Databases Final Project: Phase 1

1. Our group consists of Lucy Zhang (615) and Nader Najjar (415).

2. Our goal is to explore the various impacts of COVID-19 on Health, Economy, and Presidential Election Voting Trends in 2020. We also aim to create visualization tools for cases of COVID-19 vs. Voting Trends separated by state.

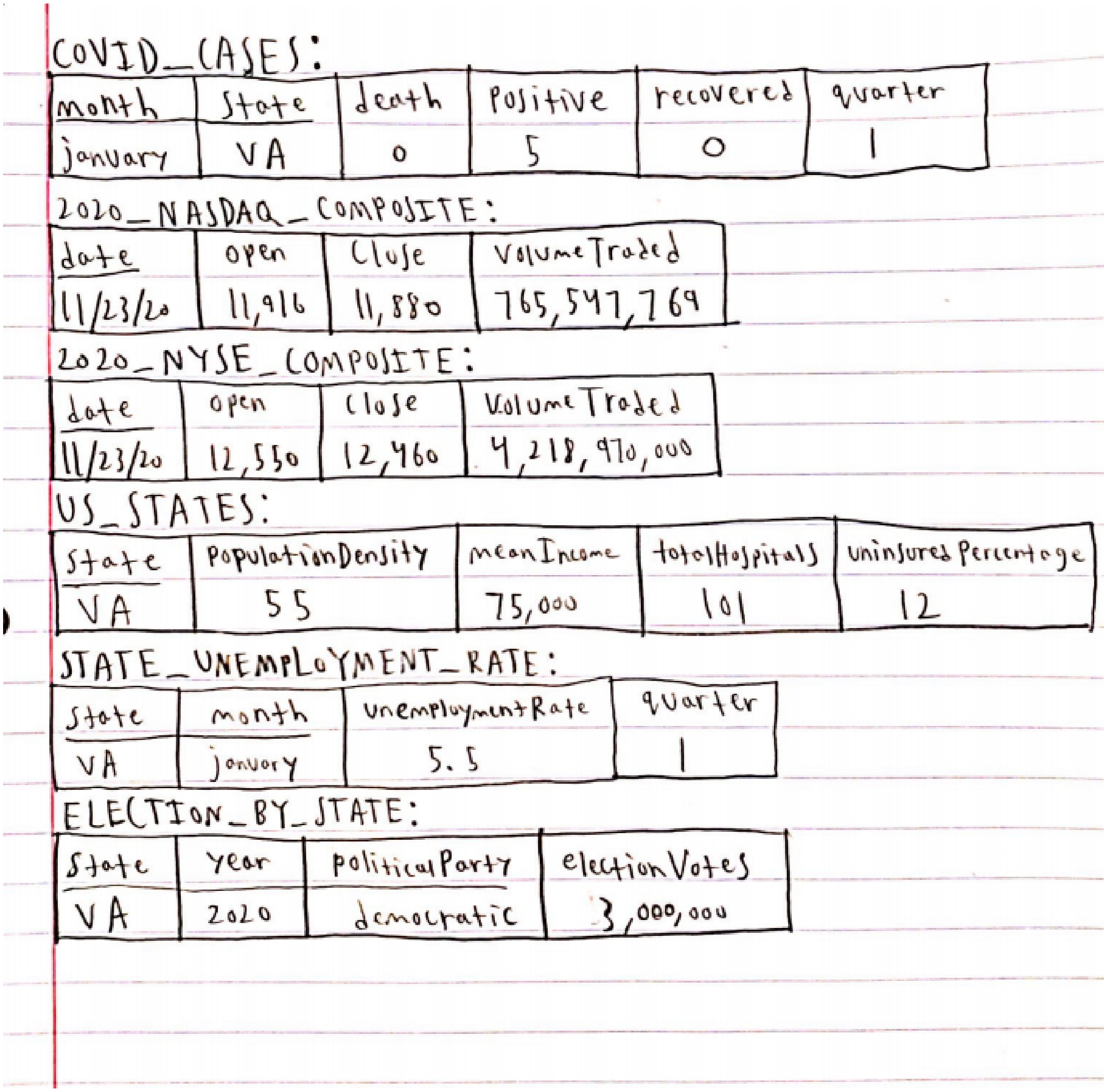
3. English Questions
    a. What state has the most COVID-19 cases in January? In the first quarter? Up until now?
    b. What percentages of vote were casted for the Democratic Party in Michigan in 2016 vs. 2020?
    c. How does the increase in COVID-19 cases each month affect the turnout of NASDAQ composite?
    d. Which industry is more impacted by COVID-19, traditional industries (NYSE) or the technology industry (NASDAQ) ?
    e. Are the COVID-19 cases affected by the state's population density?
    f. Print out the percentage of increase of unemployment rate in each state, along with the number of positive COVID-19 cases for each month.
    g. Among the top ten states with the most unemployment rate, print out the political party they voted for in 2016 vs. 2020.
    h. Does the number of hospitals affect the number of COVID-19 rated deaths in a state?
    i. Print the voting trend for ten states with the highest mean household income and the voting trend for ten states with the lowest mean household income in 2016 vs. 2020.
    j. Discover whether there is a trend of economic recovery by looking at the NASDAQ and NYSE open and closing prices, traded volumes for the first quarter vs. the fourth quarter. Which industry shows faster recovery?
    k. Does health insurance coverage percentage affect the number of COVID-19 positive cases for each state? The number of COVID-19 related deaths? The number of recovered patients?
    l. Print out the unemployment rate by quarter, aggregated by every state, along with the open and closing prices, traded volumes for NASDAQ and NYSE for every quarter. Does this show a trend of economic impact and recovery with the progression of time in 2020?
    m. Which state had the most increase in the number of overall presidential election votes in 2020, compared to the number of votes in 2016?

