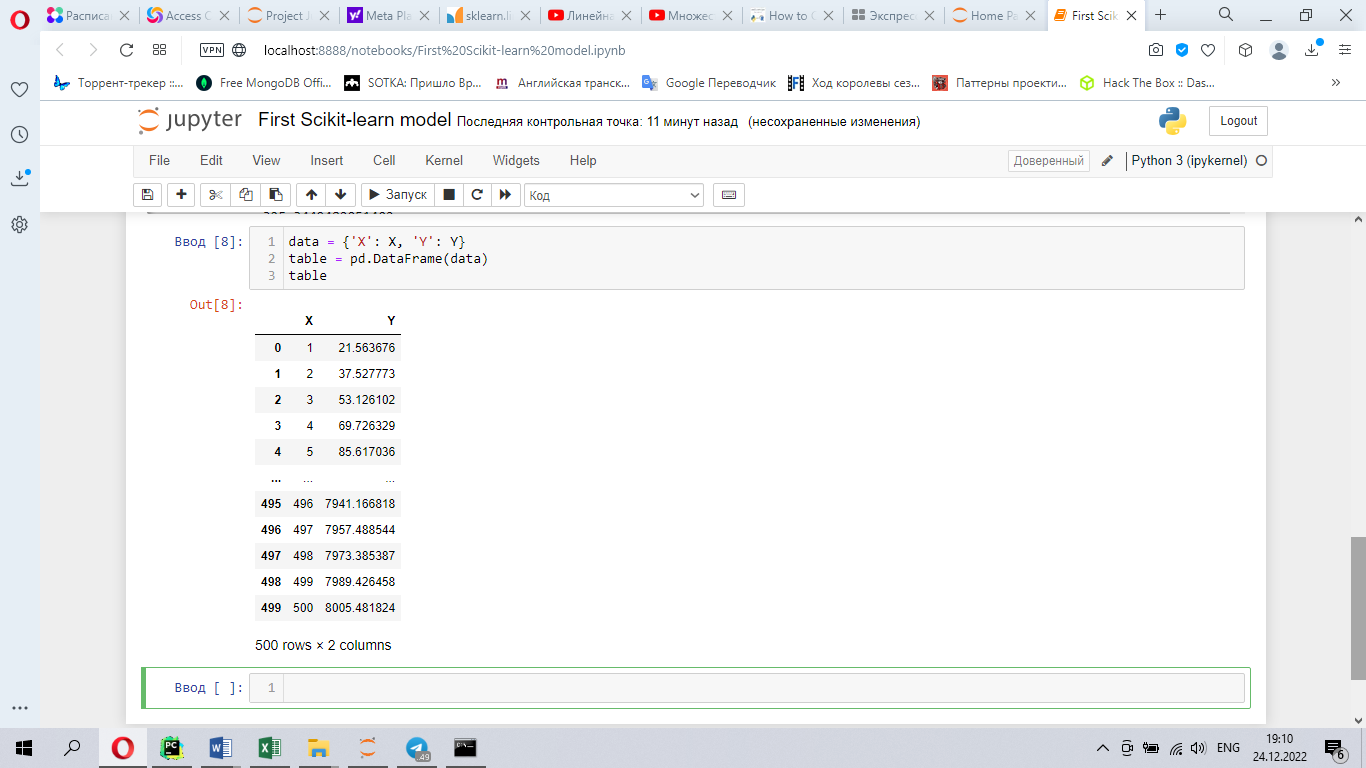
*Написав данный код, мы получаем 2 списка со значениями X и Y первого листа нашего эксель файла*



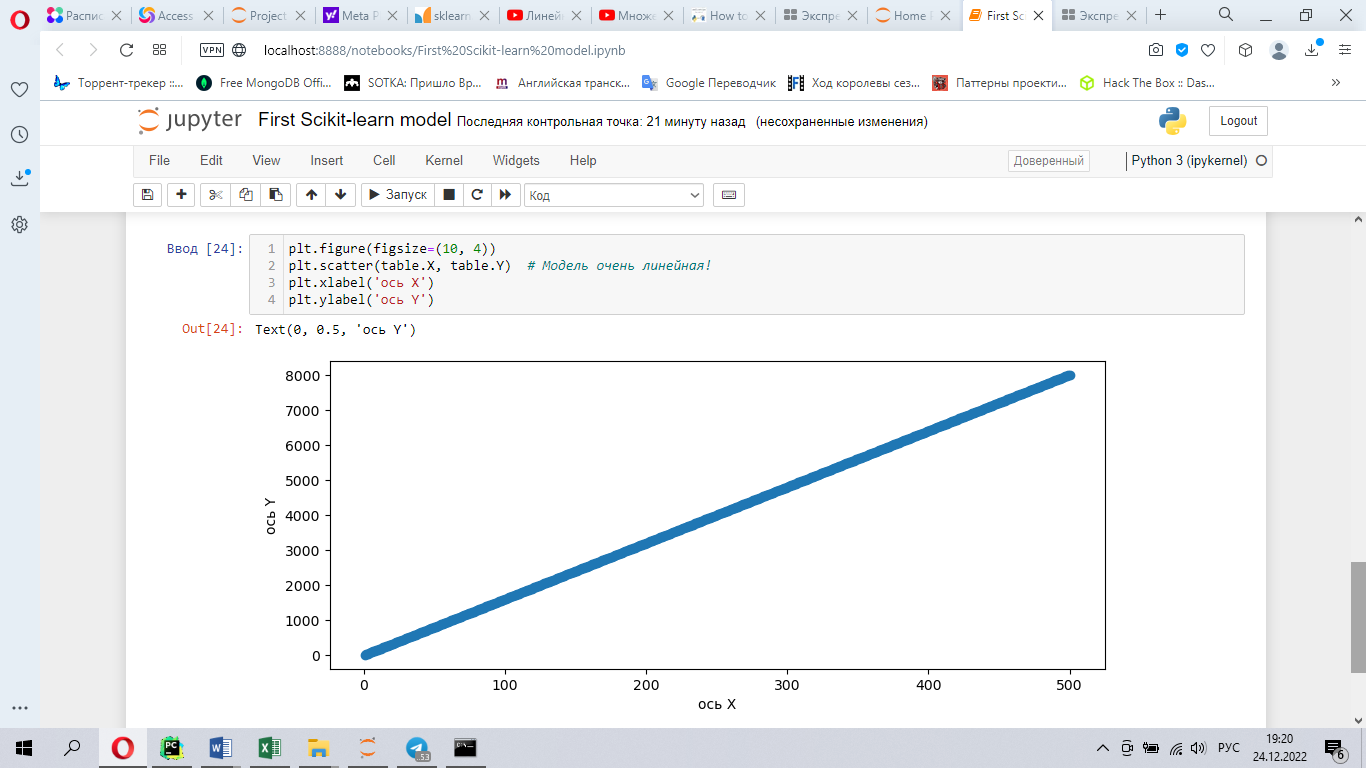
*С помощью фреймворка Pandas превратим наши ряды данных в таблицу (дата-фрейм)*



