|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_\_\_\_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ (ИУ5)\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

по дисциплине \_\_\_\_\_\_\_\_\_\_\_\_\_ Оперативный анализ данных \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

по теме \_\_\_\_\_\_\_ «зарботная плата Data science специалистов» \_\_\_ \_\_\_ \_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент ИУ5-52Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** А. Д. Зайцев

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** К. Ю. Маслеников

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2022 г.*

Аннотация

В задании необходимо проанализировать базу данных “заработная плата Data Science специалистов” и очистить от не нужных данных, также нужно выдвинуть гипотезы, проанализировать их и составить графики по данным гипотезам.

В ходе работы будут использоваться такие библиотеки как: «Pandas» и «Seaborn». Анализ будет проводиться через программу «Jupyter Python».

Также будут проводиться корреляционный анализ, агрегирование данных, оптимизация память, удаление дубликатов, очистка данных. Будет производиться анализ гипотез по известным данным и построены соответствующие графики и схемы.

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 4](#_Toc122080054)

[1. Определение данных для анализа 5](#_Toc122080055)

[2. Описание данных 5](#_Toc122080056)

[3. Формулирование гипотез 6](#_Toc122080057)

[4. Изучение общей информации 7](#_Toc122080058)

[5. Подготовка данных для работы 9](#_Toc122080059)

[6. Агрегирование данных 13](#_Toc122080060)

[7. Исследовательский анализ данных 14](#_Toc122080061)

[8. Общий вывод 21](#_Toc122080062)

[ЗАКЛЮЧЕНИЕ 22](#_Toc122080063)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 23](#_Toc122080064)

# ВВЕДЕНИЕ

Необходимо исследовать базу данных и выявить закономерности, которые потом могут быть полезны для коммерческой деятельности компаний, заинтересованных в рынке анализа данных.

Цели:

1. определение данных;
2. формулирование гипотез;
3. загрузка данных в Python;
4. проверка данных;
5. очистка данных;
6. преобразование данных;
7. выбор данных для анализа;
8. агрегирование данных;
9. визуализация данных;
10. подтверждение или опровержение поставленных гипотез;
11. формулирование ограничений и выводов.

# Определение данных для анализа

В качестве данных для анализа был выбран датасет «заработная плата Data Science специалистов». В датасете представлены данные о заработных платах Data Science специалистов и их характеристики. Собранные данные помогут нам провести глубокий анализ и обратить внимание на весомые проблемы в Data Science рынке. На данный момент существует много вопросов, касаемых этого направления. Интересно узнать в какой стране самые высокие зарплаты, какие должности получают больше всех, нужны ли навыки управления персоналом и что ждет рынок в будущем?

1. **Описание данных**

Для анализа были собраны данные обо всех проектах, включённых в список. В наборе данных содержатся:

1. Id – уникальный номер значения в таблице;
2. Work\_year – год в который работал специалист;
3. Experience level – опыт работы;
4. Employment\_type – тип трудоустройства;
5. Job\_title – занимаемая должность;
6. Salary – зарплата;
7. Salary\_currency – курс зарплаты;
8. Salary\_in\_usd – зарплата в долларах;
9. Employee\_residence – страна проживания сотрудника;
10. Remote\_ratio – коэффициент удаленной работы;
11. Company\_location – страна где расположена компания;
12. Company\_size – размер компании;

# Формулирование гипотез

Гипотеза 1: Руководящие должности зарабатывают больше.

Гипотеза 2: Фриланс – самое выгодное место работы.

Гипотеза 3: В удаленном режиме люди работают продуктивнее, так как не тратят время на поездку до работы.

Гипотеза 4: Страна с самой высокой заработной платой - США.

