Alexander Bogdanowicz and Kelly Marshall

CSCI-GA 3033 Spec Topics in Computer Sci: Intro to Big Data

Professor Anasse Bari

Final Project

Brand Mentions as a Proxy for Predicting Brand Value

I. Introduction

In the past half decade, the music industry has undergone a dramatic transformation in the direction of music streaming. Instead of purchasing specific music, listeners instead pay a monthly fee to have access to a large body of available music. As a result of this, there is now a much greater body of data which can be used to capture mass music trends. Unlike the traditional format, which only allows one to tell how many times a given work has been purchased in a single week, streaming data gives daily insight into how many times a song was listened to. This represents a significant increase in the degree to which we can study how users interact with music as well as what music trends can tell us about the behavior of those listening to it.

In this project, we analyze the daily occurrence of luxury brand mentions in Spotify's most streamed songs and model the relationship between a brand's occurrences in popular songs with the brand's popularity, as measured by Google Trends. The potential for a song to influence consumer behavior is an effect that seems rather intuitive. Many songs make mention of brands and products, resulting in the listener being exposed to these names in a relatively repetitive format. This propagation of brands, which in many cases is done in coordination with the companies in question, creates the possibility for increased name recognition and greater brand value. Most notably, this is very common with luxury brands, whose brand images are often tied to the celebrity lifestyle portrayed in many songs. Of course, it is also the case that the effect of a brand mention in a song depends on the greater context in which it appears and so the topic of the song plays an important role as well.

II. CRISP-DM

A. Business and Data understanding

In classic CRISP-DM (Cross-Industry Standard Process for Data Mining) fashion, we begin our analysis with a question often missed in academia, but which may nevertheless prove important in measuring the impact of discoveries related to fields concerned with actionable and practical applications. The benefit of the CRISP-DM framework is that it is an iterative process whereby knowledge-discovery helps confirm, deny, and build-upon previous assumptions to help polish analytical models (Shearer, Colin 13). Throughout the course of our analytics - which are by no means complete - we realized the importance of adaptability (and the drawbacks of commitment) and communication on the levels of Business and Data Understanding, as well as flexibility on the level of Data Preparation and Modeling.

Similar to the paper *"Twitter mood predicts the stock market"* by Johan Bollen and Huina Ma of Indiana University, with ambitious claims as to the efficacy of twitter mood as a feature for determining stock market prices, while suffering from sample size and over-fitting issues as well as a simplistic mechanism for causality detection, proved in inspiration in its attempt to determine the seemingly, as per the Efficient Market Hypothesis, unpredictable nature of the stock market.

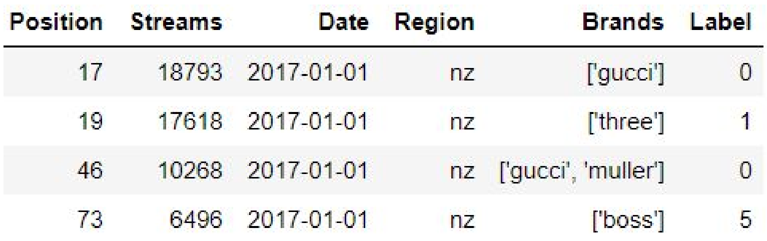
In general, markets are thought to be efficient (or semi-efficient), in the sense that market prices convey consensus estimates of the values of different assets and asset classes, and as far as prices convey the amount of information within the system. This has rarely stopped anyone from attempting to find the market's "tell" (i.e. seemingly un-correlated data that may house). Evidently, as the theory goes, once made available to the market as a whole, any newly discovered information would find itself shortly, "baked into" the price by market makers attempting to exploit arbitrage opportunities.

This is where the Business Understanding, that is, the material aspect of understanding the listening habits of the some 75 million users by the end of 2017, holds as a proxy for brand exposure, which is largely thought to correlate at least somewhat positively, with brand success and sales (Felix Richter, Statista). The effects of brand awareness on the value of luxury brands has been showcased across multiple studies, including the study "*How Brand Awareness Relates to Market Outcome, Brand Equity, and Marketing Mix"* (Huang and Sarigollu), which focuses on brand usage corresponding to brand awareness, which has a positive "association" with brand equity. We find this mechanism often used in the brand industry, whereby listeners find credibility, legitimacy, and validation from their favorite singers, songwriters, or rappers covering specific brands in their songistry. Coincidentally we find a similar technique on social media, which has been the subject of relative controversy on platforms the likes of Twitter, where device stamps have exposed certain targeted user-base marketing as being facetious (we suspect the same is true in the music industry). Identifying whether music conveys this same pattern of brand management and percolation therefore was our primary form of business understanding, as extrapolating information regarding the efficacy of marketing campaigns in a less tangible/trackable medium is extremely valuable to stakeholders.

Before delving into the Data Preparation section, we needed to identify what data we would need to acquire as a proxy for our proposed hypothesis on the utility of brand mentions in songs as a mechanism of impacting consumer behavior. To this end, we took advantage of the following datasets:

- Spotify Daily Top 200 per Country (01/01/2017 - 01/09/2018)
- Deloitte Top 100 Luxury Products Report

The logic was as follows; Spotify's Top 200 songs dataset would provide us with over a year's worth of listening data, which would give us an indication of the types of songs that listener's favored throughout the year, and coincidentally, the songs that had the largest exposure/reach. We constrained our dataset to english speaking countries (US, UK, New Zealand, and Canada), meaning we had the top 200 Songs for each day of 374 days across 4 countries, as well as the following features (includes LDA generated topic labels and extracted brands):



