**Проект 3: Role Type**

***Цели:***

1. Проанализировать и трансформировать данные
2. Удалить дубликаты и выбрать нужные столбцы для визуализации данных
3. Создать Dashboard с нужной информацией

Для этого нужно использовать Power Query + PivotTable

Шаги для выполнения данного проетка в ***Power Query:***

1. Есть csv файл “data\_Uncleaned\_DS\_jobs” с данными role type и зарплатой с которыми я буду работать
2. Создаю excel файл “Role Type Project”
3. Захожу в power query(data->get data->launch power query editor) и добавляю csv файл “data\_Uncleaned\_DS\_jobs” (new source->file->text/csv)
4. В вкладке View ставлю галочки: Column quality, Column distribution, Column profile для более понятного понимания какие столбцы нужно трансформировать
5. Так как много столбцов, выбираю только самые нужные: Job Title, Salary Estimate, Location, Size(Choose Columns)
6. Выбираю столбец Salary Estimate и нажимаю keep duplicates
7. Так как столбец Salary Estimate имеет ненужную информацию “(Glassdoor est.)” удаляю ее(Transform->Extract->Text Before Delimiter->”(“ и нажимаю OK)

И так как имеет промежутки а не конкретные значения, разделяю на 2 столбца Min Sal и Max Sal(Add Column->Column from Examples->From Selection и вписываю в 1 ряд значение 137, excel сам поймет что я хочу и следующие значения с левой стороны чтоб он вписал дальше, меняя название столбца Min Sal и проделываю тоже самое называя 2 столбец Max Sal

1. Столбец Salary Estimate убираю из списка выбранных(Choose Columns)
2. Так как в столбце Job Title много названий должностей, переделываю все под 5 категорий(Data Scientist, Data Analyst, Data Engineer, Machine Learning Engineer, Other) используя DAX формулу if Text.Contains

(Add Column->Custom Column), название стольбца Role Type выбрав столбец Job Title

= if Text.Contains([Job Title], "Data Scientist") then "Data Scientist"

else if Text.Contains([Job Title], "Data Analyst") then "Data Analyst"

else if Text.Contains([Job Title], "Data Engineer") then "Data Engineer"

else if Text.Contains([Job Title], "Machine Learning Engineer") then "Machine Learning Engineer"

else "Other"

Нужно поменять тип данных на TEXT чтоб увидеть что есть только 5 категорий

1. Столбец Job Title убираю из списка выбранных(Choose Columns) чтоб оставить только Role Type
2. Вижу что в столбце Size есть ненужная информация(-1 и unknown) и убираю их из фильтра, осталось 628 рядов из 672
3. Вижу что в столбце Location есть штат и аббревиатура его, поэтому разделяю столбец на 2 столбца(Transform->Split Column->By Delimiter->Comma).
4. После чего вижу что в столбце Location.2 есть null значение и Anne Arundel из-за того что они не имеют в конце delimiter comma(“,”).

Поэтому выбираю в фильтре Location.2 null и Anne Arundel значения для того чтобы понять в каких штатах Location.1 эти данные находятся и вижу что они в штатах(California, New Jersey, Remote, Texas, United States, Utah). Для того чтоб убрать эти null и Anne Arundel значения надо сделать новый столбец и в случае этих штатов добавить в конце “, аббревиатура штата”, тоесть:

* California – “, CA”
* New Jersey – “, NJ”
* Texas – “, TX”
* Utah – “, UT”
* Remote и United States – “, Other”

1. Удаляю последние 3 шага с APPLIED STEPS, для того убрать null и Anne Arundel значения, добавляя новый столбец(Add Column->Custom Column), название столбца Location Correction выбрав столбец Location и в Формуле пишем:

= if [Location] = "California" then ", CA"

else if [Location] = "New Jersey" then ", NJ"

else if [Location] = "Texas" then ", TX"

else if [Location] = "Utah" then ", UT"

else if [Location] = "Remote" then ", Other"

else if [Location] = "United States" then ", Other"

else [Location]

1. Столбец Location убираю из списка выбранных(Choose Columns)
2. Проделываю еще раз шаг 12 и вижу что в появившемся после Split By Delimiter, столбце Location Correction.2 значения Anne Arundel, так как отсутствует аббревиатура штата
3. Для этого захожу чтоб узнать какая аббревиатура штата у Anne Arundel и нахожу MD. Заменяю Anne Arundel на MD благодаря Replace Values
4. Из столбца Location Correction.1 убираю blank значения из фильтра и меняю название на City, а Location Correction.2 на State abbr
5. Меняю тип данных столбцов Min Sal, Max Sal на currency и умножаю значения на 1000
6. Перепроверяю все столбцы нужно ли что-то еще сделать и вижу что все готово для создания queries чтоб ответить на вопросы:

***Вопросы:***

1. Какая средняя минимальная и средняя максимальная зарплата по должностям?

