Assignment 2

DAL 371

Duncan Pienaar & Ofentse Mapheto

Contents

[Problem Definition 2](#_Toc134818591)

[Business Domain 2](#_Toc134818592)

[Resources 2](#_Toc134818593)

[Key Stakeholders 2](#_Toc134818594)

[Interviewing the Analytics Sponsor 3](#_Toc134818595)

[Initial Hypotheses 3](#_Toc134818596)

[Potential Data Sources 4](#_Toc134818597)

# Discovery Phase

The discovery phase is a critical part of the data analysis process as it lays the foundation for the subsequent stages of the analysis and helps ensure that the insights gained are accurate and meaningful.

# Problem Definition

The problem we aim to solve is improving the university's grading system using analytics and data visualization to improve learning outs. We want to analyze student data to identify patterns that can help us in recruitment and retention efforts. The current grading system is not efficient and does not provide a comprehensive view of the student's performance. We need to develop a new grading system that can provide detailed insights into student performance and help the university in decision-making.

# Business Domain

The university in question has a primary goal of improving student learning outcomes, and it recognizes that the grading system is an important aspect of achieving that goal. The university wants to ensure that the grading system accurately reflects student performance and incentivizes behaviors that lead to learning and growth.

The new grading system should provide a comprehensive view of student performance to help the university in recruitment and retention efforts. The system should be able to identify students' strengths and weaknesses to provide personalized recommendations and support. It should be scalable to handle a large volume of data and provide real-time insights. The acceptance criteria for the solution should be that it improves the university's grading system, increases student performance, and helps in decision-making.

# Resources

We will require assistance from a data engineer or DBA to assist in database management and SQL queries but sins we have done a course regarding database management no help is needed.

Shakeng Thamaga – Helped construct the Interview questions.  
Lebohang Sopazi – Helped run through the entire assignment and guided through most aspects.  
David Maila – Gave guidance as to how I can tackle the problem statement an gave his views.

# Key Stakeholders

**Students**:  
Students have a direct interest in the grading system as it affects their academic performance and progress. If the proposed improvements are successful, students stand to benefit from a more accurate and fair grading system, which would better reflect their learning and provide them with better feedback on their performance.

**Faculty**:  
Faculty members are responsible for designing and implementing the grading system. They have a direct interest in the success of any changes made to the system, as it will affect their workload and their ability to accurately assess student performance. If the proposed improvements are successful, faculty members will benefit from a more efficient and effective grading system, which will enable them to better focus on teaching and providing feedback to students.

**Administrators:**   
Administrators have a broader interest in the success of the grading system, as it affects the overall reputation and academic quality of the institution. If the proposed improvements are successful, administrators stand to benefit from improved academic outcomes, better student satisfaction, and a stronger reputation for academic excellence.

**Employers:**   
Employers have an interest in the grading system as it affects the quality of graduates produced by the institution. If the proposed improvements are successful, employers will benefit from better-trained and more competent graduates, who have been accurately assessed and received appropriate feedback during their education.

**Society:**   
Society has a general interest in the academic quality and integrity of higher education institutions. If the proposed improvements are successful, society will benefit from a better-educated workforce, which will contribute to economic growth and social development.

# Interviewing the Analytics Sponsor

* What is your vision for the university's grading system?
* What are the current challenges with the grading system?
* What data sources are currently being used for the grading system?
* How do you currently use student data for decision-making?
* What metrics do you use to evaluate student performance?
* How do you identify students who require additional support?
* What improvements do you want to see in the new grading system?
* How do you envision the new grading system improving recruitment and retention efforts?
* What are the expected outcomes of the new grading system?
* How do you plan on implementing the new grading system?

# Initial Hypotheses

* Students who perform well in one subject tend to perform well in related subjects.
* Students who attend more classes tend to perform better.
* The grading system is biased towards certain groups of students.
* Students who have a growth mindset and believe that intelligence and abilities can be developed through effort and hard work tend to perform better than those who have a fixed mindset.
* Students who receive regular feedback on their performance tend to improve more than those who don't receive feedback.
* Students who are exposed to a wide range of learning resources, such as books, videos, and interactive media, tend to have a more comprehensive understanding of the subject matter than those who rely solely on classroom instruction.