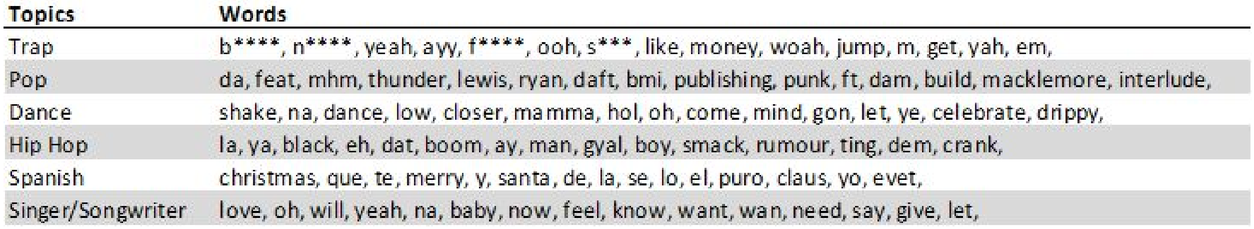
Position	Streams	Date	Region	Brands	Label
17	18793	2017-01-01	nz	['gucci']	0
19	17618	2017-01-01	nz	['three']	1
46	10268	2017-01-01	nz	['gucci', 'muller']	0
73	6496	2017-01-01	nz	['boss']	5

The Deloitte Report compiled the world's top 100 luxury goods companies based on luxury goods sales, which across the industry generated US$217 Billion for fiscal year 2016. An average luxury brand, per the report, was valued at about US$2.2 Billion in annual sales (Arienti Patrizia, Deloitte). In many cases, rappers, singers and songwriters feature "larger than life" exhibits in their songs and luxury brands have had, through this medium, a historically fruitful relationship with the music industry. The report gave us insights on the top performing brands as well as their parent and holding companies, in addition to their colloquial brand names, which proved advantageous in data preprocessing, as we were no longer forced to rely on costly Natural Language Processing mechanisms, such as Stanford Corenlp Named-Entity Recognition, or simple noun extraction, to identify most-frequently mentioned brands. From here, we proceed with further data preparation and some initial findings.

B. Data Preparation and Initial Observations

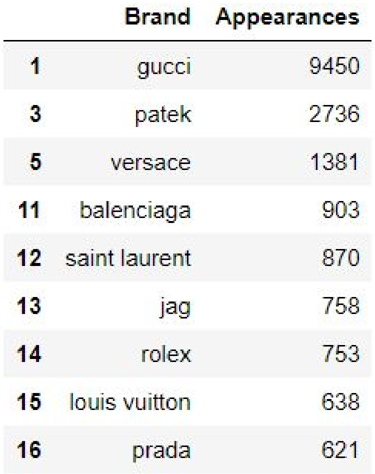
In order to make use of our Spotify dataset, we needed to figure out a way to identify which songs, on a day-to-day basis, featured any sort of brand mentions that fell into the scope of our initial list of Deloitte's top luxury brands. To that end, we needed to collect all unique songs trending over the period, and check each song's lyrics to see whether they mention a brand that fell within our dataset. We therefore ran a Python script against Genius.com's collection of song lyrics, which scraped over the website and downloaded the lyrics. Then, we proceeded to append these lyrics to the spotify dataset, performed lemmatization and stopword removal on NYU's HPC Prince Cluster, using the Stanford Corenlp library, and finally, looped through each top-200 song's lyrics, and extracted all mentions of brands in lyrics from the songs appearing in the dataset.

While this categorized our dataset of song's by which brands were featured, we also wanted to find ways to describe what "type" of songs corresponded to brand mentions in order to determine whether any specific type of song, categorized by lyrics, were perhaps more "persuasive" than others in impacting consumer brand awareness. To this end, we created a TF-IDF matrix with all of our song lyrics based on their n-grams (method for also identifying co-occurrences in sentence structures), and proceeded to apply this vector embedding in Latent Dirichlet Allocation. LDA is a classification method, used in Natural Language Processing that utilizes probabilistic latent semantic analysis (pLSA) (Wikipedia), which, in less fancy terms, considers the probabilities of specific words being associated or generated by certain topics and identifies words that help maximize topic level probabilities. The hyperparameters which specify the prior distributions were experimented with, and eventually we realized that, especially in song lyrics, there is little clear differentiation in groups, and therefore each topic is more likely to consist of a mixture of most topics, and each document is likely to consider many similar words. We therefore parameterized the distribution hyperparameters accordingly. The generated topics and their inferred topic label can be seen below:



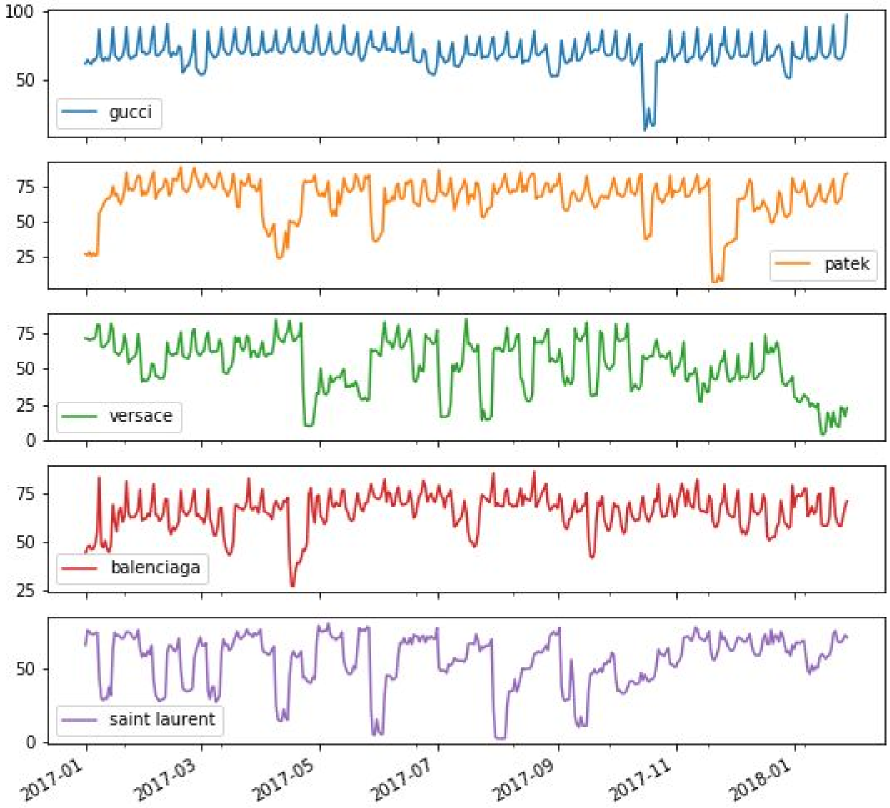
Topics	Words
Trap	b****, n****, yeah, ayy, f****, ooh, s***, like, money, woah, jump, m, get, yah, em,
Pop	da, feat, mhm, thunder, lewis, ryan, daft, bmi, publishing, punk, ft, dam, build, macklemore, interlude,
Dance	shake, na, dance, low, closer, mamma, hol, oh, come, mind, gon, let, ye, celebrate, drippy,
Hip Hop	la, ya, black, eh, dat, boom, ay, man, gyal, boy, smack, rumour, ting, dem, crank,
Spanish	christmas, que, te, merry, y, santa, de, la, se, lo, el, puro, claus, yo, evet,
Singer/Songwriter	love, oh, will, yeah, na, baby, now, feel, know, want, wan, need, say, give, let,

