Shell, Drew, Valeria, and Nicole

Blog Post 3

**VISION**

- What was your “big idea”?
    - Our initial big idea was to create a program that would, given some city as an input, suggest hot locations to go to in that city, based on some metric of popularity (e.g. likes on Instagram posts tagged at that location). Due to many obstacles in data collection, we ended up having to pivot to our big idea being: Where to go during spring break to get the most likes on your Instagram posts?
- Did you make measurable progress towards this idea or did you find other interesting things?
    - It was nearly impossible for us to continue with the original idea because scraping posts for a location such as “New York” would only yield posts tagged as “New York”; not any of the more specific places like “Top of The Rock” or “Statue of Liberty” like we intended. Therefore, we chose to scrape posts by hashtag instead; specifically #springbreak{2019, 2018, 2017}, under the assumption that posts tagged with the spring break hashtag would be at “hot” locations, and there would be a good number of places where people are going to. Indeed, we were able to collect a very reasonable sample of specific and non-specific locations (e.g. “Darucci Leather” - store in NYC, “Mexico”, respectively) that, after sorting and cleaning data, were the top popular destinations for Spring Break.
    - We found interesting trends in places that consistently ranked highly in the results, as well as places that were popular only in specific years. In 2017, there was a skew towards Ultra Music Festival & Casinos. In 2018, there was a higher concentration of USA and European metropolitan destinations. In 2019, there was a skew towards National Parks and Natural destinations.
    - While there’s some variation in the top rankings, the most popular locations remained consistent across 2017-19, usually warm and beachy, such as Florida, California, Las Vegas, with nearby islands including Cancun and Mexico.

- We were pleasantly surprised by very specific locations as contenders in the top 30, such as Papas Beer Rosarito (a restaurant in Rosarito, Mexico) and Darucci Leather (a clothing store in NYC). We suspect these places to be viral destinations that are on “must-see” lists for certain groups, and are “clout-generating”.

**DATA**

- What data did you use? How did you collect, clean, and integrate it?
  - We scraped posts and their metadata from instagram.com search results for #springbreak2019, #springbreak2018, #springbreak2017. We used multiple instagram scrapers, ran our own python scripts to parse the data, clean the data, and normalize the data.
  - Dataset Snapshot:
    - 141,105 posts over 3 years
    - 11,587 unique 2019 locations
    - 9,221 unique 2019 users
    - 58,429 unique 2019 hashtags
- Relative to its size was there enough information contained within it?
  - There was enough information for us to see shallow trends, despite the top destinations being only an incredibly small percentage of the overall number of locational tagged posts (the highest destination being 0.1% of the overall posts we scraped). We think this is because instagram has so many locational tags and such a diverse user base, that makes it impossible for any location to stand out by a landslide.
  - Nonetheless, it would be interesting to scrape more
- Were you able to find what you wanted in the data?
  - Yes. We found top locations that we were able to draw conclusions from. We also were able to scrape the data of users that posted photos in the top locations, and find an average number of likes they got per post, and make further analysis.

**METHODOLOGY**

- What did you do with your data?
    - Upon scraping the ~20k posts for each year, we transformed the json data within a directory, into posts, hashtags and locations tables within a sqlite3 database. This was done to make sql queries later on easier.
    - We then scraped the user information for each post that was posted. We wanted information on the number of followers, and the number of likes each post the user had. This was used to calculate the average number of likes each user had, which is contextual information for later analysis.
    - We combined the user and post data bases, in order to make further sql queries such as:
        - **Top locations for number of posts**
        - **Top locations for average number of likes**
        - **Top locations for Difference between average likes and likes on post, for each location that appears**
- What techniques were used to pick apart the data?
    - We used SQL queries mainly, to make quantitative rankings.
    - We also “looked at our data” (using terminology from lectures). For example, we manually grouped “miami beach”, “cancun”, “daytona beach” as sunny, beachy locations for our trend analysis. There is technically no consistency to these locations just by the name of the locations.
    - We also tried to use grep to parse the captions of posts tagged in the top locations. We wanted to see if the word “beach” came up a lot in beach location tagged photos, and “disney” in disney photos. We did not present our results on this, because we found no positive correlation between the appearance of “beach” in captions and the actual location.
    - Another analysis we did not do due to time constraints, was to cluster the top 150 locations based on lat-lon, and replace the clusters with a generic name. This could have been helpful in parsing locations that had multiple names, which would have diluted the data.
- Did you use ML? Stats?

