Databases Project Phase 1

1 Our team consists of Justin Song (jsong69) and Irfan Jamil (ijamil1). We are both in 315.

2 Our domain of interest is political data. Namely, we will be working with historical US federal election data, presidential approval data, and polling data. With respect to the US federal election data, we are interested in a variety of questions such as touching up midterm results for the party in the White House, historical battleground state results, and states that had a “split party vote” (ie: Presidential contest went in favor of the Republican candidate, yet Senate races and House races were favorable for Dems). Additionally, we are interested in exploring approval ratings between presidents over the course of their presidency and after major events in their respective presidencies (ie: Hurricane Katrina for Bush 43 and Covid-19 for Trump). Finally, and perhaps the most interesting, we intend to explore how final polling forecasts compare to actual results in House, Senate, and Presidential races in the past few election cycles. Note, for polling forecasts, we plan to use FiveThirtyEight’s forecasts (if data available or their compilation of polls by various pollsters).

3

I) List all of the Senatorial candidates who were favored to win, but lost between the years of 2000 – 2020
II) List the states which have elected a Democrat in each of the presidential contests since 1976.
III) List the states which have elected a Republican in each of the presidential contests since 1976.
IV) List the approval ratings of Trump and Obama in their first month in Office.
V) List the approval ratings of Trump and Obama in their last month in Office.
VI) Compare the approval ratings of Bush 43 and Trump during the first 2 months after each of the following events in their respective presidencies: Lehman Brothers collapse, COVID-19 declared as a national emergency
VII) For each president since 1976, how did the opposing party perform in midterm elections?
VIII) List all of the pollsters whose final polling had the eventual winner (Trump) ahead in the PA 2016 race.
IX) List all of the pollsters whose final polling had the eventual winner (Trump) ahead in WI, PA, and MI.
X) What states voted for Obama in 2012, Trump in 2016, and Biden in 2020?
XI) Were there any states whose voter turnout dropped from 1992 to 1996?
XII) Has California sent any Republicans to the Senate in the 21st century? List the names of the Senators?
XIII) What states, in 2016, elected more representatives of one of the major political parties to the House but elected the presidential candidate of the other party?
XIV) In which states is the average polling lead for the eventual winner of the state in 2016 less than the actual margin of victory?
XV) Of the Senatorial candidates who were favored to win but lost in 2016, what percentage were democrats? In 2020?

4

```
CREATE TABLE PresidentHistoricalResults (
Year        INT             NOT NULL
State       VARCHAR(30)     NOT NULL
Winner      VARCHAR(40)     NOT NULL
Party       VARCHAR(30)     NOT NULL
Candidatevotes  INT         NOT NULL
Totalvotes      INT         NOT NULL
PRIMARY KEY (Year, State, Winner));
```

```
CREATE TABLE SenateHistoricalResults (
Year        INT             NOT NULL
State       VARCHAR(30)     NOT NULL
Winner      VARCHAR(40)     NOT NULL
Party       VARCHAR(30)     NOT NULL
Candidatevotes  INT         NOT NULL
Totalvotes      INT         NOT NULL
PRIMARY KEY (Year, State, Winner));


CREATE TABLE HouseHistoricalResults (
Year        INT             NOT NULL
State       VARCHAR(30)     NOT NULL
Winner      VARCHAR(40)     NOT NULL
Party       VARCHAR(30)     NOT NULL
Candidatevotes  INT         NOT NULL
Totalvotes      INT         NOT NULL
PRIMARY KEY (Year, State, Winner));

CREATE TABLE PresidentialApproval (
Name            VARCHAR(30)     NOT NULL
Inauguration    DATE            NOT NULL
StartDate       DATE            NOT NULL
EndDate         DATE            NOT NULL
Approval        DECIMAL         NOT NULL
Disapproval     DECIMAL         NOT NULL
PRIMARY KEY (Name, StartDate, EndDate));

CREATE TABLE 2016PresidentialPolls (
Date        DATE            NOT NULL
State       VARCHAR(30)     NOT NULL
Pollster    VARCHAR(50)     NOT NULL
Clinton     DECIMAL         NOT NULL
Trump       DECIMAL         NOT NULL
PRIMARY KEY (Date, State, Pollster));

CREATE TABLE 2020PresidentialPolls (
Date        DATE            NOT NULL
State       VARCHAR(30)     NOT NULL
Pollster    VARCHAR(50)     NOT NULL
Biden       DECIMAL         NOT NULL
Trump       DECIMAL         NOT NULL
PRIMARY KEY (Date, State, Pollster));

CREATE TABLE 2020SenatePolls (
Date            DATE            NOT NULL
State           VARCHAR(30)     NOT NULL
Pollster        VARCHAR(40)     NOT NULL
Candidate       VARCHAR(40)     NOT NULL
Party           VARCHAR(20)     NOT NULL
Percentage      DECIMAL         NOT NULL
PRIMARY KEY (Date, State, Pollster, Candidate));
```

