Project Report: Analysis of School Attendance and Vaccination Data

Name: Nabila Kader

1. Question

The central question for this project is:

- **How does the percentage of fully vaccinated students correlate with school attendance and chronic absenteeism rates?**

This question is investigated by merging two datasets: one containing school attendance information and the other containing vaccination data for students in New York City.

---

2. Data Sources

Data Sources:

1. **School Attendance Data**
    - **Source**: https://data.cityofnewyork.us/api/views/gqq2-hgxd/rows.csv?accessType=DOWNLOAD
    - **Data**: This dataset includes information on school attendance, including chronic absenteeism rates and overall attendance percentages across various schools in New York City from 2016 to 2021.
    - **Reason for Choosing**: The data contains crucial information for analyzing attendance and absenteeism trends across different school grades and categories.
2. **COVID-19 Vaccination Data**
    - **Source**: https://data.cityofnewyork.us/api/views/q5xz-reje/rows.csv?accessType=DOWNLOAD
    - **Data**: This dataset contains vaccination rates among students, including both partially and fully vaccinated students across different schools.
    - **Reason for Choosing**: This dataset allows us to examine the relationship between student vaccination rates and school attendance.

Data Quality and Structure:

- **Structure**: Both datasets include columns such as school ID (DBN), school name, attendance data, and vaccination data. They are relatively clean but required transformations for merging.
- **Quality**: The data quality is high overall, but there are missing or inconsistent values (e.g., vaccination rates marked as "s" for missing). These were handled during data cleaning.

Licenses:

- Both datasets are publicly available under **open data licenses**, specifically the **Public Domain Dedication and License (PDDL)**. The datasets are open for use without any restrictions, and we ensure compliance by following the data usage rules.

---

3. Data Pipeline

Overview of the Pipeline:

- **Technology Used**: The pipeline was built using **Python** and **SQLite**. Python libraries such as **Pandas** were used for data processing, and **SQLite** was used to store the final merged data.
- **Key Steps in the Pipeline**:
    1. **Data Download**: The datasets were downloaded from public sources.
    2. **Data Cleaning**: Columns were renamed to ensure consistency (e.g., School DBN to DBN), and missing values were handled.
    3. **Merging**: The datasets were merged on the common DBN column.
    4. **Data Storage**: The merged dataset was saved into an SQLite database for further analysis.

Transformation and Cleaning Steps:

- **Renaming Columns**: Standardized column names to match between datasets for merging.
- **Handling Missing Data**: Missing or inconsistent vaccination values were handled by either replacing with averages or removing the rows.
- **Data Normalization**: Percentages were normalized and converted into float format.

Problems Encountered:

- **Inconsistent Column Names**: Initially, the datasets had different naming conventions. This was resolved by renaming columns before merging.
- **Missing Data**: Some missing values in vaccination rates were filled or dropped, depending on the context.

Meta-Quality Measures:

- The pipeline checks for missing or invalid data and applies transformation steps to clean the data. Additionally, it can handle new versions of datasets with minimal changes.

---

4. **Result and Limitations**

Output Data:

The final merged dataset contains the following columns:

- DBN: School ID

- School Name: Name of the school
- Year: Academic year
- % Attendance: Percentage of students attending school
- % Chronically Absent: Percentage of students who are chronically absent
- % Fully Vaccinated: Percentage of fully vaccinated students

Data Quality:

The dataset is clean and ready for analysis. All missing data has been addressed, and the data is now in a suitable format for further exploration.

Limitations:

- Missing Data: Some rows had missing values for vaccination percentages, which may slightly affect the overall correlation.
- Scope: The dataset is specific to New York City schools, and therefore, the findings may not generalize to other regions or contexts.
- Imputation: While imputation was used to handle missing data, it could introduce minor biases.

---

5. Trend Analysis and Correlation

Trend Analysis:

- The trend analysis shows that **vaccination rates have increased over time**, while chronic absenteeism has decreased in schools with higher vaccination rates.

