**Introduction:**

In this project Charlie Ward and Duncan Applegarth aimed to gain insight into the effects and qualities of the COVID-19 pandemic through COVID-19 trends in social groups. The social groups we were primarily concerned with were “political groups” and “cultural social groups”. The data used to accomplish this was sourced from CSVs and APIs provided by Yelp, United States’ Census Bureau, New York Times, Harvard Dataverse, and Center for Systems Science and Engineering at Johns Hopkins University.

**Data Collection and Cleaning:**

With the Yelp datasets we built a DataFrame full of dated United States restaurant reviews. These datasets contained subsets of Yelp’s business, reviews, and user data. This was partitioned into 2 JSON files that we were concerned with: business.json for data about its businesses and reviews.json for plain text and metadata about reviews. Both of these JSON files are very large (~9GB) and doing the initial collecting and cleaning in a Google Collab notebook was going to be too slow due to internet speeds. As such, we initially created a Jupyter Notebook where we downloaded these files locally. Within the Jupyter Notebook, restaurants in the business.json file were mapped to those in reviews.json in chunks of 1,000,000 reviews (9 chunks total) and then the resulting DataFrame was further cleaned to be more memory conscious. This dataframe was then exported as a compressed csv file and uploaded to Google Drive (now only 243MB).

Collecting the New York Times cases and deaths by county dataset, CSSE Johns Hopkins time-series vaccination data, and the Harvard Dataverse 2020 presidential election data was done by simply reading in the .csv files from the appropriate sources.

Collecting the United States’ Census data was done by requesting a JSON file from the Census API and parsing the returned data in order to separate names and county populations.

The US Census data was then further cleaned so that the county names matched those in the New York Times cases and deaths data (this required many cleaning operations). These two datasets were then merged on their County and State name in an outer join such that counties not in the New York Times dataset were dropped. This DataFrame was then merged with the CSSE Johns Hopkins vaccination data and columns were dropped/renamed for clarity.

**Data Exploration:**

As stated in the Introduction, there are two different “social groupings” that we are concerned with. The first section of this exploration analysis will include our findings from the “political groups,” and the second section will include our findings from the “cultural social groups”.

**Data Exploration, Section I: Political Groups**

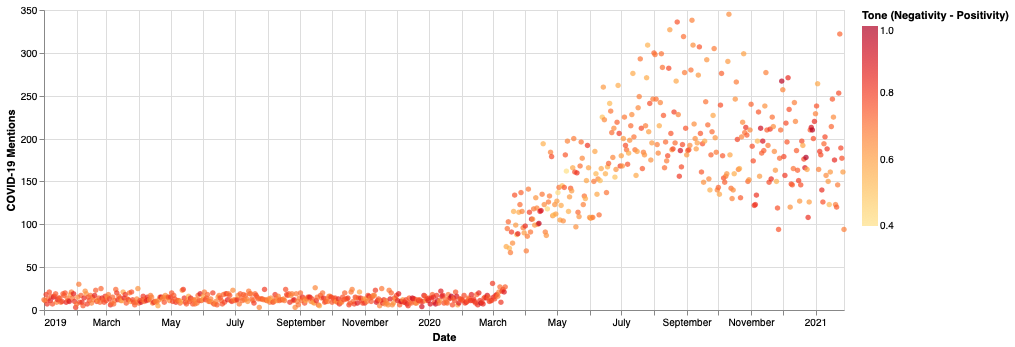
What should I write about here? Theres quite a lot …

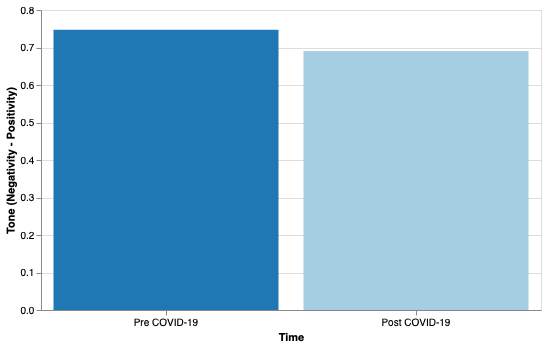
* Built choropleth, limited information to gain from it (not conclusive enough) (and bar chart), not in depth – just explored
* Line charts – discuss how these were explored (include figures for cumulative COVID cases, total vaccinations) – reference all other charts tho (didn’t prove fruitful enough)

**Data Exploration, Section II: Physical Social Groups**

In exploring the Yelp review data we aimed to learn more about how people have been reviewing restaurants during the pandemic period. To do this, we first assembled three different sets of vocabulary (COVID-19, negative, and positive language) found by searching and reading through dictionaries and related articles. A CountVectorizer object was then fitted to a dictionary composed of these three vocabulary sets and the Yelp reviews were then transformed into a sparse matrix containing the term frequencies. The vocabulary was then mapped back into counts of COVID-19 mentions, positive words, and negative words and collected in a DataFrame. A negative tone was calculated by subtracting the frequency of negative language from the frequency of positive language per review (on a daily average, the Yelp reviews are all more negative than positive; we surmise this is due to selection bias in reviewers). The number of COVID-19 mentions and average negative tone per day were then calculated and plotted as shown in **Figure 1**.

