Project Report
Maya Reddy
1.23.25

Dataset Sources

- Global Burden of Disease Collaborative Network. Global Burden of Disease Study 2021 (GBD 2021) Results. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2022. Accessed January 13, 2025. Available from https://vizhub.healthdata.org/gbd-results/.

This dataset provided information on cervical cancer incidence and mortality rates within the United States from 1990 to 2021.

- The World Health Organization / UNICEF. Papillomavirus Rapid Interface for Modelling and Economics (PRIME): London School of Hygiene & Tropical Medicine, Mark Jit, Marc Brisson, Kaja Abbas, and Han Fu. Cervical Cancer & HPV Vaccines. Cervical Cancer & HPV Vaccines. Owned by Joakim Arvidsson. Accessed January 15, 2025. Available from https://www.kaggle.com/datasets/joebeachcapital/cervical-cancer-and-hpv-vaccines.

These datasets provided information on HPV vaccination rates and estimated the number of cervical cancer cases and costs prevented due to vaccination in the United States from 2010 to 2020.

- Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health. 500 Cities: Census Tract-level Data (GIS Friendly Format), 2017 release. Created October 15, 2018. Data last updated October 15, 2018. CDC UserSA. Accessed December 16, 2024. Available from https://data.cdc.gov/500-Cities-Places/500-Cities-Census-Tract-level-Data-GIS-Friendly-Fo/kucs-wizg/about_data.

- Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health. 500 Cities: Census Tract-level Data (GIS Friendly Format), 2019 release. Created November 2, 2016. Data last updated December 5, 2019. Owned by The 500 Cities. Accessed December 16, 2024. Available from https://data.cdc.gov/500-Cities-Places/500-Cities-Census-Tract-level-Data-GIS-Friendly-Fo/k86t-wghb/about_data.

These datasets provided information on Pap smear screenings issued to women aged 21-65 years in the United States in 2014 and 2016.

- Centers for Disease Control and Prevention. Vaccination Coverage among Adolescents (13-17 Years). Created July 28, 2021. Data last updated September 12, 2024. Owned by HHS Office

of the Chief Data Officer. Accessed December 16, 2024. Available from https://healthdata.gov/dataset/Vaccination-Coverage-among-Adolescents-13-17-Years/47pk-jpce/about_data.

This dataset provided information on HPV vaccination rates and sociodemographic factors among adolescents in the United States from 2016 to 2023.

Process of Database Creation

*I frequently revised both of these Google Colab notebooks as I received feedback on how to structure my relational database and the tables it contained.

1. *First Semester Project: Data Cleaning & Preprocessing (2).ipynb*

    - For each dataset, I standardized column names, adjusted datatypes, handled null/duplicate values, added/dropped columns, and filtered the datasets to include data relevant to the United States.

    - For `data.csv`, I restructured the dataset in order to reduce the number of columns necessary for a primary key in the relational database and to allow for easier manipulation of data in my analysis.

    - I merged the smaller datasets for Pap smear screenings and HPV vaccinations into larger datasets, combining `pap_smear_2014.csv` and `pap_smear_2016.csv` into `pap.csv`, and all the files in the `hpv_vaccines` folder into `hpv.csv`.

2. *First Semester Project: Creating Relational Database (3).ipynb*

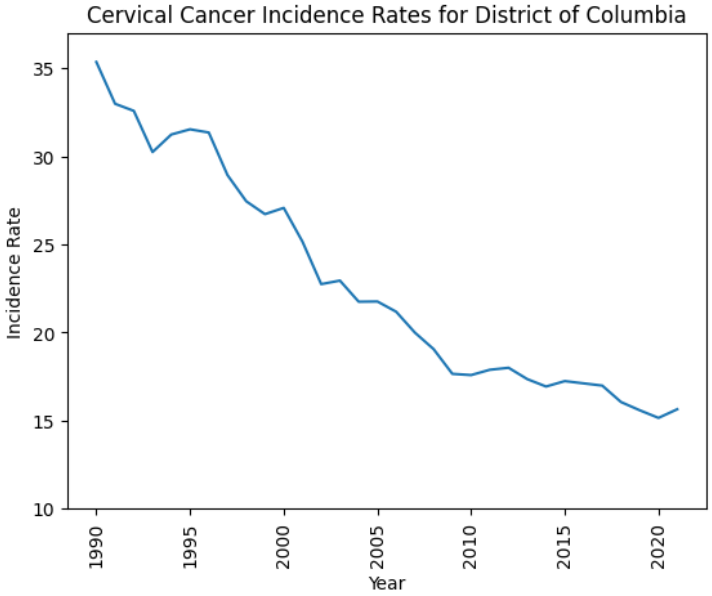
    - I loaded the datasets into the Google Colab notebook and created a connection to the `cervical_cancer.db` database. After creating tables within the database with SQLite, I transferred the preloaded DataFrames into those tables.