With our categorical feature discovered, we proceeded to focus on finding a suitable dependent variable that would capture the effect of brand features in songs over time. Initially, we began our data collection process on the premise that our initial analysis on the most popular brand mentions in songs streamed from spotify across our sample period would consist of brands either directly representing companies publicly listed on recognized stock exchanges, or at least publicly traded as a brand-portfolio, such as Richemont (owner of A. Lange & Sohne, Piaget, Cartier etc.). We include a list of the most popular individual brand mentions in songs over the dataset period below:



	Brand	Appearances
1	gucci	9450
3	patek	2736
5	versace	1381
11	balenciaga	903
12	saint laurent	870
13	jag	758
14	rolex	753
15	louis vuitton	638
16	prada	621

While the results of this initial analysis were promising, we realized that most luxury brands are in fact privately held and are often partially owned by other luxury brands, perhaps as a brand-diversification strategy. In this case, we were forced to default to another means for gauging the effect of this marketing strategy, and settled on one that more immediately captures the effects of streaming on consumer interest: Google Trends Data. Google Trends Data essentially reflects the relative volumes of searches for a particular terms over time, and can help discover "trends" in consumer interests. The benefits of using GTD were two-fold; we were able to have a less noisy gauge on the impact of song brand mentions, as they impact search levels directly, and as public markets are open only 5-days a week, while internet searches are done around the clock, we could better correlate our set of independent features and our dependent google trends data. As all features are, in the end, proxies for unobservable data, we continued with our belief that Google Trends data could stand as a proxy for "brand value" and proceeded with the same planned treatment. A time-series of our dependent variables can be found in the table below:



From the get-go, as far as time series data goes, we observe a few distinct observations regarding the 1-year search trends. It seems that certain brands, especially Saint Laurent, lack variance or mean stationarity, and appear similar to financial time-series, to exhibit random-walk tendencies. This should not come as a surprise given the intuition behind random-walk data, and the behavior of searches (both move on the basis of new information). This reinforced our beliefs that we had identified a suitable proxy dependent variable.

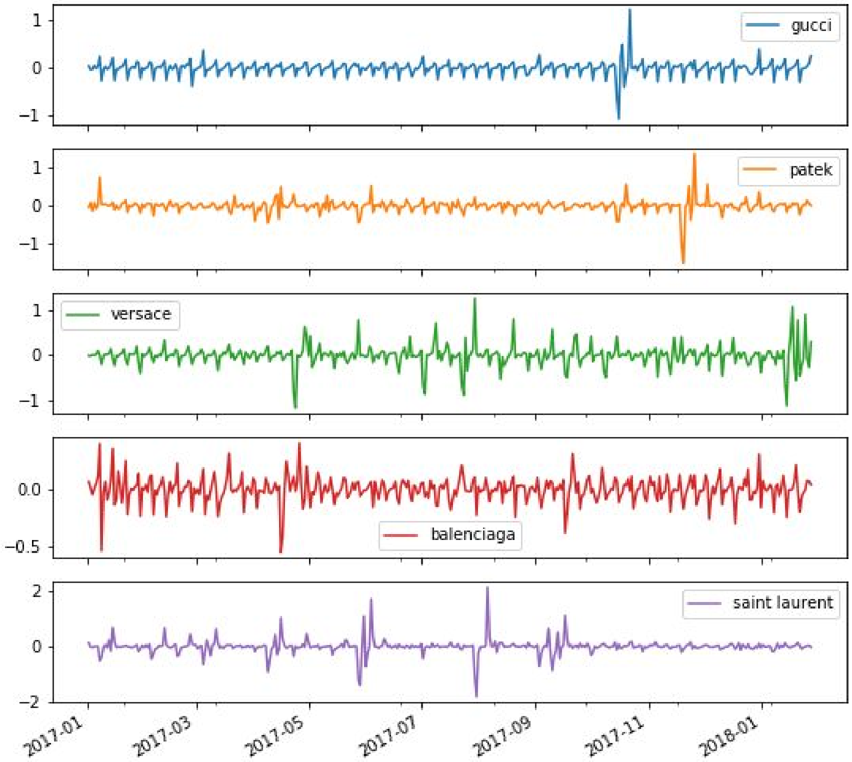
We performed an initial correlation analysis to determine the strength of streaming data as potential linear explanatory variables for the absolute values of our Google Trend Data. It was clear that our data, even lagged, correlates strongly with the daily search trends for these specific brands and therefore, without the information of a previous lag, we might determine that streaming volumes or trend momentum are good predictors of future trends. We will expand further on this line of thought in the following section.

C. Modeling

We begin our modeling with an emphasis on what we are trying to learn from our compiled dataset. Our question as to whether streaming volumes and exposure over time cause, or can be leading indicators, of shifts in google trends search volumes is fundamentally one of causality. Therefore, it was intuitive for us to turn to a popular measure of causality in economic environments, Granger Causality, which has, since its propositive by Clive Granger in 1969, has been applied to Big Data (Song and Taamouti, Oxford), expanded to high dimensional analysis fairly recently (Hecq et al, Cornell), and can likely provide us with some intuition as to our initial research question. Despite this, we voice a few concerns in applying the model to our specific dataset.