# Potential Data Sources

The university has access to student performance records, course materials, gradebooks, and surveys. These data sources will be used to understand the current grading system and identify areas for improvement. The data shall then be collected from the university's databases and is checked for completeness and accuracy.

Another data source for this project is Twitter. We can extract data related to university events, feedback from students, and academic discussions. To extract data from Twitter, we need to follow these steps:

* Create a Twitter developer account and get access to the API.
* Use a Python library such as Tweepy to connect to the API and extract data.
* Clean and preprocess the data to remove any irrelevant information.
* Load the data into a database or data warehouse for further analysis.

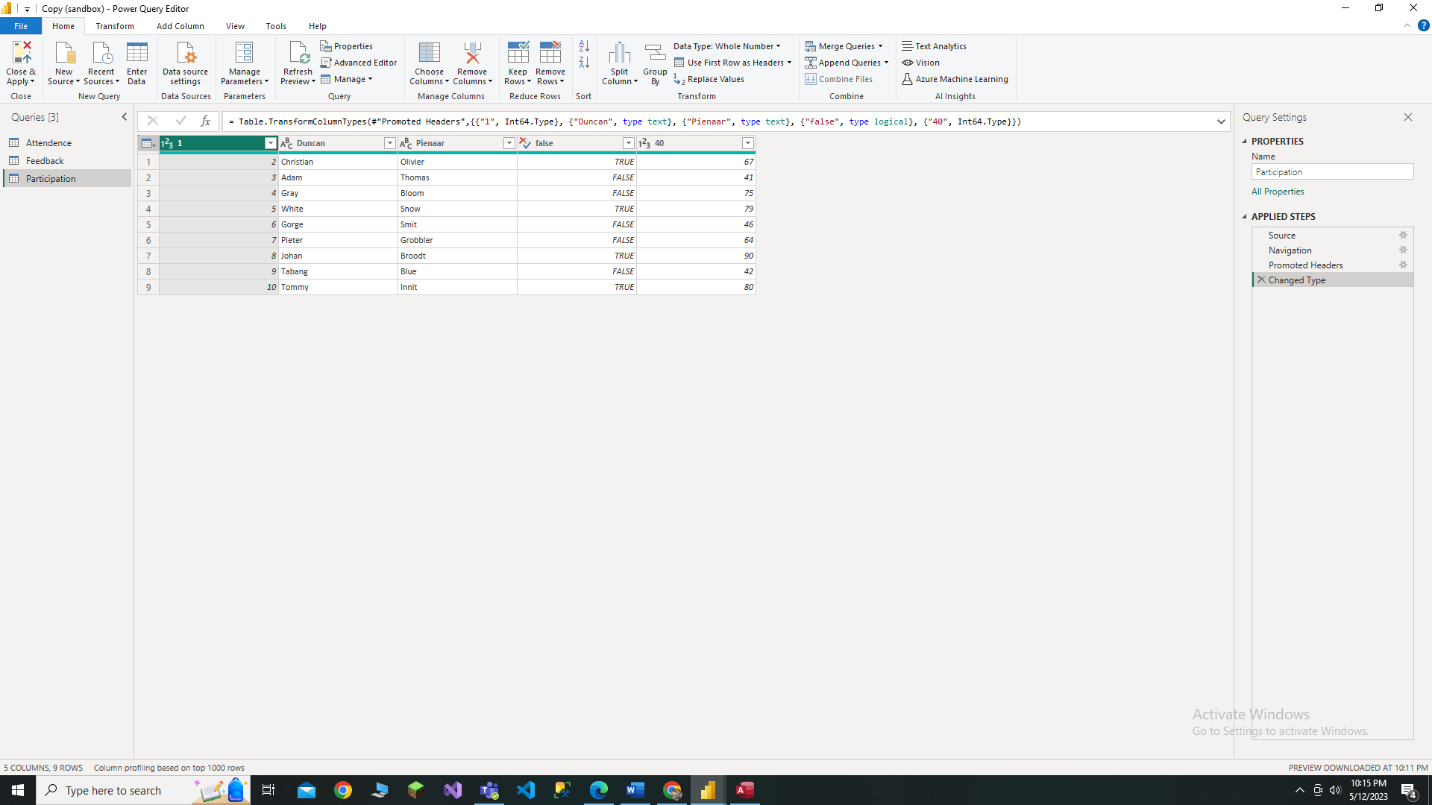
Another potential data source is the web, where datasets are already collected and can thus be used to clean up and used. To extract data from the web, we need to follow these steps:

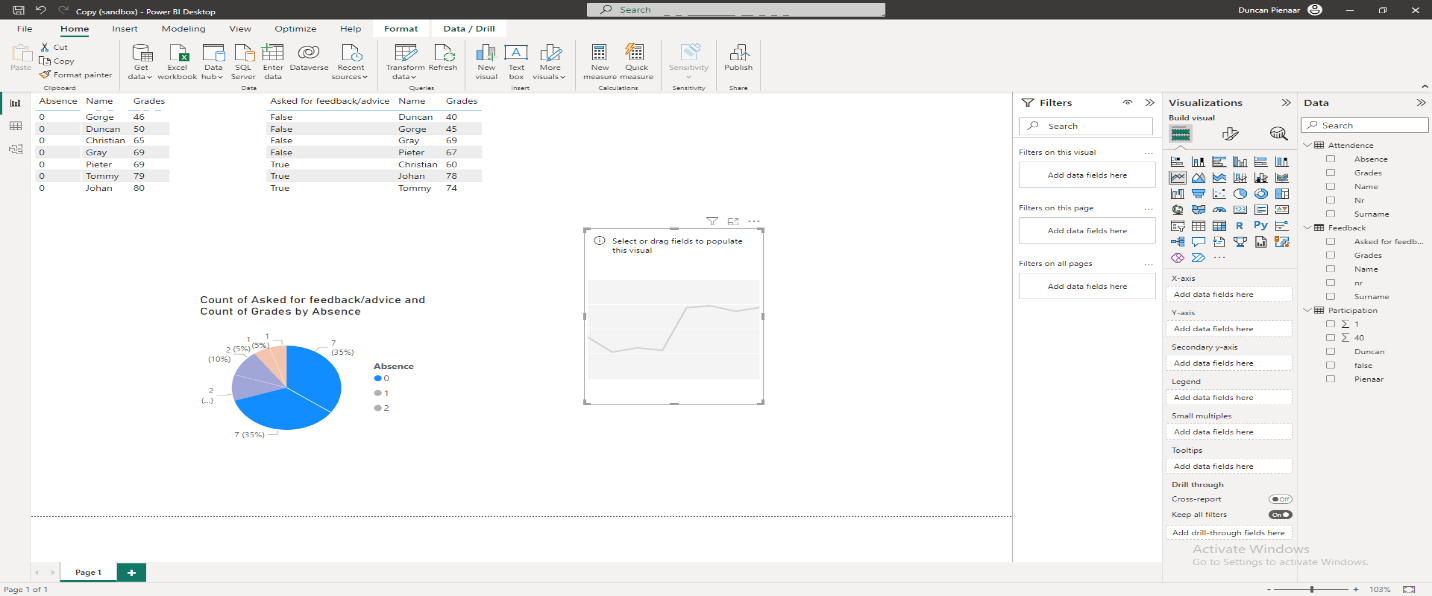
* Find a reliable dataset.
* Extract the data from the dataset.
* Clean and preprocess the data to remove any irrelevant information.
* Load the data into a database or data warehouse for further analysis.