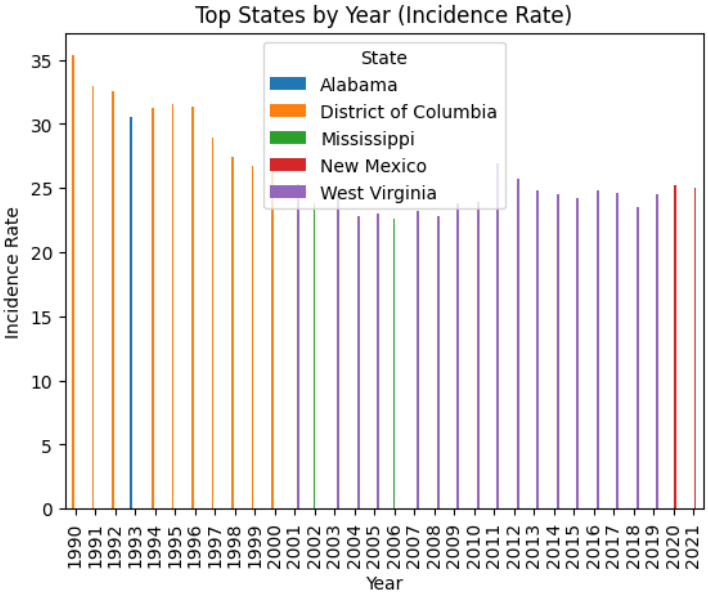
    - The `pap`, `adolescent`, and `demo` tables all referred to `data` using the states' names as foreign keys; since `hpv` did not contain state-level data, it had no foreign keys. All tables except for `hpv` used a combination of multiple columns for their primary keys to ensure uniqueness.

    - After committing the changes I made to the database file, I confirmed the existence of the tables by extracting data from `cervical_cancer.db` using SQLite. Once I verified that my datasets had been loaded into the database, I closed the connection.