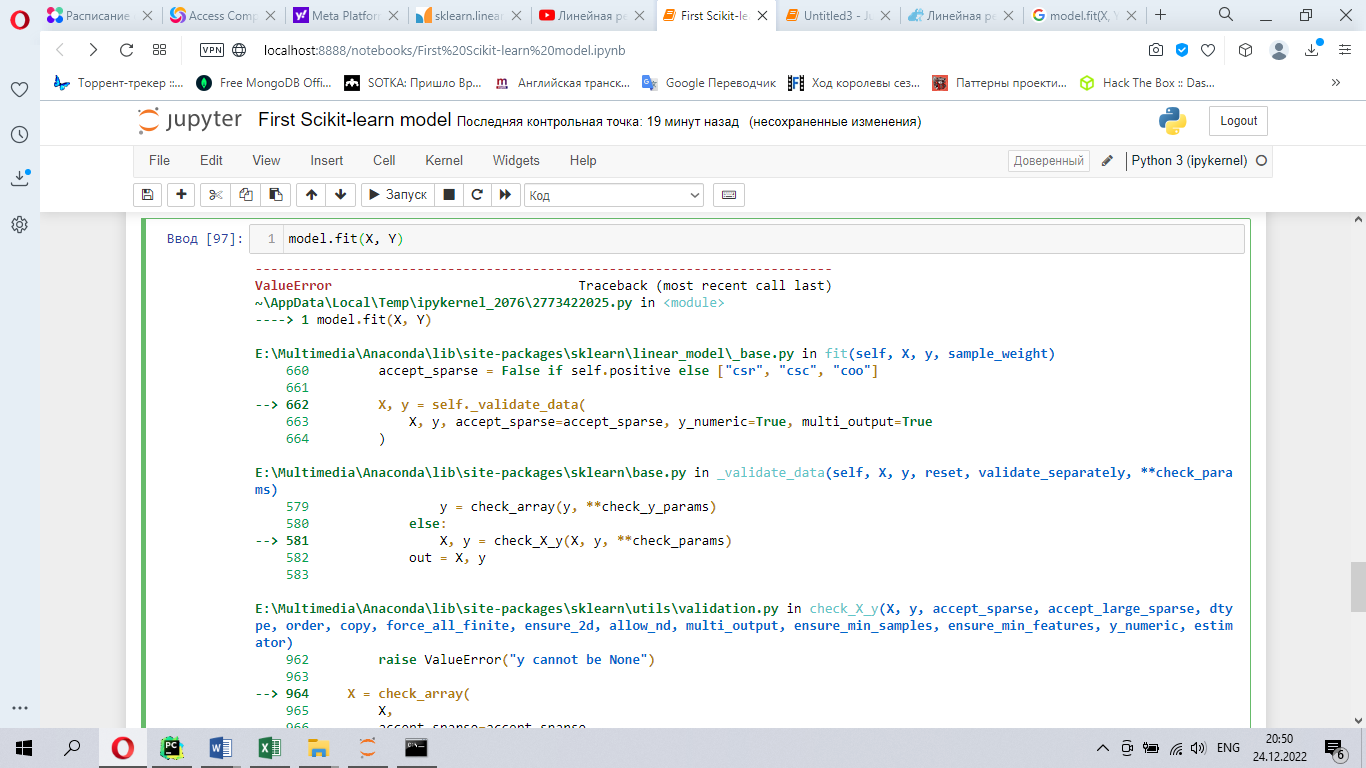
*Параллельно с PyCharm будем осуществлять те же действия с помощью более подходящей для анализа данных Jupyter notebook. Для импорта библиотек не входящих в ноутбук по дефолту используем статью* [*https://russianblogs.com/article/503462951/*](https://russianblogs.com/article/503462951/)



*Как видим всё получилось даже более нагляднее!)*



*С помощью фреймворка Matplotlib построим по точкам значения пересечений по X и Y. Если увеличить размер фигуры, то она будет в виде совокупности дискретных (точечных) значений, но сейчас она в таком масштабе выглядит как линия*

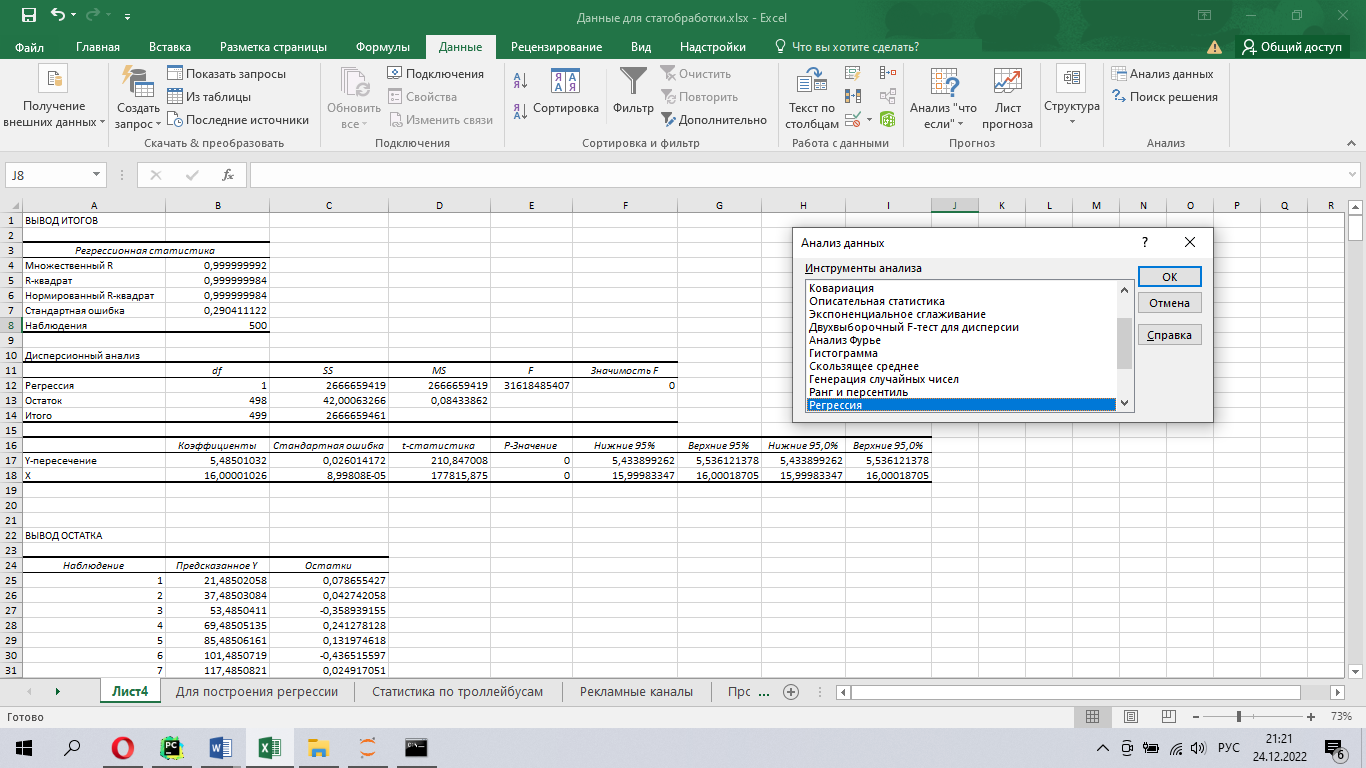


*Попытка обучить модель не удалась по неизвестной ошибке. Более 2 часов поисков не смогли её исправить.*

