Федеральное государственное образовательное бюджетное

учреждение высшего образования

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ**

**РОССИЙСКОЙ ФЕДЕРАЦИИ»**

**(Финансовый университет)**

**Институт развития профессиональных**

**компетенций и квалификаций**

**ИТОГОВАЯ РАБОТА**

|  |  |
| --- | --- |
| Название программы | **«Аналитик данных»** |
| Группа обучения | **«** 1│studoo│AD│17.06.24**»** |
| Срок обучения | **«**17.06.24 - 20.09.24 г.**»** |
| **«‎Михайлова Татьяна Алексеевна»** | |
| Номер/Название Датасета | **«Аналитика рынка труда с использованием искусственного интеллекта»** |

Москва 2024 г.

**Основные ссылки на работу:**

Ссылка на Looker studio:

<https://lookerstudio.google.com/reporting/0140743b-6384-4a57-aa76-6abb9ebde659>

Ссылка на Colaboratory:

<https://colab.research.google.com/drive/1-xF1XH8ejlhifk18VkaHOG8qYTG-1xZn?usp=sharing>

Ссылка на Loginom:

<https://drive.google.com/file/d/1NA3Ao4ym7BN_nx6JTpevH5z96ryQdmpR/view?usp=sharing>

Ссылка на очищенные данные:

<https://drive.google.com/file/d/1eA_UaYTuuW1g4n-exTzTFU1pIBAL84l-/view?usp=sharing>

Ссылка на Tableau:

<https://drive.google.com/file/d/1Dq2EF1qorU9UF8V9onZA0C1kUrbUxcqw/view?usp=sharing>

**Обзор набора данных:**

Данный набор данных я брала с сайта Kaggle (<https://www.kaggle.com/datasets/uom190346a/ai-powered-job-market-insights>).

Набор данных ‘AI-powered Job Market Insights’ предоставляет синтетический, но реалистичный снимок рынка труда, уделяя особое внимание роли внедрения искусственного интеллекта (ИИ) и автоматизации в различных областях.

**Характеристики набора данных:**

1. Job\_Title- Название должности (Тип: Категориальный)
2. Industry- Отрасль, в которой находится должность (Тип: Категориальный)
3. Company\_Size- Размер компании, предлагающей работу (Тип: Категориальный)
4. Location- Географическое расположение работы (Тип: Категориальный)
5. AI\_Adoption\_Level- Степень внедрения искусственного интеллекта в деятельность компании (Тип: Категориальный)
6. Automation\_Risk- Оценочный риск того, что работа может быть автоматизирована в течение следующих 10 лет (Тип: Категориальный)
7. Required\_Skills- Ключевые навыки, необходимые для данной должности (Тип: Категориальный)
8. Salary\_USD- Годовая заработная плата, предлагаемая за эту работу, в долларах США (Тип: Числовой)
9. Remote\_Friendly- Указывает, может ли работа быть выполнена удаленно (Тип: Категориальный)
10. Job\_Growth\_Projection- Прогнозируемый рост или спад должностной роли в течение следующих пяти лет (Тип: Категориальный)

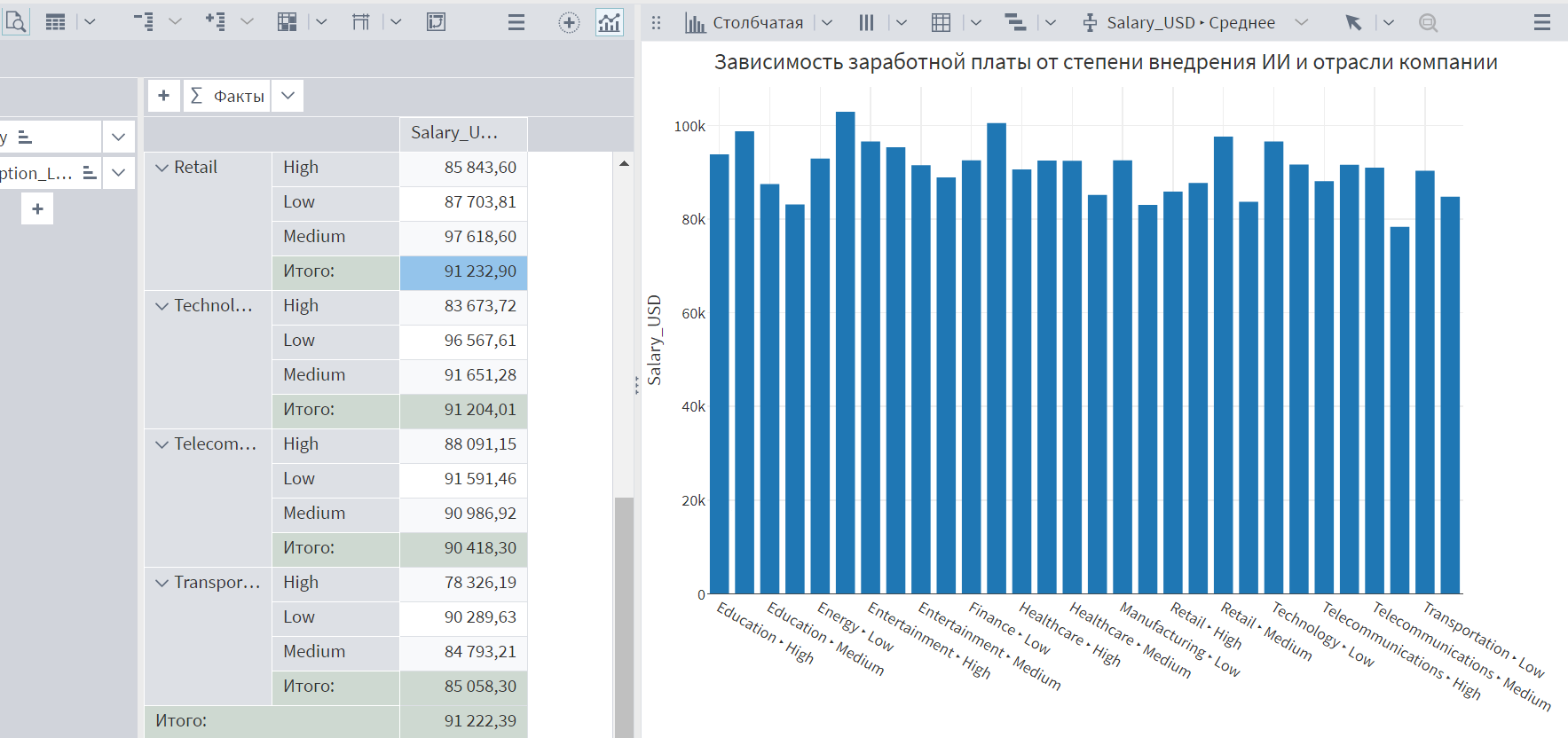
В данной работе мы исследуем взаимосвязь между внедрением ИИ и диапазонами заработных плат на разных должностях и в разных местах. А также посмотрим как заработная плата зависит от других характеристик, приведенных в нашем наборе данных.

**Основная часть работы:**

В Loginom производится очистка данных и их подготовка для дальнейшего использования (изменение форматов на верные), проверка на наличие дубликатов и противоречий (в нашем наборе данных они отсутствуют), после этого мы можем приступить к анализу данных.

