Project Report
Maya Reddy
1.28.25

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*

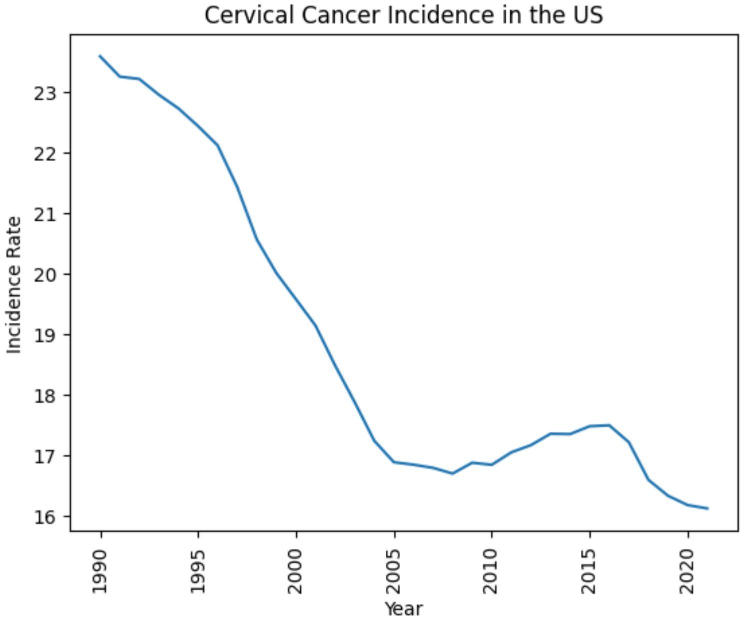
   - I loaded the datasets into the Google Colab notebook and created a connection to the `cervical_cancer.db` database. After creating tables within the database with SQLite, I transferred the preloaded DataFrames into those tables.

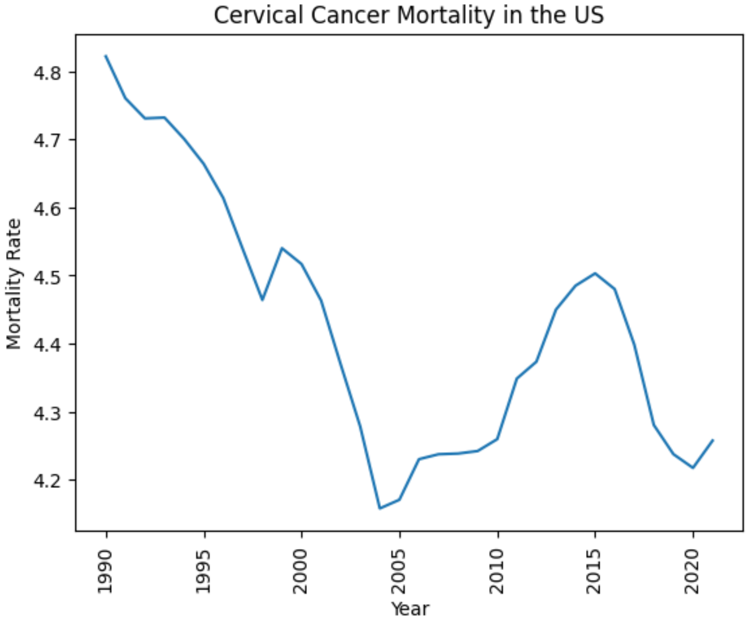
   - The `pap`, `adolescent`, and `demo` tables all referred to `data` using the states' names as foreign keys; since `hpv` and `us_data` did not contain state-level data, they had no foreign keys. All tables except for `hpv` and `us_data` used a combination of multiple columns for their primary keys to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