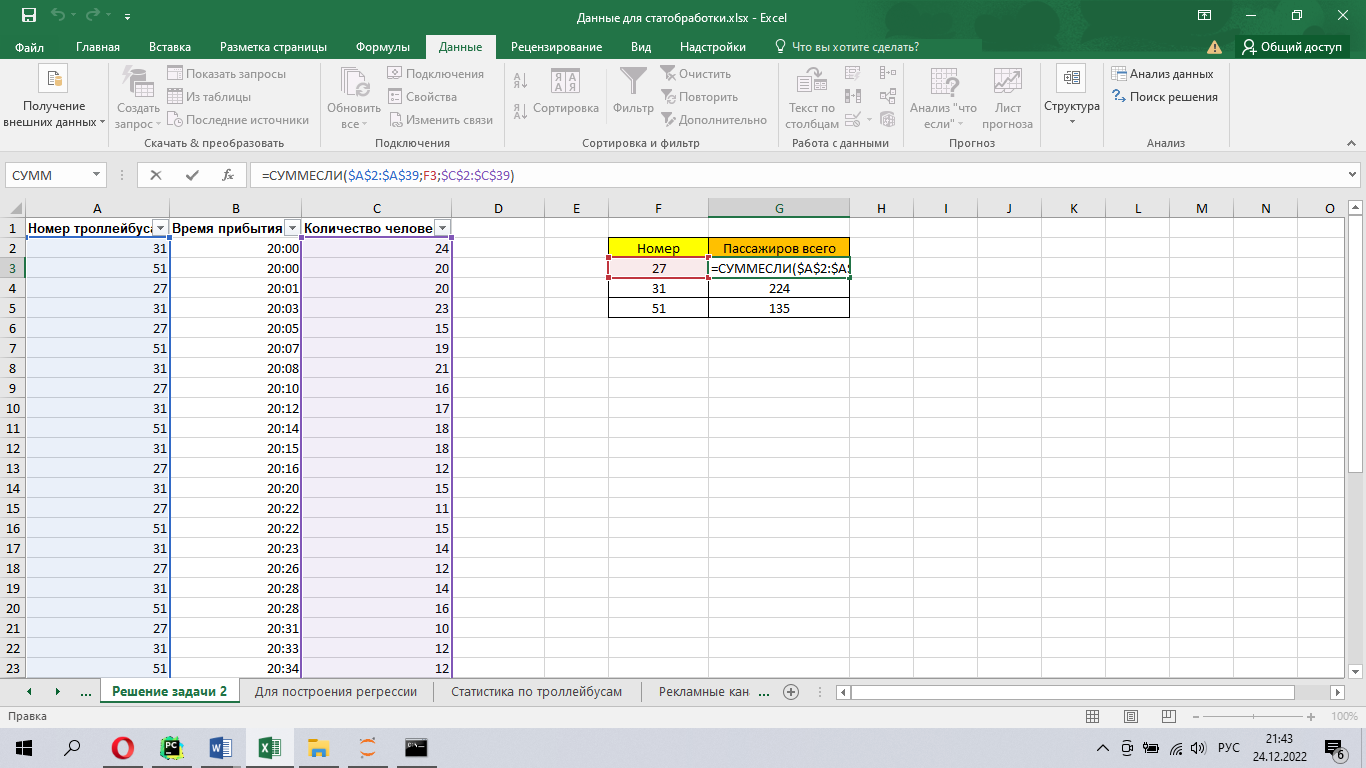
*К сожалению, придется всё строить в Excel. Скорее всего ошибка в размерности, но попытка сделать*

*.reshape(1, 500) не помогла.*

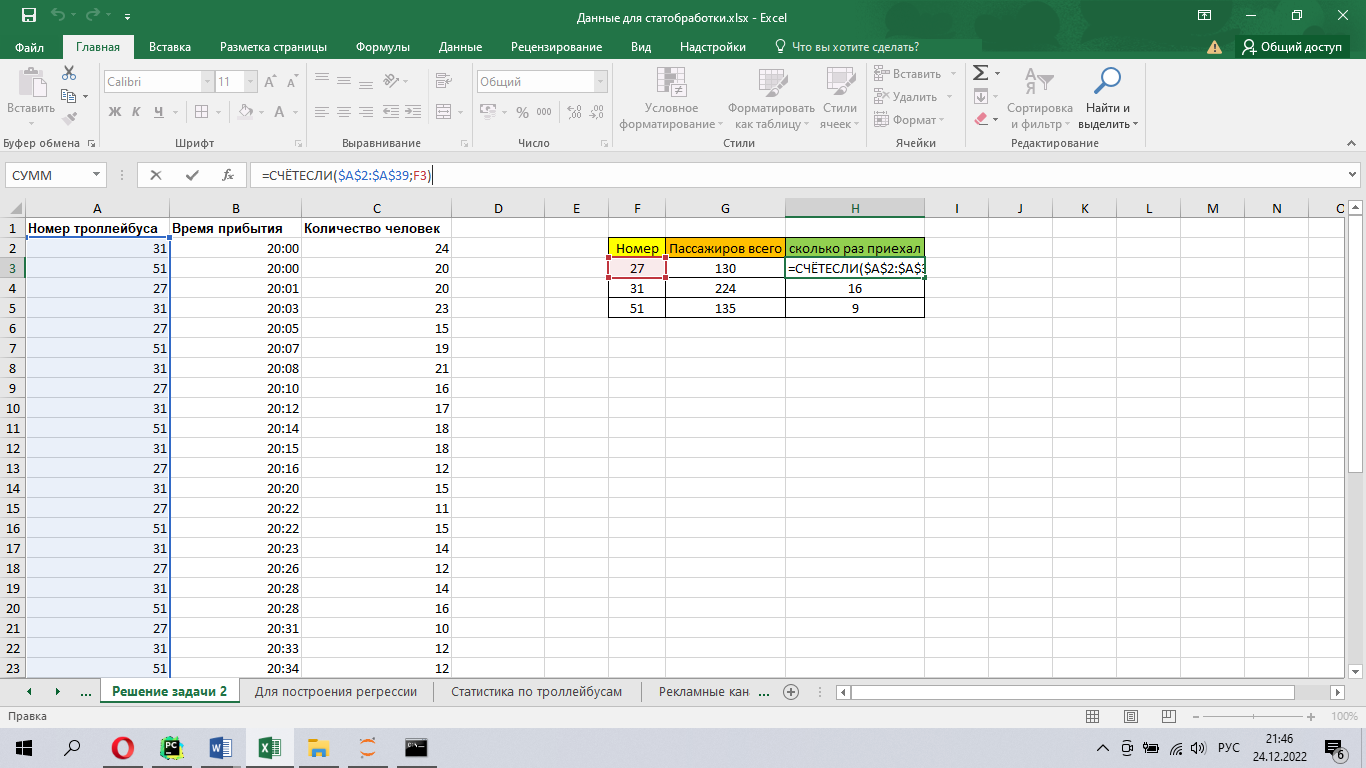
*Моё решение на Python лежит в репозитории по адресу* [*https://github.com/Ultraluxe25/PriorBank-Risk-Analyst*](https://github.com/Ultraluxe25/PriorBank-Risk-Analyst)



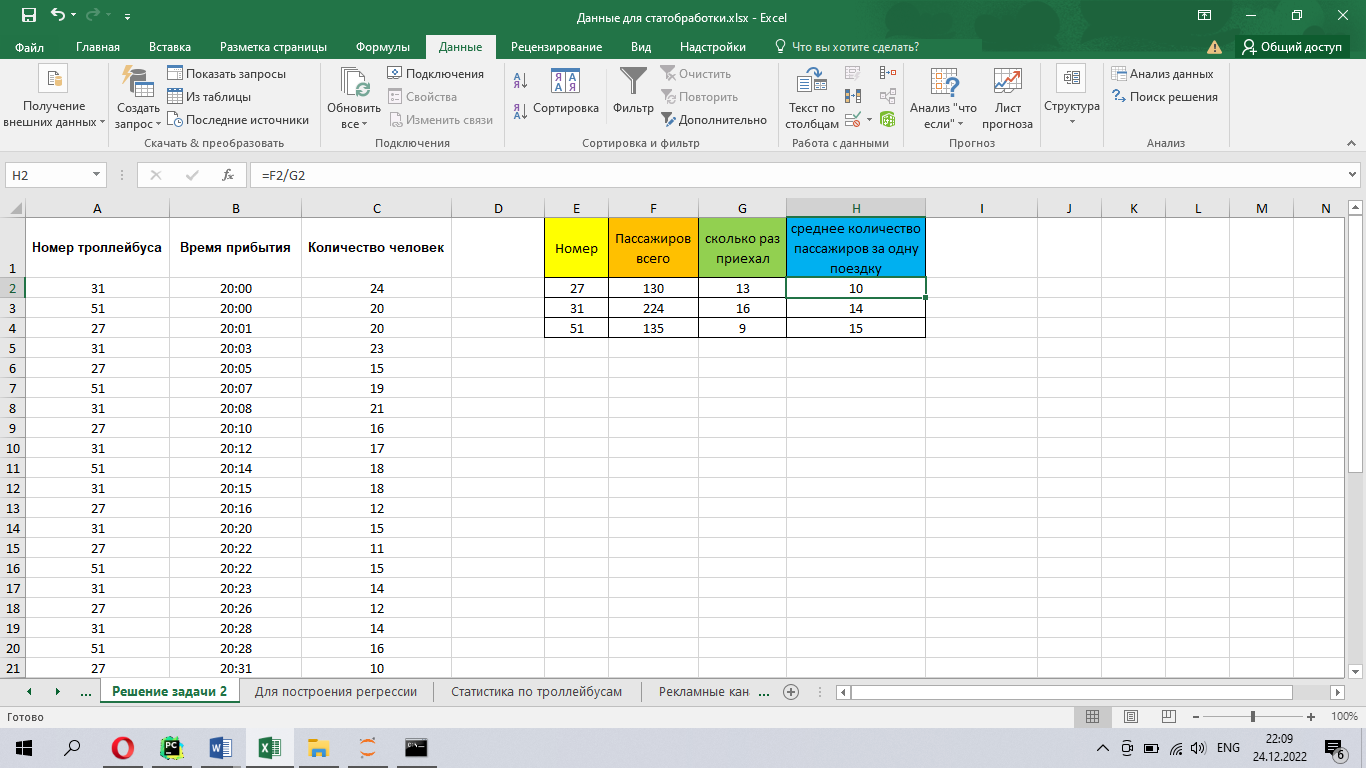
*В Excel получим следующий результат. Коэффициент A будет равен 16, а коэффициент B (y-пересечение) 5.845 Коэффициент детерминации составит 0.999 при 500 наблюдений. Уравнение Y = 5.485 + 16X парной линейной регрессии. Р-value 0, значимость высокая при 1%, отвергаем H0, принимаем H1 о значимости коэффициентов.*