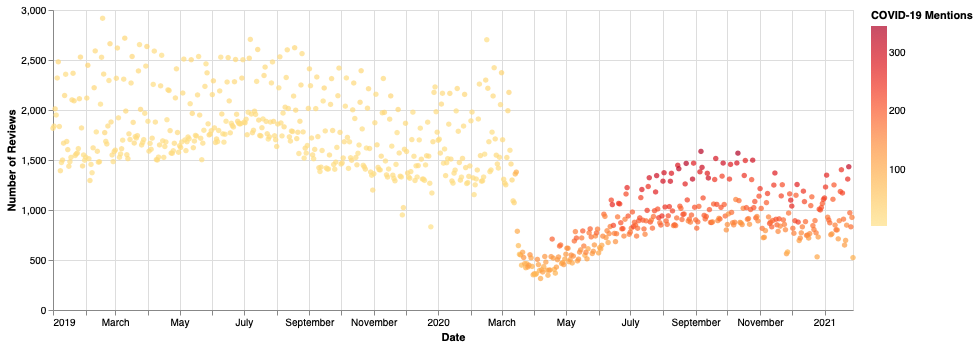
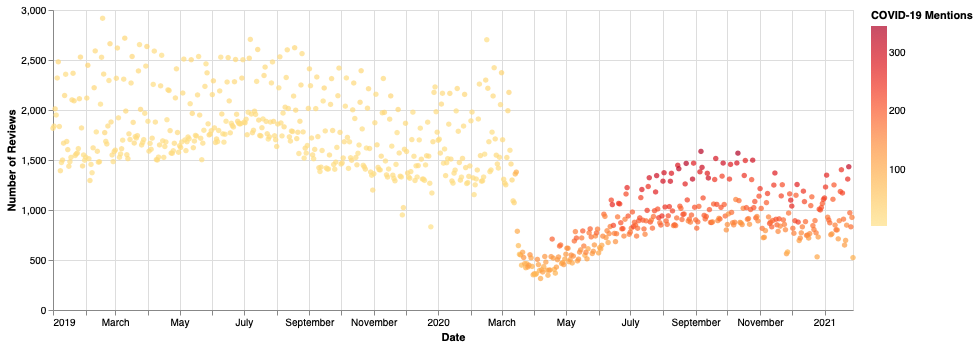
*Figure 1:*

We see here that United States Yelp reviewers didn’t begin to report much on restaurants COVID-19 preparations until mid-March 2020 when stay-at-home orders were first announced and that recent (~January 2020) Yelp reviews seem to be reporting less about COVID-19 preparations than they were in September 2020. This is likely because most of the restaurants in our dataset have already been reported on. It also looks like Yelp reviews before the pandemic were more negative than those after the pandemic. To analyze this further we developed **Figure 2** (shown below) where we see that pre COVID-19 Yelp reviews were marginally more negative (~5 percentage points) but not enough to draw any significant conclusions.

*Figure 2:*

Another question Figures 1 and 2 raise is whether or not the number of daily Yelp reviews before and after the pandemic are roughly equal. Figure 1 shows much variability in the number of COVID-19 mentions starting in mid-March 2020. Is this caused by low counts of reviews in the post COVID-19 data?

*Figure 3:*

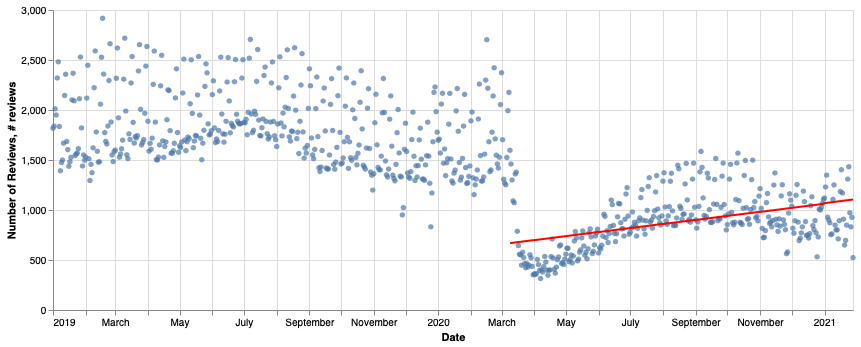
As we see in **Figure 3**, the number of daily United States Yelp restaurant reviews has dropped significantly during the pandemic period. Also, for the days with high numbers of Yelp restaurant reviews after mid-March 2020 we see large counts of COVID-19 mentions. This indicates that these days are likely during the periods of relaxed restrictions where a greater population is dining out and reporting on restaurant COVID-19 care. Another interesting detail shown here is that the variability in the number of restaurant reviews beginning in mid-March 2020 was much smaller than that in the July 2020 and onward. This is likely due to large-scale social conformation to government imposed restrictions during the first two months of the pandemic and a growing agitation within subgroups as pandemic-life wore on.

