Jonathan Burger

**Database Project**

**Goal:**

Examine the effect of the election on stock prices.

Look at daily stock data for S&P 500 companies from the week before and the week after the election.

Calculate the slopes of each based on percent change from start. Compare before & after slopes and see if they increased or decreased and by what factor.

Compare company headquarters to election data, seeing if there is a relationship between stock price change and if county of company headquarters voted from Clinton or Trump.

Additionally look at company industries to see if certain industries performed better or worse than others.

**Datasets:**

Stock data (from api):

<http://ichart.finance.yahoo.com/table.csv?s=YHOO&d=10&e=15&f=2016&g=d&a=10&b=2&c=2016&ignore=.csv>

Election results:

<https://github.com/mkearney/presidential_election_county_results_2016/blob/master/pres16results.csv>

City, State to County (Now seems to have been removed from website)

<http://download.geonames.org/export/zip/>

Company List

<https://en.wikipedia.org/wiki/List_of_S%26P_500_companies>

**Schema:**

*s\_and\_p\_500.csv*

***s\_and\_p***

*symbol,company\_name,sector,sub\_industry,city,state,date\_added,cik*

*CREATE TABLE* ***s\_and\_p*** *(  
 symbol VARCHAR(10) NOT NULL,  
 company\_name VARCHAR(50) NOT NULL,  
 sector VARCHAR(50) NOT NULL,  
 sub\_industy VARCHAR(50) NOT NULL,  
 city VARCHAR(50) NOT NULL,  
 state VARCHAR(50) NOT NULL,  
 date\_added DATE,  
 CIK INT,   
 PRIMARY KEY (symbol),  
 FOREIGN KEY (city, state)  
 REFERENCES cities(city, state)  
);*

*us\_city\_list.csv*

***cities***

*zip,city,state,state\_code,county,county\_code,latitude,longitude*

*CREATE TABLE* ***cities*** *(  
 zip CHAR(5) NOT NULL,  
 city VARCHAR(50) NOT NULL,  
 state VARCHAR(50) NOT NULL,  
 state\_code CHAR(2) NOT NULL,  
 county VARCHAR(50) NOT NULL,  
 county\_code INT(2) NOT NULL,  
 latitude VARCHAR(50) NOT NULL,  
 longitude VARCHAR(50) NOT NULL,  
 PRIMARY KEY (city, state)  
);*

*pres16\_results\_county.csv*

***pres\_results***

*county, fips, cand, st,pct\_report, votes,total\_votes, pct, lead*

*CREATE TABLE* ***pres\_results*** *(  
 county VARCHAR(50) NOT NULL,  
 fips INTEGER(10) NOT NULL,  
 cand VARCHAR(50) NOT NULL,  
 st CHAR(2) NOT NULL,  
 pct\_report NUMERIC NOT NULL,  
 votes INT(7) NOT NULL,  
 total\_votes INT(7) NOT NULL,  
 pct NUMERIC NOT NULL,  
 lead VARCHAR(50) NOT NULL  
);*

*CREATE TABLE stocks(  
symbol varchar(10) NOT NULL,  
slope\_diff NUMERIC NOT NULL  
)*

**Steps**

1. Obtain data
2. Create tables using *create.sql*
3. Clean data using OpenRefine, specifically matching up similar county & city names
4. Using Java, import csv data sets and fetch, calculate, and import stock data and corresponding slopes
5. Create new tables using queries
   1. Table results\_clean created using query *results\_clean.sql*
   2. Table companies created using query *companies.sql*
6. Run queries in Sqlite and MySQL
7. Create graphs in R

**Queries:**

1. Average slope by county, ordered by winner, pct\_lead

*query: avg\_slope\_diff\_by\_county.sql*

*output: avg\_slope\_diff\_\_county\_\_pct\_lead\_\_winner.csv*

SELECT avg(slope\_diff) as "avg\_slope\_diff",county,state,pct\_lead,winner

FROM

(SELECT companies.company\_name, companies.symbol, ROUND(companies.slope\_diff,4) AS slope\_diff, companies.city, cities.county, cities.state, results\_clean.pct\_lead, results\_clean.winner, companies.sector, companies.sub\_industry

FROM

companies

JOIN

cities

ON(companies.city = cities.city and companies.state = cities.state)

JOIN

results\_clean

ON(cities.county = results\_clean.county)