# Data Preparation

# Sandbox

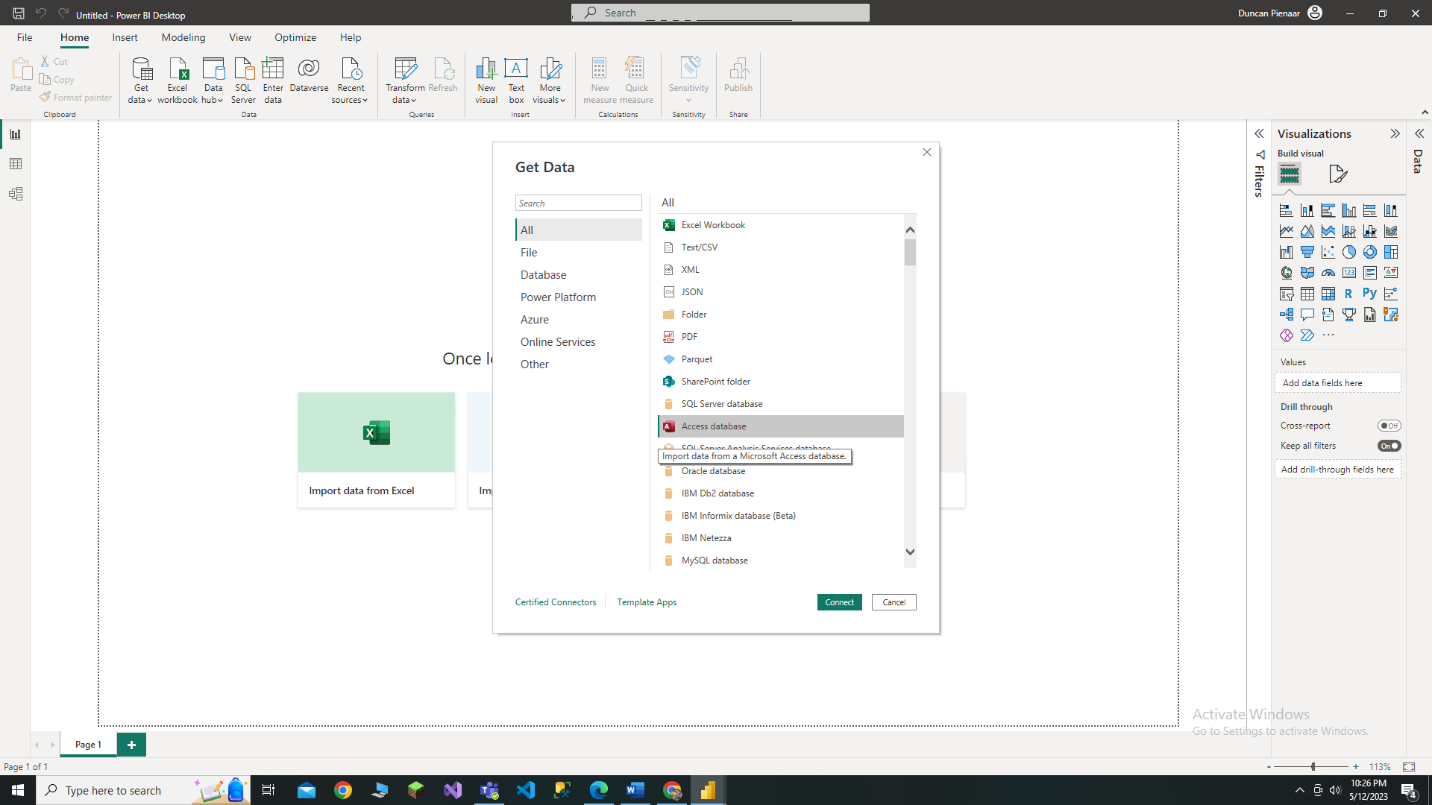
A sandbox is a testing environment that allows you to experiment with data without affecting your production environment. To create a temporary analytic sandbox, you can use a cloud-based service like AWS, Azure, or Google Cloud, or you can use an on-premise solution like VMware. The sandbox should contain both structured and unstructured data relevant to your analysis. In this example I just made a copy of my data to play with and experiment with in PowerBI.