**Machine Learning:**

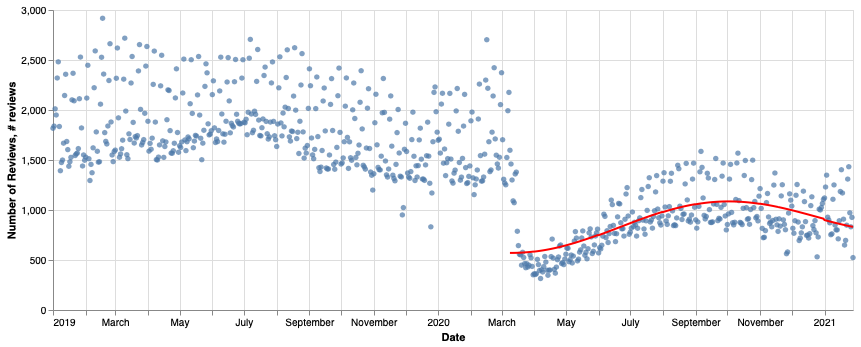
For the Yelp restaurant review data we aimed to build a Linear Regression in order to predict the number of daily reviews for restaurants within our dataset. This implies the degree to which people are getting outside and participating in social activities. If we predict the number of daily Yelp reviews for the restaurants in our data to be equal to the pre-pandemic average of 1,802 reviews/day then we can claim this is roughly when we will see a return to pre-pandemic social activity.

To build the model we first isolated the data on the reviews for the pandemic period. We chose the date March 9, 2020 to be the start of the pandemic period as this is when the Yelp data started to reflect the pandemic. We then hyperparameter tuned our Linear Regression with polynomial features using a GridSearchCV with a cross–validation of 10 and a negative-mean-squared-error scoring metric. The best estimator was found to be a Linear Regression with only a linear term. This model was fit to the pandemic period Yelp review data and used to build the plot in **Figure 4**, with the red line indicating the model’s predicted numbers of reviews over time.

*Figure 4:*

The test error of this model was estimated to be 262.077 reviews using a cross\_val\_score with a cross–validation of 10 and a negative-mean-squared-error scoring metric. This means that we estimate our predicted number of reviews to be within 262 of the real number. 

Next, a model with seasonality was built and used to create the plot in **Figure 5**. We thought a seasonal model would be sensible due to possible fluctuation trends in the number of reviews (e.g. more people go out on the weekends, so the number of reviews should be higher during the weekend and lower during the work week).

*Figure 5:*

This model was evaluated according to the same metrics described above and was found to have an estimated test error of 260 reviews. This is marginally better than the estimated test error for the model without seasonality, but after careful evaluation we see that it likely overfits to the pandemic training data and doesn’t make accurate predictions for when a post-pandemic world will be realized. It predicts we won’t see a pre-pandemic daily average number of restaurant reviews until July 23, 2024! This is obviously too large of an extrapolation to make and clashes with our recent optimism brought about by vaccine distribution that is not captured in the training data (Yelp review data ends January 28, 2021). As such, we will continue with the model without seasonality.

Our chosen model (model 1, seen in **Figure 4**) predicts the number of Yelp restaurant reviews for the restaurants in our dataset on September 1, 2021 to be 1,392 reviews. This is still less than our pre-pandemic average of 1,802 daily reviews and indicates that although schools and other social activities may be resuming, the “end” to the pandemic period’s influence on social life will not yet be fully realized. In fact, our model predicts that the total daily reviews will be equal to the pre-pandemic period average of 1,802 on July 1, 2022. As such, we predict that social-life will be back to the pre-pandemic “normal” in July of next year.

*Conclusions:*

The projected vaccination percentage model predicts we will see herd immunity in the United States sometime in July 2021 while the Yelp review model predicts we will see a return to pre-pandemic social daily life in July 2022. The difference in these two predictions comes from the fact that the Yelp review model predicts a date for the realization of a cultural shift whereas the projected vaccination percentage model predicts a time where a cultural shift may reasonably start to take shape. From this, we gather that we will likely see herd immunity in July of this year and a return to pre-pandemic cultural normalcy by July of next year.