Далее в Loginom производится исследовательский анализ данных:

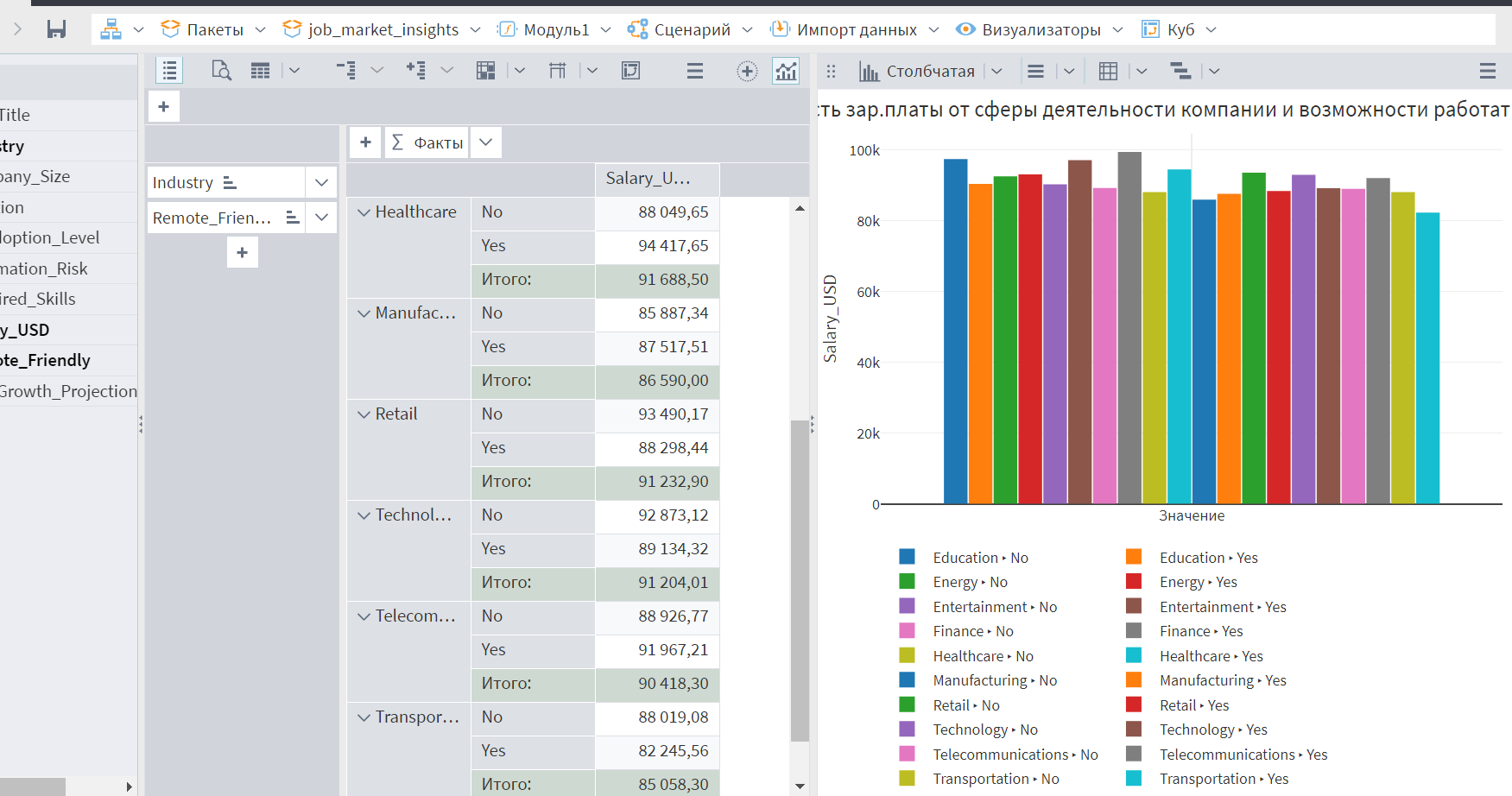
По данной диаграмме мы видим, что наибольшая годовая заработная плата, предлагается сотрудникам компаний в сфере энергетики со средней степенью внедрения искусственного интеллекта в деятельность компании, а наименьшую годовую заработную плату предлагают сотрудникам транспортных компаний с высокой степенью внедрения искусственного интеллекта в деятельность компании.



Так же исследуем зависимость заработной платы от сферы деятельности компании и возможности работать удаленно:

- Самая высокая среднегодовая заработная плата с возможностью работать удаленно в компаниях в сфере финансов.

- Самая высокая среднегодовая заработная плата без возможности работать удаленно в компаниях в сфере образования.

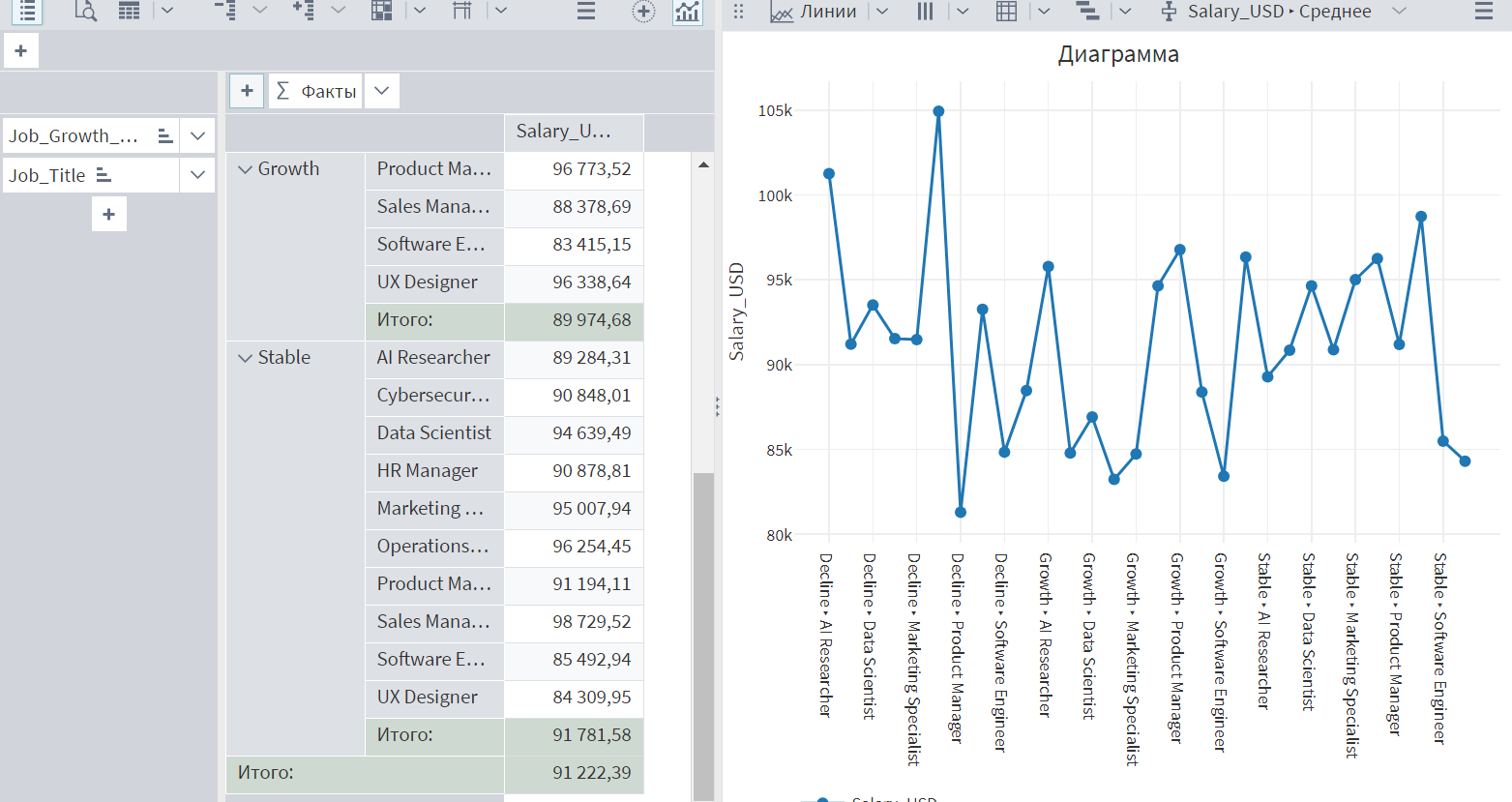


Теперь рассмотрим уровень заработной платы в зависимости от должности и прогнозируемого роста или спада должностной роли в течение следующих пяти лет

- Самую высокую заработную плату предлагают на должности операционного менеджера, но также роль этой должности пойдет на спад в течение 5 лет.

