Shell, Drew, Valeria, and Nicole

Blog Post 2

1. **Describe your progress.**
    - **Data Collection:**

      Data Collection is arguably the most challenging aspect of our project. Unlike other teams, who might have found datasets online, which would only require the step of cleaning, we are gathering data in real time, and working with a very protected data source: Instagram posts & their metadata. As a result, it took us quite a while to figure out the limits of our data gathering sources, through the API and through online scrapers. As a result, we had to pivot our project goals, and pivot our data collection criteria too. Thus, instead of gathering LOCATIONAL DATA based on locational searches in a city, we decided to use #hashtags to collect posts sharing a common tag and theme. Since pivoting to the very specific question of answering "what are the top destinations for spring break travels that will yield the most insta-likes", our data collection has been a lot smoother.

      We still are unable to get around the issue of only being able to scrape data starting from any other time other than "now". In other words, we can't just scrape for posts from April 2018 - we would have to scrape posts from today all the way back to April 2018, including everything in between. Not only is that extremely computationally intensive, but it also would leave us with too much unneeded data and take too long to accomplish. Our current plan is to sidestep this problem and look at hashtags like #springbreak2018 #springbreak2017, #springbreak2016 instead. Taking into consideration that many of the recent posts under that hashtag are actually not from 2018 at all, we would only start saving the data after the 100 first recent posts or so.

Furthermore, for each post we scraped, we're also scraping data on the user who posted it. This is necessary for us to calculate the average number of likes that the user usually gets, so that we can see if these spring break destination posts actually were more aesthetic and appealing to the followers that granted a greater level of engagement.

- **Analysing data base / qualitative factors**

  We took into consideration how to weight the value of the number of likes for a particular post. There are many variables that go into determining the number of likes for a post *other* than the actual photo. For instance, the caption could be funny or clever, the account may have a lot of followers, and so on. We hope to eliminate at least one of these variables by considering the average number of likes for the most recent 20% of posts from the account, when determining the popularity of the post. If all of the user's photos get approximately 1k likes, then this one is nothing special and we can lower the true popularity of the location accordingly. Conversely, if the user does not usually get above 50 likes, but this post got 200, then its popularity would be weighted more.

  Another aspect we considered was the lack of locational characteristics of the data we scraped. For each post, since we want to better understand what makes a post more "likeable" than others, just having the number of likes, number of comments, hashtags, isn't very useful. Thus, we decided to parse the accessibility caption of each post, which denotes whether the post is inside or outside, potentially at a restaurant, or in nature, as intrinsic variables that could influence the number of likes for each post. We manually entered the location type and continent for the top 100 locations, sorted using number of likes. We created mapping from location types to integer index: {"Beach": 0, "Nature": 1, "Nightlife": 2, "City Outdoors": 3, "City Indoors": 4, "Theme Park", 5}. For continent, we just have: {"North America": 0, "South America": 1, "Europe": 2, "Asia":3, "Australia": 4, "Antarctica": 5, "Africa": 6}

2. **ML Analysis on dataset + INTERPRETATION OF RESULTS**

- **Current Results:**

```
[0.14660406 0.27147749 0.0104422  0.00116421 0.0051193  0.56519275]
Random Forest Training R-Squared: 0.5046
Random Forest Training MSE: 1188174.98
Random Forest Testing MSE: 1699642.71
Boosted Random Forest Training R-Squared: 0.0213
Boosted Random Forest Training MSE: 2476732.26
Boosted Random Forest Testing MSE: 254036.38
   Coefficients  Standard Errors  t values  Probabilites
0    301.521465        90.449286  3.333597      0.000969
1    -26.446474         9.781220 -2.703801      0.007263
2     -2.402361        29.455699 -0.081558      0.935055
3    131.883455        71.258800  1.850767      0.065227
4    -46.987598       337.599899 -0.139181      0.889404
5   -116.112549       115.569044 -1.004703      0.315883
6     -0.022512         0.101968 -0.220773      0.825426
Regression Training R-Squared: 0.0034
Regression Training MSE: 2390564.60
Regression Testing MSE: 212345.99
```

- **Explanation:**

We've done 3 ML analyses on our data to see if we could predict the number of likes of a given post by a number of various fields: number of pictures, time of day it was posted, whether the post was had one person, a group of people, or no people, whether it was a selfie, whether it was indoors, and the length of the caption. The random forest was able to explain 50% of the variance in likes which was very interesting considering the multiple linear regression and AdaBoost regressors were unable to explain anything. Right now we're missing the number of followers someone has which is a very clear component of the likes they will get. We will need to rerun this and account for the number of followers to be able to accurately determine the effects of these components to make more precise analyses.

The first array describes the importance of each feature. Currently without any additional data, we can see that length of captions are quite important to

determining likes. The linear regression shows us that the statistically significant attributes are the number of pictures, and number of people (at an alpha of .1). Posts with less pictures tend to do better (quality over quantity) and posts with more people do better as well. We're currently scraping the user data and it'd be very interesting to see how our analysis changes after we account for number of followers.

** We asked the HTAs for more GCP credit to run the second part of our analysis.

3. Elaborate on details from Midterm report
    - In the Midterm report, we were really excited about using the Yel[and TripAdvisor APIs to scrape top locations in a city, that could act as a starting point for further instagram search. However, we've come to realize that both are not very accessible; the Yelp API's is limited and does not provide endpoints anymore for the most popularly searched places. It has one optional argument that queries for "hot and new" but that's not perpendicular to the scope of our project. For the TripAdvisor API, the application process took much longer than expected.
4. Next steps:
    - We plan on classifying the locations using Geocoder, to make better visualizations.
    - https://geocoder.readthedocs.io/?fbclid=IwAR2Vuftr_xYUZMogCaYqJ8Q8h8OqRiFMPWJh0Ic3aRPNMeBgM8E9pjd5qdk