CREATE TABLE 2018SenatePolls (

Date	DATE	NOT NULL
State	VARCHAR(30)	NOT NULL
Pollster	VARCHAR(40)	NOT NULL
Candidate	VARCHAR(40)	NOT NULL
Party	VARCHAR(20)	NOT NULL
Percentage	DECIMAL	NOT NULL

PRIMARY KEY (Date, State, Pollster, Candidate));

Omitting other polling data relations for brevity since the polling data in other cycles will have identical attributes/schema.

5) Some SQL Queries (the number the query refers to is listed before the actual SQL query itself)

- (ii) (SELECT DISTINCT ph.State FROM PresidentHistoricalResults as ph) EXCEPT (SELECT DISTINCT ph.State FROM PresidentHistoricalResults as ph WHERE ph.Party = 'democratic';
- (iii) (SELECT DISTINCT ph.State FROM PresidentHistoricalResults as ph) EXCEPT (SELECT DISTINCT ph.State FROM PresidentHistoricalResults as ph WHERE ph.Party = 'republican';
- (iv) SELECT 100End.PresidentName, AVG(pa2.Approval) as AverageApproval FROM (SELECT pa.Name as PresidentName, DATEADD(month, 1, pa.Inauguration) as 1Month FROM PresidentialApproval as pa WHERE pa.Name = 'Barack Obama' OR pa.Name = 'Donald Trump') as 1Temp, PresidentialApproval as pa2 WHERE pa2.Name = 1Temp.PresidentName AND pa2.StartDate >= pa2.Inauguration AND pa2.EndDate <= 1Temp.1Month GROUP BY pa.Name;
- (v) SELECT finaltemp.PresidentName, AVG(pa3.Approval) as AverageApproval FROM (SELECT pa2.Name as PresidentName, DATEADD(month, -1, EndOfTerm.TempEndDate) as TempStart, EndOfTerm.TempEndDate FROM (pa.Name,SELECT MAX(pa.EndDate) as TempEndDate FROM PresidentialApproval as pa WHERE pa.Name = 'Barack Obama' OR pa.Name = 'Donald Trump') as EndOfTerm, PresidentialApproval as pa2 WHERE pa2.Name = EndOfTerm.Name) as finaltemp, PresidentialApproval as pa3 WHERE pa3.Name = finaltemp.PresidentName AND pa3.StartDate >= finaltemp.TempStart AND pa3.EndDate <= finaltemp.TempEndDate;
- (vi) (SELECT 100End.PresidentName, AVG(pa2.Approval) as AverageApproval FROM (SELECT pa.Name as PresidentName, DATEADD(month, 2, '9/15/2008') as 1Month FROM PresidentialApproval as pa WHERE pa.Name = 'George W. Bush') as 1Temp, PresidentialApproval as pa2 WHERE pa2.Name = 1Temp.PresidentName AND pa2.StartDate >= pa2.Inauguration AND pa2.EndDate <= 1Temp.1Month GROUP BY pa.Name) UNION ALL (SELECT 100End.PresidentName, AVG(pa2.Approval) as AverageApproval FROM (SELECT pa.Name as PresidentName, DATEADD(month, 2, '01/31/2020') as 1Month FROM PresidentialApproval as pa WHERE pa.Name = 'Donald Trump') as 1Temp, PresidentialApproval as pa2 WHERE pa2.Name = 1Temp.PresidentName AND pa2.StartDate >= pa2.Inauguration AND pa2.EndDate <= 1Temp.1Month GROUP BY pa.Name);
- (vii) SELECT phr.Winner FROM PresidentHistoricalResults as phr GROUP BY phr.Year
- (viii) SELECT p.Pollster, MAX(p.Date) FROM 2016PresidentialPolls WHERE p.Date = (SELECT MAX(p.Date) FROM 2016PresidentialPolls as temp GROUP BY temp.Pollsters) WHERE p.Trump > p.Clinton AND p.State = 'PA' GROUP BY p.Pollster;
- (ix) (SELECT p.Pollster, MAX(p.Date) FROM 2016PresidentialPolls WHERE p.Date = (SELECT MAX(p.Date) FROM 2016PresidentialPolls as temp GROUP BY temp.Pollsters) WHERE p.Trump > p.Clinton AND p.State = 'PA' GROUP BY p.Pollster) INTERSECT (SELECT p.Pollster, MAX(p.Date) FROM 2016PresidentialPolls WHERE p.Date = (SELECT MAX(p.Date) FROM 2016PresidentialPolls as temp GROUP BY temp.Pollsters) WHERE p.Trump > p.Clinton AND p.State = 'WI' GROUP BY p.Pollster) INTERSECT (SELECT p.Pollster, MAX(p.Date) FROM 2016PresidentialPolls WHERE p.Date = (SELECT MAX(p.Date) FROM 2016PresidentialPolls as temp GROUP BY temp.Pollsters) WHERE p.Trump > p.Clinton AND p.State = 'MD' GROUP BY p.Pollster);
- (x) (SELECT phr.State FROM PresidentHistoricalResults as phr WHERE phr.Year = 2012 AND phr.Winner = 'Obama') INTERSECT (SELECT phr.State FROM PresidentHistoricalResults as phr WHERE phr.Year =

