INTRODUCTION

For this project I explore the jobs in data community across the globe from the year 2020 to 2024. Data jobs has unprecedent impact in the world and has keep on taking impact year by year in world’s companies. The dataset contains figures such as work year and salary organized by experience level, job title, employee residence, work setting, company location and job category. The aim was to drive out an insight that will provide the user a simple way to understand how salary is distributed according to experience level, employee residence, work setting, company location, work year and job category.

DATA EXPLORATION

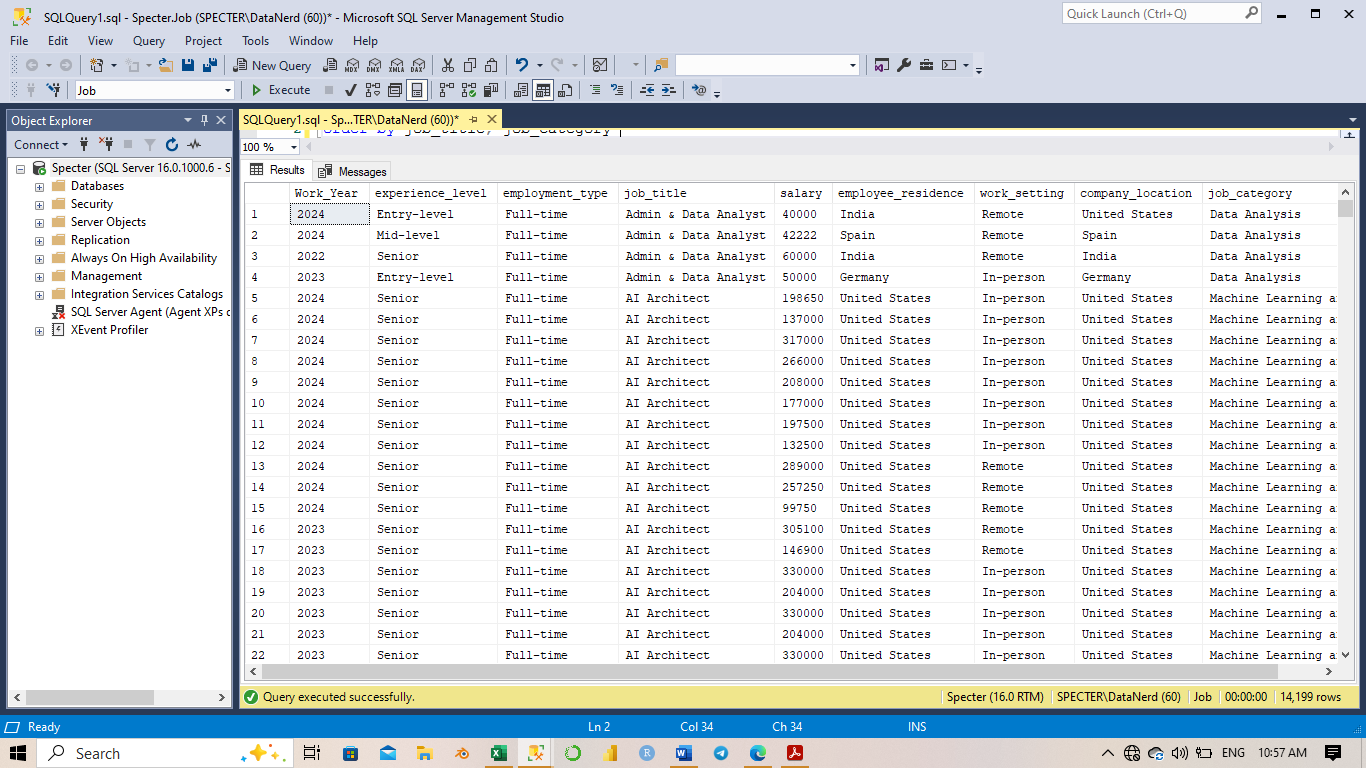
How The Data Was Collected

The dataset was collected from Kaggle Datasets. This dataset contains summary of jobs in data, from January, 2020 to December, 2024. It was downloaded into a desktop folder which was later loaded into SQL Server Management Studio (SSMS)

Feature Identify for The Analysis

This is a Data Exploration Analysis using Sql. In this Project, i use SQL to answer the questions related to dataset as well as answer specific questions about the dataset. The features identified for the analysis are total job titles, total salary, total salary by work setting, experience level, employment type and total salary by company residence. The reason for choosing these features is simply because we want to know how the salary is distributed across various different categories, companies’ residence and work setting in the world.

Screenshot of the clean Data open in SQL Server Management Studio (SSMS)



METHODS

Pre-Processing Techniques Used

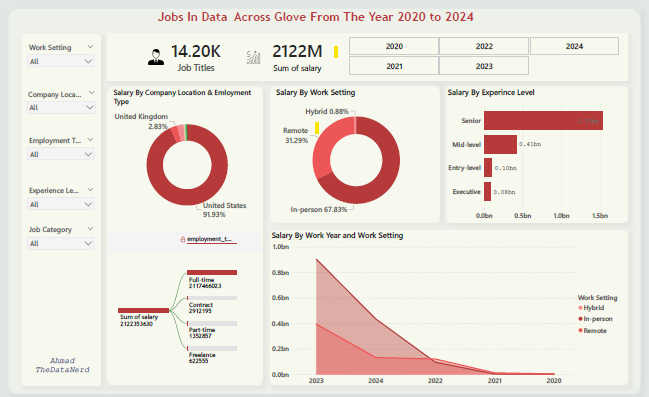
The following are some of the pre-processing techniques we carried out:

Loading the Dataset: after downloading the dataset, the first pre-processing technique was to open the dataset. This was done by SQL server Management Studio (SSMS). The dataset was downloaded and named as a xlsx file and then loaded it in to SQL server Management Studio (SSMS) for cleaning and exploratory analysis.

Understanding the Dataset: this was done by knowing the features each column stands for to avoid mistakes in data analysis, with the names of the columns, data types, the first and last few rows’ values, unique column values and statistical summary from the data dictionary.

Dataset Cleaning: the dataset cleaning was done with SQL which checked for any null value and removed Duplicates.

Visualizations Generated Using Power BI



CONCLUSION

According to the Line plot, the year with highest salary is 2023, second by 2024 and then 2022, 2021 and 2020 come with the lowest spend salary by companies. However, the senior level experience staffs have over the 70% of the whole salary, 20%. Distributed to mid-level experience staffs, meanwhile the remaining 10% is shared between Entry-level and executive Staffs. 67% of staffs work in person and 31% work remotely (from anywhere in the world) and only 1% staffs work as hybrid staffs. 91% of the companies reside in United States Of America only 2% of the Companies reside in United Kingdom and the while the rest of countries share the remaining 7%. The challenges faced in this project is clean the data in the SQL server due to complicated data sets. The initial data was too complex so I had to utilize another data set. A success of this work was that I was able to use SQL for data cleaning and Microsoft Power Bi to Generate visualization.