*Задача 2. Начнём решать вторую задачу. Для начала подсчитаем сколько всего пассажиров за час пользовались тем или иным номером троллейбуса.*

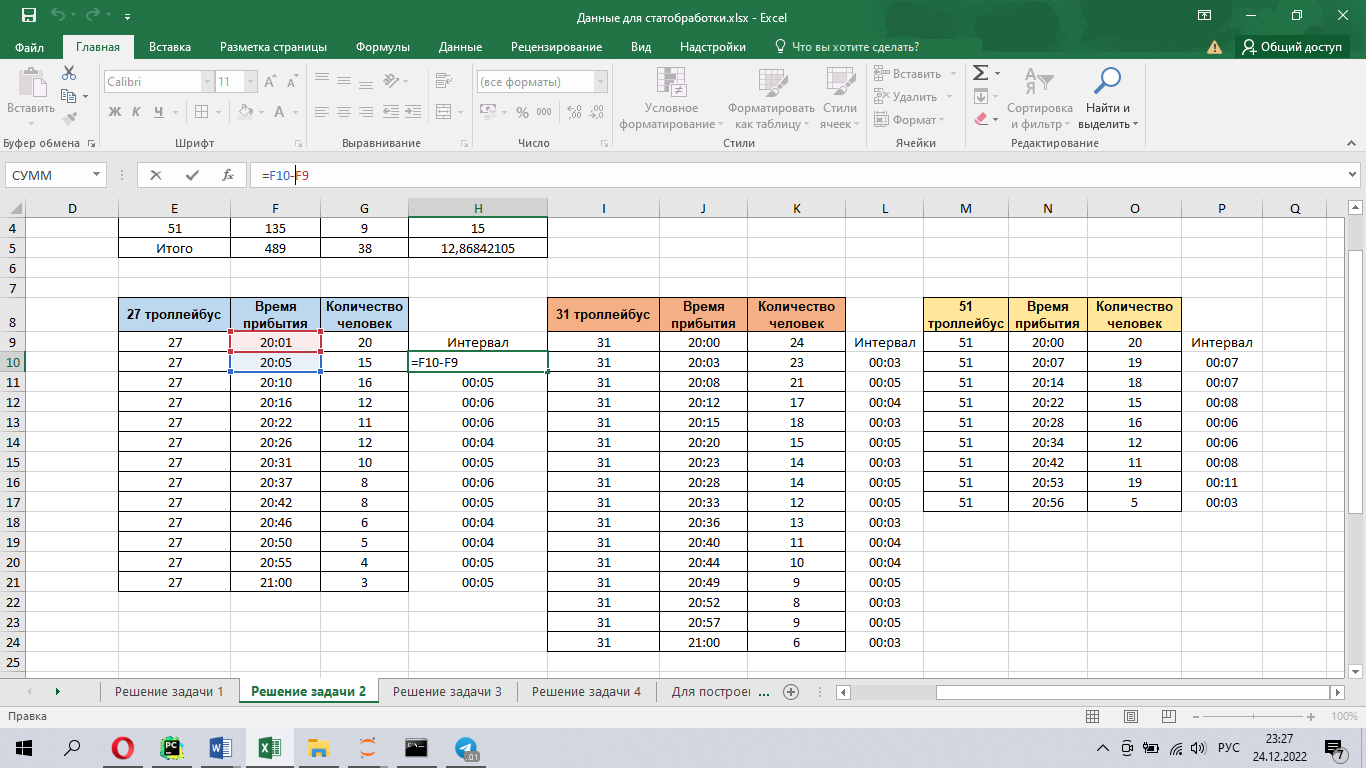


*Также мы рассчитали сколько каждый из номеров троллейбуса приехал раз за наблюдаемый час*

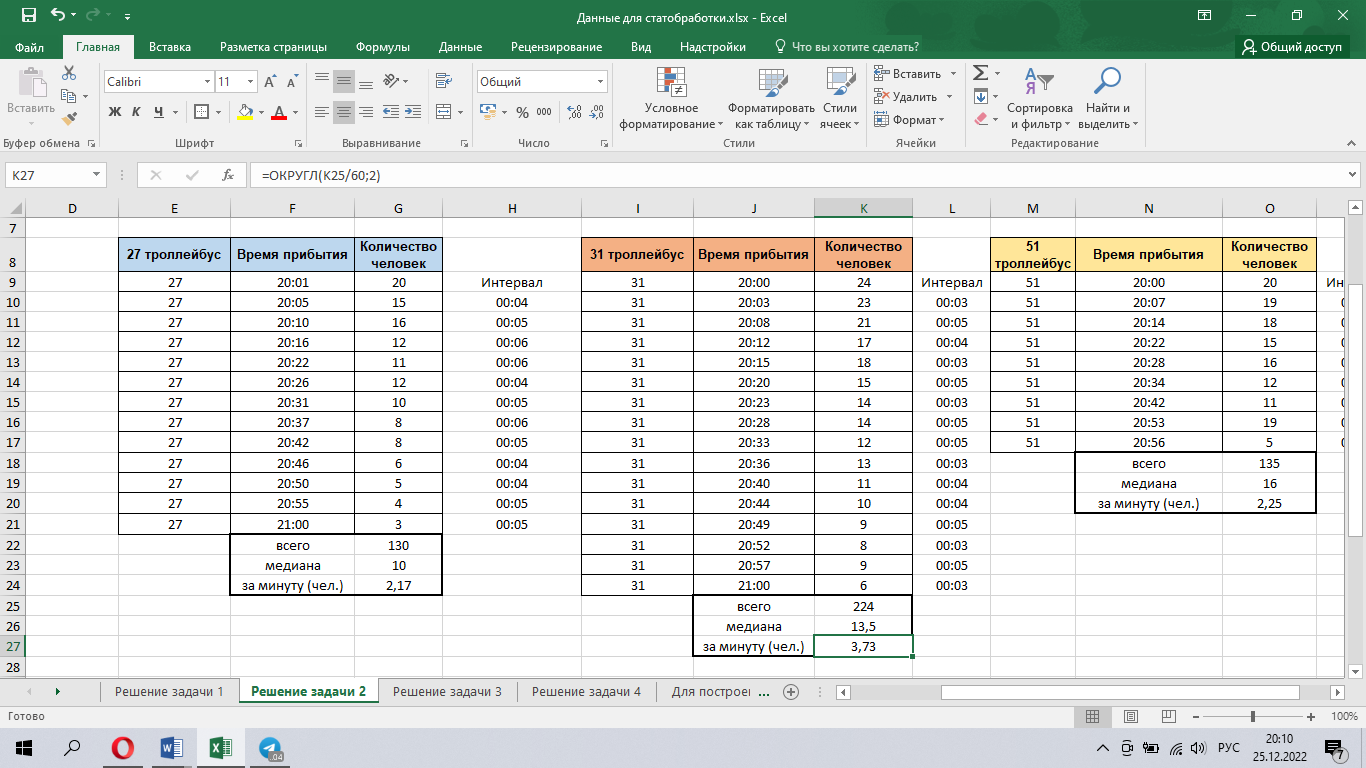


*Мы рассчитали среднее количество пассажиров за одну поездку для каждого из номеров троллейбуса. Давайте рассмотрим на сколько больше людей или меньше заходит в транспорт от среднего значения. Ведь там, где заходит больше чем в среднем, интервал между троллейбусами следует сократить, а там, где меньше людей, наоборот увеличить.*