Для этого нужно создать дубликат файла data\_Uncleaned\_DS\_jobs с названием query “sal by role type dup” и выбрать столбцы min sal, max sal, role type. Группирую столбцы по role type(Transform->Group by->Advanced, выбрать группировку по role type), далее столбцы:

* allrows – нажав на таблицу покажет всех работников с выбранной должностью и зарплатой
* avg min sal – Average по столбцу Min Sal, средняя минимальная зарплата
* avg max sal – Average по столбцу Max Sal, средняя максимальная зарплата

Меняю тип данных avg min sal, avg max sal на currency. ***Query 1 готов(таблица)***

1. Какая средняя минимальная и средняя максимальная зарплата по количеству работников в компании?

Для этого нужно создать референс файла data\_Uncleaned\_DS\_jobs с названием query “sal by size” и выбрать столбцы size, min sal, max sal. Столбец size представляет собой промежуток разденный на 7 от 1 до 10000+ сотрудников: 1-50, 51-200, 201-500, 501-1000, 1001-5000, 5001-10000, 10000+

Группирую столбцы по size и далее столбцы для вывода:

* count - количество работников из разных промежутков
* avg min sal - Average по столбцу Min Sal, средняя минимальная зарплата
* avg max sal - Average по столбцу Max Sal, средняя максимальная зарплата

Меняю тип данных avg min sal, avg max sal на currency. ***Query 2 готов(таблица)***

1. Добавить новый столбец в data\_Uncleaned\_DS\_jobs с полным названием штата по аббревиатуре

Для этого скачиваю excel файл states abbr в котором есть полное название штата и его аббревиатура, и добавляю его в power query.

Далее объединяю данные благодаря Merge Queries(оператор слияния) по “State abbr” из data\_Uncleaned\_DS\_jobs и “2-letter USPS” из states abbr. Так как показывает объединение 0 из 605 то нужно проверить length аббревиатур(Add Column->Extract-Length). Показывает что там 3символа вместо 2(space лишний), поэтому удаляю последний шаг и удаляю лишний space благодаря Trim(Transform->Format->Trim)

Объединяю еще раз и показывает 605 из 605 значений, нажимаю ок. Появляется новый столбец и меняю его название на State. ***Query 3 готов***

1. Какая средняя минимальная и средняя максимальная зарплата разных штатов (полное название штата которое создали шаг ранее в data\_Uncleaned\_DS\_jobs)?

Для этого нужно создать референс файла data\_Uncleaned\_DS\_jobs с названием query “sal by state” и выбрать столбцы state, min sal, max sal.

Группирую столбцы по state и далее столбцы для вывода:

* count - количество работников из разных штатов
* avg min sal - Average по столбцу Min Sal, средняя минимальная зарплата
* avg max sal - Average по столбцу Max Sal, средняя максимальная зарплата

Меняю тип данных avg min sal, avg max sal на currency. ***Query 4 готов(таблица)***

1. Какая средняя минимальная и средняя максимальная зарплата по количеству работников и их должности?

Для этого нужно создать дубликат data\_Uncleaned\_DS\_jobs с названием query “sal by size & role type” и выбрать столбцы size, role type, min sal, max sal.

Группирую столбцы по size и role type(сразу по двум столбцам) и далее столбцы для вывода:

* count - количество работников с должностью которые входят в те 7 промежутков категорий
* avg min sal - Average по столбцу Min Sal, средняя минимальная зарплата
* avg max sal - Average по столбцу Max Sal, средняя максимальная зарплата

Меняю тип данных avg min sal, avg max sal на currency. ***Query 5 готов(таблица)***

Я создал 4 таблицы: sal by role type, sal by size, sal by state, sal by size & role type, с которыми я буду работать в power pivot

Манипулирование и трансформирование данными в power query завершено, далее нужно анализировать данные в power pivot

Шаги для выполнения данного проекта в ***Power Pivot + PivotTable***

Создаю папку в power query editor и перемещаю data\_Uncleanead\_DS\_jobs и states, так как они не нужны для анализа и визуализации данных

Нажимаю Close & Load, после чего появляются все 4 таблицы созданные ранее

* sal by size
* sal by role type
* sal by size & role type
* sal by state

1. Вижу что значения зарплат в avg min sal и avg max sal нужно изменить, для этого выделяю полностью столбец avg min sal и в вкладке Home->Number->Comma Style->Decrease Decimal(убираю нули после “,”). Проделываю так и с столбцом avg max sal во всех 4 таблицах
2. Создаю new sheet с названием PivTables где будут находиться 4 pivot таблицы
3. Захожу в Power Pivot->Manage Data Model->PivotTable->PivTables
4. Появляется меню PivotTable Fields с 4 таблицами которые я беру и создаю каждой по pivot таблице и убираю grand total

***Анализирую PivotTables:***

1. Из таблицы sal by size & role type, понимаю что в компаниях у которых от 1001-5000 и от 5001-10000 работников отсутствуют сотрудники с должностью такая как Machine Learning Engineer
2. Понимаю из таблицы sal by role type, что самая прибыльная зарплата у Data Scientist
3. Вижу из таблицы sal by state, что есть 1 человек у которого самая большая средняя зарплата 331,000
4. Вижу из таблицы sal by size, что в основном на работу берут больше Data Scientist нежели других работников исходя из количества их в компании

***Далее создаю визуализации таблиц на новом листке с названием Dashboard***