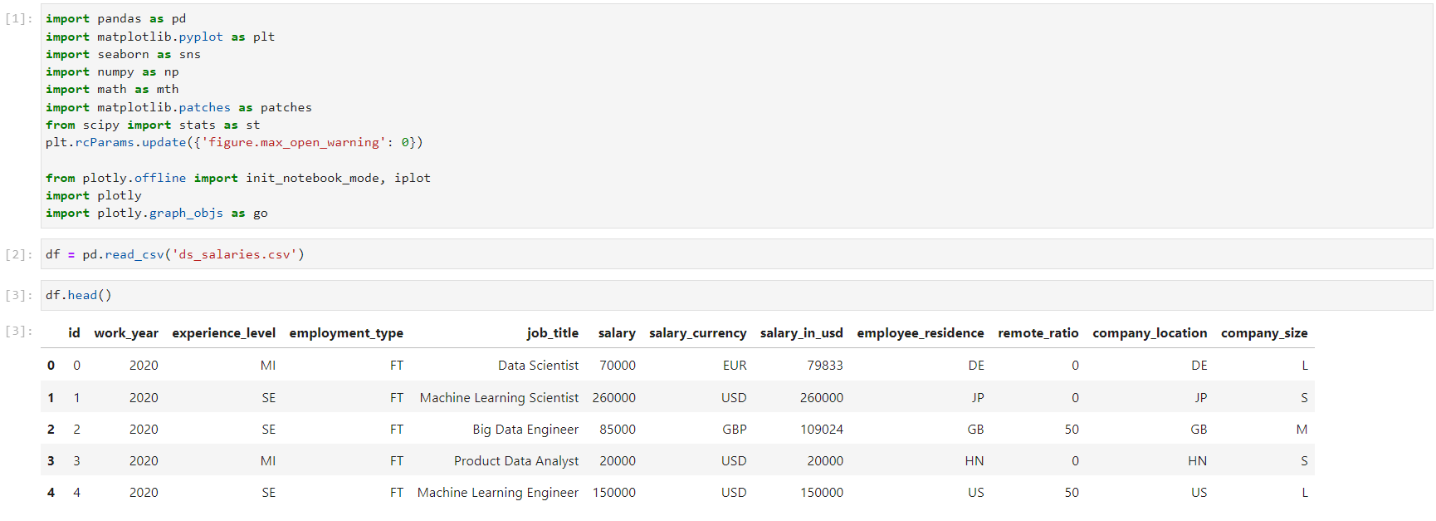
Гипотеза 5: В больших компаниях платят больше.

Гипотеза 6: Рынок Data Science растёт.

Данные формулировки помогут нам проанализировать рынок Data Science специалистов и сделать нужные выводы.

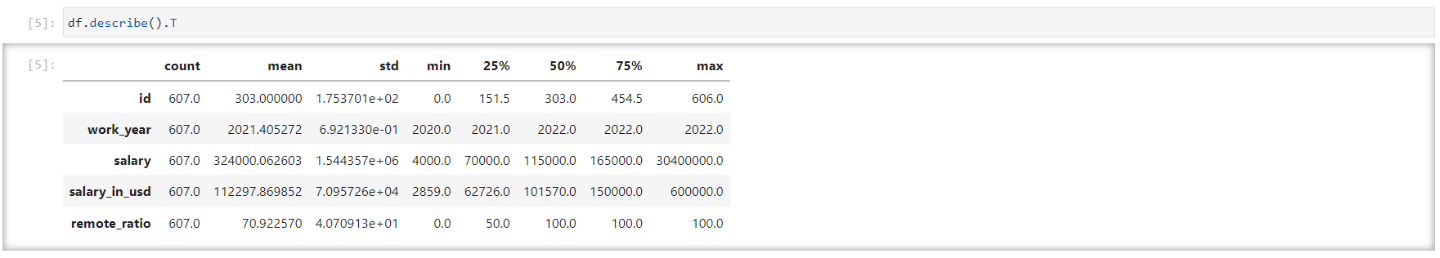
# Изучение общей информации

Загружаем датасет и подключаем необходимые библиотеки: Pandas, matplot, seaborn, numpy, scipy, plotly:

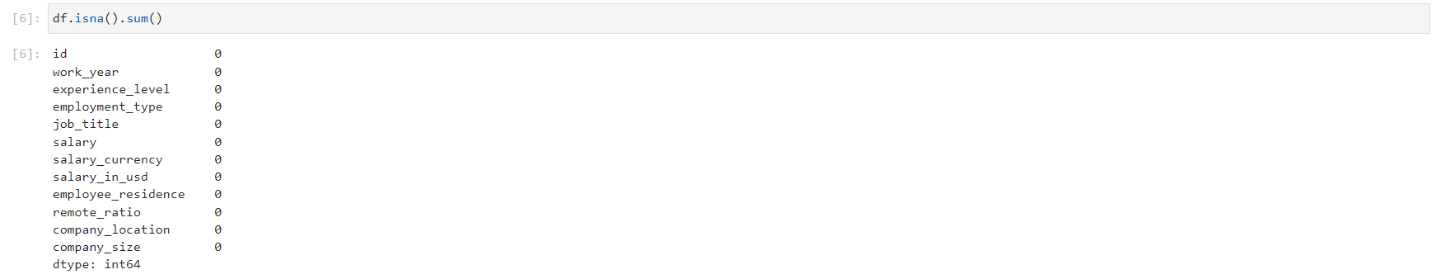


Получим информацию о датасете, на которой увидим использование памяти и количество переменных по их типам, заметим, что используется 57 КВ памяти, типов int-64 используется 5 раз, типов object используется 7 раз, в дальнейшем придется оптимизировать использование памяти для типов данных int-64 и object:





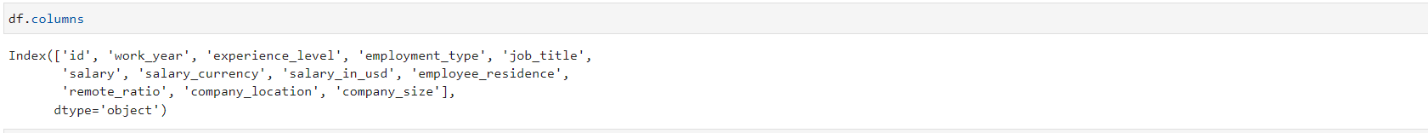
Посчитаем количество пустых значений:



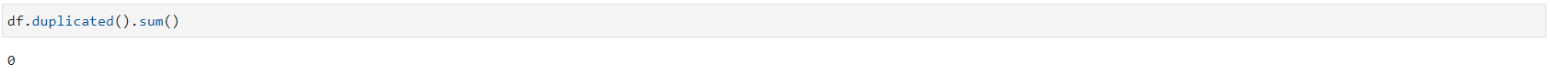
Пустые данные отсутствуют, значит никаких дополнительных обработок нам не потребуется. На этом этапе мы изучили общую информацию о нашем датасете и готовы перейти к подготовке данных для анализа.

1. **Подготовка данных для работы**

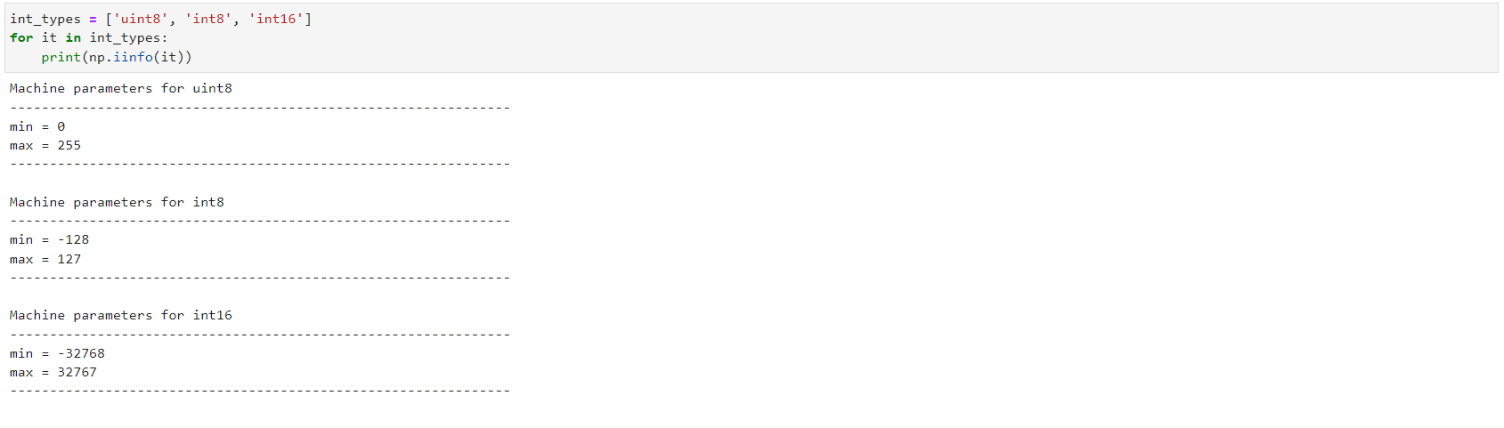
Проверим названия столбцов:



Посчитаем количество дубликатов:



Оптимизируем память и изменим типы данных на более корректные:



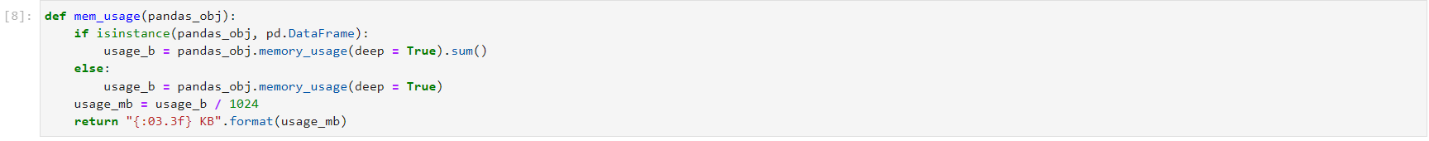


Так как данные разных типов хранятся раздельно, мы исследуем использование памяти разными типами данных. Начнём со среднего показателя использования памяти по разным типам данных:



Эти сведения дают нам понять то, что большая часть памяти уходит на столбцы, хранящих объектные значения.

Создадим функцию, которая поможет нам сэкономить немного времени:

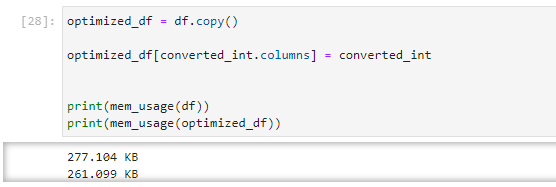






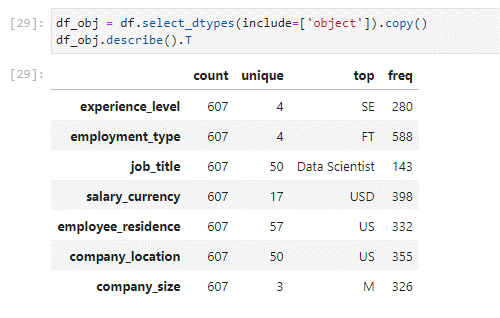
В результате можно видеть падение использования памяти int-данными с 23.836 KB до 7.831 KB, с float-данными мы видим повышение с 0.125 KB до 11.980 KB, то есть — мы снизили потребление памяти для int однако увеличили для float, соответственно для float использовать не будем.

Создадим копию исходного объекта Data Frame, используем эти оптимизированные числовые столбцы вместо тех, что присутствовали в нём изначально, и посмотрим на общий показатель использования памяти после оптимизации:

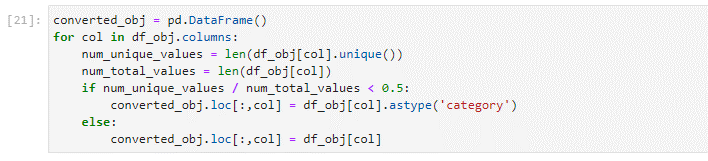


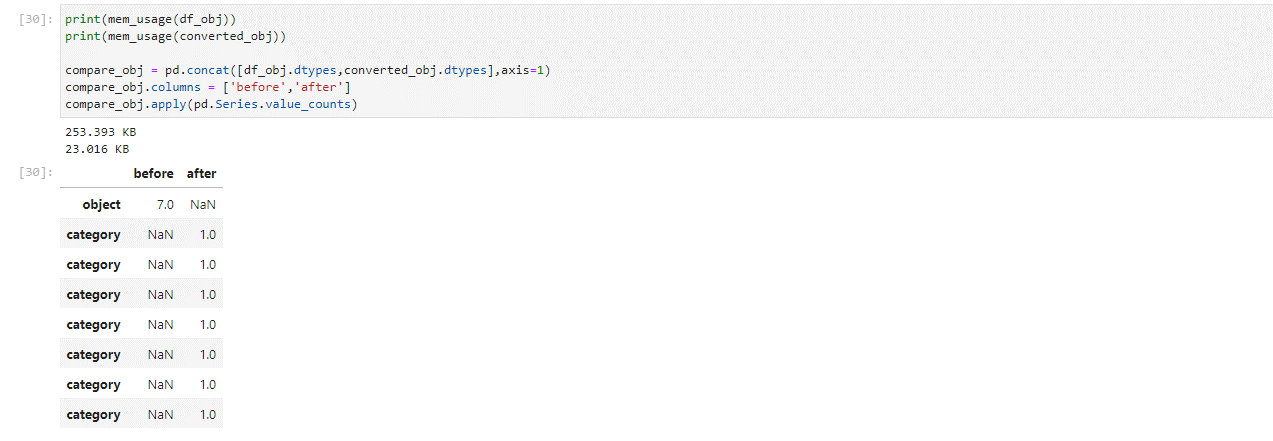
После уменьшения потребления памяти столбцами, хранящими числовые данные, нам удалось уменьшить потребление памяти. Для достижения еще большего результата оптимизируем хранение объектных типов.