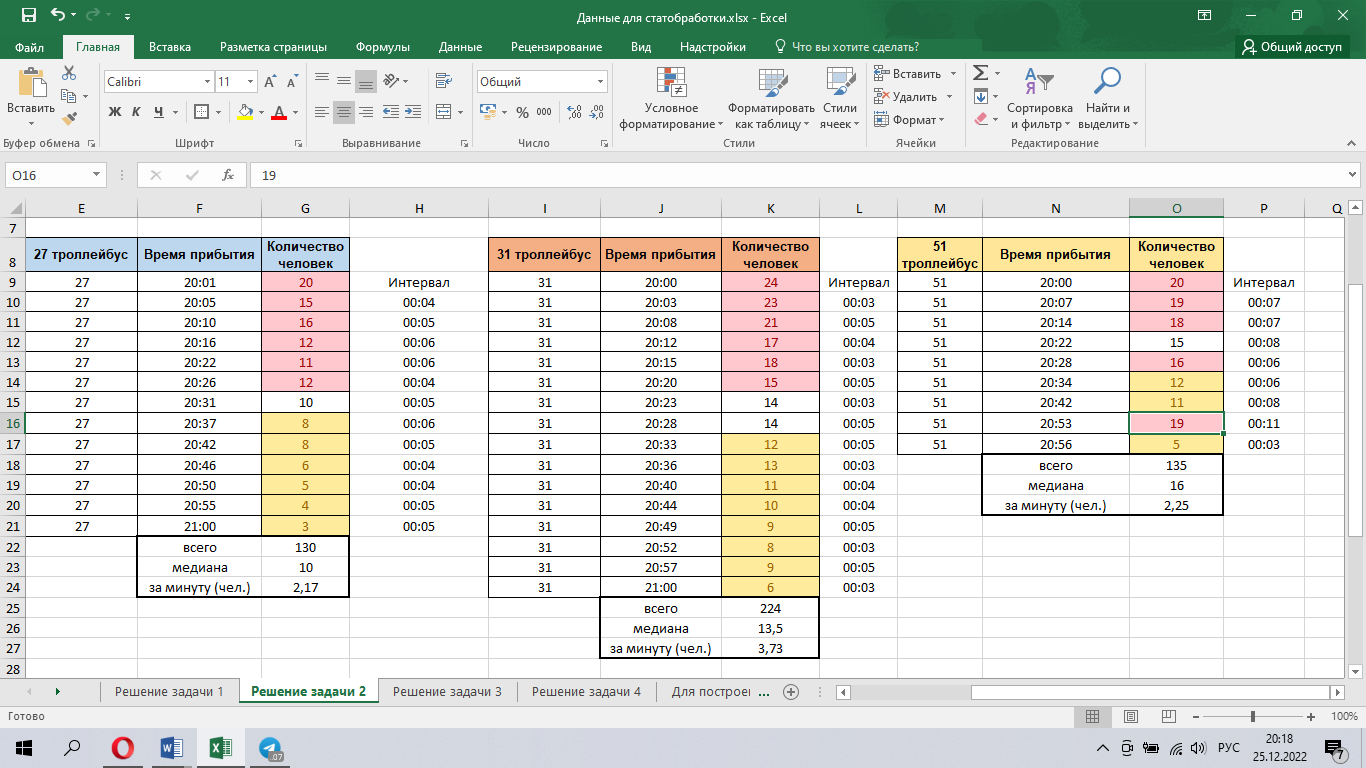
*Воспользуемся фильтром, чтобы отдельно отобразить каждый из троллейбусов и посмотрим в какое время, сколько человек ездило на них.*



*С помощью фильтров и вставки отдельно соберём расписание по каждому из троллейбусов. Интервалы позволят нам определить время ожидание. Рассчитаем в среднем сколько в минуту человек должно уезжать из остановки, а сколько уезжает фактически. Это поможет найти расхождения.*

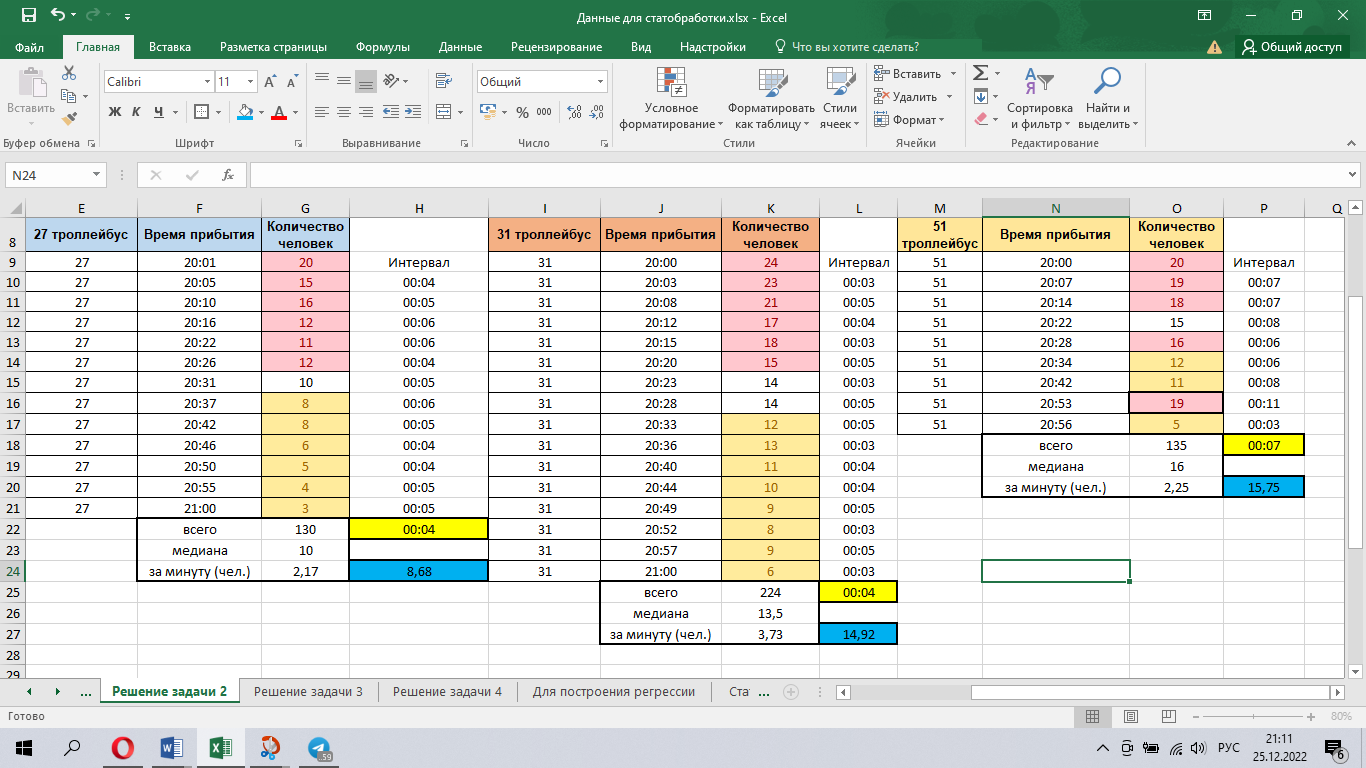


*Подсчитали медиану, чтобы нивелировать выбросы крайних 5-10%, дополнительно вычислили сколько человек в среднем за минутный интервал перевозит каждый из троллейбусов.*



