Deliverable 1

All the Python file(s) used in the project. If you have other code files, include them too.

At minimum, your Python files should consist of:

1. The Python code you used to dynamically generate your INSERT statements that are in your .sql file
2. The code that you used to pull data from the database and the subsequent handling of that data to tabulate your results and create the figures used in the final research paper.
3. The Python file should contain the SQL queries that you have written to pull data from the database.
    a. The SQL queries should be as specific to your needs as possible. That is, do not write queries to generically pull all the data to your database and then process the data after. Your queries should be as specific as possible for your needs.

Deliverable 2

A README file that contains information someone would need to know if they were to run your code. The README should also list your research questions and the SQL queries that were used to answer the questions. Please include line numbers to make it easier for your TA to find the particular queries in your code.

Please list out each of your files and provide a short description of each so your TA can quickly reference the files they need to look at when grading.

Deliverable 3

Your research paper which includes the following sections. You are free to reuse text from your proposal/midway checkpoint.

NOTES: points 1 - 3 have been covered, we have kind of covered 4 in our original question

1. Introduction

1. Give a bit of background on the topic to give context to your research question. Imagine that this paper is being read by someone totally new to your project (e.g., someone outside of the class). What would they have to know to understand your research topic/research questions?
2. State your research questions.
    a. WARNING: If you are in a situation where your project TA has concerns about the suitability/scope/feasibility of your project, we expect that you have taken steps to resolve those issues prior to the submission of your research paper and that you have double checked your changes with your TA prior to submission. We warmly recommend that you take advantage of the weekly meetings with your TA to ensure that your research questions (and project) meet expectations around scope an suitability.
    b. In the event that you do not resolve the outstanding concerns, we reserve the right to apply a 60% penalty to your project. The penalty will be calculated based on the full value of the deliverable.
3. What is the impact of knowing this information?

2. Related work/work that has been done in this area by others. Don't forget to cite!

1. Based on prior work that has happened in the area of interest, what is the background information that suggests that your research question is important/valid?
2. Discuss some other related work/projects people have done in the area.

3. Data

1. What datasets are you using?

2. Why did you choose these particular datasets?
3. What data cleaning steps did you perform and why do you think these steps were sufficient. (**WILL HAVE TO CHANGE THIS IN RESEARCH MIDWAY POINT**)
4. What are some things people need to be aware of about your data? This
5. is where you would discuss issues around outliers or missing data or any other significant findings from your EDA process.

4. Methodology

1. What methodology did you employ to answer your research questions?
2. Why did you choose your particular methodology?

5. Results and Discussion

1. What were your results?
2. For each of your research questions, produce a graph that depicts your results.
3. Validity refers to how generalizable your results are to the overall population/situation. Comment on the following:
    a. Internal validity: How well does your measure reflect what your research question is measuring? How do you know?
    b. External validity: How well does your measure reflect the truth in real life? How do you know?
4. Remember that it's fine if this is not the most perfectly designed study. Even published research studies often have shortcomings - sometimes due to the design but often due to constraints outside of their control.
5. We are looking to see that you can critically reflect upon your work to understand what is good about it and what shortcomings it has.

6. Discussion of results

1. Explain and interpret your results by putting it in context based on your research questions.

    a. In cases where you do not have clear/meaningful results, explain why this might have been the case.
    b. What impact does this have on the interested community
2. If there are other limitations to your results that you have not discussed in the validity section, discuss them here.

7. Future work

1. How can people use your results to benefit them?
2. What are some things that you would like to explore but did not have time to? Another way to think about this question is if you had an extra four months to work on the project, what else would you do? This could be extensions to your original research question or a follow up on your results or using another method to analyze the data.

8. References

1. References should be given in ACM format

Deliverable 4

If you have used an AI tool to help refine your work, state which tool and what prompts were given. You can also provide a URL of the chat log instead (please make sure the URL is accessible). If you have not used an AI tool to help refine your work, please explicitly state so.