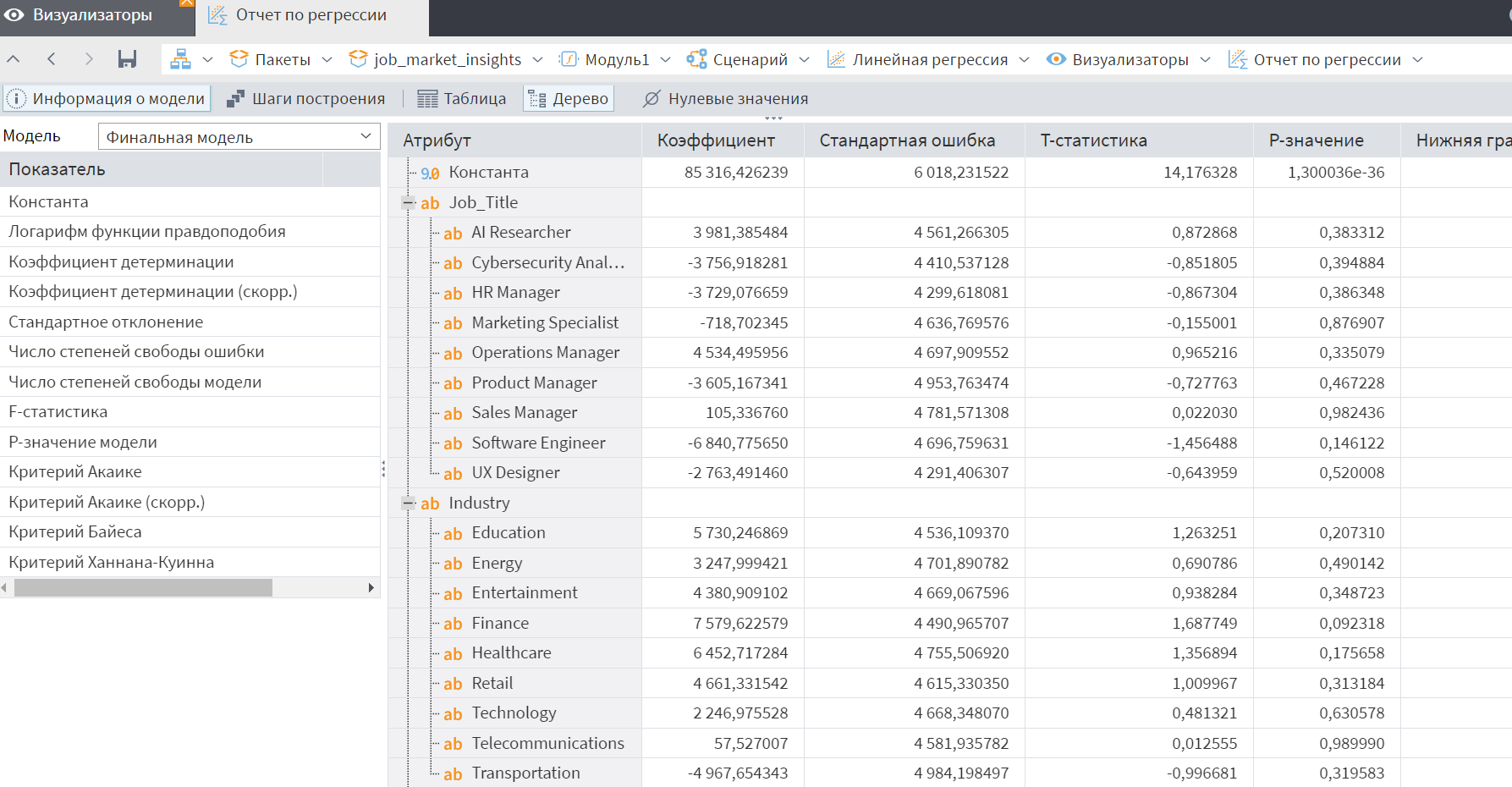
- Если рассматривать должности с прогнозируемым ростом их роли, то самую высокую заработную предлагают продуктовому менеджеру.

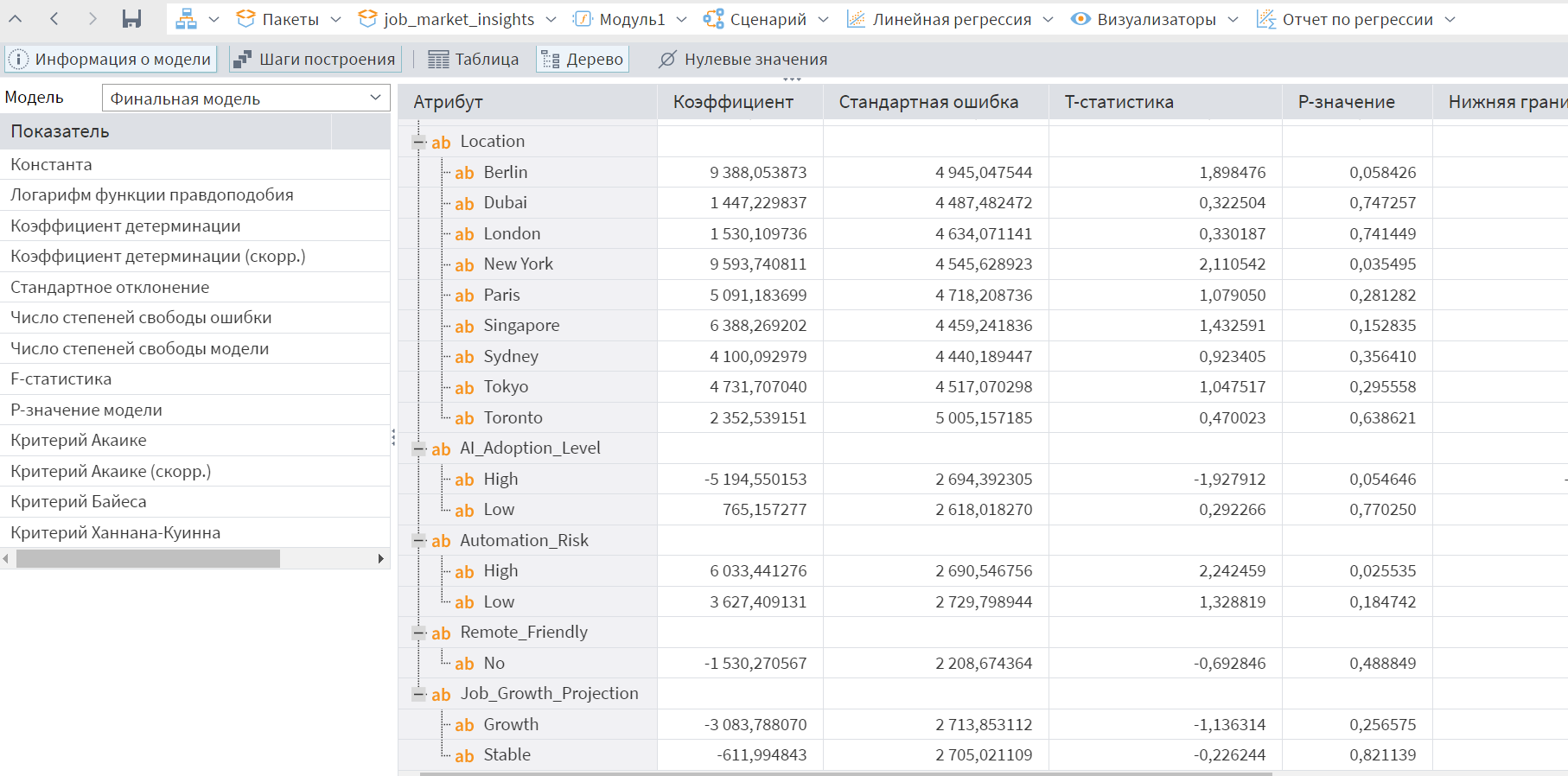
- Стабильную роль в течение 5 лет будет занимать должность менеджера по продажам с самой высокой заработной платой в этой категории.