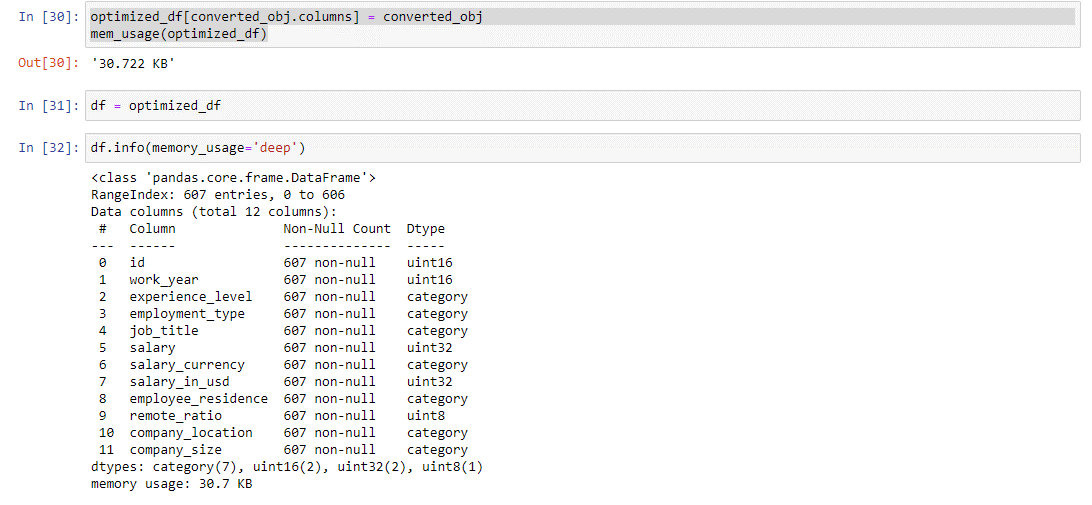
Для того чтобы понять, где именно мы сможем воспользоваться категориальными данными для снижения потребления памяти, выясним количество уникальных значений в столбцах, хранящих значения объектных типов:

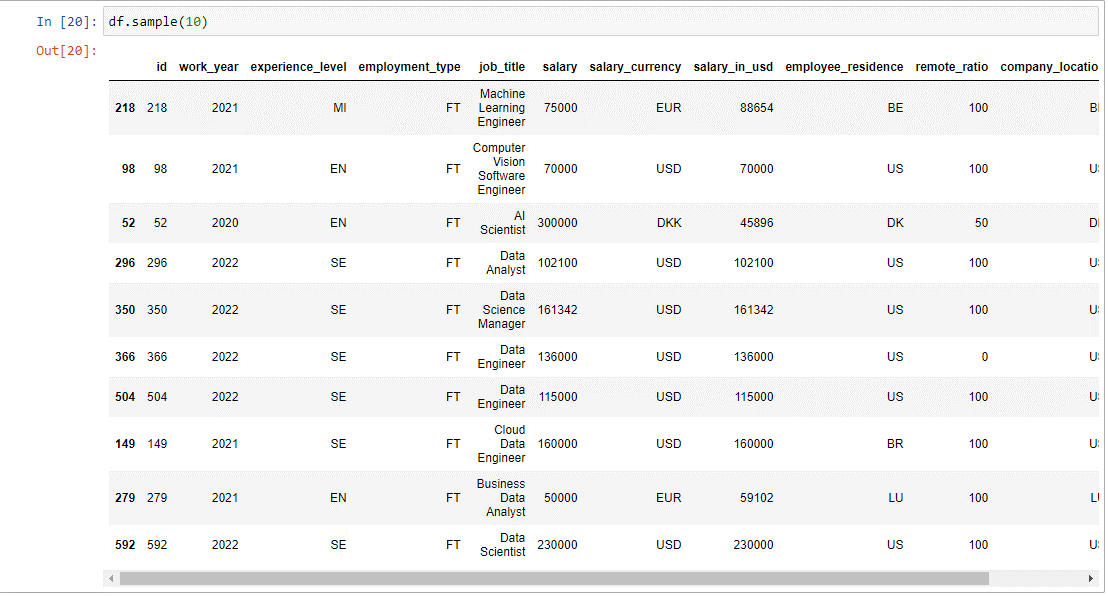


Создадим цикл, который перебирает все столбцы, хранящие данные типа object, выясняет, не превышает ли число уникальных значений в столбцах 50% и, если это так, преобразует их в тип category.:



Теперь сравним то, что получилось после оптимизации, с тем, что было раньше: 





1. **Агрегирование данных**

Добавим новый столбец: показатель работает ли человек в родной стране или нет. Для этого напишем функцию checkEqual(), которая сравнивает два значения. После этого добавим столбец “WorkAtMotherland” путем сравнения страны проживания и страны работы. В итоге получим новую колонку в которой 0 – сотрудник работает на родине, 1 – сотрудник работает не на родине. Данная информация пригодится нам в дальнейшем анализе.



**Вывод:**

1. Была произведена оптимизация памяти и изменение типов данных для дальнейшего анализа;
2. Пропущенных значений не имеется;
3. Дубликатов не имеется.
4. **Исследовательский анализ данных**

**Руководящие должности зарабатывают больше**



Рисунок 7.1 – Самые прибыльные должности

**Вывод:**

По графику на рисунке 7.1 можно заметить, что должности ,включающие в себя руководство, возглавляют список. Это значит, что на рынке труда ценятся не только “hard skills”, но также “soft skills”.

Гипотеза 1 – подтверждена.

**Зависимость заработной платы от типа трудоустройства**



Рисунок 7.2 – Диаграмма зависимости заработной платы от типа трудоустройства

**Вывод:**

Из графика на рисунке 7.2 видно, что люди, которые работают по контракту получают в среднем в 2 раза больше чем люди, которые работает по системе “full time job”. Скорее всего это связано с тем, что за рубежом система работы по контракту больше распространена. Также видно, что “фрилансеры” получают больше чем люди, работающие на пол ставки. Это связано с тем, что они уделяют больше времени работе.

Гипотеза 2 – опровергнута.

**Эффективность работы в удаленном режиме**



Рисунок 7.3 – Диаграмма зависимости зарплаты от типа работы

**Вывод:**

Из графика на рисунке 7.3 видно, что от режима работы не сильно зависит зарплата сотрудников, это может быть связано с глобальной эпидемией, когда людям приходилось работать удалённо. Однако наибольшая средняя зарплата всё равно у людей работающих на своём рабочем месте. Из чего можно сделать вывод, что максимальная эффективность достигается непосредственным присутствием на работе.

Гипотеза 3 - опровергнута

**Страна с самой высокой заработной платой – США**



Рисунок 7.4 – Средняя зарплата в зависимости от страны компании

**Вывод:**

Из графика на рисунке 7.4 видно, что самые высокие зарплаты у компаний в России.

Гипотеза 4 – опровергнута.

**В больших компаниях платят больше**



Рисунок 7.5 – Диаграмма зависимости средней заработной платы от размера компании

**Вывод:**

Из графика на рисунке 7.5 видно, что маленькие компании не могут предоставить такую оплату труда как средние или большие. Скорее всего это связано с задачами, которые преследует маленькие компании. Большой компании намного больше необходимы специалисты Data Science , так как у них больше ресурсов, которые необходимо анализировать.

Гипотеза 5 – подтверждена.

**Анализ рынка Data Science**



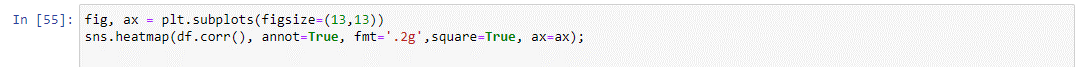
Рисунок 7.6 – Диаграмма выплат по годам

**Вывод:**

Из графика на рисунке 7.6 виден резкий скачек примерно на 20% в 2022 году. Это говорит о том, что сфера Data Science набирает обороты и становится всё более востребованной. Также на это повлияли глобальная эпидемия и мировой экономический кризис, поэтому виден застой на 2021 году, когда зарплаты почти не изменились по сравнению с предыдущим годом.

Гипотеза 6 – подтверждена.