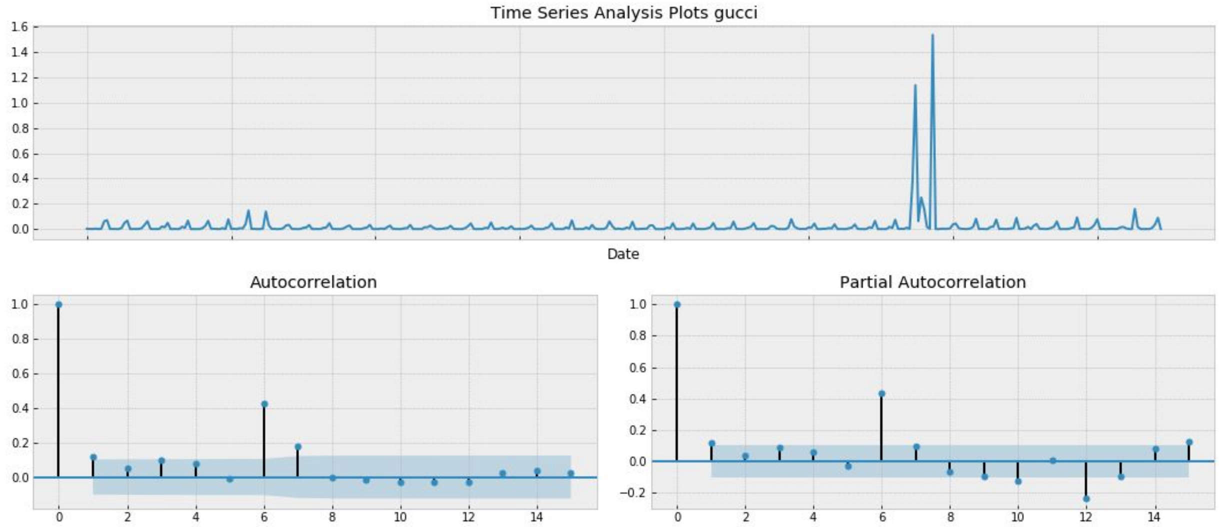
As a result of the time-scale of our time-series data, it was likely that we would find that the streaming data of the previous day is “baked into” the google trend data of the previous day, meaning that essentially, the information gain from including a lagged term in any sort of linear-regression would simply create problems of multicollinearity amongst the lagged trends and other explanatory variables. In this regard, Granger Causality will answer the question, which is that while we are rather convinced of the trend on a sub-daily level, on a daily basis the value of using the previous days streaming to understand the current days web-traffic or brand-value is minimal. As we already know that google trend data correlates extremely well with streaming volume in relation to songs, we can almost certainly make the logical jump that the absolute values of each will reveal strong correlation, but no granger-causality.

This theory would then pose, that an ARIMA (Autoregressive Integrated Moving Average) model of some orders AR, I, and MA would well approximate the Google Trends brand data, with little statistical significance warranted to our other daily attributes. Therefore, we pivoted the granger causality question to see whether any of our variables can be useful, not in describing the absolute trend data, but the variance (i.e. explaining when there are disturbances). By design, in using the residual values of our ARIMA model, that is, the error resulting from the difference between the predicted and actual values, as dependent variable in a linear regression over lagged values of our feature set, we might discover whether previous days streaming data or perhaps a specific categorical variable, was predictive of future spikes in variance (positive or negative). This is similar to the intuition behind GARCH models and is premised on the data exhibiting heteroskedasticity (which it does).

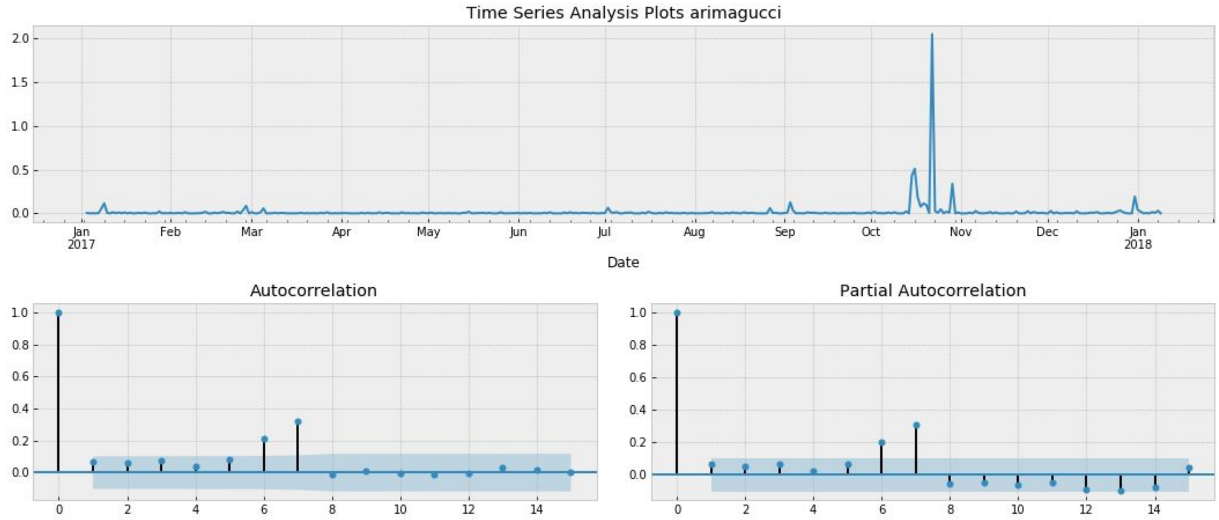


We therefore proceed to calculate the logged returns of our Google Trends Data (to normalize and prepare the data for ARIMA analysis) and the initial generated figures display trends of time-sensitive variance and likely require integration treatment.