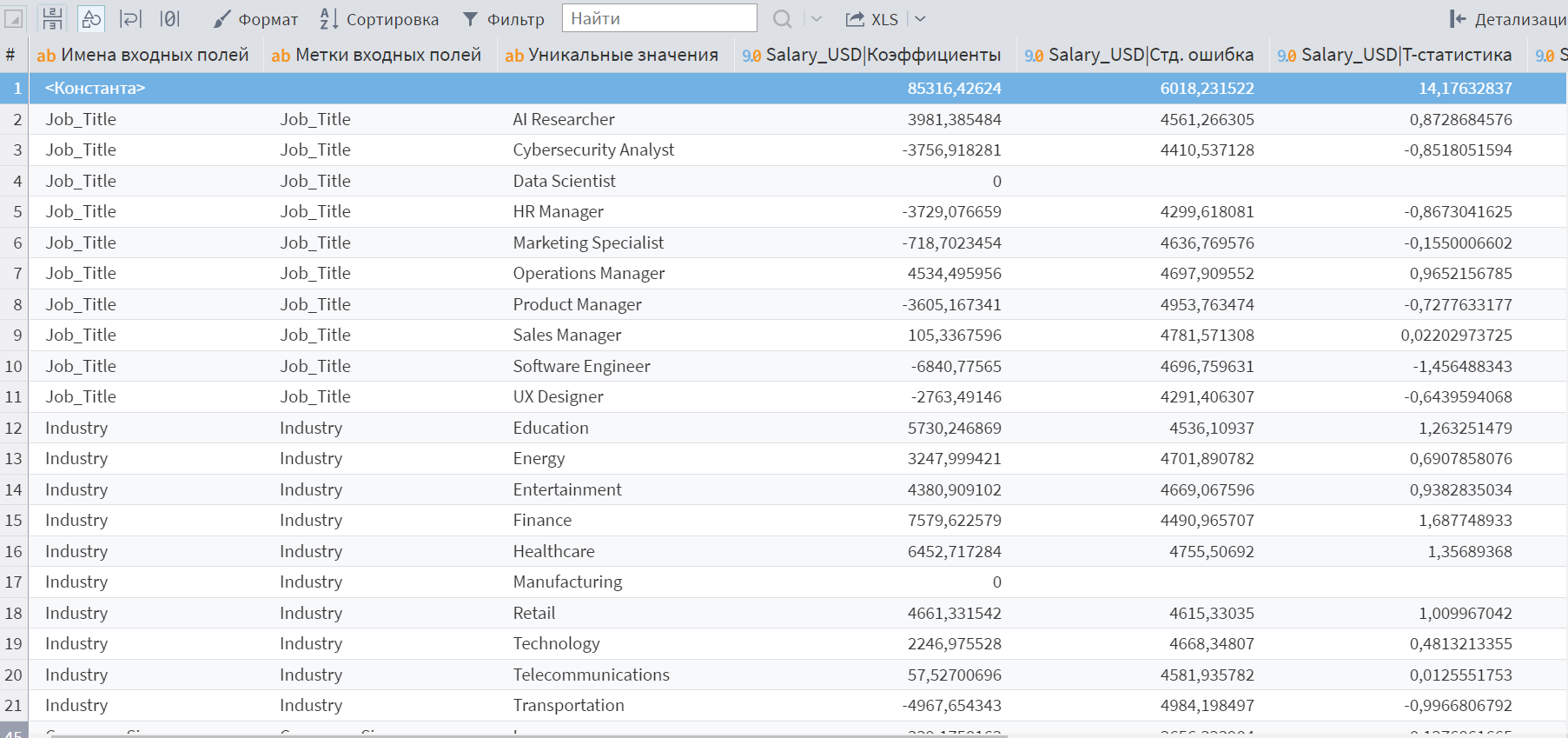
В Loginom была произведена линейная регрессия, искомый признак линейной регрессии – заработная плата (Salary\_USD).

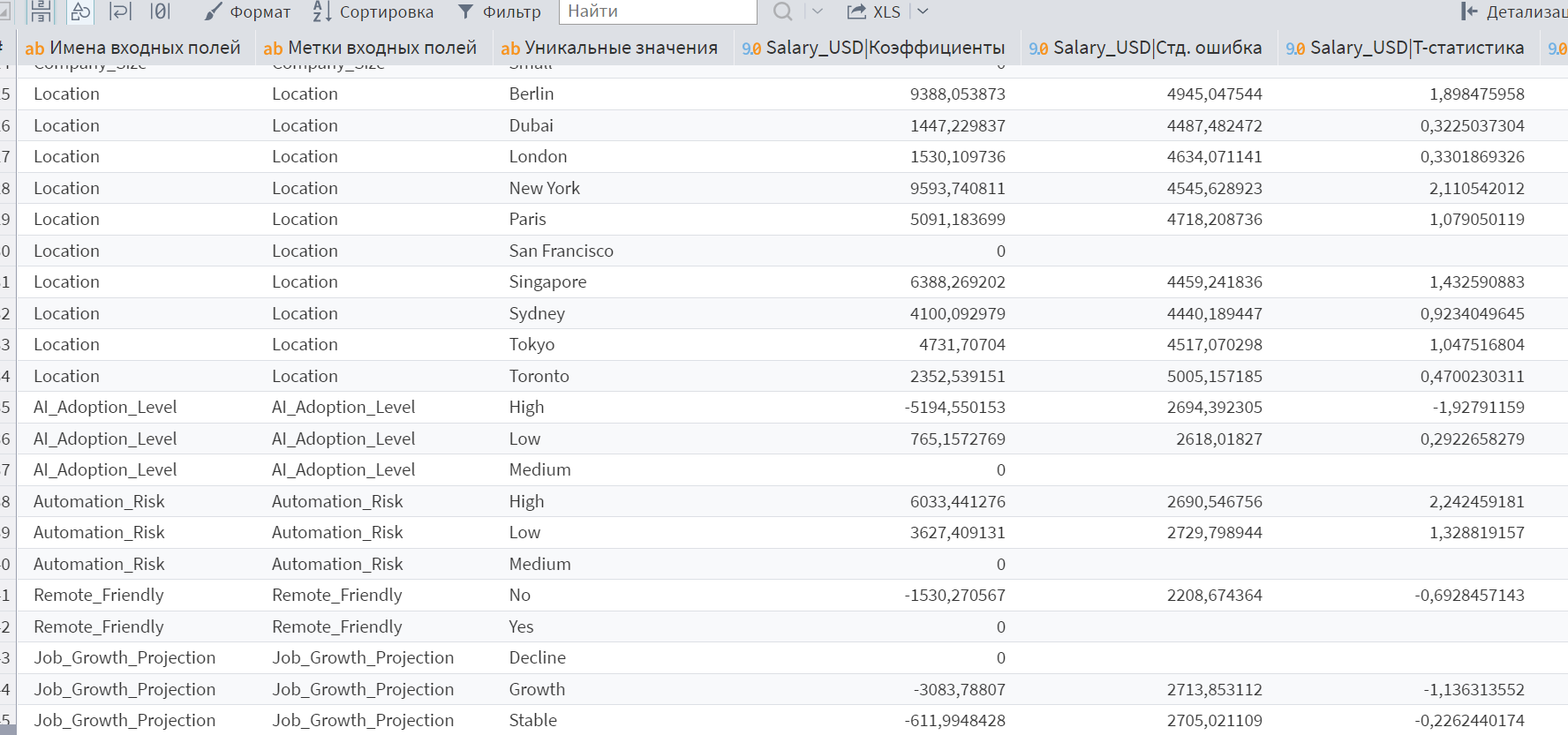
Результаты (отчет по регрессии) приведены ниже:

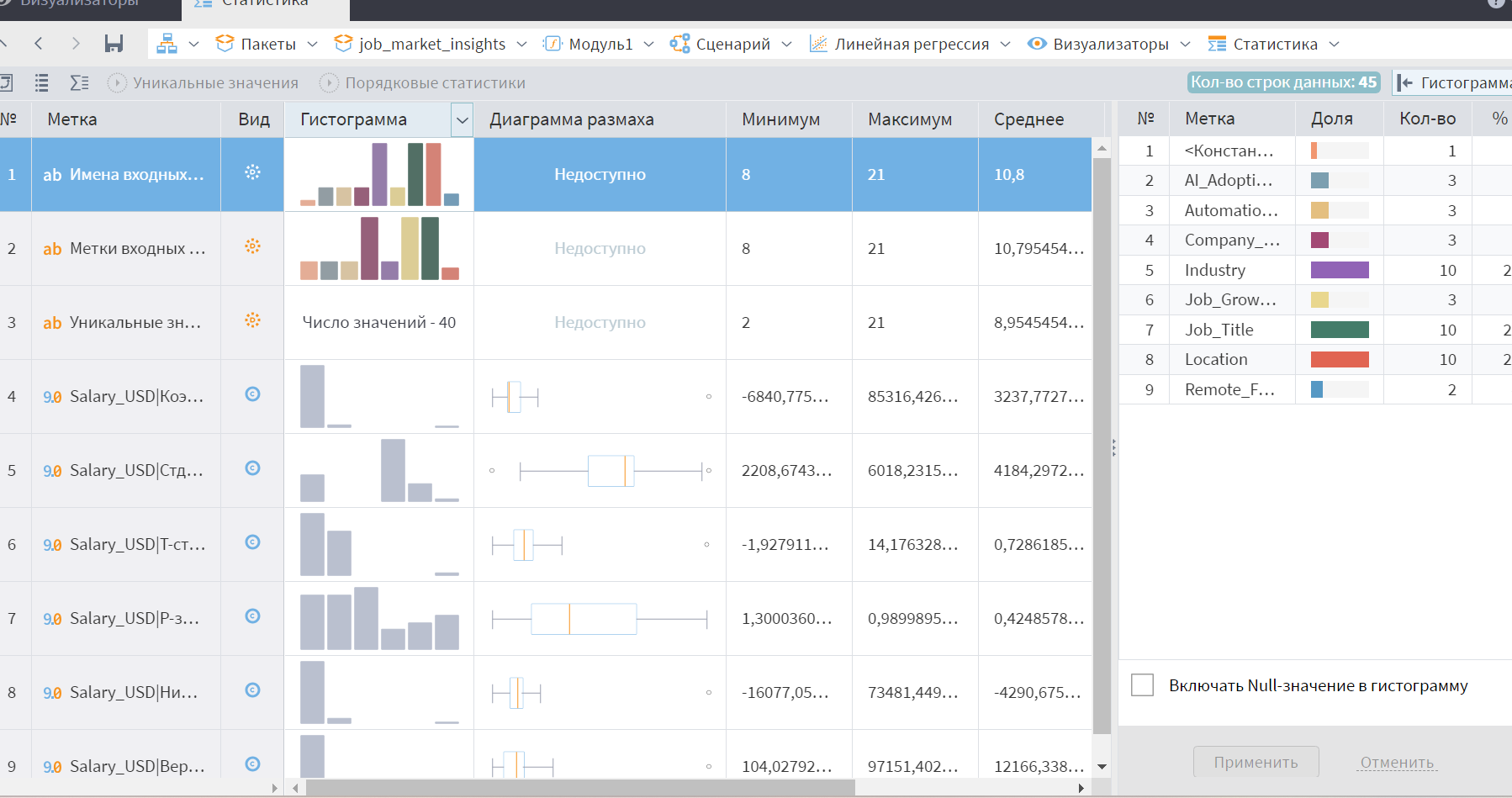




Коэффициенты регрессионной модели:



  
Статистика по регрессии:

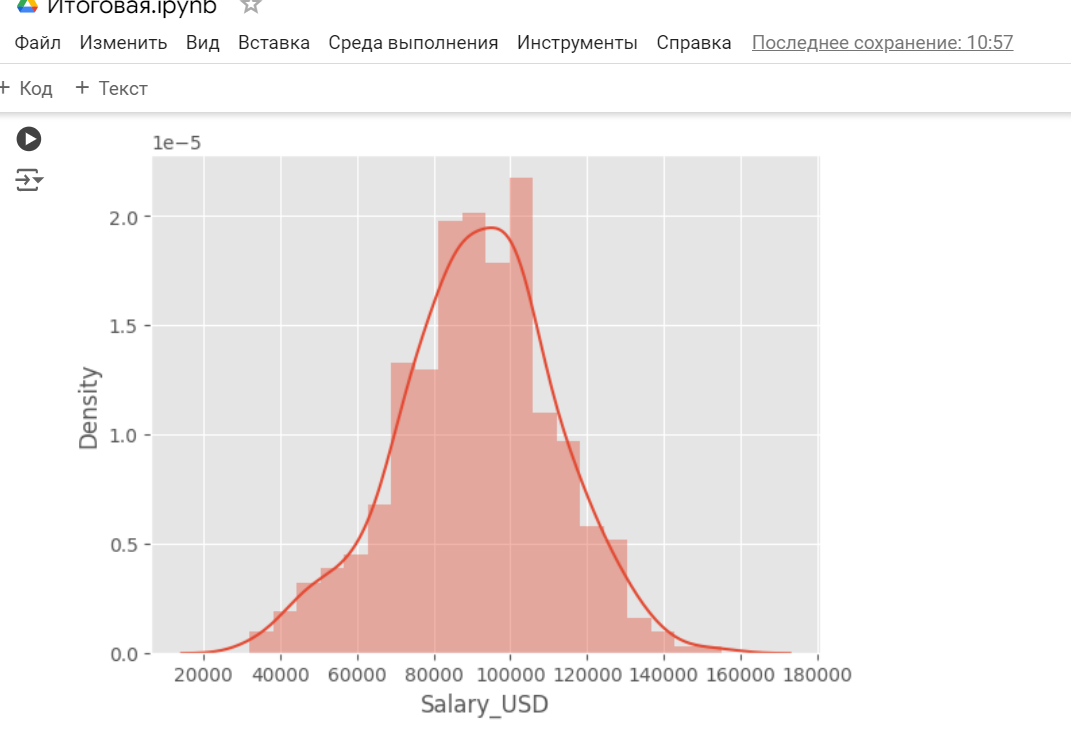


Можно сделать вывод, что на заработную плату влияет в основном Предполагаемый риск автоматизации (Automation\_risk), Location (местоположение) и Job title (профессия)

**Работа в Colaboratory**

Далее проведем работу в Colaboratory (ссылка на файл находится выше):

- Построим график плотности по заработным платам и увидим, что большая часть находится в диапазоне от 80000 до 110000 USD/год.



Далее посмотрим на график зависимости заработных плат от степени внедрения Искусственного интеллекта в деятельность компании.

****

Можно заметить, что заработная плата выше в тех компаниях, где степень внедрения ИИ низкая (в высокой степени внедрения ИИ есть выбросы(экстремально высокие значения)), так же видно, что прогнозируемая должностная роль в таких компаниях в большинстве остается стабильна.



На данной диаграмме мы можем видеть зависимость заработных плат от требуемых навыков, возможности работать удаленно, локации и степени внедрения искусственного интеллекта в деятельность компании. Допустим, с навыком UX/UI дизайн, в локации Дубай заработная плата выше в компаниях с возможностью работать удаленно и со средним внедрением ИИ в работу компании.

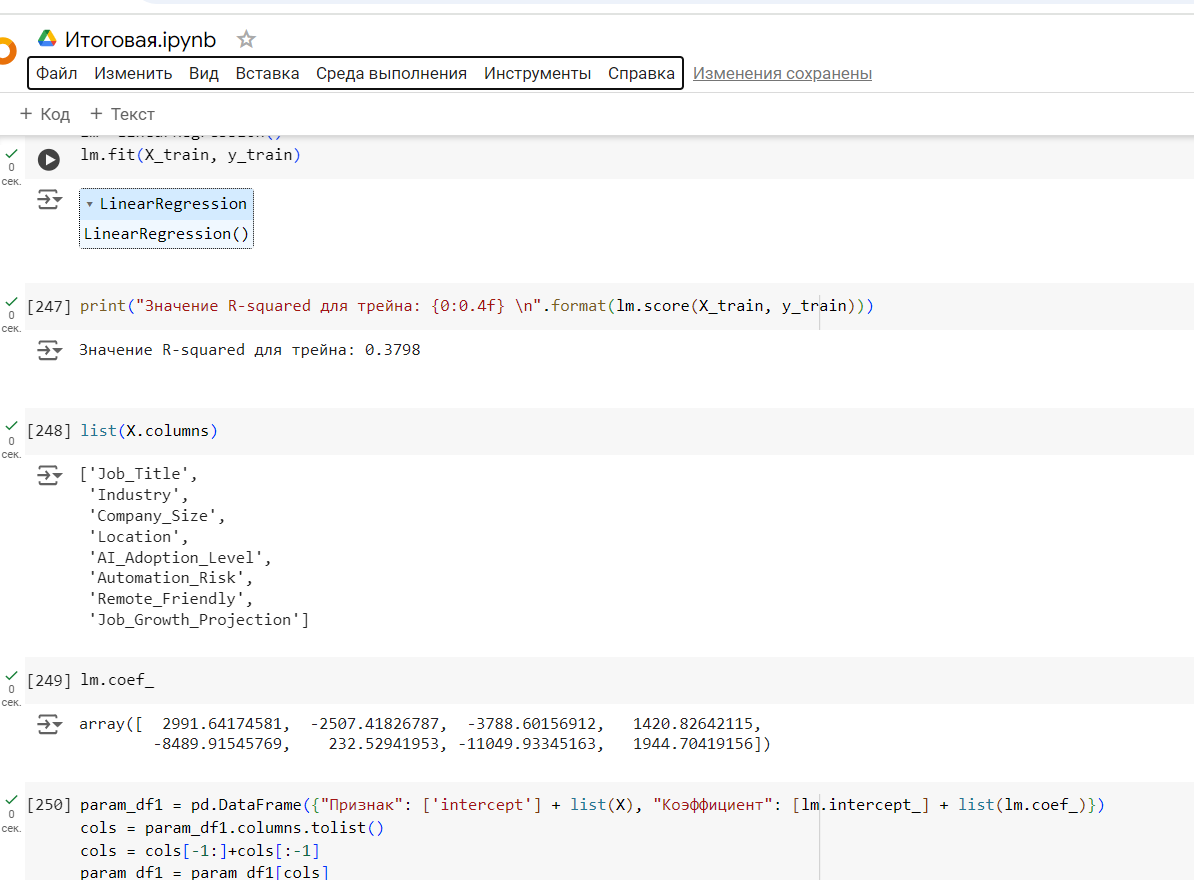
Далее приступим к машинному обучению в Colaboratory:

Первой моделью будет Линейная регрессия:

Искомый признак – заработная плата.