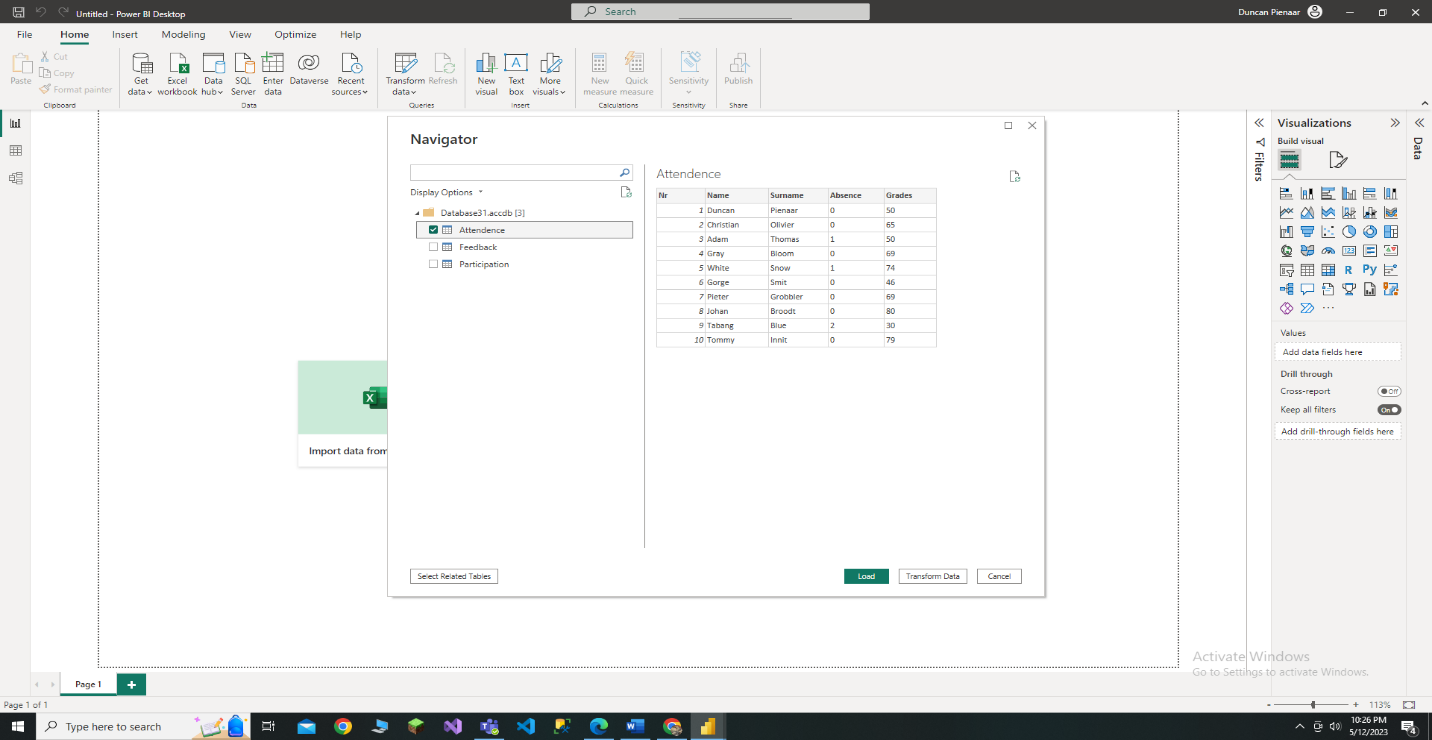




# Performing ETLT

The ETLT process involves extracting data from the source, transforming it into a format suitable for analysis, and loading it into the data warehouse. In Power BI, you can use Power Query to perform ETLT operations. Power Query is a data connection and transformation tool that allows you to extract data from various sources, transform it using a range of built-in functions, and load it into Power BI for analysis. You can also use Power BI's Query Editor to perform data transformations. I extracted the data from using PowerBI’s Microsoft Access.



