horizontal line

**TrueDataAnalysis**

123 Street Court

City, ST 12345

(123) 456-7890

**State-level Utilization of Telemedicine Prior to and During COVID-19**

# 

# Data Analysis conducted by:

**TrueDataAnalysis**

**Analysis Conducted by:**

**Kenny**

**Jason**

**Zainab**

# Table of Contents

1. Introduction 3-4
   1. Summary 3
   2. Scope 3-4
   3. Technologies and resource contributions 4
   4. Definitions, Acronyms and Abbreviations 4
2. ETL Details 5-7
   1. Data Import/Extract Sources and Methods 5
   2. Data Acquisition 5
   3. Data Transform 6
   4. Data Integrity 6
   5. Data Refresh Frequency 6
   6. Data Security 7
   7. Data Loading and Availability 7
3. Data Quality 7

# 

# Introduction

**1.1 Summary**

The ongoing outbreak of COVID-19 in the United States caused a severe reduction in utilization of healthcare services due to perceived risk of getting infected by the disease. Moreover, this disease also led to widespread household-level economic devastation, which may further reduce access to healthcare due to severance from employer-sponsored plans due to job losses, reduced incomes, and reduced discretionary spending (including on healthcare) due to an uncertain economic climate. Interactions between healthcare providers and patients may change substantially as states continue to ease restrictions on movement and other related measures (e.g., mask-wearing, social-distancing, etc.). Telemedicine provides an effective way for patients to communicate with providers safely and increase their access to care during this pandemic. This project utilized publically available data sources to examine the changes of people’s search interests in telemedicine before and during the breakout of COVID-19 at state level in the United States.

**1.2 Scope**

# The Johns Hopkins University (<https://coronavirus.jhu.edu/us-map>) aggregates data from multiple credible sources to track the spread of COVID-19. Their data can be used to track the trend of COVID-19 confirmed cases, deaths, and fatality rate. We mainly looked at state level COVID-19 infection rates (per 100,000 population) and stratified all 50 states into three categories: low infection states, middle infection states, and high infection states.

# The google trend data provides a largely unfiltered sample of actual search requests made to Google. In our case, we used the keyword, i.e. “telemedicine”, to examine the changes of search interests before and during COVID-19 pandemic.

# Specifically, google trend results were normalized to the time and location of a query. Each data point was divided by the total searches of the geography and time range. Therefore, it represented relative popularity. The resulting numbers were then scaled on a range of 0 to 100 based on a topic’s proportion to all searches on all topics. Different regions that showed the same search interest for a term don't always have the same total search volumes.

# 1.3 Technologies and resource contributions

In order for the team members to perform the necessary research, we used google trend and John Hopkins github repository to obtain the datasets. We also performed data transformation using the python applications to clean up the data. We used the SQL Database to perform merging of both datas and also queried the database for necessary information we needed from the tables compiled.

All three analysts (Kenny, Jason, Zainab): Discussed, extracted data, transformed data, and contributed to every aspect of the research, final report write-up.

**1.4 definitions, Acronyms and Abbreviations**

|  |  |
| --- | --- |
| JHU - Johns Hopkins University | SQL - Structured Query Language |
| CSV - Comma-Separated Values | ETL - Extract, Transform and Load |
| PD - Pandas |  |

# ETL Details

**2.1 Data import/Export Sources and Method**

The COVID-19 data was retrieved from the JHU github repository, i.e. <https://github.com/CSSEGISandData/COVID-19.git>, which is publically available. We used the state level number of infections per 100,000 people as the key parameter, and stratified states into three groups by infection rates: low ⅓ quartile, middle ⅓ quartile, and high ⅓ quartile.

The google trend data was obtained from the website: <https://trends.google.com/trends/?geo=US>, which is also publically available. We defined 01/01/2019 - 12/31/2019 as the pre-COVID-19 period and 03/01/2020 - 05/31/2020 as the time period during COVID-19.

**2.2. Data Acquisition**

To do our research we collected our data using Google Trends for people’s search interests on “telemedicine” and JHU repository for COVID-19 Data. Because we narrowed down on behavioral interactions, we used a date range. We compared the general trend from last year (2019) to the trend following the early outbreak of COVID-19 (3/1/20 - 5/31/20). Because of constant fluctuation of cases and increasing knowledge of telemedicine, the data should be updated on a biweekly to monthly basis. With this in mind the data should be taken as it is, a trend not a definite indication of usage. It is important to understand the state which is being considered, as well as the likelihood that the market will remain following the resolution of this pandemic. It is with these factors in mind that we made our statements. There are a few things that seem to be consistent. States which are more knowledgeable of telemedicine, remained close to their original ranking. If some more time passes the data should become more closely spread as more information is passed along. The scores from the three month spread are more closely related compared to the trend of 2019.

**2.3. Data Transformation**

Prior to the data merge we had to filter the data from the JHU’s repository. Their data set included the U.S. territories which we will not be accounting for. So once we dropped those values then we proceeded with the merge. The merge took the two datasets (prior to and during COVID-19) and placed them in one dataframe. Following the data merge the created table represented all 50 states’ search trend based on the keyword “telemedicine”, with a relativity score compared to the District of Columbia (score = 100 in both time periods). Because there were no missing values at state level, we correlated the state level COVID-19 infection rate with google trend data, stratified by low, middle, and high infection rates.

**2.4. Data integrity**

The COVID-19 data is from a reliable source which is updated and validated constantly, both of the data sets are available in CSV format which allows integration and manipulation to be. However while the Google Trend data is accurate, it is also just an indication of trends. Meaning while people could be using the keywords which we have indicated, that doesn’t necessarily mean that they are acting on these searches. The data sources are constantly updated, however it is not necessary to update the local data at the same frequency. It would be more beneficial to update the information on a biweekly to monthly basis as stated previously. This would be more telling also due to the COVID-19 case fluctuations.

**2.5. Data refresh frequency**

The state level COVID-19 infection rates can be updated daily; although the google trend data can be updated daily too, we expect little changes from day to day google searches at state level. Therefore, we will update the google search research of “telemedicine” bi-weekly.

**2.6. Data Security**

Because both of our data sources are publically available, i.e. the JHU’s github repository and google trend results, we did not have to build additional procedures in data privacy, encryption, or data masking etc. Moreover, we only examined the state level data, therefore, our analyses can not be used to identify any specific individuals, and therefore, was exempted from IRB reviews. We stored our data on google drive, and shared a copy with our client. We backuped the data using google cloud.

**2.7. Data Loading and Availability**

The data schema was in the form of merging COVID-19 data with google trend data for 50 states and generated two heat-maps. These maps were readily available for download and interaction upon pulling the data. Because the data will be stored in the google cloud it can be accessed at any point in time.

1. **Data Quality**

To evaluate the success of this project, following criteria should be considered:

1. Observe statistically significant changes in search interests of “telemedicine” before and during COVID-19 at state level;
2. Identify states that with search interests changed significantly higher compared to other states; also use t-test to compare the means of relative search scores by low, middle, and high groups of states;
3. Visualize the changes of search interests on telemedicine at state level;
4. Make informed decision about how telemedicine is affecting patients and caregivers in different states, and may associate this pattern with state level health policy, e.g. reimbursement for telemedicine in Medicaid.