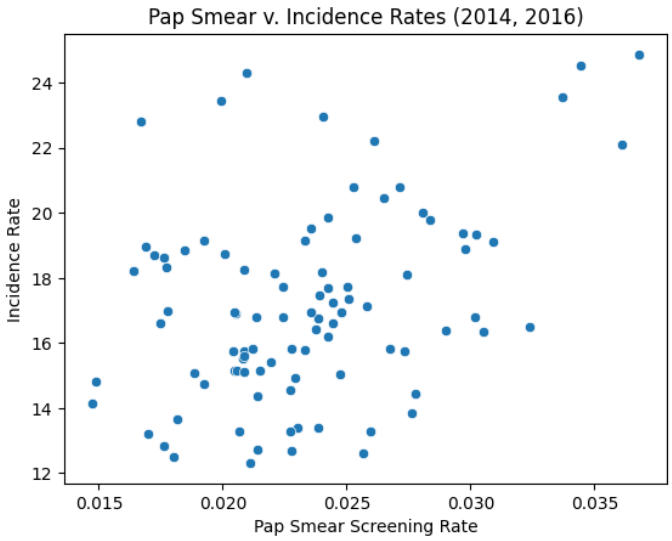
Key Insights from Analysis

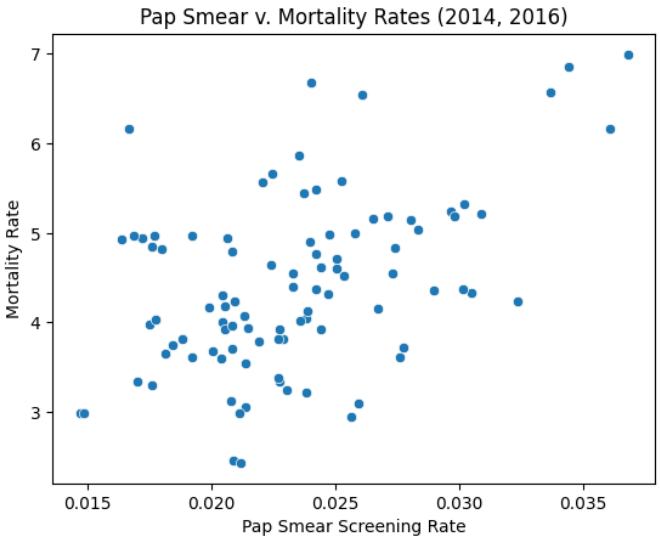
- The District of Columbia had the highest incidence/mortality rates in the United States during the 1990s, but managed to rapidly decrease its cervical cancer rates by 2021. A further investigation into the factors contributing to this decline could yield helpful methods for managing cervical cancer rates.



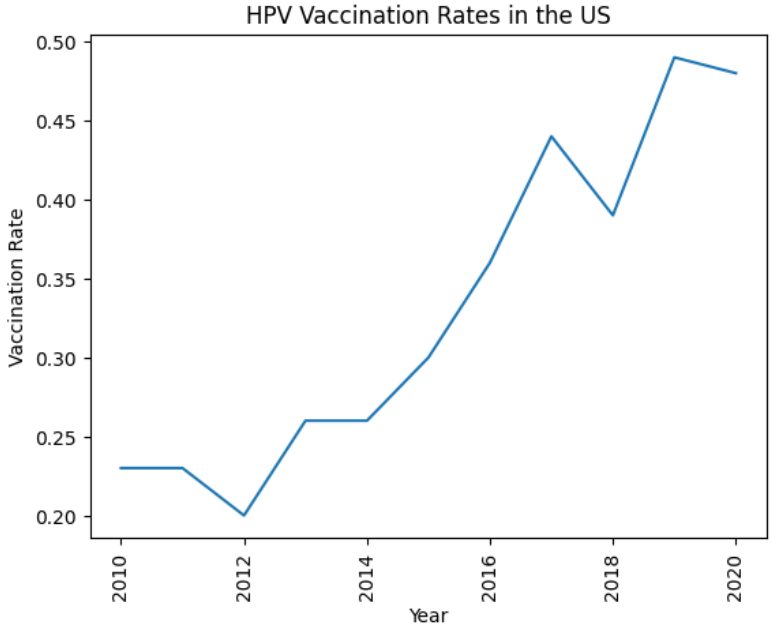


- Pap smear screenings seemed to have little effect on cervical cancer rates in this analysis: the correlation between Pap smear screening rates and incidence rates was 0.381, while the correlation between Pap smear screening rates and mortality rates was 0.469. Since these datasets only spanned a short timeframe between 2014 and 2016, though, this section of the analysis does not offer an accurate depiction of Pap smears' impact on cervical cancer.