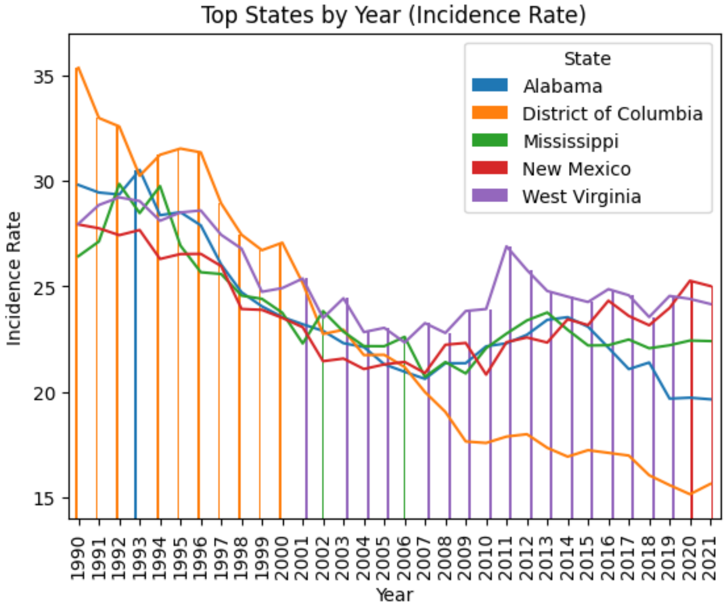
Key Insights from Analysis

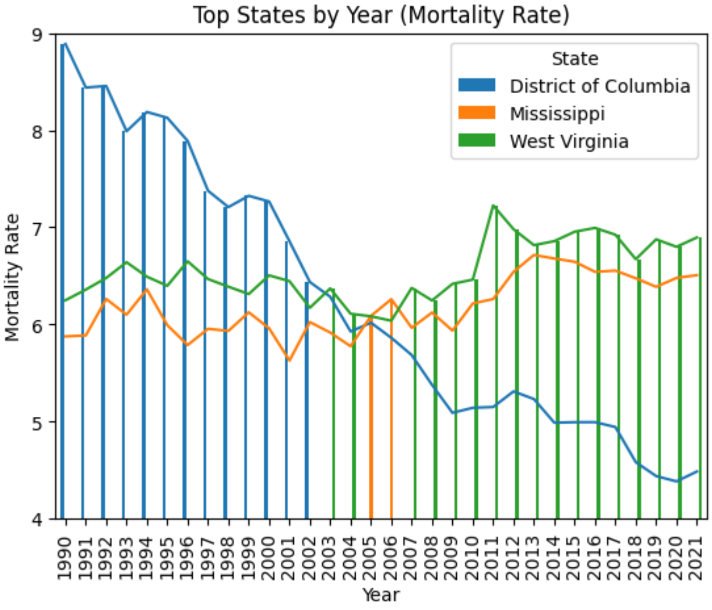
- The United States experienced a steep decline in cervical cancer (CC) incidence and mortality rates during the 1990s, with the lowest rates of the past four decades found in 2005. However, there was a slight spike in cervical cancer rates during the 2010s.



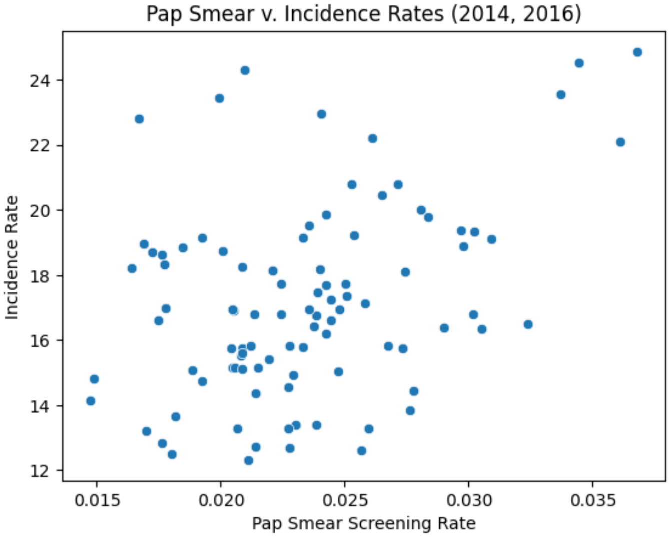


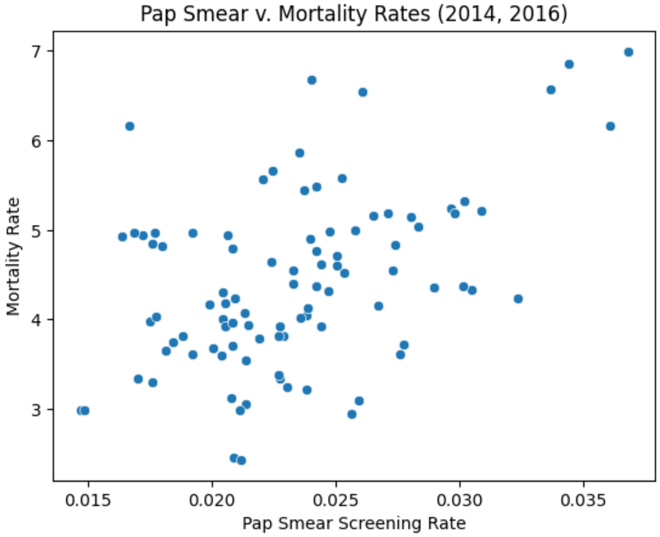
- The District of Columbia had the highest cervical cancer incidence and mortality rates in the United States during the 1990s, but managed to rapidly decrease its cervical cancer rates by 2021. A further investigation into the factors contributing to this decline could yield helpful methods for managing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



