KOAA-Google Analytics: Final Report

Spencer Leonardi, Matthew Koo, Sahaj Somani, Tanmay Thakkar

Introduction:

Kids Out and About (KOAA) is a website that aggregates all local events within communities into one place. It serves as an easy way to find out what events and things there are to do in an area for kids, teens, and families. The website has separate hubs for about 45 locations all around the United States and some cities in Canada. KOAA tracks their site visits through Google Analytics (GA) which is a platform that tracks and reports website traffic data for any website.

The platform can determine the number of users in a given time period, the bounce rate, how much time people spend on the website per visit, locations of all the users, and many more different features. Out of all of the measures that Google Analytics calculates, the most relevant one is the number of pageviews. KOAA keeps track of many different categories of events that happen within communities, such as music, sports, exercise, and many more. The activity categories for each URL is recorded and was provided as a csv file from Drupal. Since each event web page is tied to a different category or a set of categories, we can use the amount of pageviews each web page receives in a given time period and determine how each category trends over time.

Unfortunately, Google Analytics doesn't have an easy way to export all of the data from its platform into a usable Excel spreadsheet or CSV file. There is an export button, but it can only export up to about 5,000 of the most recent rows for each location. Furthermore, visualization is provided directly from Google Analytics but the data on there was not in a usable form initially. The overall motivation of this project is to implement a Google API to manually extract the data from Google Analytics, and with that data analyze how each category of event performs over time.

Data Set Preparation:

The data set that we used consists of 35 separate csv files (34 from Google Analytics and 1 from Drupal). By writing a Python script shown below, three key features from each location on Google Analytics were extracted for our analysis (34 locations were used). Raw data from Google Analytics on the left were turned into csv files on the right. The three key features on the csv files created were:

- **'year_month'** = the year and month representing when a particular web page path has received a particular number of pageviews
- **'page_path'** = URL of a web page link from a particular location

- **'pageviews'** = the number of pageviews the web page URL obtained in a given time period

Figure 1: Python Script for data extraction

The second data set we used was the Drupal dataset. This is an SQL relational database that our sponsor had provided which tied each URL to a set of categories that could be used to describe the different types of events each URL would fall under.

Figure 2: Merging Drupal data with GA data

This csv file was combined with each Google Analytics csv file through an inner join on URL to produce 34 csv files (shown above). As an inner join on URL was done, the rows and subsequently the URLs that did not have any category label were omitted from the final datasets as they would not be helpful in the analysis.

Following the merging process, the 'Categories' feature was one hot encoded as shown below. This took each new individual category from a list and made it its own column. If the category appeared in the URL's list, the value in that category's own column would be represented as 1. If the category didn't appear in the list, then the value would be represented as 0.

	year_month	page_path	pageviews	Arts-Participate	Arts-Watch	Arts-Visual	Outdoor-general	STEM	Active - Participatory:Exercise	Participatory
0	2013-09-01	westchester.kidsoutandabout.com/content/hudson...	1	0.0	1.0	1.0	0.0	0.0	0.0	
1	2013-09-01	westchester.kidsoutandabout.com/content/ecolog...	5	0.0	0.0	0.0	1.0	1.0	1.0	
2	2013-09-01	westchester.kidsoutandabout.com/content/chappa...	3	1.0	0.0	0.0	0.0	1.0	0.0	
3	2013-09-01	westchester.kidsoutandabout.com/content/beary-...	2	0.0	0.0	1.0	1.0	1.0	1.0	
4	2013-10-01	westchester.kidsoutandabout.com/content/spooky...	4	1.0	1.0	0.0	0.0	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	
38673	2020-02-01	westchester.kidsoutandabout.com/content/victor...	1	0.0	0.0	1.0	0.0	0.0	0.0	
38674	2020-02-01	westchester.kidsoutandabout.com/content/school...	1	0.0	0.0	0.0	0.0	0.0	1.0	
38675	2020-02-01	westchester.kidsoutandabout.com/content/winter...	2	0.0	0.0	0.0	1.0	1.0	1.0	
38676	2020-02-01	westchester.kidsoutandabout.com/content/garry-...	7	0.0	1.0	0.0	0.0	1.0	0.0	
38677	2020-02-01	westchester.kidsoutandabout.com/content/winter...	3	0.0	0.0	0.0	1.0	0.0	1.0	