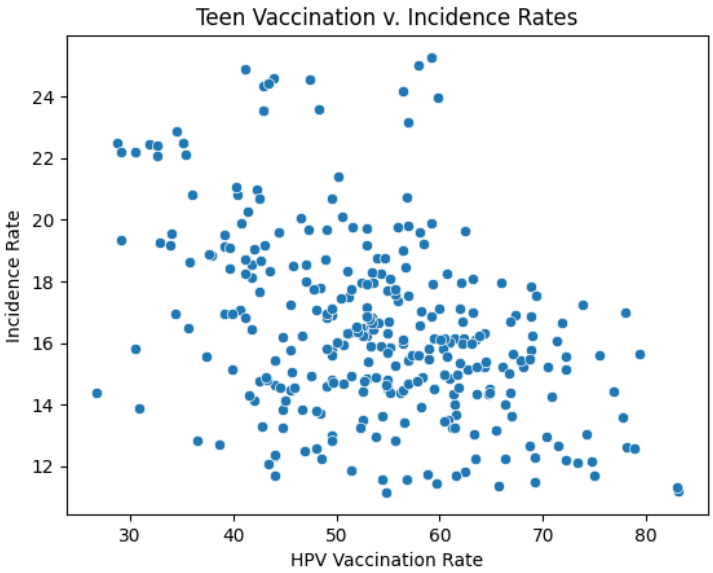


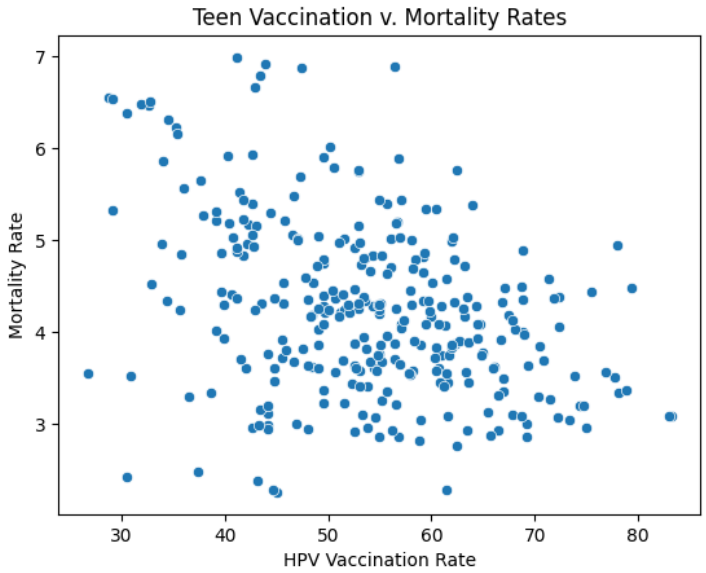


- HPV vaccination rates steadily increased in the United States during the 2010s. According to the dataset's estimates, though, thousands of cervical cancer cases and the spending of millions of dollars could have been averted under a projected 90% vaccination rate, demonstrating the instrumental impact of HPV vaccinations on cervical cancer rates.

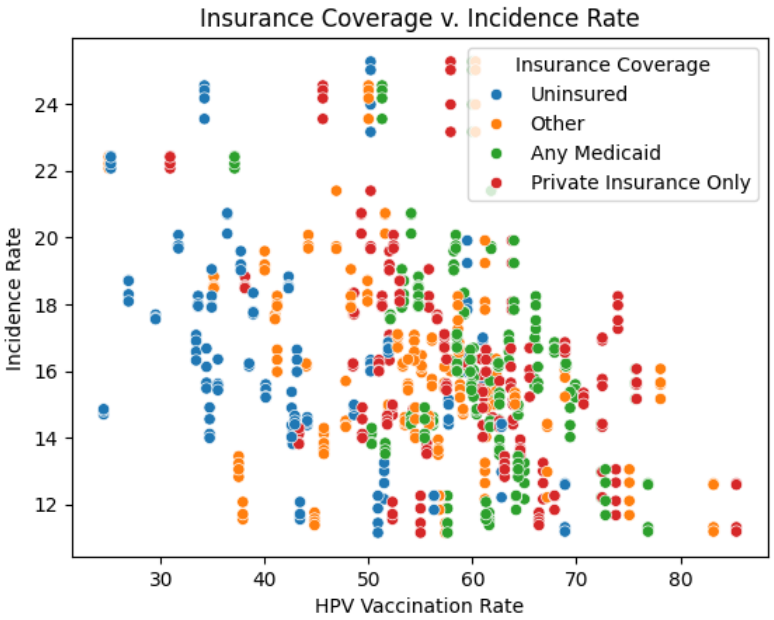


- HPV vaccination rates among adolescents had a moderate negative correlation of -0.414 with cervical cancer incidence rates, and a weak negative correlation of -0.375 with cervical cancer mortality rates. Since I had more data on HPV vaccinations, these findings are likely more reliable compared to the results of the analysis on Pap smear screenings.