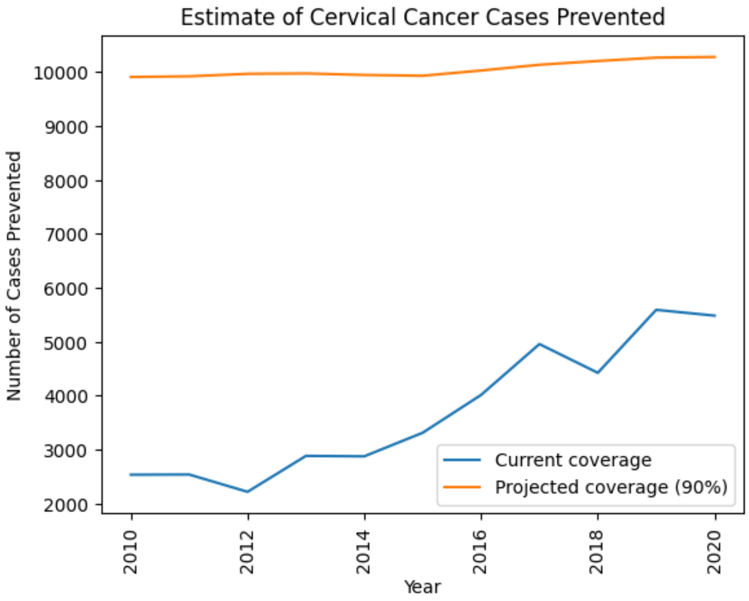


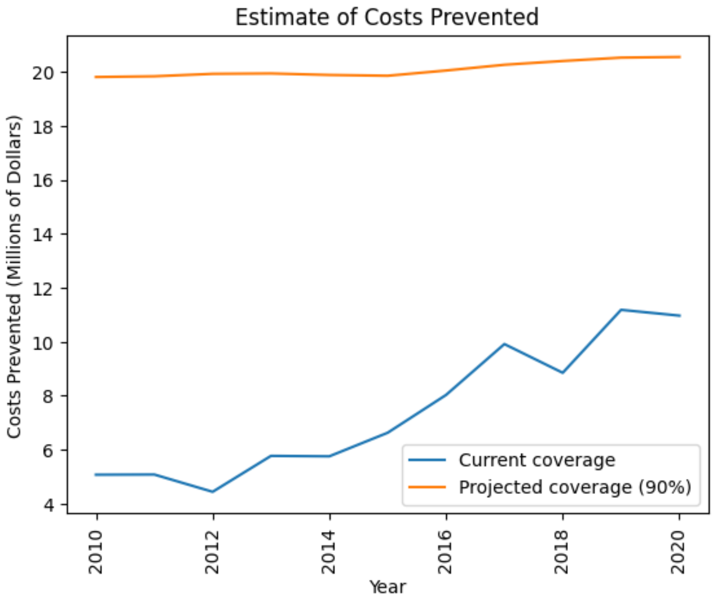
- The correlation coefficient between Pap smear screening rates and incidence rates was 0.381, while the correlation coefficient between Pap smear screening rates and mortality rates was 0.469. While a positive correlation would not be unusual for incidence rates — as Pap smears detect cancerous cells rather than prevent them — a positive correlation with mortality rates did not agree with previous studies on the benefits of Pap smears. However, since these datasets spanned a short timeframe, during which there was a noticeable increase in cervical cancer rates, this section of the analysis does not offer a reliable or accurate depiction of Pap smears' impact on cervical cancer.