38678 rows × 11 columns

Figure 3: Hot-encoding categories

At the end of the hot encoding process, there were 78 additional features, each representing an activity category. Although some locations did not have certain activity categories, all 78 category columns were added to all 34 csv files to keep the number of features in each consistent. The regions were also one hot encoded as shown below.

	year_month	page_path	pageviews	Northeast	Southeast	Midwest	Mountain	Texas	Canada
0	2013-09-01	westchester.kidsoutandabout.com/content/hudson...	1	1	0	0	0	0	0
1	2013-09-01	westchester.kidsoutandabout.com/content/ecolog...	5	1	0	0	0	0	0
2	2013-09-01	westchester.kidsoutandabout.com/content/chappa...	3	1	0	0	0	0	0
3	2013-09-01	westchester.kidsoutandabout.com/content/beary-...	2	1	0	0	0	0	0
4	2013-10-01	westchester.kidsoutandabout.com/content/spooky...	4	1	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
38673	2020-02-01	westchester.kidsoutandabout.com/content/victor...	1	1	0	0	0	0	0
38674	2020-02-01	westchester.kidsoutandabout.com/content/school...	1	1	0	0	0	0	0
38675	2020-02-01	westchester.kidsoutandabout.com/content/winter...	2	1	0	0	0	0	0
38676	2020-02-01	westchester.kidsoutandabout.com/content/garry-...	7	1	0	0	0	0	0
38677	2020-02-01	westchester.kidsoutandabout.com/content/winter...	3	1	0	0	0	0	0

Figure 4: Hot-encoding regions

Exploratory Analysis:

With the hot encoded data set the relative pageviews for each category over time needed to be calculated. To do this, all of the hot encoded category values were multiplied by the pageviews

column. Next, the entire data set was grouped by month, giving the total pageviews for each category in each month. Lastly, the relative page views needed to be calculated by creating a proportional value of the number of pageviews a specific category received in that month divided by the total pageviews received in that month across all categories.

When initially visualizing this data the issue was that there were too many categories. Visualizing 78 different lines on one plot and trying to analyze their trends all at once was incomprehensible. Furthermore, a lot of the categories were redundant. Some examples of this include Ski and Ski/Snowboard being two different categories, along with there being separate categories for different musical instruments that followed nearly the same trend (Guitar, Drums, Bass, Horns, String Instruments, Vocals, etc.). The sponsor was made aware of these issues and was also given a solution: to cluster these categories together to about 8-10 different subcategories. She agreed that it would be best and also decided which categories belonged to which cluster.

In team meetings when the idea of clustering was brought up before the sponsor was made aware of the issue, one of the TAs had suggested using a machine learning model to cluster the data. However, the manual approach was used instead. The rationale was that it would be better to group categories that had similar subject matter manually in logical ways. Furthermore, a machine learning algorithm could mistakenly group together dissimilar categories such as sports and art together. The sponsor also had more knowledge about what categories should be grouped together so the decision for which categories to cluster together was left to her. After clustering the values from all 78 categories down to 8, some initial visualizations were made. These plots highlight how each activity cluster's relative pageviews perform over time on a quarterly basis in each region

In all regions the trends for each category have a lot of high variance and are generally noisy until 2015 when all of the clusters calm down and start to look like more realistic trend lines. Additionally, the Arts-Visualize category in teal becomes the top category by 2018 for all regions. Also, in general the blue and green categories which are Active-Participatory:Exercise and Arts-Watch are the next two most popular categories. There also seems to be a trend that all locations follow a general hierarchy in which clusters trend more compared to others among all regions.

