Введение

Файл «layoffs.csv» – это база данных, содержащая информацию об увольнениях в мире в период с 2020 по 2023 годы.

Цель данного проекта – произвести очистку данных (Data Cleaning) и исследовательский анализ данных (Exploratory Data Analysis).

Процесс «Data Cleaning» включает в себя удаление или исправление ошибок, дубликатов, неполных данных и других проблемных аспектов в наборе данных для обеспечения их точности и целостности перед анализом.

Процесс «Exploratory Data Analysis» (EDA) — это процесс анализа и изучения набора данных для понимания их основных характеристик. Цель EDA состоит в том, чтобы выявить закономерности, тенденции, аномалии, отношения между переменными и другие интересные аспекты данных, которые могут быть полезны для дальнейшего анализа.

Метаданные

Набор данных представляет собой файл в формате CSV и представлен следующими столбцами:

- company (text) – компания
- country (text) - страна
- date (date) – дата
- funds_raised_millions (int) – привлеченные средства в миллионах
- industry (text) – отрасль
- location (text) - местонахождение
- percentage_laid_off (text) –процент уволенных
- stage (text) – стадия развития компании
- total_laid_off (int) – общее число уволенных

Файл содержит 2361 строку и 9 столбцов. Все операции были выполнены в MySQL Workbench. Все файлы доступны по ссылке: enupilov/WorldLayoffs (github.com).

Очистка данных (Data Cleaning)

Очистка данных включает в себя 4 этапа:

1. Удаление дубликатов (Removing Duplicates);
2. Стандартизация данных (Standartizing the Data);
3. Проверка пустых значений (Null Values or Blank Values);
4. Удаление столбцов или строк, которые не требуются для анализа (Removing Any Columns or Rows).

1. Удаление дубликатов (Removing Duplicates)

Перед началом очистки данных была создана копия данных в виде таблицы layoffs_staging для избежания случайного изменения или удаления данных.

Для нахождения дублирующихся строк было решено использовать общую таблицу выражений (Common Table Expression, CTE) для поиска дубликатов в таблице layoffs_staging на основе столбцов (company, location, industry, total_laid_off, percentage_laid_off, date, stage, country, funds_raised_millions).

Далее была создана таблица layoffs_staging2, в которой был создан столбец row_num для отображения числа повторений определенной строки. В случае отображения числа больше 1 означает, что данная строка повторяется в базе данных.

Таким образом, было удалено 22 строки.

2. **Стандартизация данных (Standartizing the Data)**

При выводе значений было отмечено наличие ненужных пробелов в столбце company. Соответственно, было решено избавиться от ненужных пробелов с помощью TRIM().

К тому было отмечено, что отрасль Crypto имеет наименований. Для единого формата все отрасли, начинающиеся на Crypto были переименованы в Crypto.

В столбце country страна United States была записана либо с точкой на конце, либо без точки. Соответственно, значения, где на конце имелась точка были заменены на значения без точки.

Формат даты столбца date был изменен с формата 8/29/2022 на формат 2022-08-29. Также был изменен тип данных столбца date на DATE.

3. **Проверка пустых значений (Null Values or Blank Values)**

При выводе значений было отмечено наличие пустых строк в столбце industry. При анализе было отмечено, что по названию в одной записи данные столбца industry имеются, а в другой записи нет. Чтобы исправить пустые значения был использован оператор JOIN, чтобы связать между собой две копии и заменить значения в одной таблице значениями другой.

Некоторые пустые значения невозможно заполнить, так как они не имеют других записей как в первом варианте.

4. **Удаление столбцов или строк, которые не требуются для анализа (Removing Any Columns or Rows).**

Также были обнаружены строки, которые не имеют чисел в столбцах total_laid_off и percentage_laid_off, а значит они не дадут нам никакой информации. Соответственно, было решено их удалить.

Исследовательский анализ данных (Exploratory Data Analysis)

Для беглого анализа было интересно узнать наибольшее количество увольнений, а именно в 1 день. Оказалось, что максимальное количество увольнений составило 12000 человек, что несомненно очень много. Наибольший процент же составляет 1 % от общего количества сотрудников.