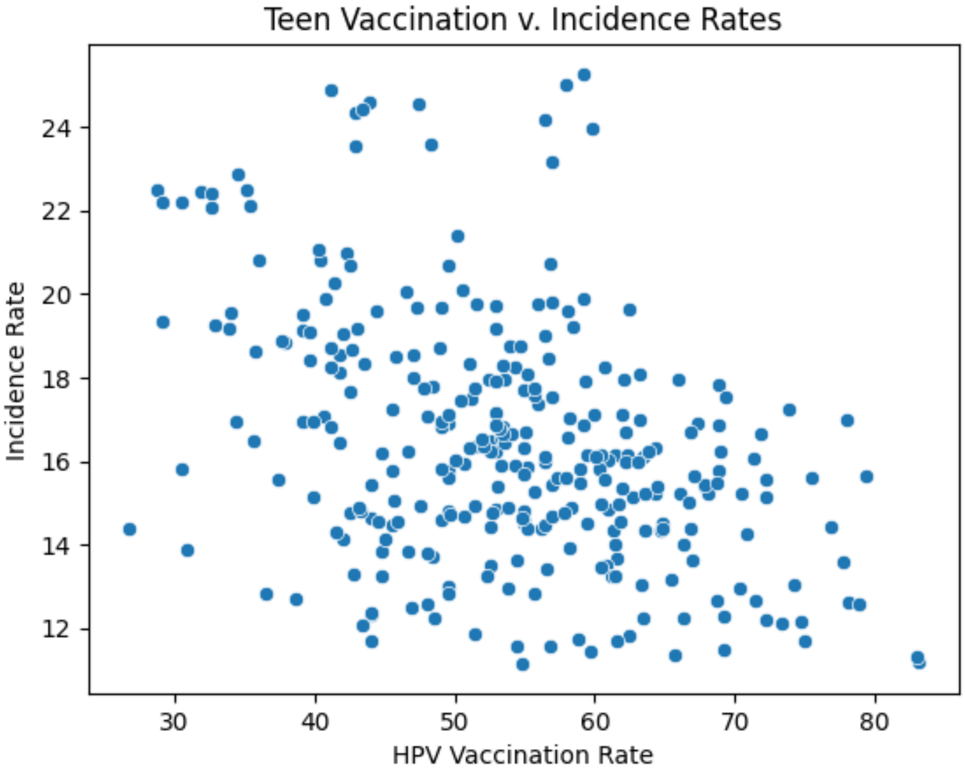


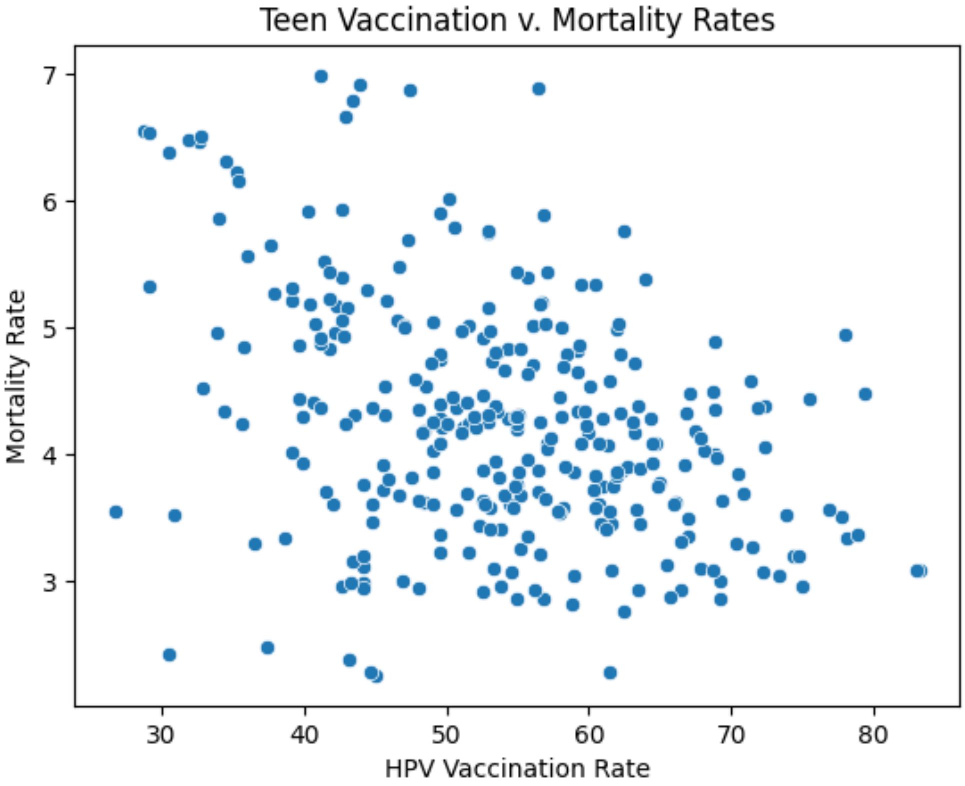
- HPV vaccination rates steadily increased in the United States during the 2010s. According to the dataset's estimates, thousands of cervical cancer cases and the cost of millions of dollars could have been averted under a projected 90% vaccination rate, demonstrating the instrumental impact of HPV vaccinations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