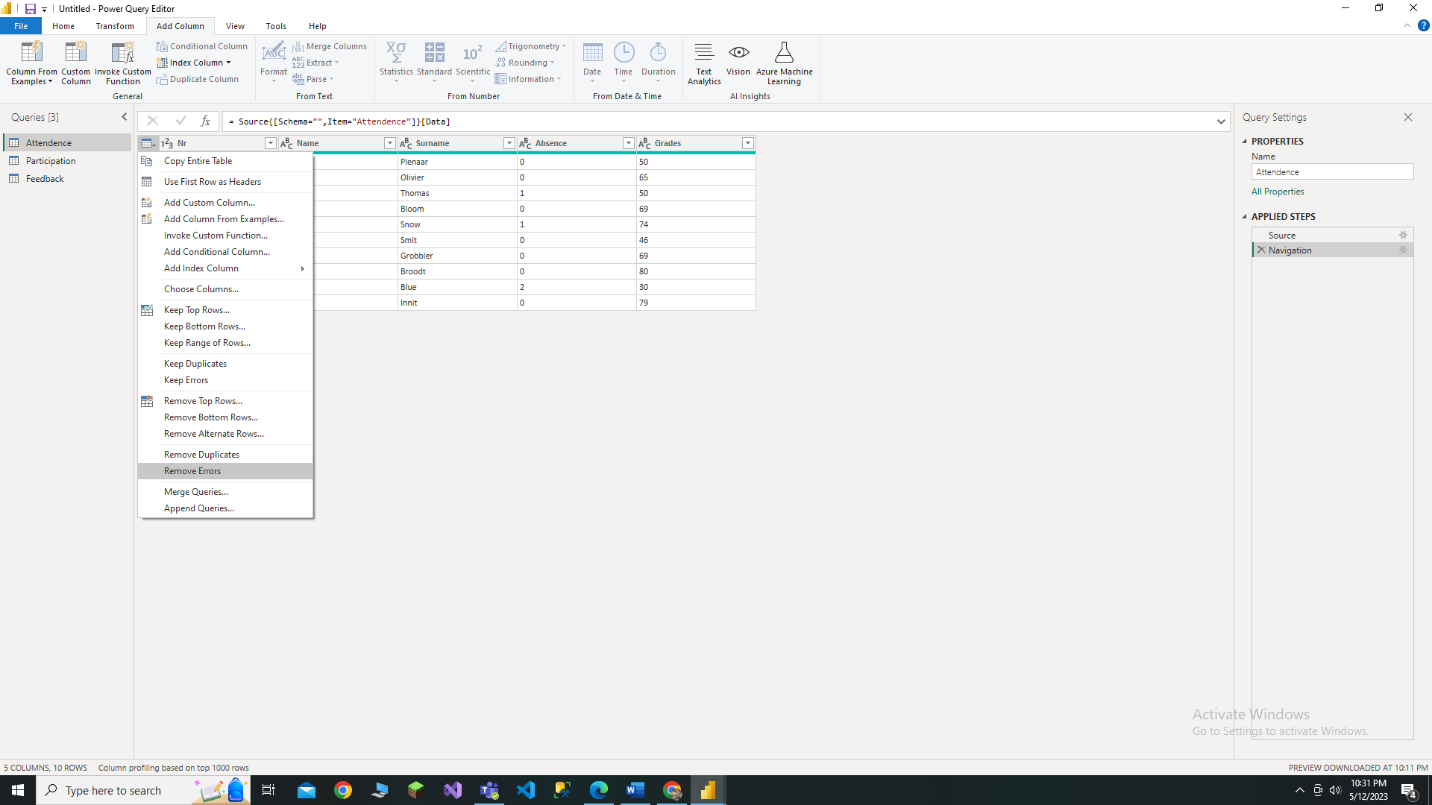


# Data Conditioning

Data conditioning involves cleaning and preparing data for analysis. This includes tasks like removing duplicates, correcting errors, filling in missing values, and normalizing data. Power BI provides a range of tools for data conditioning, including the Query Editor, which allows us to perform data transformations, and the Data View, which allows us to view and edit data.