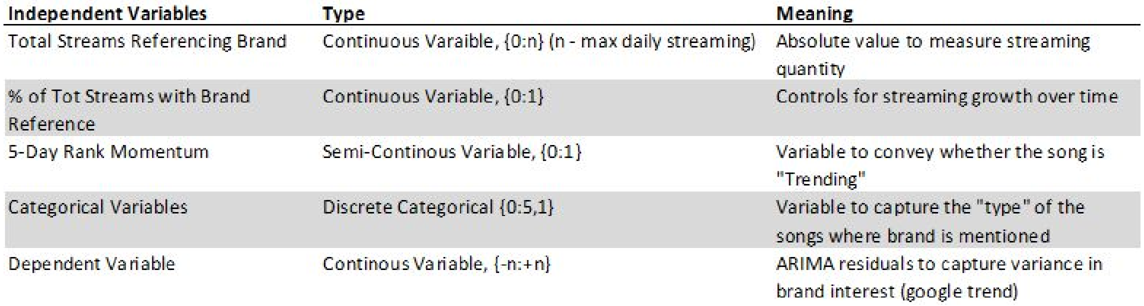
In order to generate reliable residual variables we hope to fit these time series as best we can with the proper hyperparameters of the ARIMA model, which will determine the extent that the model factors in lagged terms. Traditionally, ARIMA models can be optimized in two ways, either intuitively by interpreting the time-series auto and partial-auto correlation figures to determine the optimal amount of lags for the AR and MA components, and the optimal amount of integration via differencing, or by optimizing a cost function (such as AIC) (Duke University). We display a sample figure of the logged returns squared, auto, and partial autocorrelation plots for Balenciaga.



Based on the correlation functions, the Gucci Google Trend Brand data seems indicative of an ARIMA process of rank (6,1,6) and the arima residuals. We plot the residual values squared of the model and note that they follow a similar, albeit less pronounced path, and this is the pattern that we are looking for (the table below is of residuals squared, and, although we still see patterns of autocorrelation, these are really indicative of autocorrelation features on the level of non-stationary variance, which we have not yet been able to control for). We hope that further regressions on our feature set will help to explain this, as of yet, unexplained pattern of variance.

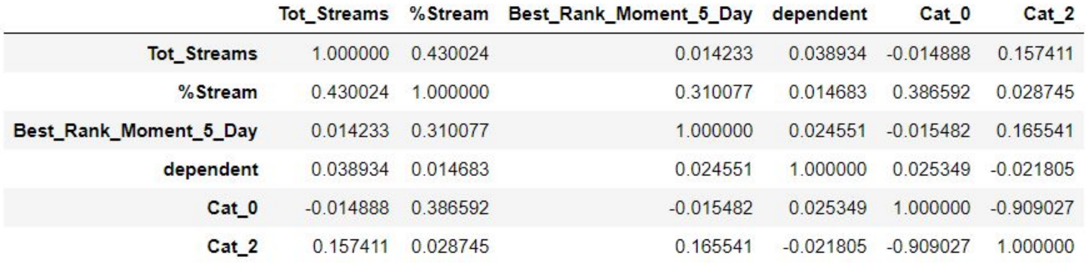


We note that similar results were achieved in using ARIMA models with the rest of our brands. We proceed to compile the residuals into dependent variables, and comment briefly on the state of our dataset.



Independent Variables	Type	Meaning
Total Streams Referencing Brand	Continuous Varaible, {0:n} (n - max daily streaming)	Absolute value to measure streaming quantity
% of Tot Streams with Brand Reference	Continuous Variable, {0:1}	Controls for streaming growth over time
5-Day Rank Momentum	Semi-Continous Variable, {0:1}	Variable to convey whether the song is "Trending"
Categorical Variables	Discrete Categorical {0:5,1}	Variable to capture the "type" of the songs where brand is mentioned
Dependent Variable	Continous Variable, {-n:+n}	ARIMA residuals to capture variance in brand interest (google trend)

At this stage, we have settled on 4 potential types of explanatory variables, with the categorical variable capturing music type adding between 1 or 2 extra dependent variables depending on whether brands are featured in any of the 6 categories. We also generate a sample correlation table of our data in order to depict the current relationships within one of our brands, Gucci, as seen below:

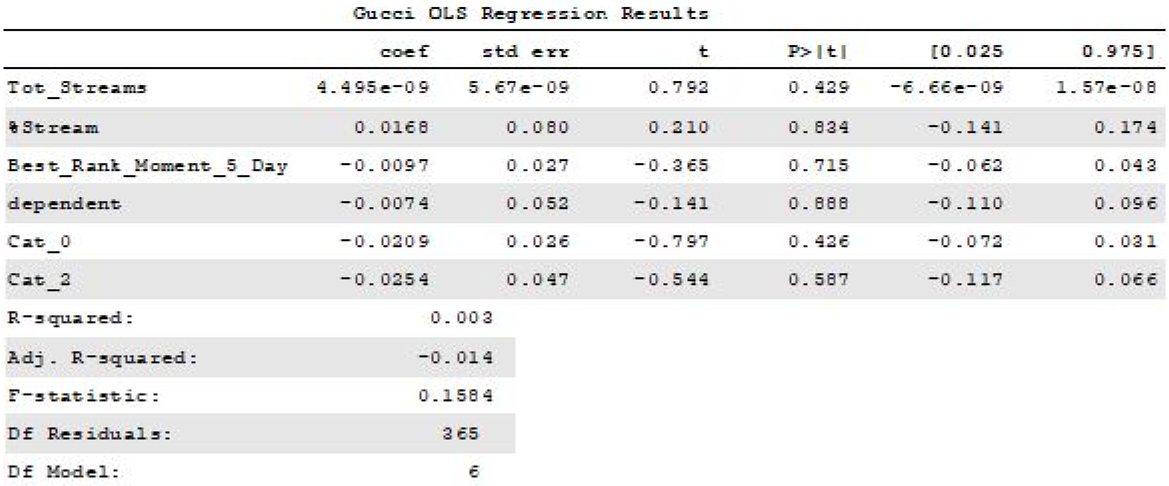


	Tot_Streams	%Stream	Best_Rank_Moment_5_Day	dependent	Cat_0	Cat_2
Tot_Streams	1.000000	0.430024	0.014233	0.038934	-0.014888	0.157411
%Stream	0.430024	1.000000	0.310077	0.014683	0.386592	0.028745
Best_Rank_Moment_5_Day	0.014233	0.310077	1.000000	0.024551	-0.015482	0.165541
dependent	0.038934	0.014683	0.024551	1.000000	0.025349	-0.021805
Cat_0	-0.014888	0.386592	-0.015482	0.025349	1.000000	-0.909027
Cat_2	0.157411	0.028745	0.165541	-0.021805	-0.909027	1.000000

An initial hypothesis at this stage is that our feature set does not correlate well with our dependent variable given any isolated regression, however we are unable to see the effects of multicollinearity, as they may increase the explanatory value of our feature set. Therefore we proceed with the final step of our analysis.