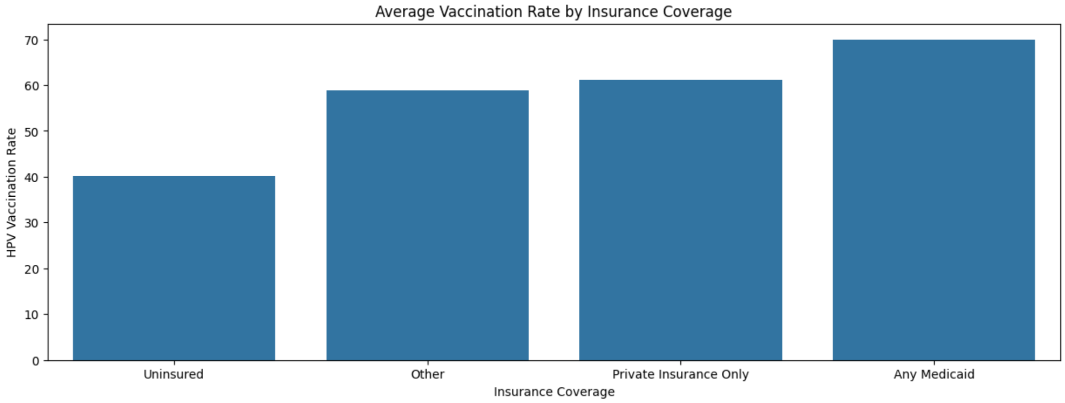


- HPV vaccination rates among adolescents had a moderate negative correlation of -0.414 with cervical cancer incidence rates, and a weak negative correlation of -0.375 with cervical cancer mortality rates. Since I had access to more data on HPV vaccinations, these findings are likely more reliable compared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

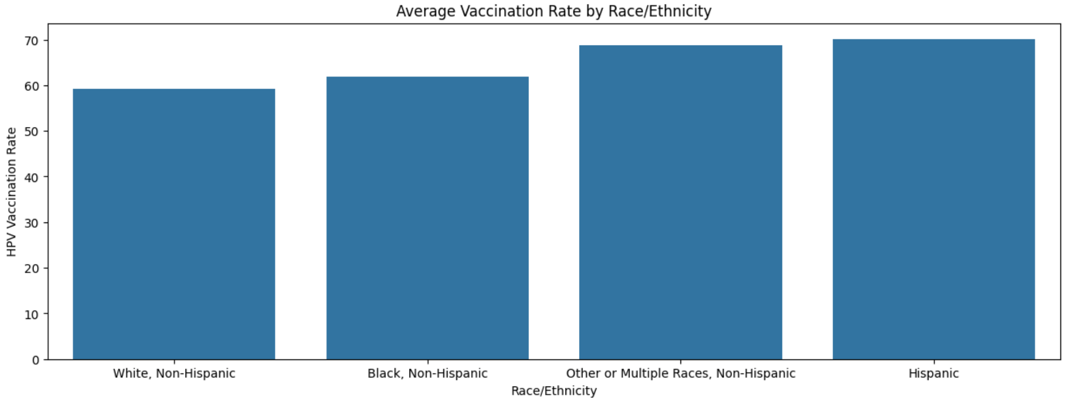




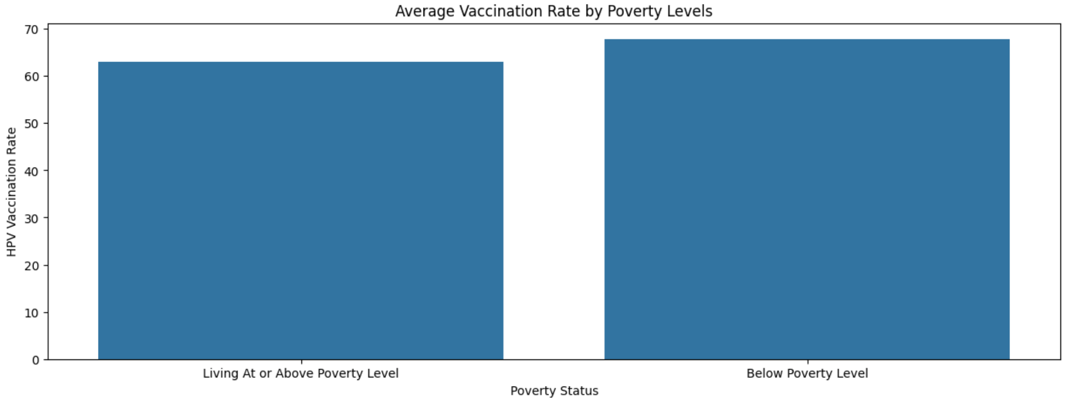
- Regarding insurance coverage, adolescents with Medicaid had higher HPV vaccination rates on average. Teenagers who were uninsured experienced the lowest rates of vaccination. Even with private insurance, vaccines cost a substantial amount of money, so adolescents with access to Medicaid and, by extension, free vaccines would likely have higher vaccination rates.