Figure 5: Initial activity trend analysis visualisations

***Model Development*:**

While the initial visualizations show a lot about what's happening in each region across the entire lifespan of each region, they weren't perfect. There were more questions that came up about the trends of these activity categories. Also, the number of clusters was expanded from 8 to 13 and the final 13 categories were decided by the sponsor. The first topic addressed was how each category performed on a seasonality basis. Perhaps there were trends relating to how popular each activity cluster was.

To implement this into the analysis, a time series decomposition had to be done. The typical time series model consists of a trend, a seasonality value, and added noise. However, the data set did not entirely satisfy all of the assumptions necessary to create a rigorous additive or multiplicative time series model. Similar to decisions made in previous steps of this project, a more data-driven approach of developing a time series model was implemented as opposed to a more traditional and rigorous model-driven way of performing a time series decomposition. Here is a representation of the time series decomposition done:

$$y_t = S_t + T_t + R_t$$

Because the original trend for all regions was very noisy for all categories in years prior to 2015, smoothing based on seasonality would help clean and remove a lot of the noise that was found here. An explanation for why the noise averages out when comparing the seasonality among years can be explained by the Law of Large Numbers. This law explains that repeating the same calculation many times will average out to the expected value. Meaning that since noise follows the distribution N(0, σ^2), over a lot of observations it would average out and wouldn't bias the seasonality or trend of the time series.

Our data followed the additive model more than the multiplicative one so we decided to use the additive model process to decompose our data. The process we used is as follows:

- Detrend the time series by subtracting the 13 period moving average
- Extract seasonality by averaging the monthly data for each year past 2015

Following these two steps, we were able to decompose the time series separating the noise. This approach is not perfect, but works best given the data we have and more importantly since we are not trying to forecast, rather just analyze, the data driven approach does not seem to have any upright issues.

Another question that needed to be addressed was how the visualizations could be improved. Also, the initial trends of relative page views were influenced by the locations themselves too

much. To eliminate any bias that the regions may have had on the observations in how each activity category performed over time, the sponsor suggested normalizing the relative pageviews for each activity per region. This meant that instead of visualizing how all of the categories perform in each region, to instead visualize how each region performs over time for each category. The regional trends were normalized around 0. Values above 0 demonstrates an overperformance of a region towards a particular category, while a value below 0 demonstrates a region underperforming for a particular category.

Performance and Results:

Before highlighting any visualizations, it's important to note that the means of how the data was visualized was different. The initial visualizations of how each category trended over time were performed using Tableau. To improve visualization capabilities and the ease of obtaining different graphs per regions and per category, interactive plots were created using Plotly in Python. The visualizations below can be altered by the user with a few clicks. The user can zoom in and out of whatever sections of the graph they specify, or even click on a specific point on a line and obtain the value at that point. Additionally, the user can select which lines they want to visualize at a time with ease by clicking on the names of the categories in the legend. Below is an example of a visualization of how each activity cluster performs in each region on a seasonality basis after performing the time decomposition:

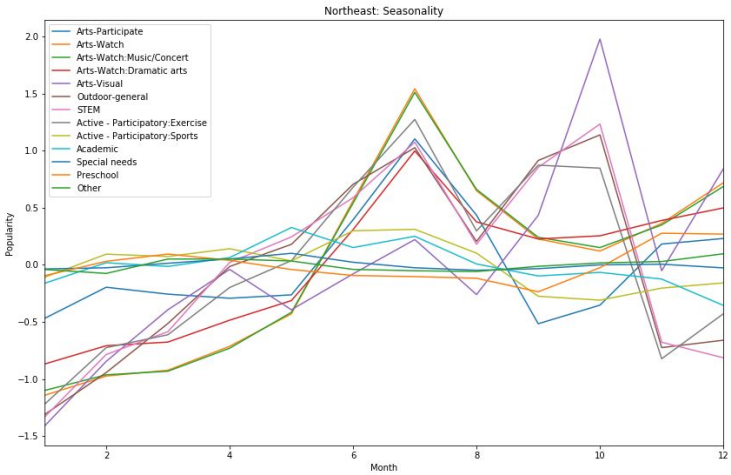

Figure 6: Seasonality analysis for each activity in the Northeast region

For the Northeast, it can be observed that there are general spikes in seasonality for all trends in the month of July and October. It's also interesting to note that even though Arts-Visual is one of the highest trending categories in general, it is not always the most popular category every

month. It should be pointed out that not every region has the same seasonality trends for each category. Different regions have different seasonality trends as we can in Figure 7. Some of the differences between the Southeast region and the Northeast region include the fact that the Southeast peaks in December while the Northeast peaks in October. Another observation that can be made between both regions is that KOAA in general is more popular from January to August in the Northeast when compared to the Southeast. Differences in seasonality like the ones mentioned between the Northeast and the Southeast can be observed between all region's graphs. These graphs can be found in the updated final presentation and Google Drive.

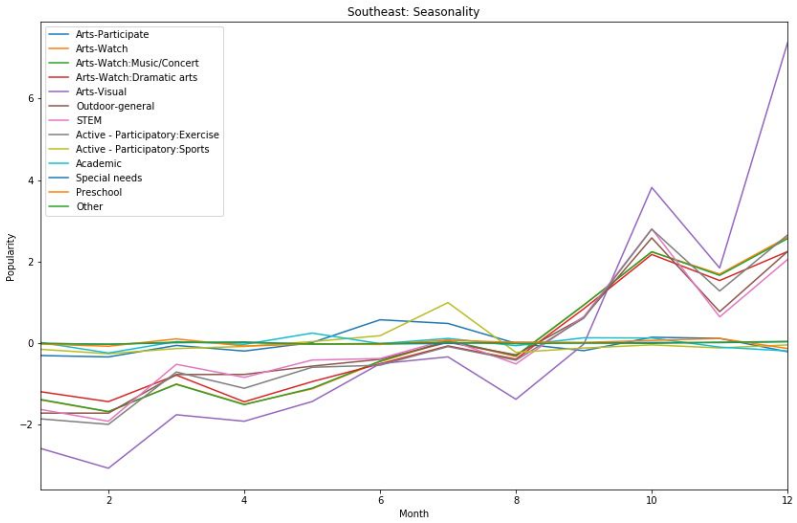

Figure 7: Seasonality analysis for each activity in the Southeast region

Furthermore, we also plotted how each region performs for each category over time after normalizing the trend of the category with respect to the trend of the region. The normalization results in a curve oscillating around 0. One can look at this plot and answer the question 'have academic activities gained or lost popularity over time' without worrying about the increasing/decreasing popularity of the region and the increasing popularity of the business. Here is an example of the visualization of one category:

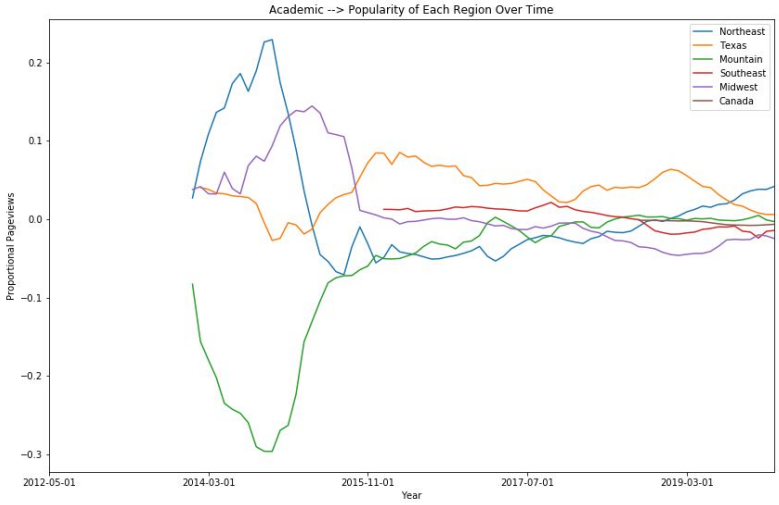

Figure 8: Popularity of the "Academic" category overtime and across regions

For the academic category, the Northeast and Midwest generally overperform in the beginning and then near the end of 2015 normalize to a relatively consistent value. The opposite is true for the Mountain region. It's also interesting to see that Texas has a relatively more constant value for the academic category when compared to other regions, but it also is the highest grossing region from the end of 2015 to sometime in 2019. Like with the seasonality visualizations, these visualizations of how each region performs for each category over time are different depending on which category is observed.

For instance, looking at the exercise category below, the Mountain region grosses the highest in popularity overall while the Northeast region grosses as the lowest region in popularity overall. This makes sense since the cultural differences in what people would want to search in each region such as skiing, snowboarding, and hiking are very popular in the Mountain region. Like with the academic category, the data also seems to normalize towards the end of 2015 to relative consistency in how each region performs for the exercise category. Similarly to the seasonality visualizations, differences in regional influence over categories can give different results and insights when observing all of the different graphs for each category. All 13 of these graphs can be found in the updated final presentation and Google Drive.

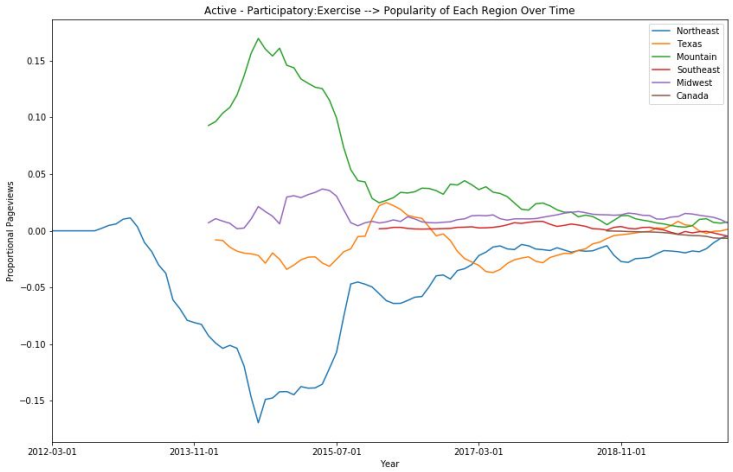

Figure 9: Popularity of the "Exercise" category overtime and across regions

Conclusion and Next Steps:

At the end of the analysis, the sponsor Debra was provided with a set of deliverables. Outside of the visualizations created, it was very important to be able to give the sponsor the code and a tutorial on how to run it all so that she could perform this analysis herself with data from Google Analytics beyond February 2020. The script that we included provides a tutorial on how to set up the Google API, calculations among CSV files to form trends of seasonality and normalization of how each region performs per category, and the creation of the interactive plots. A demo was included of how to set up her local machine to be able to run the script and perform this analysis on her own. And with that, it has been a pleasure working as a group with Debra to provide an analysis that will help Kids Out And About long-term. We'd like to thank her for allowing us to work for her and learn what it's like to work on a team to deliver a product for a client. We'd also like to thank Ajay and PJ for organizing the Data Science Capstone course and meeting with us every week prior to quarantine giving us advice on how to meet deadlines, tackle each problem step-by-step, presenting ourselves professionally, and providing us the ability to work for a real client.

References:

1. Medium: A Visual Guide to Time Series Decomposition, Thalles Silva
2. Plotly.com, Dash Documentation and User Guide
3. Statsmodels.org, Notes, Seasonal Decompose