- We used ML and statistical analysis tools to understand where people were going to find *clout* in terms of spring break, and if we could predict *clout* using a given location.
    - This model would allow us to understand the impact of a location on the total number of likes.
- We utilized chi squared, random forest, and linear regression to perform our analysis.
- Chi squared provided us with a rigorous way to understand how the top destinations were changing between the years.
    - This analysis was bogged down by how infrequent post of specific locations appeared in the greater sample of posts. A given location made up less than 1% of the total posts.
    - Chi squared was unable to distinguish these distributions of the top destinations from 0, so it gave us a p-value=1. I.E. there was no possibility of a difference between the years.
- We used a random forest regressor to predict the number of likes given 9 features: # Followers, # of pictures, time posted, person, selfie, indoors, beach, disney, length of caption
    - Our model had an R^2 of .9 which signals that it is able to explain 90% of the variability of a post using these features.
        - 80% of the weight in determining the number of likes was dependent on followers
        - Location specific features such as disney and beach each made up less than 1% of the weight.
    - We had a training MSE of ~45 thousand and testing MSE of ~63 thousand
- Linear regression provided us with a simpler model to understand what was going on behind the random forest.
    - The only feature that was statistically significant was the number of followers
        - Meaning that the number of followers has a definitive impact on the number of likes
            - However, this is a very obvious observation

            - People with more likes generally have more followers
        - Location specific features—disney and beach—were not statistically significant in increasing the number of likes.
            - We cannot determine statistically whether or not taking a picture at disneyland or disneyworld or at the beach leads to more likes.
    - Although our more quantitative analyses failed to provide us with results, looking at the data qualitatively we were able to glean more insight.
- How did you visualize your data?
    - We visualized our results in a few different ways:
        - In order to get across the distribution of popular posts, we created a map of the top 30 locations of 2019, based on the average "bump" in likes they gave users who tagged pictures there. All were in North America with the majority clustered around warm locations such as Florida.
        - We also wanted to visualize the most popular locations in terms of how many posts were tagged there. We looked at the top 30 locations across the three years we collected data for, and were able to group them into broader categories to see how trends of spring break destinations changed over time.
        - Lastly, we made a word cloud to visualize the frequency of hashtags used in the posts we scraped. This visualization gives us a better picture of the user base we were gathering data from as well as confirms trends we saw from the location tags, as hashtags such as "travel" and "Florida" and "beach" were very popular.

**OVERALL RESULTS**

- Fully explain and analyze the results from your data, i.e. the inferences or correlations you uncovered, the tools you built, or the visualizations you created.
    - From our data, we were able to extract the most popular spring break destinations over the years, namely Disneyland and beaches such as Panama City Beach. We also found the destinations that had the highest average number of likes each year, such as South Beach, Miami. However, as this metric is heavily influenced by other factors such as follower count, et cetera, we also looked to find how much location played into the number of likes a post would get.

- As explained above, using our ML analyses, we found that we were unable to determine how big a role location plays in the number of likes a user might get on a post. While we were able to find the locations that gave users on average the biggest bump in likes, we could not prove that location was the defining factor. These locations were the Dominican Republic, which gave users an average of 70% increase in likes, and New Orleans with 51% increase.
- From our visualizations and analyses, we found that the majority of spring break destinations clustered in warm, beachy locations as expected, along with popularity spikes in metropolis and nature destinations. We also found that although Florida beaches attract the most posters, similar but slightly less popular locations are where users get the most likes compared to average. We discovered in addition a few very specific locations that stood out among entire cities, returning us full circle to our initial idea of finding specific spots for users to post in order to gain the most likes.

**OTHER REQUIREMENTS**

- **Data:** Our uncleaned data is uploaded as .db files on github.
- **ML/Stats:** Use at least ***two*** machine learning or statistical analysis techniques to analyze your data, explain what you did, and talk about the inferences you uncovered.
  - Chi squared
  - Random forest
  - Linear regression
- **Visualization:** Provide at least ***two*** distinct visualizations of your data or final results. This means two *different* techniques. If you use bar charts to analyze one aspect of your data, while you may use bar charts again, the second use will not count as a *distinct* visualization. Note: we would like for you to avoid using Tableau or similar programs. We would prefer you write your own D3 et al code. Although we will not take off points for using Tableau, we will hold you to a **much** higher standard.
  - On github as mostusedhashtags.png, locationtrends.png and springbreakmap.pdf.
- **Additional work:** In addition to the requirements in the ML and viz sections above, we would like to see at least one extra from either category. That means a total of five deliverables.

- We have three visualizations and three ML analyses.