Для того, чтобы преобразовать категориальный тип данных в числовой воспользуемся технологией one-hot encoding.

После обучения модели видим, что результат точности модели составляет почти 38%.



Далее посмотрим какие признаки влияют на уровень заработной платы больше всего:



Так же как и в Loginom видно, что большее влияние оказывают такие признаки как:

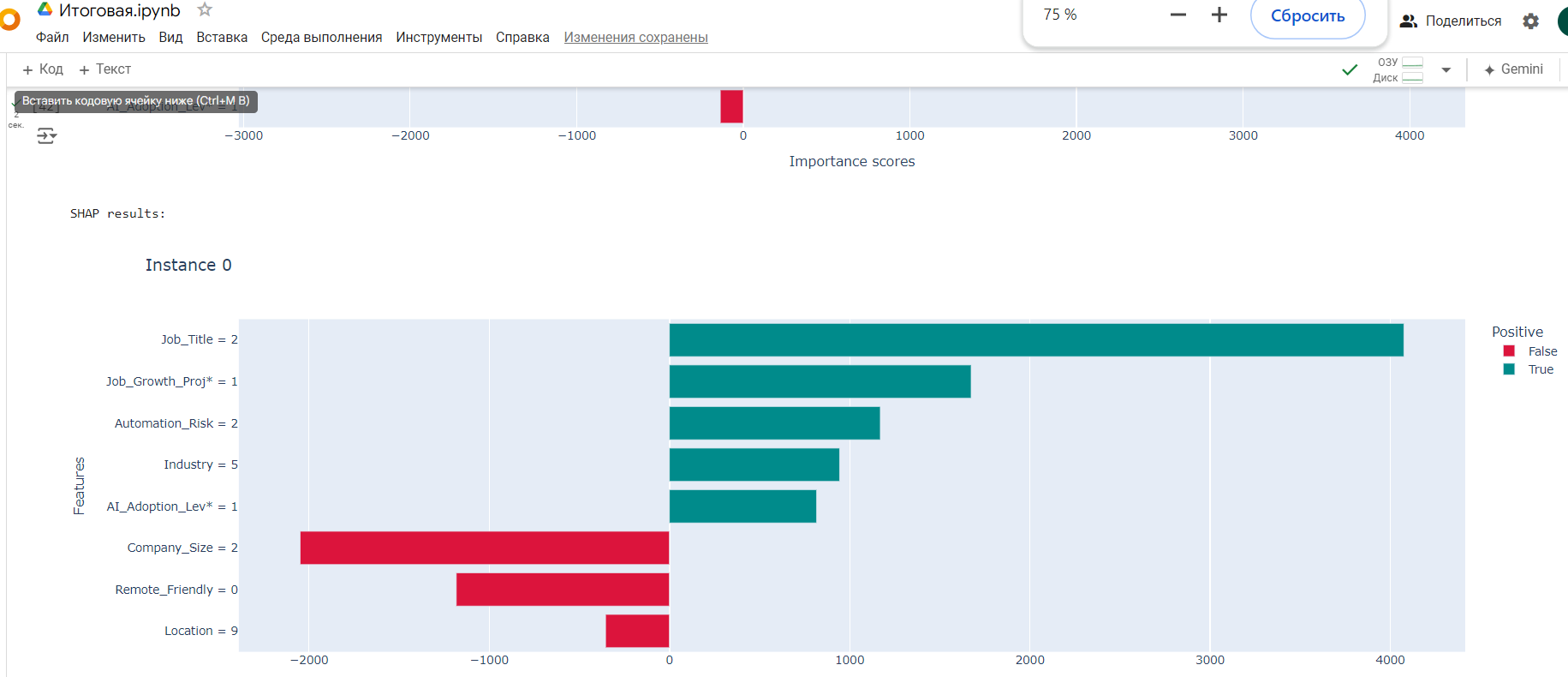
Предполагаемый риск автоматизации (Automation\_risk), Location (местоположение) и Job title (профессия).

Далее построим модель Random Forest:



Данная модель оказалась точнее Линейной регрессии, ее точность 53,54%.

Далее используем библиотеку OmniXAI:

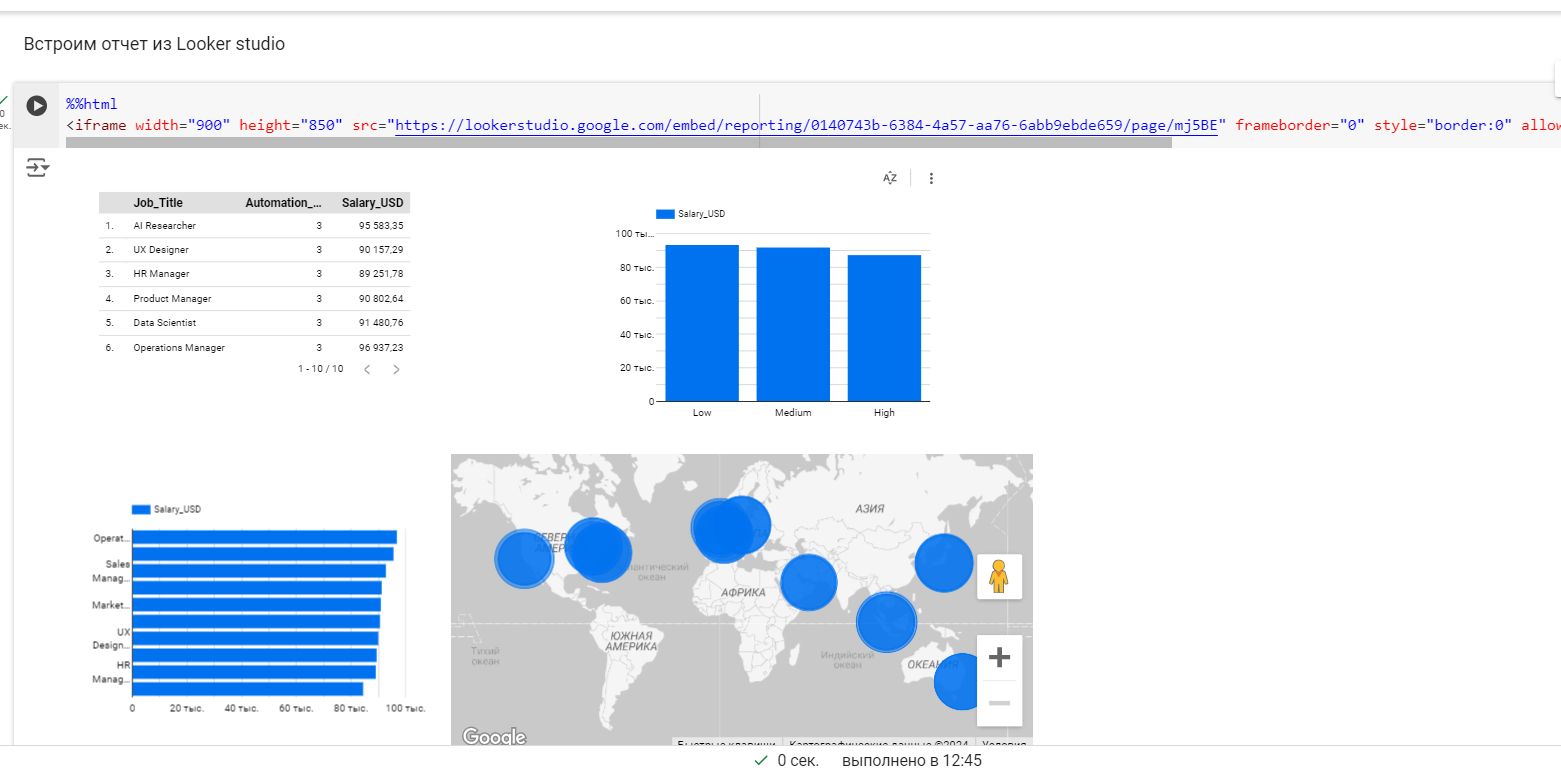






На основании данных из этой библиотеки мы так же видим влияние двух признаков Automation Risk и Job title, а также Job\_Growth\_Projection (данный признак влиял в модели линейной регрессии, но не в Loginom).