III. Results

After the data cleaning, feature selection, and feature generation processes, we proceeded to regress our dependent brand Google Trends variables on our chosen features. As the correlation analysis in the previous section had depicted, while some of our features correlated with each other some of the time, none of them were particularly correlated with our dependent variable. We draw inspiration from GARCH processes, specifically their use of models that suffer from auto-correlative, heteroskedastic time series trends, and their use of variance as a theoretical base. Our ARIMA generated residual feature, which act in place of the true "variance" therefore have a similar interpretation. The results of our regression analysis can be found in the *Results Appendix*, and we include a table on pg. 9 for our Gucci results for further interpretations:



Gucci OLS Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
Tot_Streams	4.495e-09	5.67e-09	0.792	0.429	-6.66e-09	1.57e-08
%Stream	0.0168	0.080	0.210	0.834	-0.141	0.174
Best_Rank_Moment_5_Day	-0.0097	0.027	-0.365	0.715	-0.062	0.043
dependent	-0.0074	0.052	-0.141	0.888	-0.110	0.096
Cat_0	-0.0209	0.026	-0.797	0.426	-0.072	0.031
Cat_2	-0.0254	0.047	-0.544	0.587	-0.117	0.066

R-squared:	0.003
Adj. R-squared:	-0.014
F-statistic:	0.1584
Df Residuals:	365
Df Model:	6

It’s rather clear from the Gucci Regression results, with an almost zero valued R-squared, that our feature set does not perform well in explaining the variance in the dataset. We included all of the variables in our final feature set as we concluded that an independent regression would not have fared well as a result of poor initial correlation, and were depending on patterns of multicollinearity to generate a more statistically significant model.

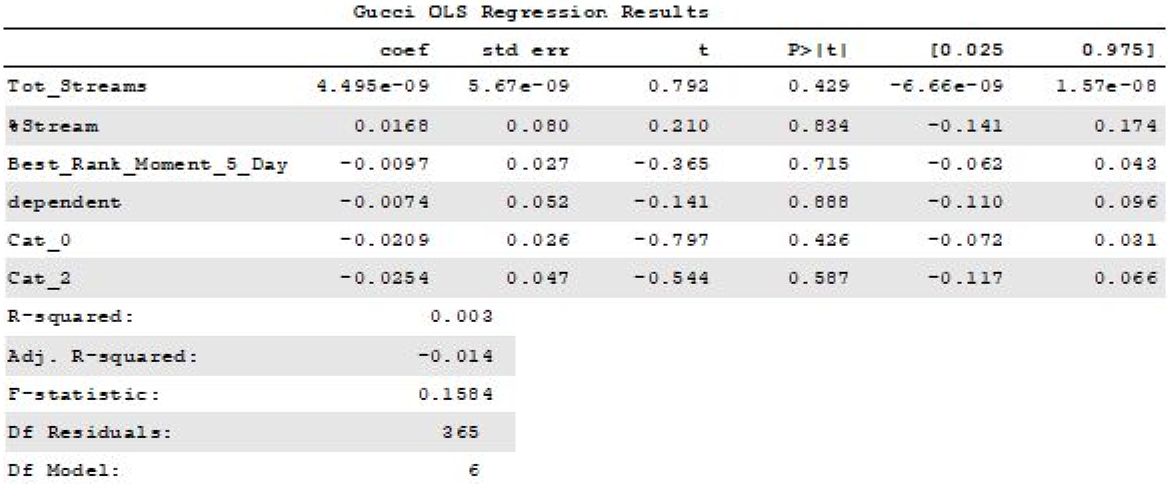
There are many possible interpretations from these results. While initially we considered the possibility that the variables were poor descriptors, remembering the initial correlation analysis, as well as the example of the Saint Laurent music release impact on the google trends search results, we instead prefer the explanation that it is in fact a flaw of the time-scale of our dataset that promoted such results. As a result of the immediacy of google searches (i.e. individuals do not “sit-on” queries, but instead are motivated to search immediately via smart-phone, tablet etc.), it is likely that streaming patterns on a finer time-scale, perhaps hours or minutes, would yield better time-lagged results for predicting search results, and therefore generate actual value for predictive models.

IV. Conclusion

Despite out lackluster results, the process of generating our dataset, of features selection, of creating categorical variables and grouping songs with the help of LDA, and even the intuition behind our use of ARIMA were invaluable, even in as far as their capacity to direct further research with a better understanding of user streaming habits, brand search tendencies, and the influence of different song types. From our understanding, the brand value hypothesis that backs modern day advertising campaigns through artists is likely much deeper than our initial work. Trends in song longevity, repeatability, artist following, and singer/songwriter credibility are likely responsible for the impact that brand features in songs make on user interest and actual brand value, and likely factor into the sponsorship decisions of the majority of brands. While our google trends dependent variable had presented challenges given its time scale, we acknowledge that its ability to act as a proxy variable for actual value is not certain, and perhaps the cumulative listening habits may impact value, in the form of shares or market capitalization in a more deterministic capacity. In addition, if we consider the problem from a simple micro-economic perspective, we find that a grand use of models might be not in identifying which luxury brands are purchased, but perhaps which brands are potential substitutes and which brands receive potential residual value from song features. Perhaps brand mentions are a zero-sum phenomenon of constant brand competition.