n. Print out how each state voted for the 2020 Presidential Elections, with the amount of vote casted with the winning Party and the amount of vote casted for all other parties

o. Which state has the most difference in the percentages of votes casted for Democratic Party vs. Republican Party in the 2020 Election?

p. Compare the greatest drop in both the NASDAQ and NYSE indices during January/February to the greatest drop during March/April

4.



COVID_CASES:

month	State	death	Positive	recovered	quarter
january	VA	0	5	0	1

2020_NASDAQ_COMPOSITE:

date	open	Close	VolumeTraded
11/23/20	11,916	11,880	765,547,769

2020_NYSE_COMPOSITE:

date	open	Close	VolumeTraded
11/23/20	12,550	12,460	4,218,970,000

US_STATES:

State	PopulationDensity	meanIncome	totalHospitals	uninsuredPercentage
VA	55	75,000	101	12

STATE_UNEMPLOYMENT_RATE:

State	month	unemploymentRate	quarter
VA	january	5.5	1

ELECTION_BY_STATE:

State	year	politicalParty	electionVotes
VA	2020	democratic	3,000,000

DDL:

```
CREATE TABLE COVID_CASES(
    Month VARCHAR(20)  NOT NULL,
    Quarter INT NOT NULL,
    State VARCHAR(20) NOT NULL,
    Death INT NOT NULL,
    Positive INT NOT NULL,
    Recovered INT NOT NULL
    PRIMARY KEY(Month, Quarter, State)
);

CREATE TABLE ELECTION_BY_STATE(
    State VARCHAR(20) NOT NULL,
    Year INT NOT NULL,
    Political_Party VARCHAR(20) NOT NULL,
    Election_Votes INT
        PRIMARY KEY(State, Year, Political_Party)
);



CREATE TABLE 2020_NASDAQ_COMPOSITE(
    date DATE NOT NULL,
    Open INT NOT NULL,
    Close INT NOT NULL,
    Volume_Traded INT NOT NULL
    PRIMARY KEY(date)
    );

CREATE TABLE 2020_NYSE_COMPOSITE(
    date DATE NOT NULL,
    Open INT NOT NULL,
    Close INT NOT NULL,
    Volume_Traded INT NOT NULL
        PRIMARY KEY(date)
    );

CREATE TABLE US_STATES(
    State VARCHAR(20) NOT NULL,
    Population_Density DOUBLE NOT NULL,
    Mean_Household_Income INT NOT NULL,
    Total_Hospitals INT NOT NULL,
    Uninsured_Percentage DOUBLE NOT NULL
    PRIMARY KEY(State)
    );

CREATE TABLE STATE_UNEMPLOYMENT_RATE(
    State VARCHAR(20) NOT NULL,
    Month VARCHAR(20) NOT NULL,
    Quarter INT NOT NULL,
    Unemployment_Rate DOUBLE NOT NULL
    PRIMARY KEY(State, Month)
    );
```

5. SQL Statements

Query for Part 3 Question b:

```
SELECT round(D_Vote_2016.Votes_2016/Votes_2016.All_Votes_2016,3) as percent_vote_2016,
round(D_Vote_2020.Votes_2020/Votes_2020.All_Votes_2020,3) as percent_vote_2020
FROM
(SELECT SUM(Election_Votes) as All_Votes_2016
FROM Election_by_State as E
WHERE E.State = "Michigan" AND E.Year = 2016) as Votes_2016,
(SELECT Election_Votes as Votes_2016
FROM Election_by_State as E
WHERE E.State = "Michigan" AND E.Year = 2016 AND E.Political_Party = "Democratic") as D_Vote_2016,
(SELECT SUM(Election_Votes) as All_Votes_2020
FROM Election_by_State as E
WHERE E.State = "Michigan" AND E.Year = 2020) as Votes_2020,
(SELECT Election_Votes as Votes_2020
FROM Election_by_State as E
WHERE E.State = "Michigan" AND E.Year = 2020 AND E.Political_Party = "Democratic") as D_Vote_2020
```