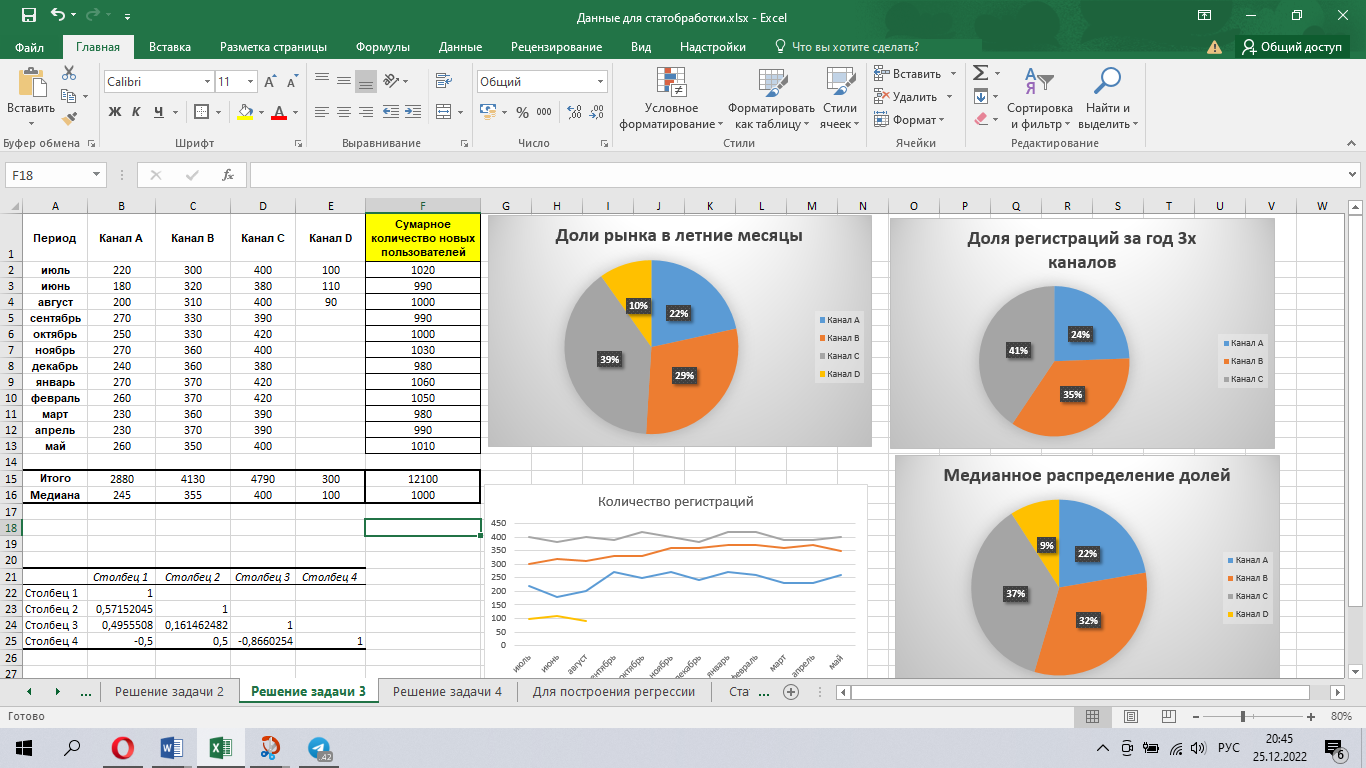
*Использовали условное форматирование. Ячейки, помеченные красным, время ожидания между ними следует снизить, а в оранжевых наоборот увеличить временной интервал между приездом.*

*Особенно выделяется предпоследний 51-й в ячейке O 16, проходящий в 20:53 и загружающий 19 человек, хотя до этого времени загружалось всего 11 человек, а после всего 5.*

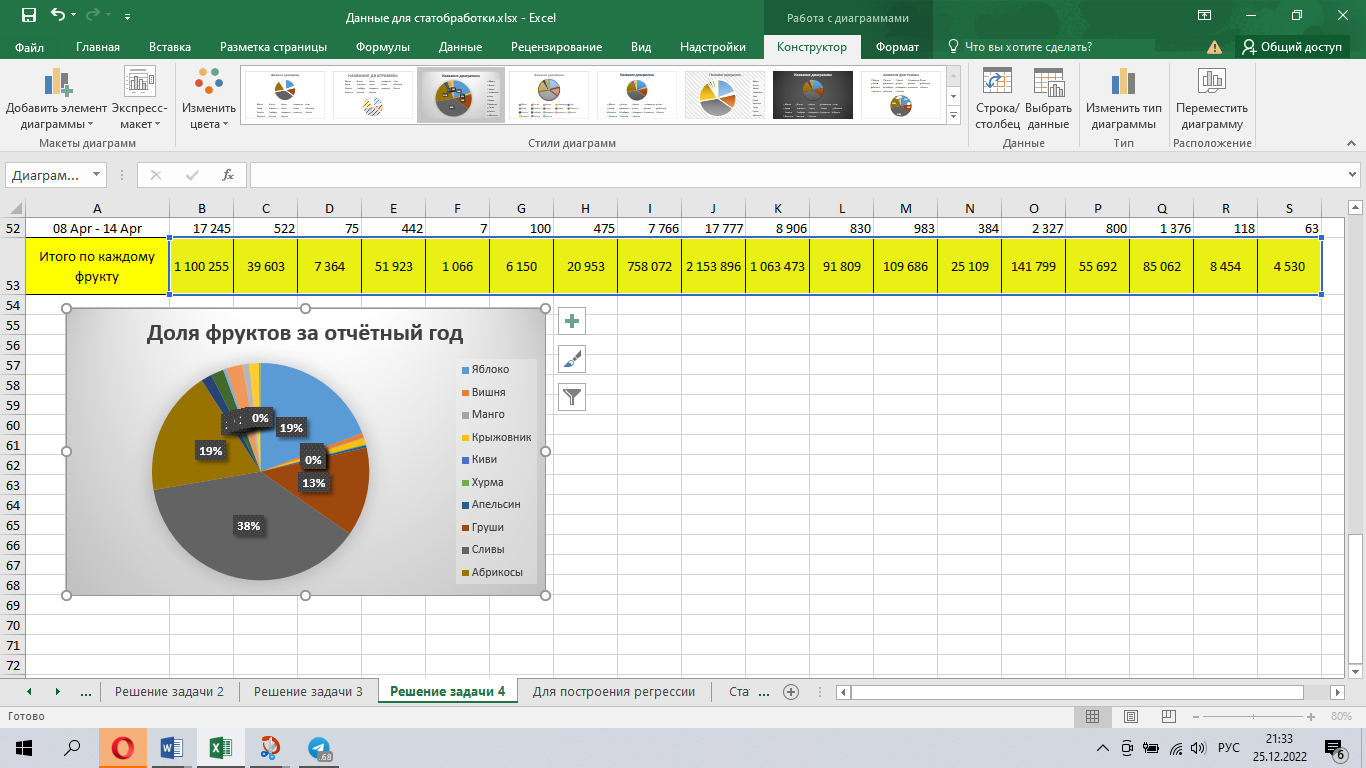


*На последок рассчитал, что интервал движения 27 и 31 троллейбуса в среднем 4 минуты, а 51 целых 7 (это не честно, особенно учитывая, что 27 троллейбус перевозит людей на 5 меньше, но ездит 13 раз против 9 для 51 троллейбуса и интервалы у него меньше). Желтые ячейки показывают средний интервал (медиану), а синие ячейки то, сколько в среднем следовало бы перевозить людей за интервал.*