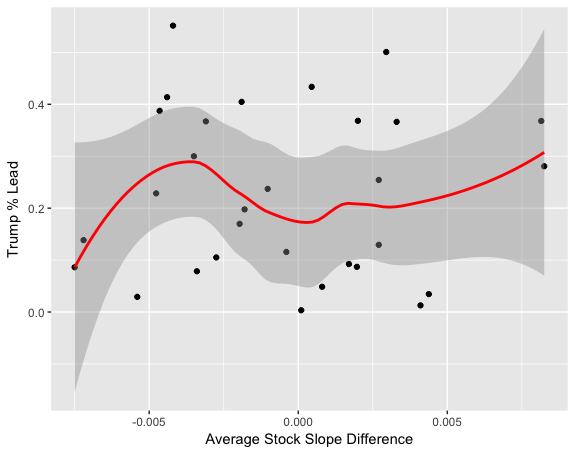
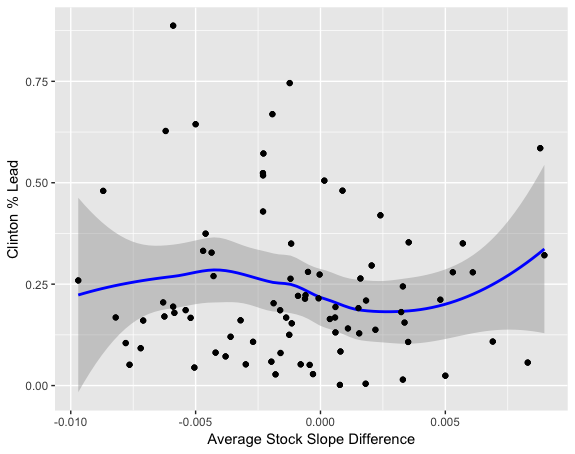
GROUP BY companies.symbol, companies.city, companies.state

ORDER BY cities.county,cities.state)

GROUP BY county

ORDER BY winner,avg\_slope\_diff;

Using this query I was able to look at if there was a correlation between a candidate's percent lead in a county and the average stock slope difference of S&P 500 companies residing in those counties. To analyze this I split up the data by county winner and created 2 separate scatter plots against each other.



While the included lines of best fit may show otherwise, neither graph has a correlation with statistical significance.

2. Compare how stocks from different sectors performed on average

SELECT companies.company\_name, ROUND(companies.slope\_diff,4) AS slope\_diff, companies.city, cities.county, cities.state, results\_clean.pct\_lead, results\_clean.winner, companies.sector, companies.sub\_industry

FROM

companies

JOIN

cities

ON(companies.city = cities.city and companies.state = cities.state)

JOIN

results\_clean

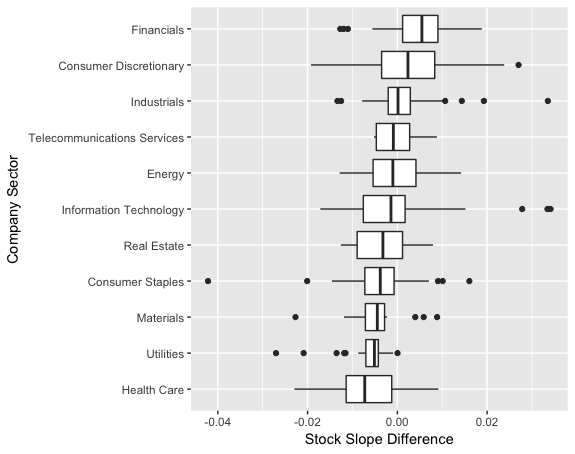
ON(cities.county = results\_clean.county)

GROUP BY companies.symbol, companies.city, companies.state

ORDER BY winner,pct\_lead

*query: companies\_results.sql*

*output: company\_slope\_diff\_coords.csv*



Looking at this boxplot, one is able to see to slope difference distribution by sector allowing one to compare the difference in sectors stock performance before and after the election.

3. Put results on a map

This map shows companies stock performance overlaid on the election results map.

Counties that are blue voted for Clinton while counties that are red voted for Trump, the opacity of the color indicated the percent lead.

Green circles represent companies whose stocks increased from the week before the election to the week after the election, while black triangles represent companies whose stock decreased in that time period. The size of the shape represents the magnitude of slope difference (larger means a bigger difference)