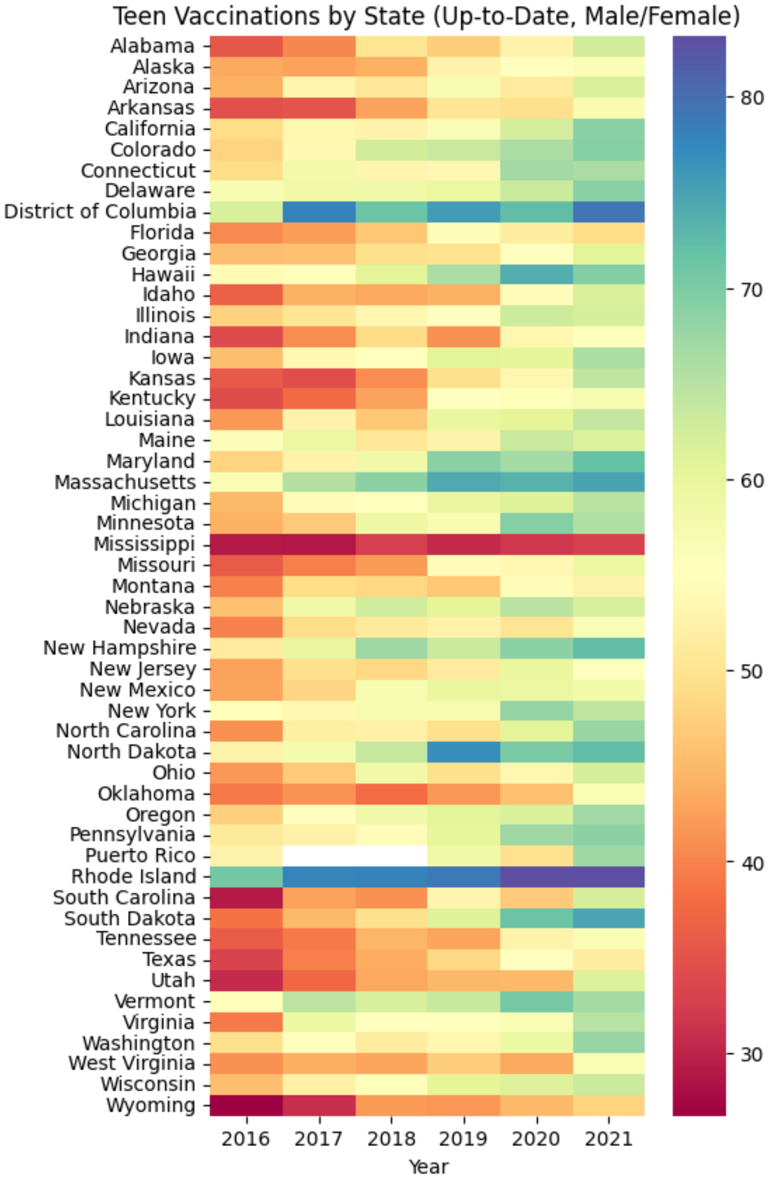
- Regarding race and ethnicity, adolescents who identified as Hispanic had the highest HPV vaccination rates on average, while those who identified as white had the lowest HPV vaccination rates on average.