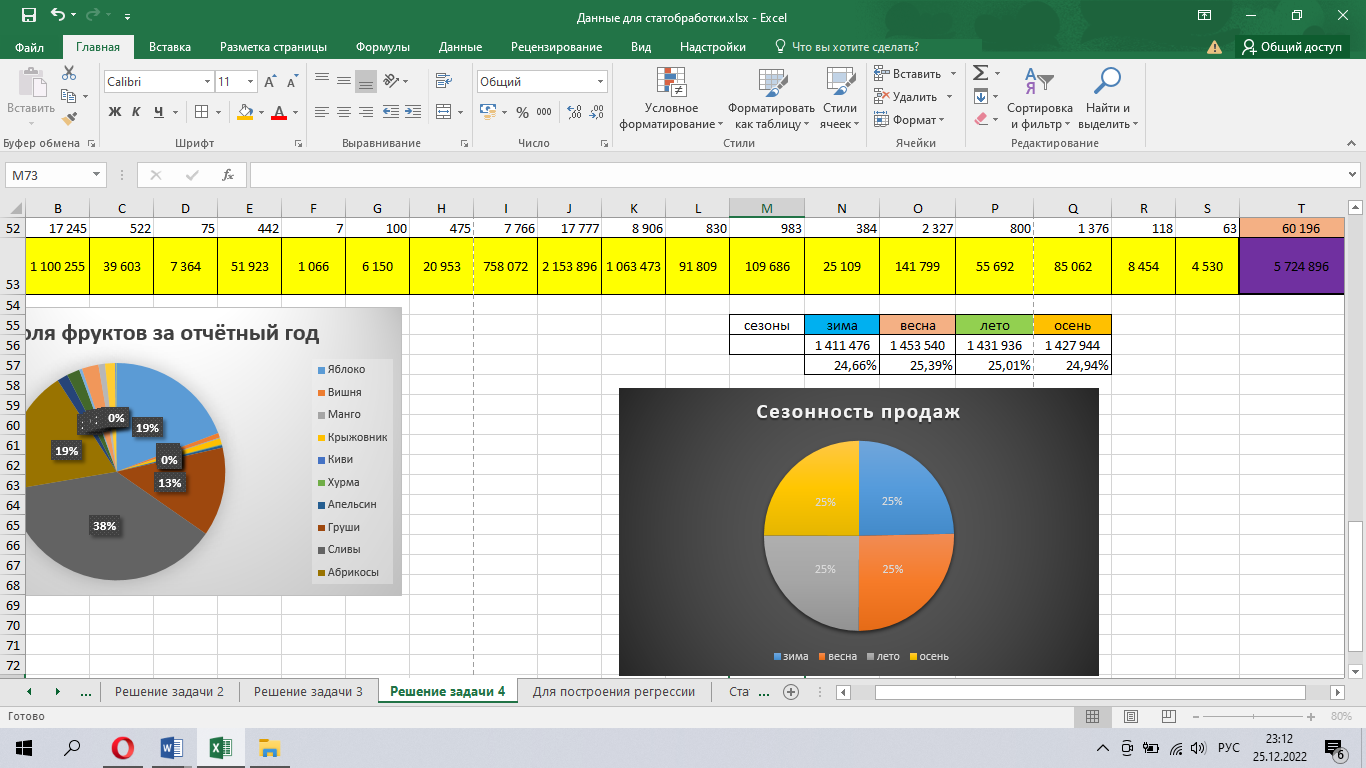
*По итогу предлагаю уменьшить количество рейсов 27 (увеличить интервал) и увеличить количество рейсов 51-го.*



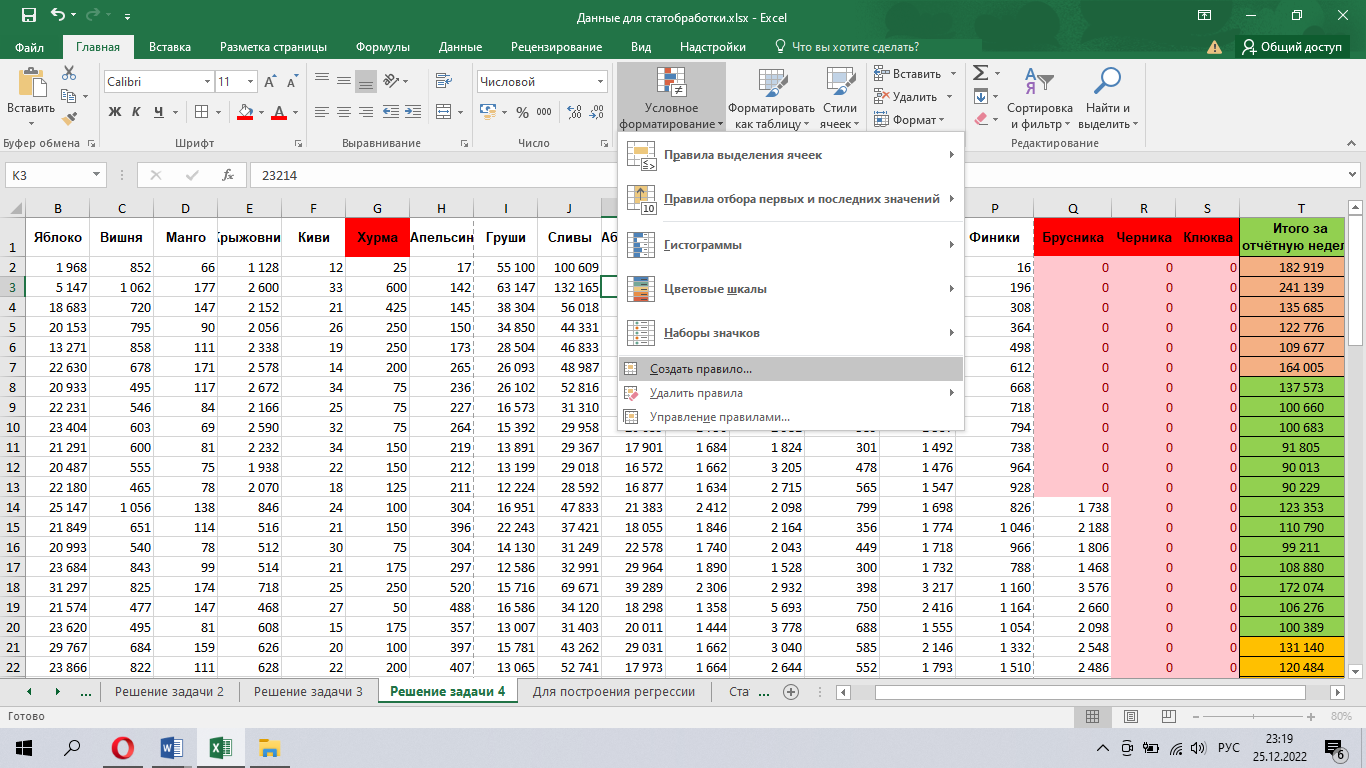
*Задача 3. В третьей задачи я подсчитал количество пользователей за год 12100. Медианное значение (абстракция от 5-10% крайних) составило +1000 регистраций в каждый месяц. Круговые диаграммы долей (в летние месяцы, а также медианное распределение долей) показывают, что доля регистраций на канале D составляет в среднем около +100 регистраций в месяц (около 9-10%). Также представлена таблица корреляции (зависимости каналов) друг от друга. В целом, после ухода канала D на 9 месяцев из рынка, доли других компаний не особо поменялись (только доля компании B выросла с 29 до 35%). Не строя уравнений регрессии можно предположить, что возвращение на рынок канала D позволить вернуть свои небольшие рыночные доли и собирать +100 +/- 10% новых регистраций.*