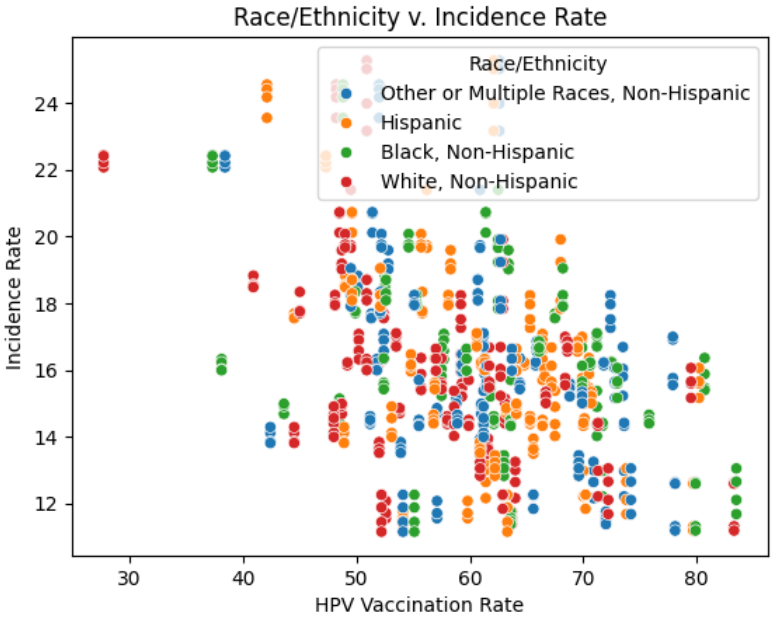




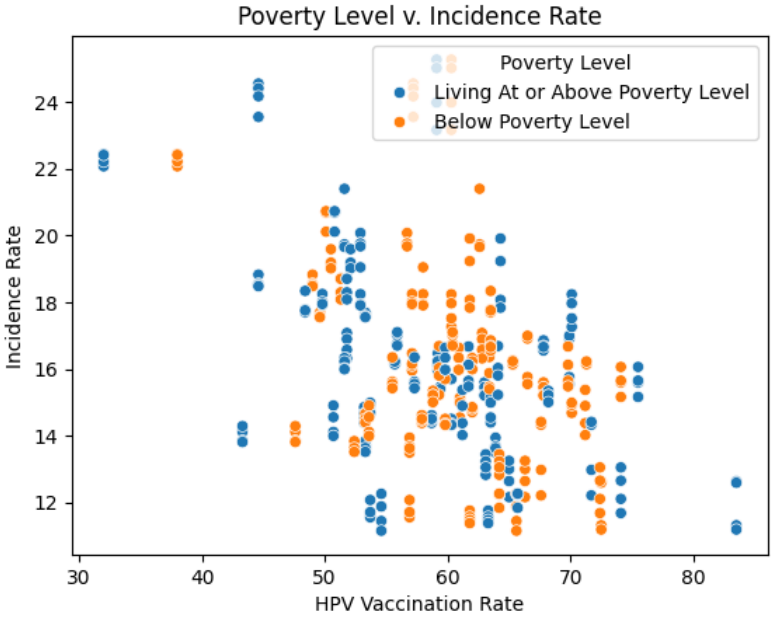
- Regarding insurance coverage, adolescents with Medicaid or private insurance generally had higher HPV vaccination rates than those who were uninsured in 2018-2022. Private insurance had the highest negative correlation of -0.479 with cervical cancer incidence rates, while adolescents with no insurance or unidentified insurance coverage had weaker correlation coefficients.



