**Summary of steps:**

*What is the relationship between the size of stores, number of employees and revenue?*

1. To solve this question , I firstly visited the Sales.Store table within the AdventureWorks database. I noticed that all the relevant information that i needed was stored in the Demographics column of the table.
2. Upon further exploration however , I noticed that individual demographic records were stored in an xml file. I had two options to be able to access the data that i needed:
   1. to create a view that would separate data from the Demographics column into a table and then analyse.
   2. to parse the xml file for each record in python using the 'xml.etree.ElementTree' module.
3. After exploring the database, I found a view like my option (1) already existed, it was called Store.vStoreWithDemographics. It contained the data for store size, number of employees, and total revenue.
4. The next step was to load in the data in python, using pydocb to connect to the database , pandas to load the query and save the table as a dataframe, used matplotlib and seaborn to visualise the data.

**SQL code used initially in SSMS:**

SELECT \* FROM Sales.Store 🡪 Table

Data stored in Demographics column

A screenshot of a computer

Description automatically generated

Data was stored in xml file like this

A screenshot of a computer

Description automatically generated

SELECT \* FROM Store.vStoreDemographics 🡪 View

A screenshot of a white table

Description automatically generated

Got access to data that I needed through this view

**Libraries used in Python**

1. **pyodbc** - Library through which we can connect to the AdventureWorks database
2. **pandas** - To be able to store the relevant tables as a dataframe for data cleaning , manipulation and analysis
3. **matplotlib** - Plotting library used for data visualisations
4. **seaborn** - Another data visualisation library , built on top of matplotlib and used in conjunction with it

**Plots used in python ( SquareFeet, AnnualRevenue, NumberEmployees)**

* **Scatter plot –** Signify relationships between two continuous variables
* **Violin plot –** Showcase the distribution of variables
* **Heatmap –** To show correlation between said variables
* **Bar chart –** To compare variables to a categorical variable