A screenshot of a computer

Description automatically generated

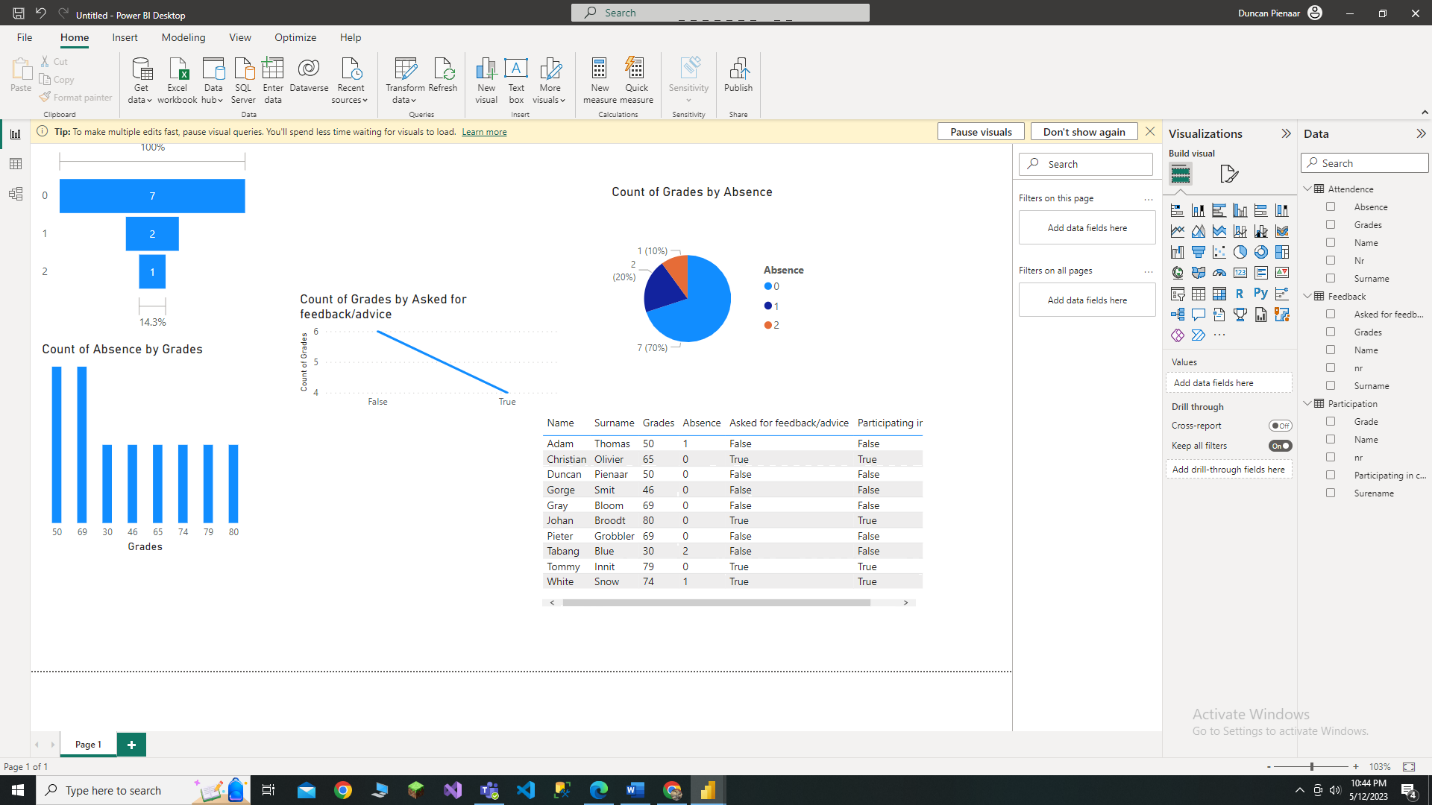


# Datasets

To identify datasets outside the organization that may be useful for your analysis, we searched for open data sources like data.gov, Kaggle, and Google Dataset Search. We also searched for industry-specific data sources relevant to my analysis but unfortunately found nothing.

# Visualizations

To create visualizations of our data, we use Power BI's built-in visualization tools, including charts, tables, and maps. These tools allow you to create interactive visualizations that can be used to explore and analyze your data.



# Tools for the data Phase

Other tools that can be used for data preparation include Python, R, SQL, and Excel. These tools provide a range of functions and capabilities for data transformation, cleansing, and analysis. However, Power BI is a powerful and flexible tool that provides a range of features and capabilities for data preparation and analysis, and it is well-suited to the needs of most organizations, although we may not be an organization we used Microsoft Access and PowerBI mainly.