Встроим отчет из Looker studio:

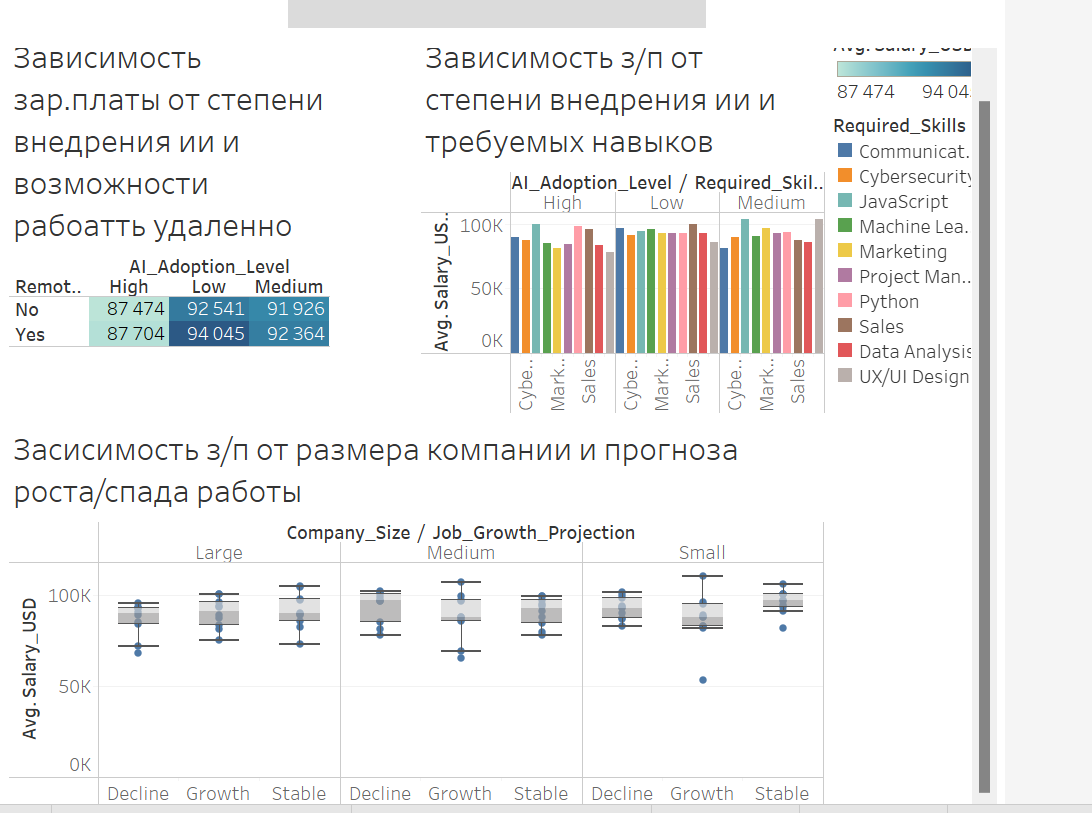


На данном отчете можно увидеть, что самыми высокооплачиваемыми должностями являются – Data Scientist и UX Designer, далее идет AI Researcher, так же в компаниях с низким уровнем внедрения Искусственного интеллекта зарплата выше, что было видно и на графике в Colaboratory.

**Tableau**

Мною были построены дашборды в Табло, рассмотрим их:

Дашборд1:



Можно сделать некоторые выводы:

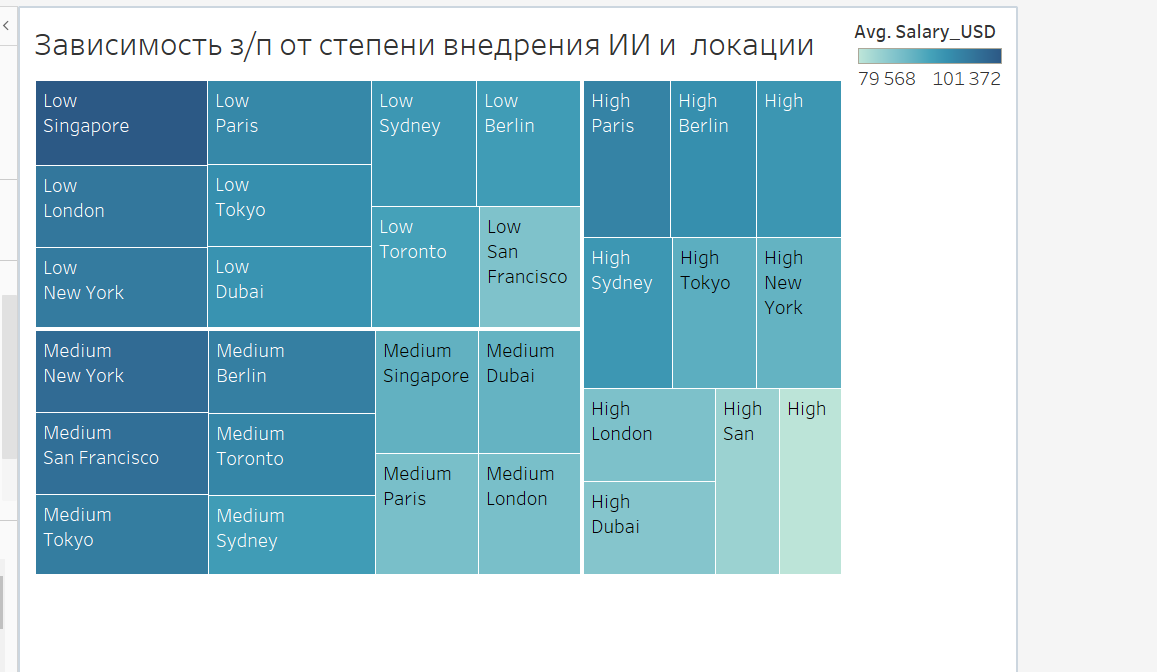
- В компаниях с низкой степенью внедрения Искусственного интеллекта и без возможности работать удаленно самые высокие заработные платы.

- В компаниях с высокой и средней степенью внедрения ИИ возможность работать удаленно не сказывается в такой мере.

- Навыки совершенно по-разному оплачиваются в разных группах компаний (по степени внедрения ИИ), допустим навык Java Scripts больше всего оплачивается (103270 USD) в компании со средней степенью внедрения ИИ, а меньше всего с малой степенью внедрения ИИ в деятельность компании (94178 USD).

- В маленьких компаниях должности, роль которых будет расти в течение 5 лет имеют заработную плату меньше, чем в средних и крупных компаниях.