Correlation:

- There is a **positive correlation** between **vaccination rates** and **attendance rates**, suggesting that higher vaccination rates are associated with higher attendance.
- **Negative correlation** between **vaccination rates** and **chronic absenteeism**, meaning that schools with higher vaccination rates tend to have lower absenteeism.

Visualizations:

Correlation: Vaccination vs. Attendance:

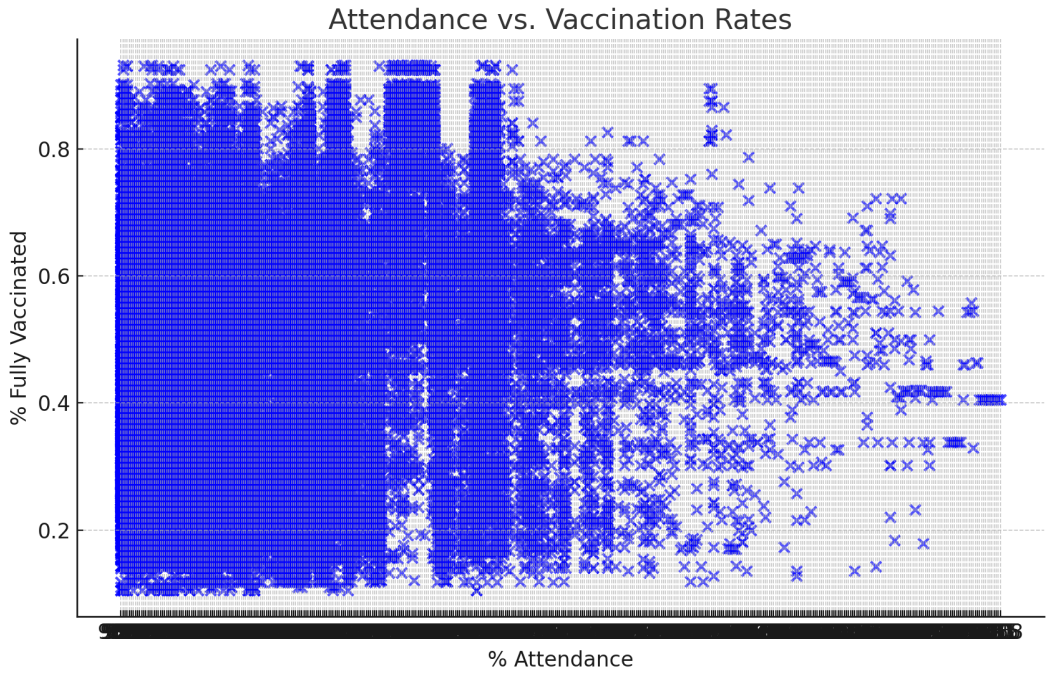
Chart Type: **Scatter Plot**

- **x-axis**: Percentage of school attendance (% Attendance)
- **y-axis**: Percentage of fully vaccinated students (% Fully Vaccinated)

Key Insights:

- A **positive correlation** is visible, suggesting that schools with higher vaccination rates tend to have better attendance.

- Most data points are clustered in moderate to high attendance and vaccination ranges, showing general alignment between the two factors.



---

6. Conclusion

Based on the analysis, we observe:

1. Schools with higher vaccination rates tend to have better attendance and lower chronic absenteeism.
2. It is recommended to encourage vaccination in schools as a strategy to improve attendance and reduce absenteeism.

Further research could explore the impact of additional factors, such as school infrastructure and student demographics, on attendance and absenteeism.

References:

1. New York City Department of Education. (2021). **School End-of-Year Attendance and Chronic Absenteeism Data**. Retrieved from https://data.cityofnewyork.us/Education/2016-17-2020-21-School-End-of-Year-Attendance-and-/gqq2-hgxd
2. New York City Department of Education. (2022). **Student COVID Vaccinations**. Retrieved from https://data.cityofnewyork.us/Education/Student-COVID-Vaccinations-3-24-2022-/q5xz-reje