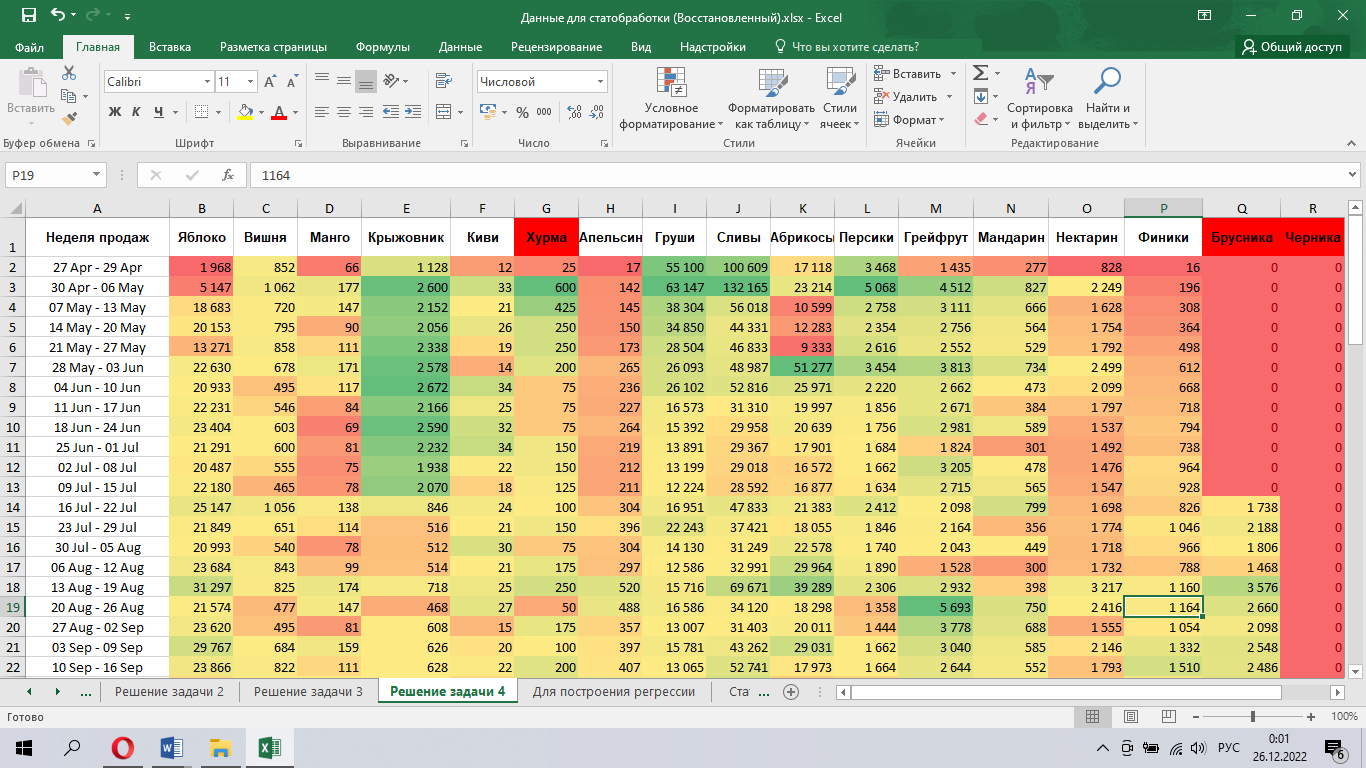
*Задача 4. Построили круговую диаграмму долей каждого из фруктов. Оказывается, большее количество приходится на: сливы (38%), яблоки и абрикосы (по 19% каждый) и груши (13%). Остальные 11% занимают оставшиеся виды фруктов.*



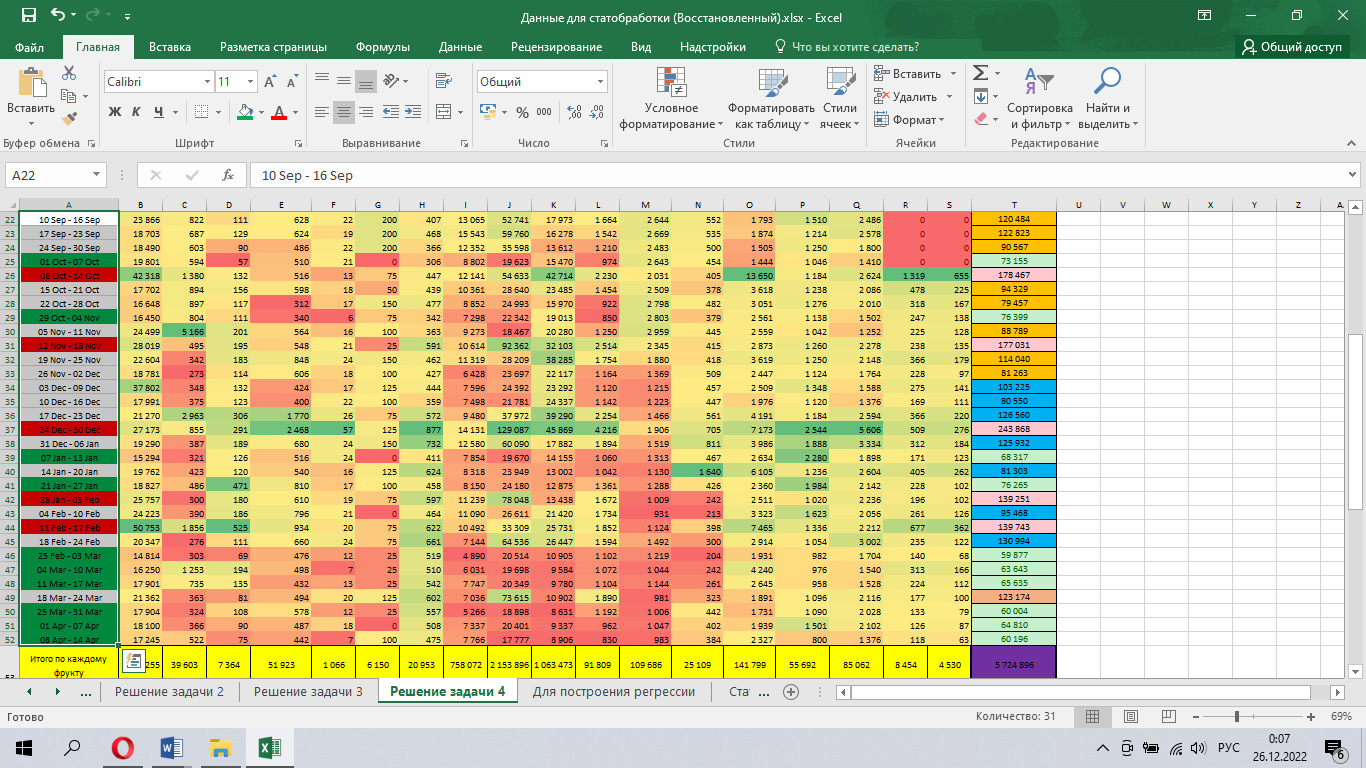
*На следующей круговой диаграмме мы проверили сезонность. В целом ни один из сезонов не выделяется среди других (примерно 25% всех продаж приходится на каждый из сезонов. Зимой продают меньше (24.66%), а весной (25.39%) больше. Абсолютные значения каждого из сезонов также представлены.*



*С помощью условного форматирования мы выделили те ячейки (продажи фруктов по неделям), в которых объём продаж равен 0. Оказывается, что периодически не продаётся хурма, а для брусники, черники и клюквы на протяжении долгого времени вообще отсутствовали продажи долгое время.*



*Также мы применили условное форматирование для каждого столбца, чтобы выделить те недели где продаж было больше (зелёные) и те недели, где продаж было меньше (красные).*



*Мы выделили те недели, в которые объём совокупных поставок минимальны (10 недель выделены красным фоном) и наоборот максимальны (10 недель выделены зелёным). Столбец таблицы А.*

*В целом хотелось бы поблагодарить авторов заданий за интересный досуг на выходные дни (25-26.12.2022). Все задания были выполнены за 2 дня (до этого не было возможности приступить к задачам из-за начавшейся сессии в университете). По часам точно не сориентирую, но делал перерывы.*

*Больше всего понравилась первая задача, т.к. наиболее подходящим инструментом для её решения стал мой любимый Python. Удалось поработать также с разными фреймворками для анализа данных (Pandas, Matplotlib и др.).*

*В 2-4 задачах комфортнее для наглядности было решать в Excel. Убедился, что в некоторых случаях он не уступает по функциональности полноценным языкам программирования, заточенным для работы с большими данными.*