Congratulations!

You've completed the research project!

CPSC 368 Research Project

Technology and Healthcare - by Aarav, Pushya, Suryansh

1. Introduction

Research Question

We will explore the following Research Question through this project:

How did socioeconomic and demographic factors influence telehealth adoption patterns during the COVID-19 pandemic in the United States (2020-2022)

The impact on the telehealth adoption patterns will be explored via the following factors:

1. ***Racial*** and ***ethnic disparities*** in telehealth utilization rates
2. ***Geographic differences*** in adoption patterns (privileged and under-privileged)
3. The relationship between ***educational attainment*** and telehealth usage

Note: In our study, we will create consistent geographic units for rural/urban classification: Our assumption will classify "metropolitan" cities and "urban" areas as one classification which will represent more connected and privileged residential areas. Similarly, we will classify "non-metropolitan" and "rural" areas as one classification to represent less connected and generally under / less privileged residential areas.

Motivation for the Research Question:

The COVID-19 pandemic catalyzed a dramatic shift toward telehealth services, potentially exacerbating existing healthcare disparities in the United States. While telehealth offered a crucial lifeline for continued healthcare access, preliminary studies suggest uneven

adoption patterns across different demographic groups. We aim to use this research to shine a brighter light on the following areas:

- Quantify disparities in telehealth adoption across different populations
- Identify specific barriers to telehealth access for underserved communities
- Contribute to evidence-based policy recommendations for more equitable healthcare delivery systems
- Examine how existing healthcare inequities were either mitigated or amplified by telehealth expansion

2. Related work

The COVID-19 pandemic accelerated telehealth adoption in the U.S., but research highlights persistent disparities shaped by socioeconomic and demographic factors. **Racial and ethnic inequities** emerged as a critical concern, with studies showing Black patients were more likely to rely on audio-only telehealth compared to White patients, who disproportionately used video visits (*Luo, Jake, et al.*). Latino and Asian populations faced compounded barriers, including language access and lower broadband connectivity, leading to reduced overall telehealth engagement 25. These disparities reflect systemic inequities in digital literacy and technology access, which risk exacerbating existing health gaps, particularly in chronic disease management and preventive care (*Chen, Evan M., et al*). For example, non-English speakers were significantly less likely to complete video visits (OR 0.49), underscoring the need for culturally tailored telehealth solutions (*Chen, Evan M., et al*).

Geographic and socioeconomic disparities further stratified telehealth access. ZIP code-level analyses revealed that areas with higher college education rates had stronger video visit adoption (coefficient 1.41), while low-income regions relied more on telephone-based care (*Luo, Jake, et al.*). Rural-urban divides were initially pronounced due to limited broadband infrastructure, though later studies (2022) suggested narrowing gaps as telehealth became normalized, with rural patients increasingly using messaging platforms for care (*Spaulding, Erin M., et al*). However, regional inequities persisted; Midwestern states lagged in telehealth utilization (aPR 0.65), likely due to policy variability and provider density (*Spaulding, Erin M., et al*). These patterns highlight how structural factors like education, income, and regional resource allocation create uneven access to high-quality telehealth, disadvantaging marginalized communities.

Educational attainment and technology access emerged as pivotal determinants of telehealth use. College-educated individuals were 24% more likely to engage in telehealth than those with lower education, reflecting advantages in navigating digital platforms and understanding insurance coverage (*Spaulding, Erin M., et al*). Lower educational attainment correlated with perceived risks (e.g., privacy concerns, technical complexity), reducing adoption willingness (*Wu T., Ho C*).

Importantly, socioeconomic status intersected with race and geography; for instance, Black patients in low-income urban neighborhoods faced dual barriers of limited tech access and systemic underinsurance (*Chen, Evan M., et al*). These findings underscore telehealth's dual role: it can democratize care but also reinforce inequities if not paired with investments in digital literacy and infrastructure (*Spaulding, Erin M., et al*).