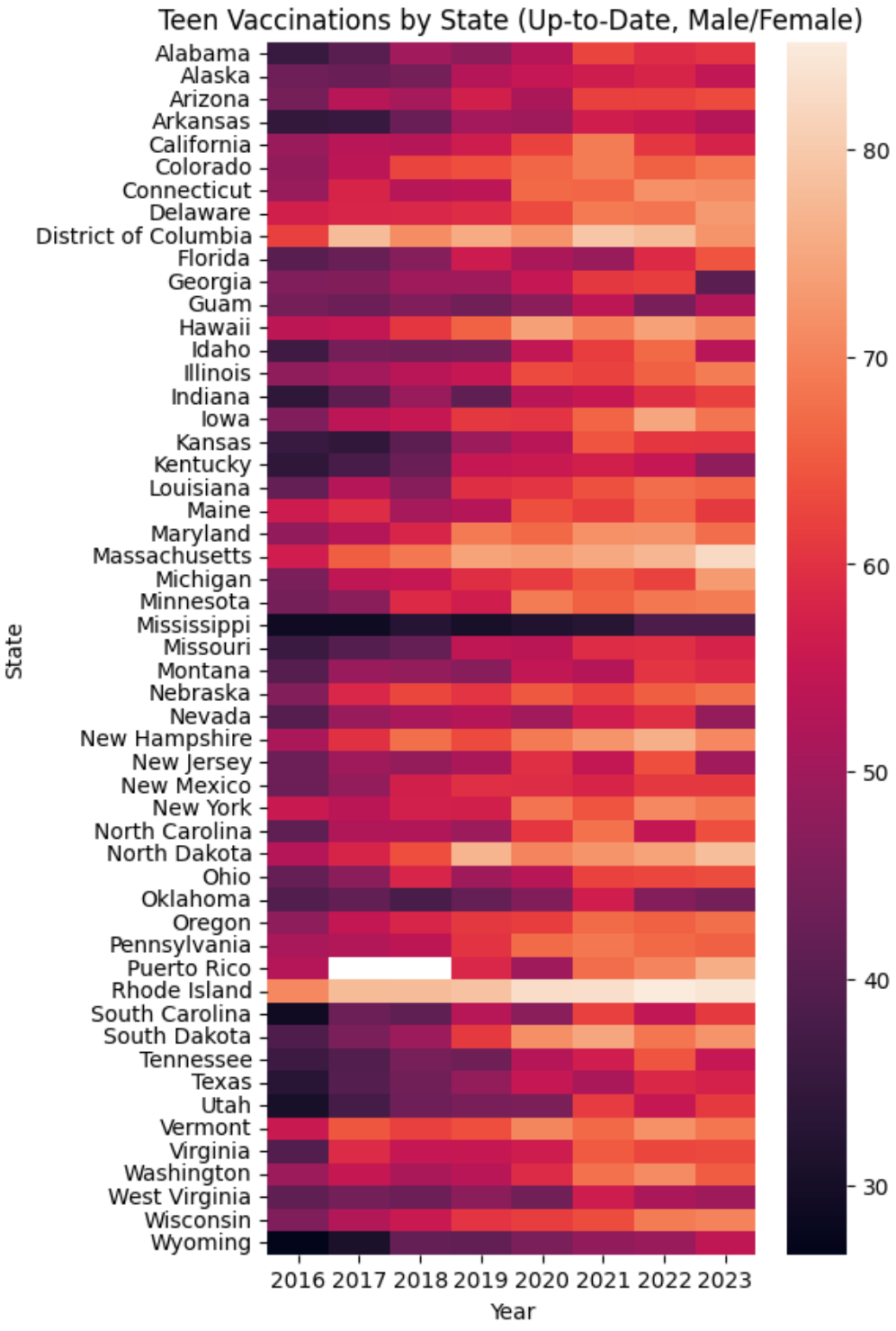
- Regarding race and ethnicity, adolescents who identified as white or Hispanic had stronger negative correlations with cervical cancer incidence rates (-0.508 and -0.482, respectively), while those who identified as Black or another race had weaker negative correlations (-0.409 and -0.400, respectively).



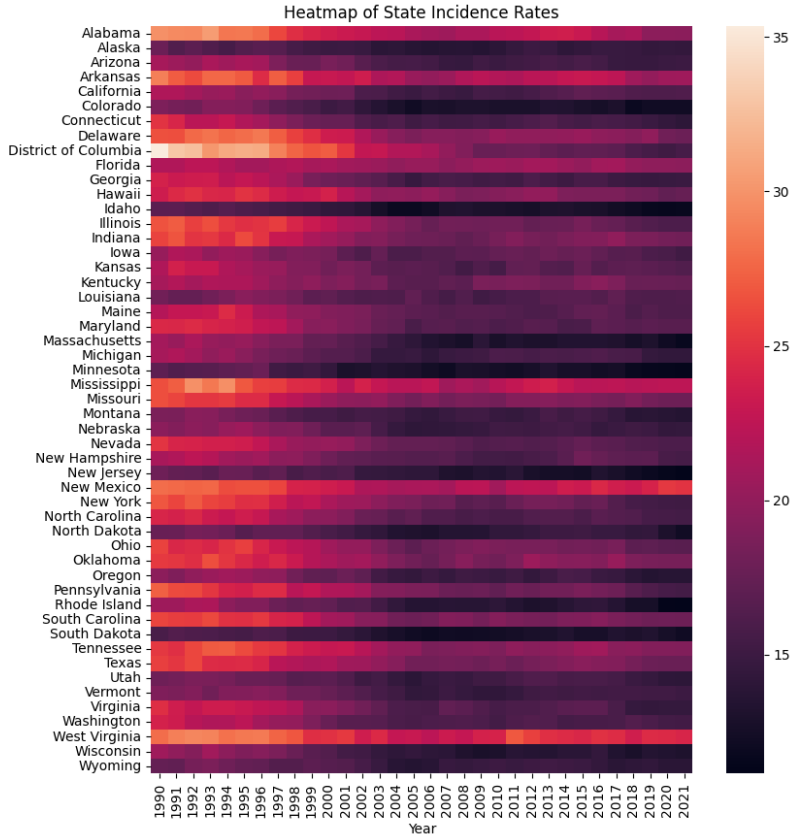
- Regarding poverty levels, adolescents who were living at or above the poverty level in 2018-2022 had a higher negative correlation of -0.483 with cervical cancer incidence rates than those who were living below the poverty level (their correlation was -0.407).