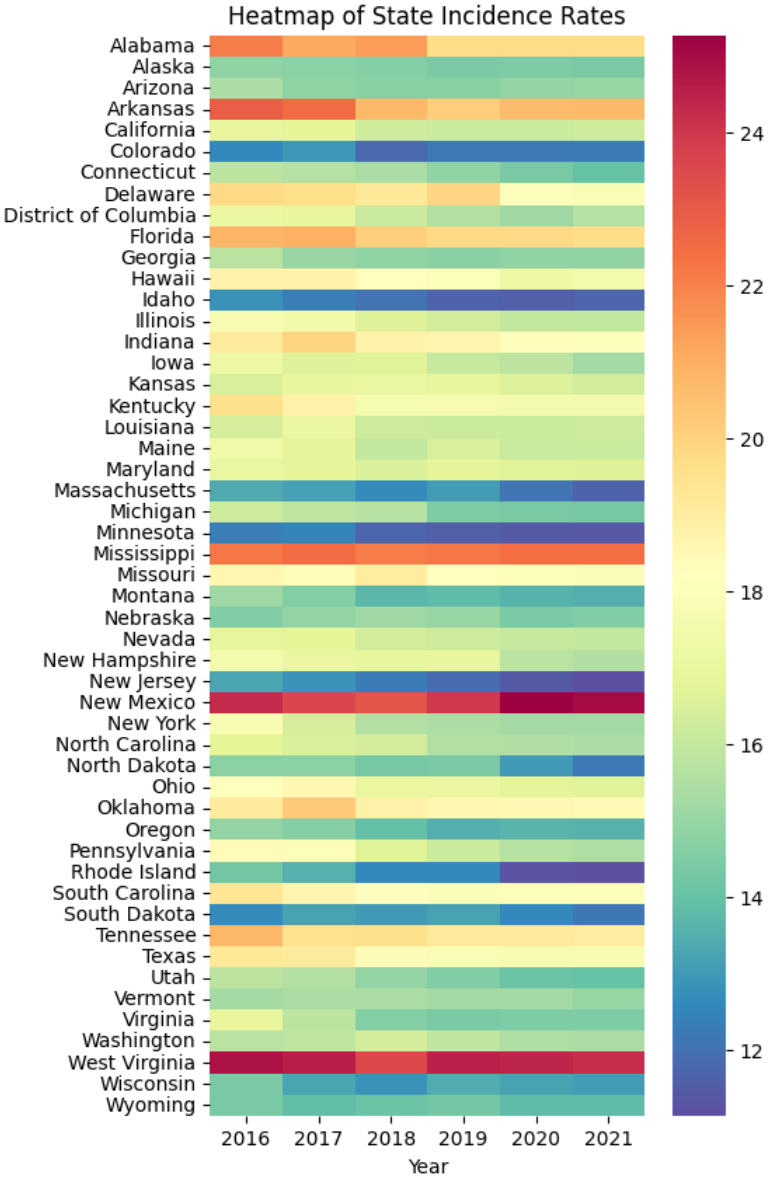


- Regarding poverty levels, adolescents who live below the poverty level had higher vaccination rates on average than those who were living at or above the poverty line. Since most adolescents who live below the poverty line are eligible for Medicaid, they have access to free vaccines and consequently have higher vaccination rates.