This research is vital because telehealth is now a permanent care modality. Without addressing these disparities, the shift to virtual care risks deepening health inequities, particularly for chronic disease management and preventive services. Studies emphasize the need for policies targeting broadband expansion, multilingual platforms, and provider training to ensure equitable access (*Chen, Evan M., et al*).

3. Datasets

We will get the publicly available datasets from the US government for the adoption of telehealth services which could be broken down by different attributes of the communities such as race, gender and educational background. Both of the data sources are accredited US government sites and focus on different issues and we will combine data from both of them to notice the impact of COVID-19 and accessibility to healthcare (via telehealth) for different groups.

1. RANDS (Research and Development Survey) data:
   Access and use of telemedicine during COVID-19 | HealthData.gov. (2021, February 25).
   https://healthdata.gov/dataset/Access-and-Use-of-Telemedicine-During-COVID-19/c835-etjt/about_data

2. Medicare Telehealth Trends dataset:
   Data.gov. (2025, February 3). *U.S. Department of Health & Human Services - Medicare Telehealth Trends*.
   https://catalog.data.gov/dataset/medicare-telemedicine-snapshot?utm_source=chatgpt.com

Trustworthiness of the data sources:

- The RANDS (Research and Development Survey) data explicitly acknowledges its experimental nature and limitations, which demonstrates transparency. It uses probability-sampled commercial survey panels and has documented its

methodology in technical notes. We notice such limitations and focus on healthcare datasets from several other publicly available health datasets for different countries.
- The Medicare Telehealth Trends dataset comes from actual Medicare service utilization records, providing concrete behavioural data rather than self-reported information.

Known limitations of the data sources:

- RANDS documentation openly acknowledges potential biases from different response patterns and sampling frames
- RANDS data has increased variability due to lower sample sizes compared to traditional NCHS surveys
- Medicare data primarily represents older Americans and those with disabilities, potentially underrepresenting other populations

Data cleaning

We used a jupyter notebook to clean, analyse and visualise (EDA) the data. Present in the attached jupyter notebook file.

In both datasets, aggregated categories such as “Total” are being removed as it was checked in “Midway checkpoint” and preliminary research that the the subgroups forms a total set of the larger group.

Furthermore, in both datasets Unkowns or NULLS are replaced with “Uknown” and or “Other” category for the attribute as this will be used to classify missing or unknown data.

For cleaning the ***“Access_and_Use_of_Telemedicine_During_COVID-19.csv***”:

1. Remove all the irrelevant columns such as “Suppression, Significant 1 and Significant 2”, which were given in the Research proposal.
2. We will also convert 'High school graduate or less' to ‘highschool or less’, 'Some college' to ‘college’ and "Bachelor's degree or above" to ‘bachelor or above’ for convenience.
3. We will also ignore the “Urbanisation” (different from point b) and other subgroups besides the one mentioned above as it does not provide any useful information or any further subgrouping required for our research question.

4. Finally, add a column called “size” that shows not only the size instead of the percentage of the population which will make further calculation easier. This will be calculated per SubGroup of the relevant Groups per Indicator.
5. The dataset is **unpivoted** to ensure that it can be matched and joined with the second dataset which includes converting ‘Group’ turn into columns and their values be taken by ‘Subgroups’. This would also include splitting Response type: ['Yes', 'No', 'Do not know', 'No usual place of care', 'No telemedicine available']

For the “***TIMEDTREND_PUBLIC_241126.csv***”:

1. We first filter only the COVID years which is 2020-2022 (inclusive)
2. Look at all the Summary statistics for the numerical rows:
    a. The summary statistics for all the numerical rows make sense as they are non-zero and meet expectations. This is done by comparing it to the values seen in the csv and don’t indicate any major outliers.
3. Look at all the NULL values in the rows:
    a. Almost all columns have complete data and are not NULL so no rows are removed.
4. We then aggregate the data for our relevant purposes which include ‘Bene_Race_Desc’, ‘Bene_RUCA_Desc’, ‘Bene_Mdcd_Mdcr_Enrl_Stus’ for each numerical column. While aggregating, every absolute value is taken as a sum except percentage which is taken as a mean as the summation would not make much sense for a relative metric.
5. We also renamed the column to ‘Race’, ‘Urbanization’, and ‘Enrollement_Status’ for ease of understanding.
6. As during the Midway checkpoint, TA reminded us that one table needs to be have a foreign key to the other table, we have decided to maps NaN missing values to unknown as they fit there the best and during the analysis, these values are most likely to be ignored or treated as missing in both datasets.

After cleaning the dataset, the following Database Schema was created:

For the cleaned version of the ***“Access_and_Use_of_Telemedicine_During_COVID-19.csv***” dataset:

```
CREATE TABLE telemedicineprovider (
  round INT,
  indicator VARCHAR(255),
```

```
  race VARCHAR(255),
  urbanization VARCHAR(255),
  education VARCHAR(255),
  samplesize FLOAT,
  sizevalue FLOAT,

  -- Percent columns
  percent_do_not_know FLOAT,
  percent_no FLOAT,
  percent_no_telemedicine_available FLOAT,
  percent_no_usual_place_of_care FLOAT,
  percent_yes FLOAT,

  -- Standard Error columns
  standarderror_do_not_know FLOAT,
  standarderror_no FLOAT,
  standarderror_no_telemedicine_available FLOAT,
  standarderror_no_usual_place_of_care FLOAT,
  standarderror_yes FLOAT,

  -- Foreign key to beneficiarydata
  FOREIGN KEY (race, urbanization)
    REFERENCES beneficiarydata(race, urbanization)
    ON DELETE SET NULL
    ON UPDATE CASCADE,

  -- Primary Key
  PRIMARY KEY (
    round, indicator, race, urbanization, education
  )
);
```

For the cleaned version of the "***TIMEDTREND_PUBLIC_241126.csv***" dataset file:

```
CREATE TABLE beneficiarydata (
  race VARCHAR(255),
  urbanization VARCHAR(255),
  enrollment_status VARCHAR(255),
  total_bene_th_elig FLOAT,
  total_partb_enrl FLOAT,
  total_bene_telehealth FLOAT,
  pct_telehealth FLOAT,
```

```
  PRIMARY KEY (
    race, urbanization,
  )
);
```

4. Methodology

In this project, we aimed to explore telehealth usage patterns during the COVID-19 pandemic by analyzing two datasets: the RANDS survey, which included self-reported telemedicine usage, and Medicare telehealth utilization data. We specifically examined telehealth adoption rates across racial groups, urban versus rural regions, and different educational backgrounds. To facilitate analysis, we first cleaned and structured the data in Python. After cleaning, we loaded the datasets into an Oracle SQL database by dynamically generating INSERT statements.

Once the data was stored in the database, we used targeted SQL queries with specific filters and JOIN operations to combine relevant information from both datasets efficiently. For instance, one of our key queries combined telehealth usage by race and urbanization level with Medicare beneficiary data:

```
SELECT tp.race, tp.urbanization, AVG(tp.percent_yes), bd.total_bene_telehealth
FROM telemedicineprovider1 tp
JOIN beneficiarydata1 bd ON tp.urbanization = bd.urbanization
WHERE tp.indicator = 'Scheduled one or more telemedicine appointments'
GROUP BY tp.race, tp.urbanization, bd.total_bene_telehealth;
```

The research methodology employed hypothesis testing through carefully designed SQL queries executed in SQLPlus. Results were systematically captured using the “spool <filename>” and then “spool off” after the query, generating consistently formatted text files. These output files were subsequently parsed using a custom Python script that extracted the relevant data elements. The structured extraction approach facilitated comprehensive hypothesis testing on the accumulated data, enabling statistical analysis and evidence-based evaluation of the research questions under investigation.