Query for Part 3 Question k:

```
SELECT T.death, T.positive, T.recovered, T1.Uninsured_Percentage, T.State
FROM
(SELECT SUM(C.death), SUM(C.positive), SUM(C.recovered), C.State
FROM COVID_CASES as C
GROUP BY C.State) as T,
(SELECT U.Uninsured_Percentage
FROM US_STATES as U) as T1
WHERE T1.State = T.State
```

Query for Part 3 Question i

```
create view VOTING_TREND_BY_STATE AS
SELECT E.State, M.Max_Vote, M.Year
FROM
(select MAX(E.ElectionVotes) as Max_Vote, E.State, E.Year
from ELECTION_BY_STATE AS E
GROUP BY E.State) as Max_Vote_Cnt as M, ELECTION_BY_STATE AS E
WHERE M.Max_Vote = E.Election_Votes AND M.State = E.State AND E.Year = M.Year;

SELECT V.State, V.Year, V.Winning_Party
FROM
((SELECT U.State
FROM US_STATES as U
ORDER BY ASC
LIMIT 10)
UNION
(SELECT U.State
FROM US_STATES as U
ORDER BY DESC
LIMIT 10)) as States, VOTING_TREND_BY_STATE as V
WHERE V.State = States.State
```

Query for Part 3 Question o

```
SELECT DIFF.State, DIFF.percent_diff
FROM
(SELECT MAX(ABS(T.Percent_D-T.Percent_R)) as max_diff
FROM
(SELECT E1.State, E1.Election_Votes/E3.Election_Votes as Percent_D, E2.Election_Votes/E3.Election_Votes as
Percent_R
FROM ELECTION_BY_STATE as E1, ELECTION_BY_STATE as E2,
(SELECT SUM(E3.Election_Votes) as All_Votes, E3.State
FROM ELECTION_BY_STATE as E3
GROUP BY E3.State
) as E3
WHERE E1.State = E2.State AND E1.Political_Party = "Democratic" AND E2.Political_Party = "Republican") as
T) as MAX,

(SELECT T.State, ABS(T.Percent_D - T.Percent_R) as percent_diff
FROM
(SELECT E1.State, E1.Election_Votes/E3.Election_Votes as Percent_D, E2.Election_Votes/E3.Election_Votes as
Percent_R
FROM ELECTION_BY_STATE as E1, ELECTION_BY_STATE as E2,
(SELECT SUM(E3.Election_Votes) as All_Votes, E3.State
FROM ELECTION_BY_STATE as E3
GROUP BY E3.State
) as E3
WHERE E1.State = E2.State AND E1.Political_Party = "Democratic" AND E2.Political_Party = "Republican") as
T) as DIFF

WHERE DIFF.percent_diff = MAX.max_diff
```

6. For COVID-19 data, we are using the data provided by the government at healthdata.gov, with link https://healthdata.gov/dataset/united-states-covid-19-cases-and-deaths-state-over-time. We are looking to download available data into CSV, and conduct basic data aggregation and pre-processing with excel/python/php to create the dataset needed for our database, and then insert into SQL. The Election and Voting Trends data could be extracted from Kaggle. The NASDAQ and NYSE Composite datasets are extracted from finance.yahoo.com, and we will also look for datasets on Kaggle to cross-validate the reliability of our data sources. Finally, US states information regarding health and

wellness are extracted from Kaiser Family Foundation, and the rest of the data are extracted from local government sites.

7. Our goal is to first create a personalized visualization tool that would provide information on basic and COVID-19 related information. For example, a user can select Maryland, and view the information about hospital numbers, COVID-19 cases, impact to economies, voting trends, etc. and would be able to view the visualization of such information. A user can also select, for example, the first quarter of 2020, and see changes with regards to NASDAQ and NYSE statistics, as well as COVID-19 infection rate for that quarter. This tool will obtain tables directly from SQL databases and have plotting functions. We hope that by implementing such visualization tools, discoveries of relations between COVID-19 to statewide and nationwide economic impact are more recognizable, and we would be potentially be able to connect such impact with the outcome of the 2020 Presidential Election by providing easy access for questions such as the outcome of swing state votes for 2016 Election vs. 2020 Election, whether the economic impact in 2020 had any correlation with the outcome of the Election, etc.
8. We are not 100% sure which focuses we will prioritize yet, but we came up with a list of potential options:
   possible major focus:
   - Advanced GUI interface: We are planning on creating specific visualization tools for our data/queries (such as hotmaps, graphs,....) to be displayed on the web interface, as well as some other web-specific features.
   - Complex data extraction: We may decide to create a script that automatically retrieves updated data related to COVID, the stock market, etc. from an online source and inserts it into the database once a day (ensuring that the procedures/queries utilize this updated data as well).
   - Advanced SQL topics: We may implement specific triggers and JDBC to make the database more seamless and interactive. In our stored procedures, we may also utilize cursors, exceptions, packages, and other advanced topics.

   possible minor focus:
   - Data mining: We plan to implement potential data models such as Linear Regression or Random Forest to discover the important features related to, for example, the outcome of the 2020 Presidential Election.
   - Any of the major focuses above that we do not select to be the major focus