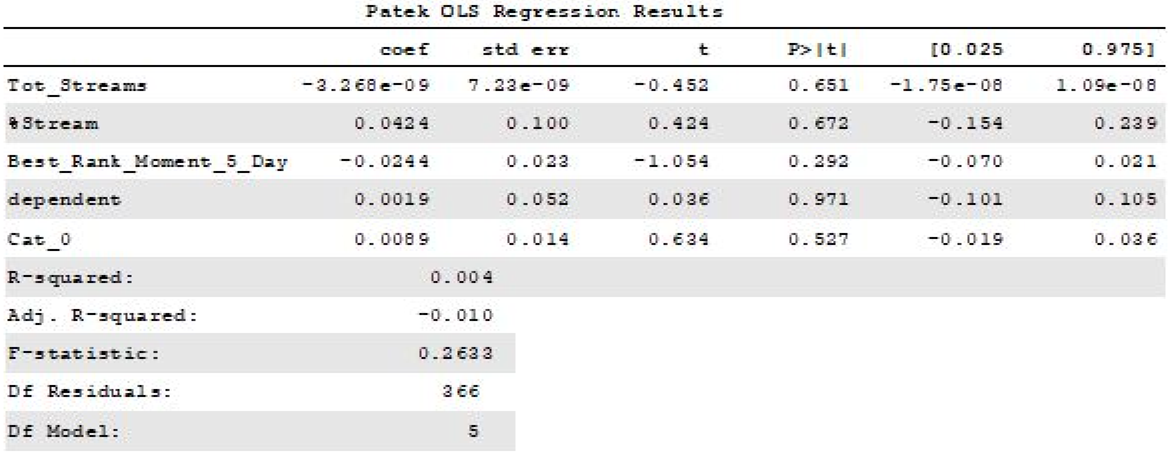
Irregardless of the results, from the get-go, the CRISP-DM framework which we followed throughout the course of the project is fundamentally built on the idea that there is an underlying need for communication and circularity of model design, to which we undeniably subscribe to following the culmination of the research. We now possess a more refined business and data understanding to further improve upon potential future models, in a likewise iterative fashion.

Results Appendix



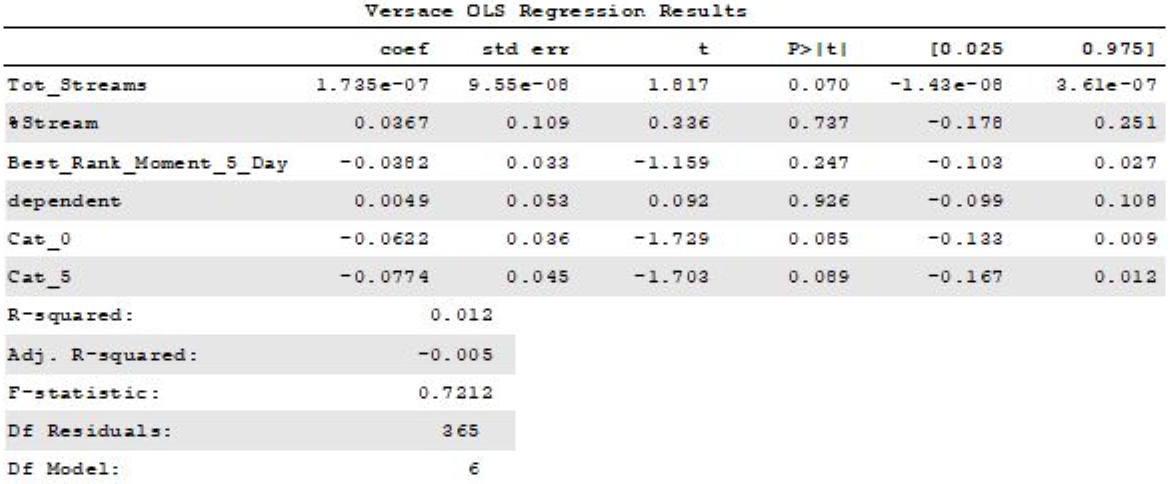
Gucci OLS Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
Tot_Streams	4.495e-09	5.67e-09	0.792	0.429	-6.66e-09	1.57e-08
%Stream	0.0168	0.080	0.210	0.834	-0.141	0.174
Best_Rank_Moment_5_Day	-0.0097	0.027	-0.365	0.715	-0.062	0.043
dependent	-0.0074	0.052	-0.141	0.888	-0.110	0.096
Cat_0	-0.0209	0.026	-0.797	0.426	-0.072	0.031
Cat_2	-0.0254	0.047	-0.544	0.587	-0.117	0.066
R-squared:	0.003					
Adj. R-squared:	-0.014					
F-statistic:	0.1584					
Df Residuals:	365					
Df Model:	6					



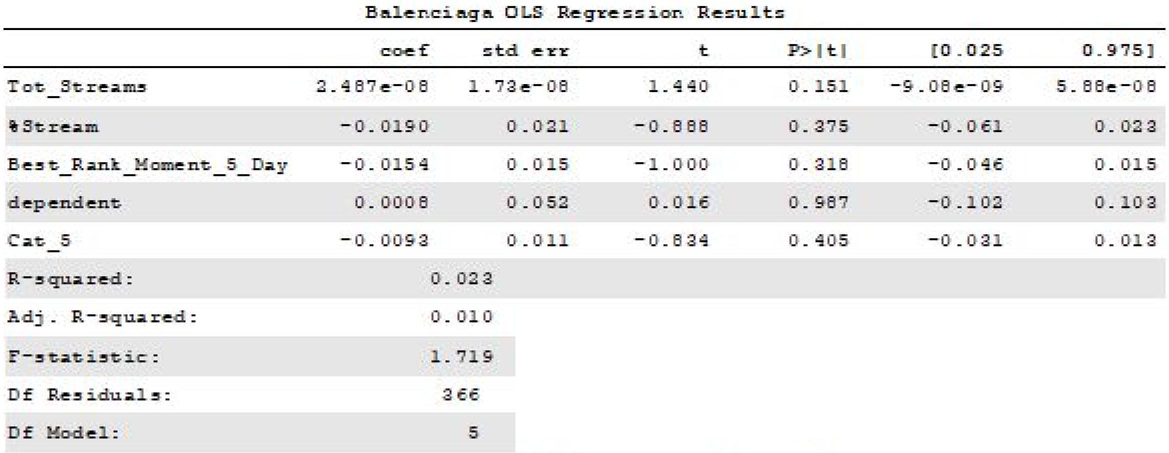
Patek OLS Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
Tot_Streams	-3.268e-09	7.23e-09	-0.452	0.651	-1.75e-08	1.09e-08
%Stream	0.0424	0.100	0.424	0.672	-0.154	0.239
Best_Rank_Moment_5_Day	-0.0244	0.023	-1.054	0.292	-0.070	0.021
dependent	0.0019	0.052	0.036	0.971	-0.101	0.105
Cat_0	0.0089	0.014	0.634	0.527	-0.019	0.036
R-squared:	0.004					
Adj. R-squared:	-0.010					
F-statistic:	0.2633					
Df Residuals:	366					
Df Model:	5					