**Корреляционный анализ**



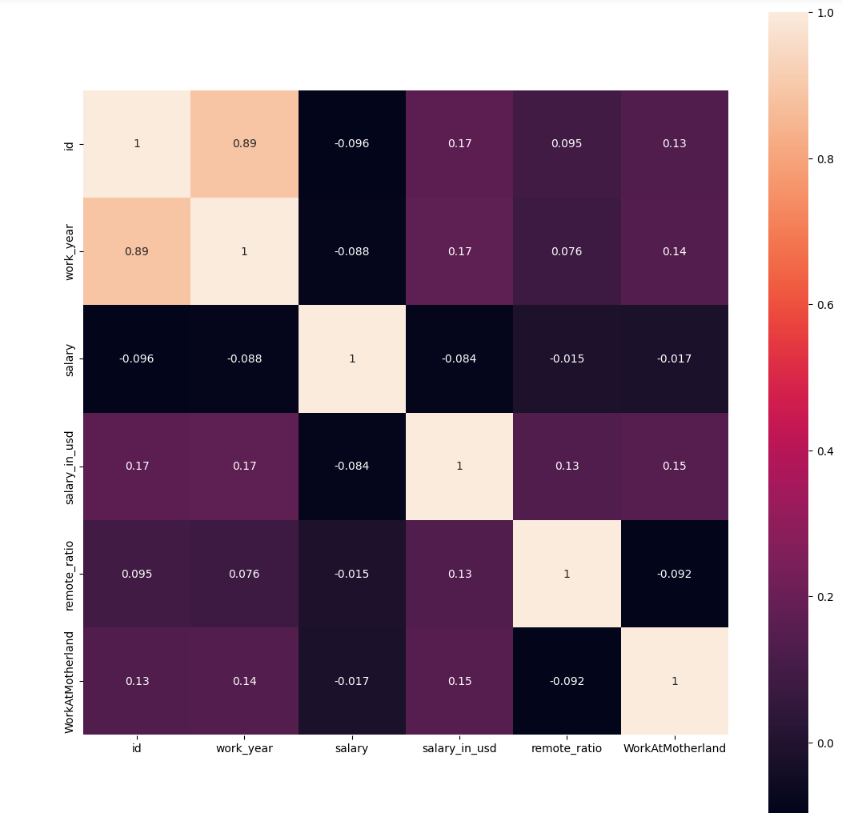


Рисунок 7.7 – Корреляционный анализ

По данной матрице корреляций на рисунке 7.7 можно проследить, что весомых зависимостей не было обнаружено.

# Общий вывод

Перед анализом было оптимизированно использование памяти путем преобразования данных к более корректным типам. Также были добавлены необходимые столбцы при помощи агрегирования данных, что в дальнейшем позволило провести необходимые в ходе работы исследования и получить ценные знания в предметной области «заработная плата Data Science специалистов».

В ходе анализа было выявлено:

1. руководящие должности зарабатывают больше остальных;
2. фриланс – самый невыгодный тип трудоустройства;
3. в удаленном режиме работать не эффективно;
4. русские компании имеют самую высокую оплату труда;
5. в больших компаниях работать выгоднее, чем в менее крупных;
6. рынок Data Science прогрессирует с большой скоростью;

# ЗАКЛЮЧЕНИЕ

В ходе научно-исследовательской работы был проведен анализ базы данных «заработная плата Data Science специалистов», выявлены важные статистические данные, которые хорошо описывают нынешнее положение рынка Data Science и тенденции развития.

Во время выполнения работы были получены необходимые навыки для выполнения анализа данных над большими объёмами информации. Также были получены знания по визуализации данных с использованием крутого программного обеспечения, а именно крутой новомодный язык “Питон”.

Для выполнения данной работы было изучено новое программное обеспечение: «Pandas»[1][3] и «Seaborn»[2][4] на базе языка программирования Python. Мы изучили основные его функции и принципы, научились визуализировать информацию, посредствам построения диаграмм.

С помощью данного функционала мы оптимизировали и подготовили данные к анализу. После этого опровергли и подтвердили гипотезы. Из них сделали важные выводы.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Методические указания по программному обеспечению «Pandas»;
2. Методические указания по программному обеспечению «Seaborn»;
3. «Pandas. Работа с данными» (2020), Автор: Абдрахманов М. И.;
4. «Python. Визуализация данных: Matplotlib, Seaborn, Mayavi»;