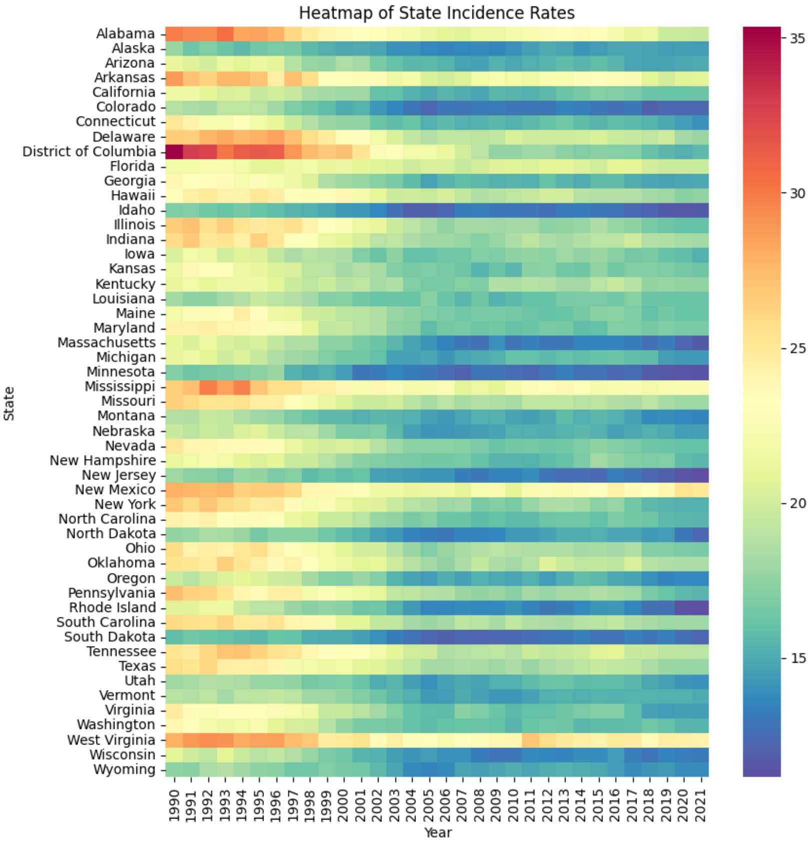


- Rhode Island, Massachusetts, and the District of Columbia managed to maintain high HPV vaccination rates among adolescents in 2016-2023. Perhaps an investigation of their state government funding or healthcare policies could offer insights into their widespread vaccination coverage. Meanwhile, Mississippi had the lowest HPV vaccination rates among adolescents in 2016-2023.
- When comparing these two heatmaps below, I noticed that most states with low vaccination rates have high cervical cancer incidence rates. Since my dataset on HPV vaccinations supplied information on teenagers rather than adults, any variations in cervical cancer incidence rates will be delayed in the left-hand graph. A more reliable analysis requires data that spans multiple decades, not just five years.

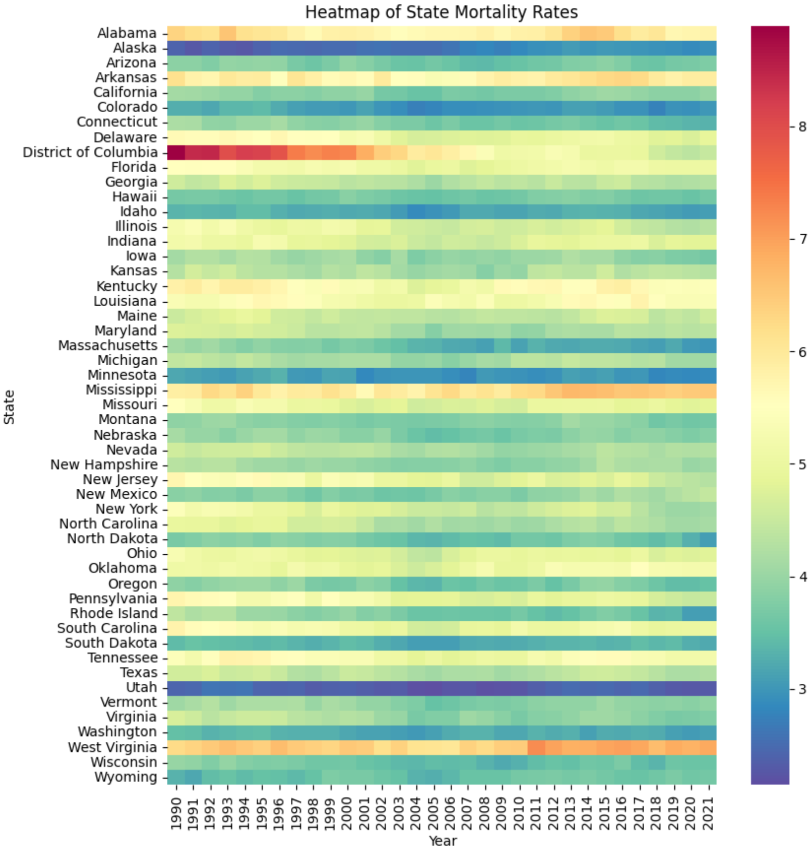




- States like West Virginia, Mississippi, Arkansas, and New Mexico have recently experienced high rates or upticks of cervical cancer. Meanwhile, states like Colorado, Minnesota, and South Dakota have experienced consistently low cervical cancer rates.



Heatmap of State Incidence Rates



Heatmap of State Mortality Rates

Conclusions & Recommendations

Although 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 strategies on how to lower their own cervical cancer rates. Unfortunately, the results of my analysis on Pap smears' impact on cervical cancer rates were unreliable because of insufficient data. However, negative correlation coefficients demonstrated how increases in HPV vaccination rates among teenagers are associated with decreases in cervical cancer incidence rates. The prevalence of vaccinations among adolescents depended upon their insurance coverage, race/ethnicity, and economic status. More specifically, the results indicate that the availability of free HPV vaccines is associated with higher vaccination rates. However, an in-depth analysis on the effects of racial/economic factors on HPV vaccination rates and, more broadly, cervical cancer rates would yield more informative conclusions.