MAX(total_laid_off)	MAX(percentage_laid_off)
12000	1

Если рассматривать увольнения по годам, то можно заметить, как сильно увеличилось количество увольнений за 2022 и 2023. Самым стабильным выдался 2021 год.

YEAR(`date`)	SUM(total_laid_off)
2023	125677
2022	160322
2021	15823
2020	80998

В контексте страны лидерами по числу увольнений являются США, а также Индия и Нидерланды. Данные страны как минимум 2 раза являлись лидерами по увольнениям в период с 2020 по 2023 годы.

country	years	total_laid_off	ranking
United States	2020	50385	1
India	2020	12932	2
Netherlands	2020	4600	3
United States	2021	9470	1
India	2021	4080	2
China	2021	1800	3
United States	2022	106381	1
India	2022	14024	2
Netherlands	2022	5120	3
United States	2023	89684	1
Sweden	2023	9100	2
Netherlands	2023	7500	3

В контексте отрасли последние 2 года в топ 5 попадают такие сегменты как Потребительский сегмент (Consumer), здоровье (Healhcare).

industry	years	total_laid_off	ranking
Transportation	2020	14656	1
Travel	2020	13983	2
Finance	2020	8624	3
Retail	2020	8002	4
Food	2020	6218	5
Consumer	2021	3600	1
Real Estate	2021	2900	2
Food	2021	2644	3
Construction	2021	2434	4
Education	2021	1943	5
Retail	2022	20914	1
Consumer	2022	19856	2
Transportation	2022	15027	3
Healthcare	2022	14999	4
Finance	2022	12684	5
Other	2023	28512	1
Consumer	2023	15663	2
Retail	2023	13609	3
Hardware	2023	13223	4
Healthcare	2023	9770	5

В контексте компаний последние 2 года в топ 5 находится компания Amazon. Примечательно, что компании в основном принадлежат США.

company	country	years	total_laid_off	ranking
Uber	United States	2020	6700	1
Booking.com	Netherlands	2020	4375	2
Groupon	United States	2020	2800	3
Swiggy	India	2020	2250	4
Airbnb	United States	2020	1900	5
Katerra	United States	2021	2434	1
Zillow	United States	2021	2000	2
Instacart	United States	2021	1877	3
Bytedance	India	2021	1800	4
Bytedance	China	2021	1800	4
WhiteHat Jr	India	2021	1800	4
Better.com	United States	2021	900	5
Meta	United States	2022	11000	1
Amazon	United States	2022	10150	2
Cisco	United States	2022	4100	3
Peloton	United States	2022	4084	4
Carvana	United States	2022	4000	5
Philips	Netherlands	2022	4000	5
Google	United States	2023	12000	1
Microsoft	United States	2023	10000	2
Ericsson	Sweden	2023	8500	3
Amazon	United States	2023	8000	4
Salesforce	United States	2023	8000	4
Dell	United States	2023	6650	5

Как уже ранее сообщалось, США являются лидерами по увольнениям среди стран.

В рамках отрасли наибольшее количество увольнений в США приходится на потребительский сегмент () и розничную торговлю (Retail).

industry	years	total_laid_off	ranking
Transportation	2020	10262	1
Retail	2020	6808	2
Consumer	2020	5482	3
Travel	2020	4317	4
Real Estate	2020	2847	5
Real Estate	2021	2900	1
Construction	2021	2434	2
Food	2021	2057	3
Retail	2021	1043	4
Other	2021	515	5
Consumer	2022	17775	1
Retail	2022	13522	2
Transportation	2022	9269	3
Healthcare	2022	8619	4

Real Estate	2022	8496	5
Other	2023	15757	1
Consumer	2023	14823	2
Retail	2023	12217	3
Hardware	2023	12193	4
Sales	2023	9271	5

Заключение

Таким образом, можно отметить, что США, Индия и Нидерланды являются странами с наибольшими увольнениями. Потребительский сегмент и Розничная торговля отличаются большим количеством увольнений. Последние 2 года являются кризисными для работников. Стоит отметить, что в 2023 году увольнений все же меньше, чем в 2024 году, что дает нам надежду на позитивный 2024 год.

Данный анализ не является полным, а проведен с целью беглого исследования данных с помощью SQL.

Обнаруженные идет могут помочь принять решение выбора компании для трудоустройства, страны для проживания, отрасли для создания бизнеса.