Далее рассмотри Дашборд 2:

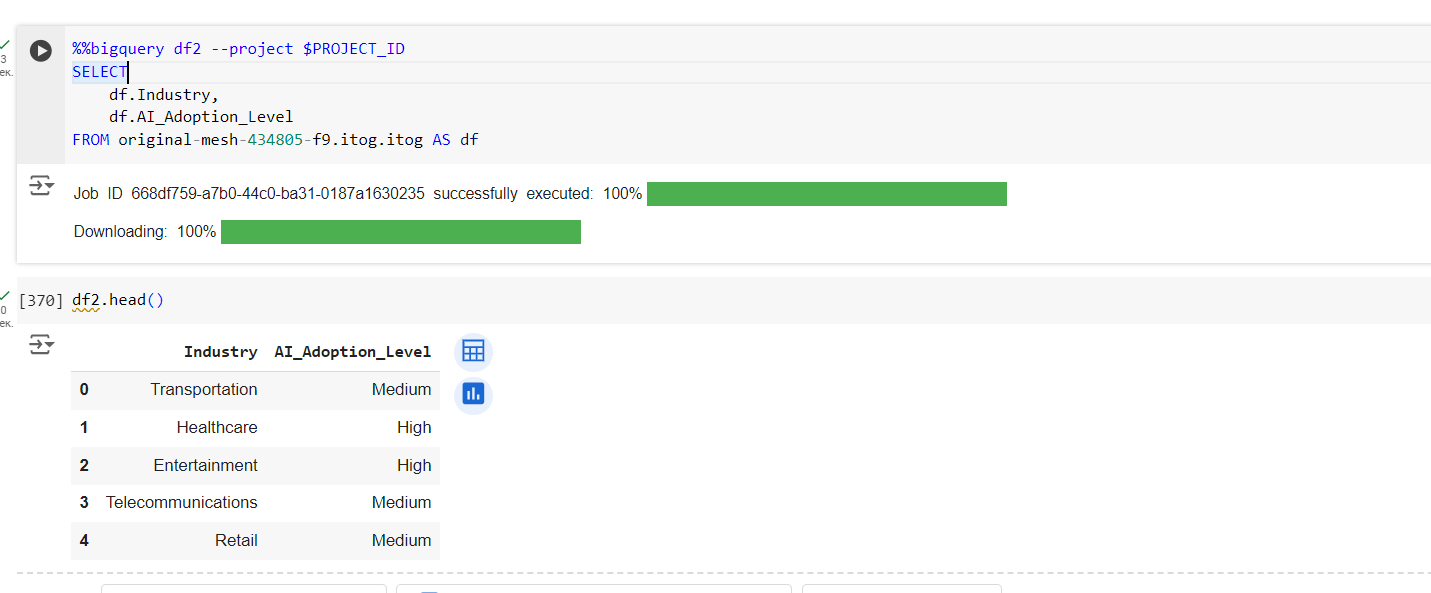


- Самые высокие заработные платы в компаниях с низким внедрением ИИ в Сингапуре, а так же со средним в Нью-Йорке и Сан-Франциско.

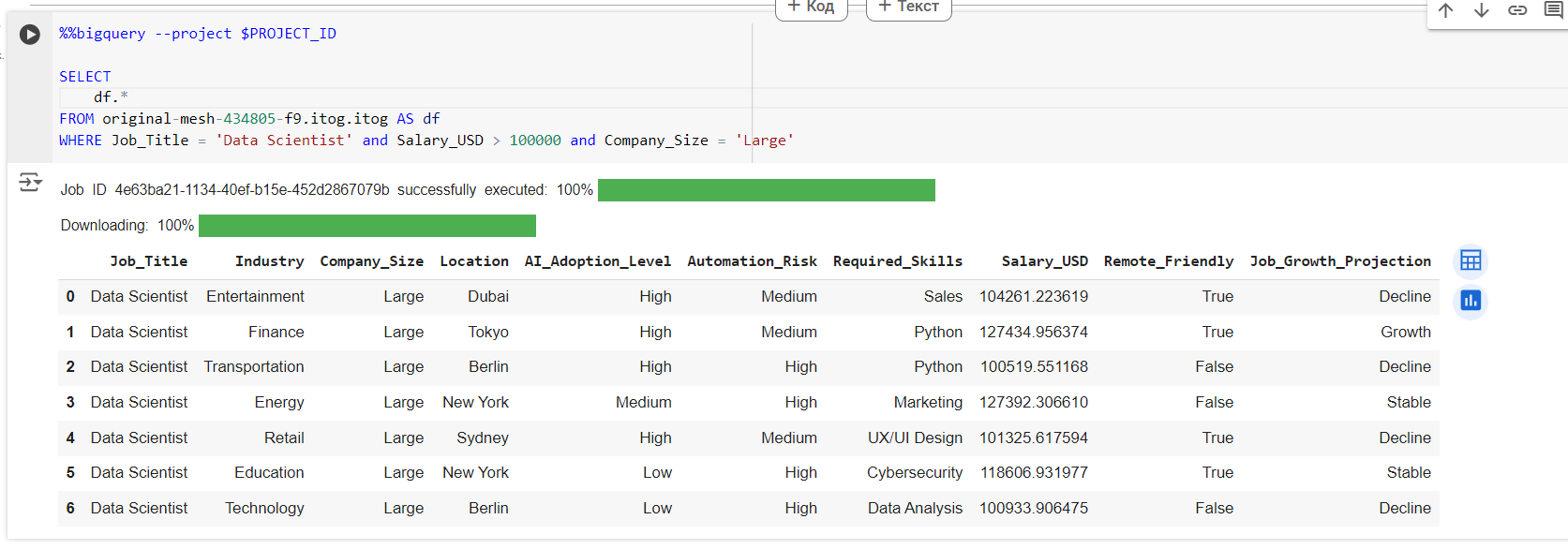
- Если рассматривать компании с высокой степенью внедрения ИИ с свою деятельность, то они находятся в Берлине и Париже.

**BigQuery и запросы на sql.**

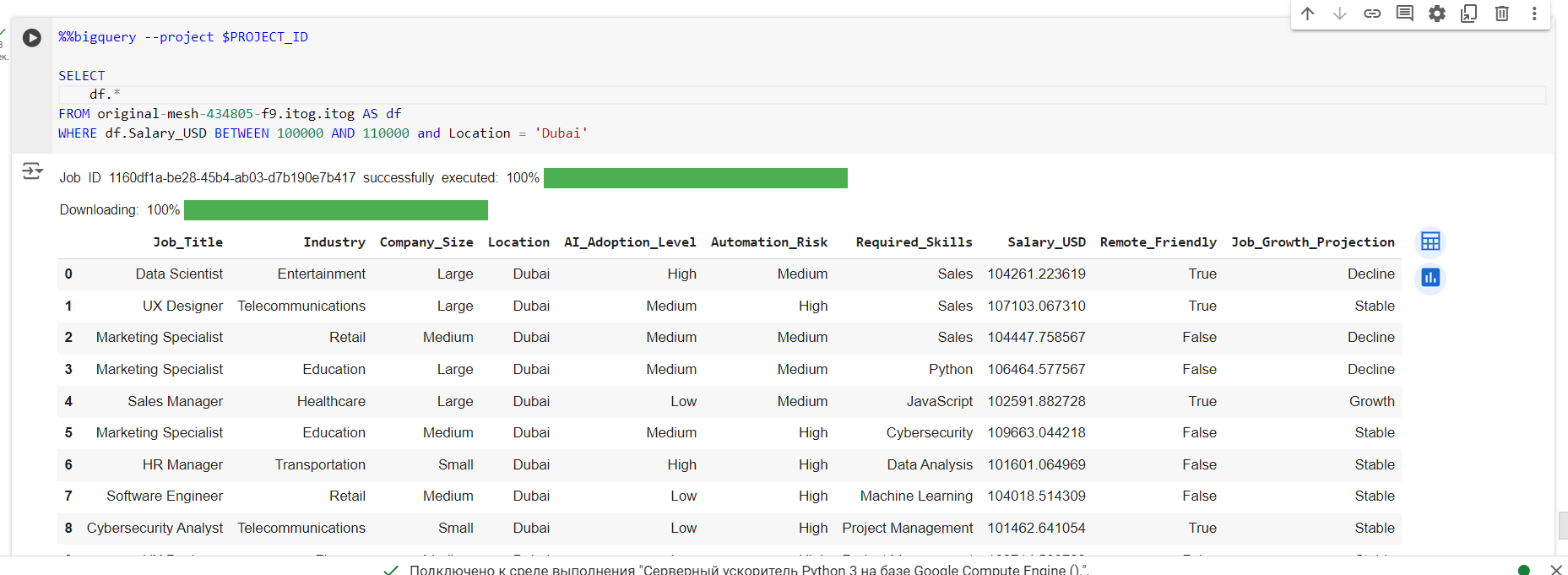
Создадим новый датафрейм df2 состоящий из Industry и AI\_Adoption\_Level и выведем его.



Выберем в датасете профессии Data scientist, с заработной платой больше 100000 usd и с работой в большой компании.



Выберем профессии с заработной платой между 100000 и 110000 usd и с локацией в Дубае.



**Заключение:**

В данной работе мною был проведен анализ, с применением моделей машинного обучения, заработных плат в зависимости от различных факторов. Были сделаны основные выводы зависимостей, а так же приведены их графики, обучены модели Линейной регрессии, Random Forest, использовалась библиотека OmniXai, была проведена работа в Tableau и Looker Studio, выполнено подключение к BigQuery и выполнено несколько запросов.