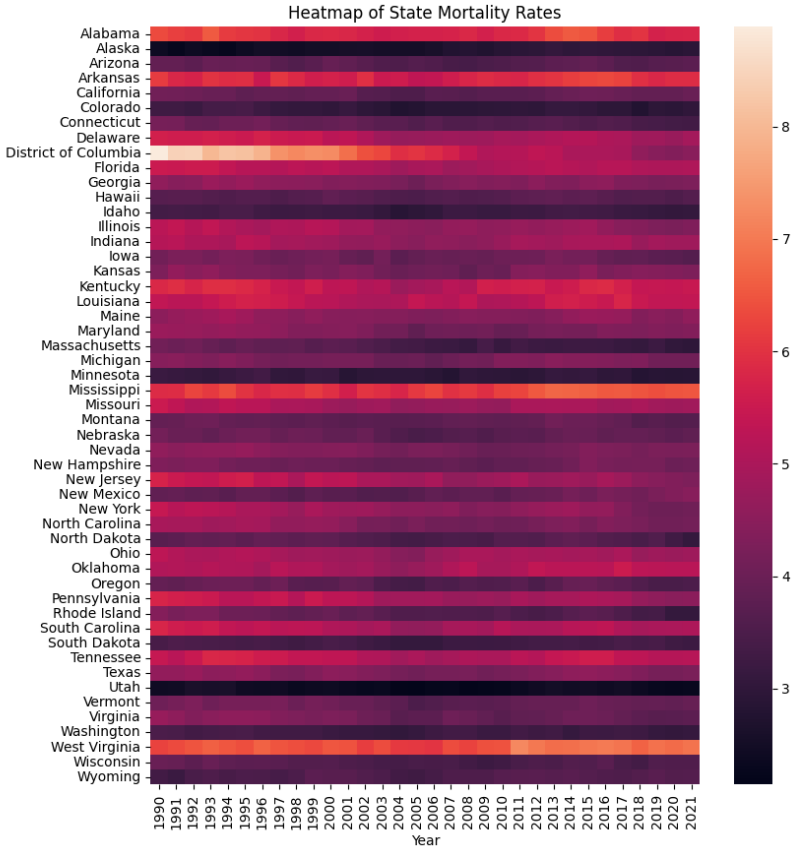


- Rhode Island, Massachusetts, and the District of Columbia managed to maintain high HPV vaccination rates among adolescents from 2016 to 2023; perhaps an investigation of their state-level governmental funding or healthcare systems could offer insights into their widespread vaccination coverage.
- Mississippi had the lowest HPV vaccination rates among adolescents in 2016-2013; additionally, it had higher incidence/mortality rates compared to other states in recent years.



- Cervical cancer incidence rates have generally decreased in most states. However, mortality rates have remained relatively consistent in most states for the past four decades.





Conclusions & Recommendations

Though cervical cancer incidence rates have declined in many parts of the United States, few states have managed to achieve rates below 15%, and several still suffer from rates above 20%. Intervention efforts would be most beneficial in states like West Virginia, New Mexico, Mississippi, and Arkansas; a further investigation into how the District of Columbia managed to drastically decrease both incidence and mortality rates could provide these states with methods and measures to lower their own cervical cancer rates. Unfortunately, the analysis of Pap smears' impact on cervical cancer rates was ineffectual because of insufficient data. However, moderate negative correlation coefficients demonstrated how HPV vaccinations in teenagers are correlated with a decrease in cervical cancer rates. The effectiveness of vaccinations in adolescents depended upon their insurance coverage, race/ethnicity, and economic status; an in-depth analysis on the effects of sociodemographic factors on cervical cancer incidence would yield more informative conclusions, though.