2016 AND phr.Winner = 'Trump') INTERSECT (SELECT phr.State FROM PresidentHistoricalResults as phr WHERE phr.Year = 2020 AND phr.Winner = 'Biden');

- (xii) SELECT s.Winner FROM SenateHistoricalResults as s WHERE s.Year >= 2000 AND s.State = 'CA' AND s.Party = 'republican';
- (xiii) SELECT phr.State (SELECT hhr.State, hhr.Party FROM (SELECT temp.State, COUNT(temp.Party) as PartyCount FROM (SELECT h.State, h.Party FROM HouseHistoricalResults as h WHERE h.Party = 'republican' OR h.Party = 'democratic' AND h.Year = 2016 GROUP BY h.State ORDER BY PartyCount DESC LIMIT 1) as temp2, HouseHistoricalResults as hhr WHERE hhr.Party = 'republican' OR hhr.Party = 'democratic' AND hhr.Year = 2016 GROUP BY HAVING COUNT(hhr.Party) = temp2.PartyCount) as finaltemp, PresidentHistoricalResults as phr WHERE phr.State = finaltemp.State AND phr.Party != finaltemp.Party);

6 Our datasets are formatted in csv files. We plan on using a PHP/MySQL interface similar to what was implemented in HW 3. We have found that there are a variety of methods to import a csv file into a MySQL table, so we don't expect much trouble importing the data into a SQL environment. **********TODO********

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/42MVDX (historical POTUS election data)

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/IG0UN2 (historical House election data)

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PEJ5QU (historical Senate election data)

https://www.presidency.ucsb.edu/statistics/data/presidential-job-approval (historical POTUS approval/disapproval ratings)

https://projects.fivethirtyeight.com/polls/senate/ (2020 Senate polls FiveThirtyEight)

https://projects.fivethirtyeight.com/2020-election-forecast/ (2020 Presidential Polls FiveThirtyEight)

https://projects.fivethirtyeight.com/2016-election-forecast/national-polls/ (2016 Presidential Polls FiveThirtyEight)

https://elections.huffingtonpost.com/pollster/api/v2 (We will be using this API to extract polling data for past elections which FiveThirtyEight hasn't released data on).

https://drive.google.com/file/d/1PB9e1-gPTFcVuI-MLpxnEhrJTVK9Ebn0/view?usp=sharing (2000-2010 US State Populations)

https://drive.google.com/file/d/10B5V1CmpI6hENNm4jMOx09BgkPIwe-_K/view?usp=sharing (2010-2019 US State Populations)

7 I think that we would like to visualize the data by using some kind of visual of the USA and allowing users to click specifically on locations to see graphs and specific data corresponding to the location the user is looking for. In general, I think that some views can be created such as a view with some MAJOR events that happened in the US the past few decades alongside dates and the President's name. This would make it easier to find interesting information corresponding to special historical events in the US.

8 A special, advanced topic we plan to focus on is data mining/visualization. Political data lends itself well to charts, graphs, and the like. Some simple examples/scenarios of how and when we would actually use data mining include: plotting presidential approval ratings between various presidents in a time-series line-graph, plotting FiveThirtyEight's final polling forecasts against the actual results for a race in a bar graph for multiple races at a time (ie: could be a user input to determine which races/forecasts to compare), and plotting a pie chart of the political parties and the percentage of races that they have won since some baseline year for a given state. These are just a subset of the myriad of data mining and visualization routes that we can explore. Another topic we will likely explore is the advanced SQL topics of cursors, stored procedures, and views derived from the raw data.