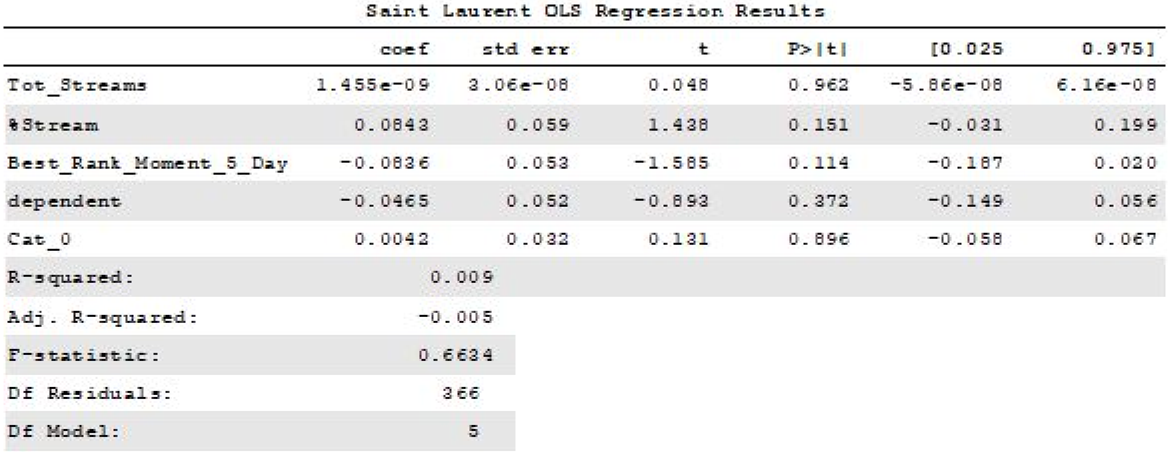
Versace OLS Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
Tot_Streams	1.735e-07	9.55e-08	1.817	0.070	-1.43e-08	3.61e-07
%Stream	0.0367	0.109	0.336	0.737	-0.178	0.251
Best_Rank_Moment_5_Day	-0.0382	0.033	-1.159	0.247	-0.103	0.027
dependent	0.0049	0.053	0.092	0.926	-0.099	0.108
Cat_0	-0.0622	0.036	-1.729	0.085	-0.133	0.009
Cat_5	-0.0774	0.045	-1.703	0.089	-0.167	0.012
R-squared:	0.012					
Adj. R-squared:	-0.005					
F-statistic:	0.7212					
Df Residuals:	365					
Df Model:	6					



Balenciaga OLS Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
Tot_Streams	2.487e-08	1.73e-08	1.440	0.151	-9.08e-09	5.88e-08
%Stream	-0.0190	0.021	-0.888	0.375	-0.061	0.023
Best_Rank_Moment_5_Day	-0.0154	0.015	-1.000	0.318	-0.046	0.015
dependent	0.0008	0.052	0.016	0.987	-0.102	0.103
Cat_5	-0.0093	0.011	-0.834	0.405	-0.031	0.013
R-squared:	0.023					
Adj. R-squared:	0.010					
F-statistic:	1.719					
Df Residuals:	366					
Df Model:	5					



Saint Laurent OLS Regression Results

	coef	std err	t	P>|t|	[0.025	0.975]
Tot_Streams	1.455e-09	3.06e-08	0.048	0.962	-5.86e-08	6.16e-08
%Stream	0.0843	0.059	1.438	0.151	-0.031	0.199
Best_Rank_Moment_5_Day	-0.0836	0.053	-1.585	0.114	-0.187	0.020
dependent	-0.0465	0.052	-0.893	0.372	-0.149	0.056
Cat_0	0.0042	0.032	0.131	0.896	-0.058	0.067
R-squared:	0.009					
Adj. R-squared:	-0.005					
F-statistic:	0.6634					
Df Residuals:	366					
Df Model:	5					

Citations

Shearer, Colin. "The CRISP-DM Model: The New Blueprint for Data Mining." *Journal of Data Warehousing*, vol. 5, no. 4, 2000, pp. 13–22.

Bollen, Johan, and Huina Ma. "Twitter Mood Predicts the Stock Market." *Indiana University*, 2010, pp. 1–8.

Richter, Felix. "Infographic: Spotify Reaches 100 Million Premium Subscribers." *Statista Infographics*, 2019, www.statista.com/chart/15697/spotify-user-growth/.

Arienti, Patrizia. "Global Powers of Luxury Goods 2018." *Deloitte.com*, 2018, www2.deloitte.com/content/dam/Deloitte/at/Documents/consumer-business/deloitte-global-powers-of-luxury-goods-2018.pdf.

Song, Xiaojun, and Abderrahim Taamouti. "A Better Understanding of Granger Causality Analysis: A Big Data Environment." *Oxford Bulletin of Economics and Statistics*, Dec. 2018.

Hecq, Alain, and Stephan Smeekes. "Granger Causality Testing in High-Dimensional VARs: a Post-Double-Selection Procedure." *Cornell University*, Feb. 2019.

University, Duke. "Summary of Rules for Identifying ARIMA Models." *People.Duke.Edu*, people.duke.edu/~rnau/arimrule.htm.