Hypothesis testing

Our research was driven by three main hypotheses. Firstly, we hypothesized there would be significant disparities in telehealth adoption rates among racial groups compared to their respective shares of the overall population. Secondly, we sought to identify whether telehealth usage differed significantly between urban and rural populations. This relied more on percentage differences and EDA for inference, as we did not have access to more parameters such as the standard error of the population. Lastly, we hypothesized educational attainment would significantly influence telehealth adoption rates.

For the racial comparison, we conducted a series of z-tests. Specifically, we tested whether the proportion of telehealth users in each racial group significantly differed from their representation in the U.S. population according to the 2020 census. Our null hypothesis was that there would be no difference between telehealth usage proportions and population proportions. The z-tests were possible as we had access to the standard error for all the relevant racial demographics, as well as a large enough sample size for the population test to be relevant.

5. Results and Discussion

Our analysis revealed significant racial disparities in telehealth usage compared to population proportions. Notably, Black non-Hispanic respondents had significantly higher telehealth usage (26.00%) than their population share (12.4%), with a z-score of 5.30 and a highly significant p-value (<0.0001). Similarly, the Other non-Hispanic group also showed significantly higher usage (18.63%) compared to their smaller population proportion (7.1%), with a z-score of 4.68. In contrast, White non-Hispanic respondents showed a significantly lower telehealth usage (21.97%) relative to their substantial population share (61.6%), reflected in a large negative z-score of -33.03. Hispanic individuals' telehealth usage (20.27%) was roughly proportional to their population share (18.9%), with no statistically significant difference (p-value = 0.5345).

When analyzing telehealth adoption by urbanization, we observed slightly higher usage rates in urban areas (~24.8%) compared to rural areas (~21.1%). However, this difference was not statistically significant due to limitations in the data. Additionally, education level analyses indicated higher telehealth adoption among those with higher educational attainment, though statistical significance could not be conclusively determined from the available data due to small sample sizes in some education subgroups.

These results highlight critical disparities in telehealth adoption during the pandemic, reflecting broader patterns of healthcare access and equity. These results may be limited in their scope due to the limited data and as it is based on secondary evidence (lack of knowledge on the details of the surveys and tests conducted), it still agrees with the consensus of the other results measured in relevant studies (*Chen, Evan M., et al)*.

6. Discussion of Results

7. Future work

8. References

Luo, Jake, et al. “Telemedicine Adoption During the COVID-19 Pandemic: Gaps and Inequalities.” *Applied Clinical Informatics*, vol. 12, no. 04, Aug. 2021, pp. 836–44. https://doi.org/10.1055/s-0041-1733848.

Chen, Evan M., et al. “Socioeconomic and Demographic Disparities in the Use of Telemedicine for Ophthalmic Care During the COVID-19 Pandemic.” *Ophthalmology*, vol. 129, no. 1, July 2021, pp. 15–25. https://doi.org/10.1016/j.ophtha.2021.07.003.

Spaulding, Erin M., et al. “Prevalence and Disparities in Telehealth Use Among US Adults Following the COVID-19 Pandemic: National Cross-Sectional Survey.” Journal of Medical Internet Research, vol. 26, May 2024, p. e52124. https://doi.org/10.2196/52124.

Wu T., Ho C. “Telemedicine Adoption During the COVID-19 Pandemic: Gaps and Inequalities.” Applied Clinical Informatics, vol. 12, no. 04, Aug. 2021, pp. 836–44. https://doi.org/10.1055/s-0041-1733848.

9. AI Declaration

We have not used any AI or GenAI tools in this research project so far, All the steps have been completed with the materials learned in class or researched online (Links provided above).

- Aarav, Pushya, Suryansh