

Data Warehousing and Business Intelligence Project

on

Historical Brand Value and Public Opinion Analysis of Most
Valuable Brands in 2019

Ricardo S. da Silva Junior
18147607

MSc in Data Analytics – 2018/9

Submitted to: Noel Cosgrave

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing



| | |
|-----------------------------|------------------------------------------------------------------------------------|
| Student Name: | Ricardo S. da Silva Junior |
| Student ID: | 18147607 |
| Programme: | MSc Data Analytics |
| Year: | 2018/9 |
| Module: | Data Warehousing and Business Intelligence |
| Lecturer: | Noel Cosgrave |
| Submission Due Date: | 05/01/2020 |
| Project Title: | Historical Brand Value and Public Opinion Analysis of Most Valuable Brands in 2019 |

I hereby certify that the information contained in this (my submission) is information pertaining to my own individual work that I conducted for this project. All information other than my own contribution is fully and appropriately referenced and listed in the relevant bibliography section. I assert that I have not referred to any work(s) other than those listed. I also include my TurnItIn report with this submission.

ALL materials used must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is an act of plagiarism and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

| | |
|-------------------|-----------------|
| Signature: | |
| Date: | January 5, 2020 |

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| Office Use Only | |
|----------------------------------|--|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

Table 1: Mark sheet – do not edit

| Criteria | Mark Awarded | Comment(s) |
|--------------|--------------|------------|
| Objectives | of 5 | |
| Related Work | of 10 | |
| Data | of 25 | |
| ETL | of 20 | |
| Application | of 30 | |
| Video | of 10 | |
| Presentation | of 10 | |
| Total | of 100 | |

Project Check List

This section capture the core requirements that the project entails represented as a check list for convenience.

- ☒ Used L^AT_EX template
- ☐ Three Business Requirements listed in introduction
- ☐ At least one structured data source
- ☐ At least one unstructured data source
- ☐ At least three sources of data
- ☐ Described all sources of data
- ☐ All sources of data are less than one year old, i.e. released after 17/09/2017
- ☐ Inserted and discussed star schema
- ☐ Completed logical data map
- ☐ Discussed the high level ETL strategy
- ☐ Provided 3 BI queries
- ☐ Detailed the sources of data used in each query
- ☐ Discussed the implications of results in each query
- ☐ Reviewed at least 5-10 appropriate papers on topic of your DWBI project

Historical Brand Value and Public Opinion Analysis of Most Valuable Brands in 2019

Ricardo S. da Silva Junior
18147607

January 5, 2020

Abstract

In the business market, Brand Value is known as the financial worth of the brand and it is believed that popular brands tend to have higher brand value because all this publicity can attract many customers and investors. This project aims to develop a data warehouse for analysis of the fifty most valuable brands worldwide in 2019 combining information about Geolocation, brand value and related twitter user's opinion sentiment analysis during the last five years. The findings of this report shows that USA is the country with more leading brands and, consequently, highest total sum of brand value, but in the other hand China was the country that achieved the highest growth rate of total sum of brand value with more than 200% rise. Not very surprisingly, Amazon, the most valuable brand in 2019, was also the most popular brand during the twitter data capture, with more mentions and relatively good overall sentiment analysis.

1 Introduction

With the advent of e-commerce and big data, new business intelligence techniques were developed to implement different approaches to obtain customer retention and also to attract new customers to the companies using the web platforms like social media networks.

The free publicity generated by good customer reviews and public satisfaction in social media can be a factor of good marketing generator for a given company or brand. In other hand, this "free publicity", if created by bad reviews, is a factor of negative impact to the company resulting in big losses of customers and consequently money loss and, if business decisions are not taken straight away it could generate permanent damage to the business.

The use of analytical technologies for extracting sentiment analysis of comments in the web platforms of customers about a certain product or service of a company are an important tool for business intelligence applications.

This paper proposes a creation of a data warehouse built upon the analysis of the currently top 50 brands worldwide, collecting data of brand value, twitter sentiment analysis and mentions based in geolocation and time variation of the leading brands.

The aim of this data warehouse is to understand the effects of customer comments and mentions with relation to the brand value of leading brands. This study can be used to support future business decisions for creating and follow the impact of marketing campaigns based in web reviews to enhance customers satisfaction and consequently increase brand value. For that, the 50 most valuable brands worldwide in 2019 are collected from Brand Finance Global 500 Report, the information about geolocation of each company is then obtained from dataset in kaggle and the sentiment analysis is performed using the reviews of customers available in the twitter platform.

The three requisites for this project are:

- (Req-1) The data warehouse shall provide for each year in analysis the sum of brand value of the leading brands grouped by country;
- (Req-2) The solution shall provide the distribution of the most valuable brands in each sector by continent;
- (Req-3) The data warehouse shall provide the daily popularity of each brand and one or more scores of types of sentiment analysis.

2 Data Sources

The datasets used to populate the presented data warehouse were obtained from three different sources: Brand Finance with the annual brands value report, Kaggle platform and the microblogging social media Twitter.

| Source | Type | Brief Summary |
|----------------------------------------------------|-----------------|------------------------------------------------------------------------------------------------------------------|
| Brand Finance Global 500 Reports from 2015 to 2019 | Semi-structured | The list of five .pdf reports to extract information of the 50 leading brands of each year between 2015 and 2019 |
| Kaggle | Structured | Provides information of geolocation (country, continent and sub region) |
| Twitter | Unstructured | Provides a set of tweets related to the brands in study for later sentiment analysis |

Table 2: Summary of sources of data used in the project

2.1 Source 1: Brand Finance Global 500 Annual Reports

Brand Finance is an enterprise with expertise in brand valuation and strategy consultancy. The webpage of the company provides a set of financial reports about brand valuation available online to the public.

The datasets containing brand information and the last 5 years of brand value of the 50 currently most valuable brands worldwide were obtained from brand finance web platform: <https://brandfinance.com/knowledge-centre/reports/> from a list of annual reports of the 500 leading world brands from past years.

Brand Finance Global 500 (USD m).

Top 500 most valuable brands 1-50

| 2019 Rank | 2018 Rank | Brand | Country | Sector | 2019 Brand Value | Brand Value Change | 2018 Brand Value | 2019 Brand Rating | 2018 Brand Rating |
|-----------|-----------|-------------|---------------|--------|------------------|--------------------|------------------|-------------------|-------------------|
| 1 | 1 | ← Amazon | United States | Tech | \$187,905 | +24.6% | \$150,811 | AAA- | AAA- |
| 2 | 2 | ← Apple | United States | Tech | \$153,634 | +5.0% | \$146,311 | AAA | AAA+ |
| 3 | 3 | ← Google | United States | Tech | \$142,755 | +18.1% | \$120,911 | AAA | AAA+ |
| 4 | 6 | ↑ Microsoft | United States | Tech | \$119,595 | +47.4% | \$81,163 | AAA | AAA+ |

Figure 1: Brand Finance: Global 500 2019

The five reports are: 1.Global 500 2019 (<https://brandfinance.com/knowledge-centre/reports/brand-finance-global-500-2019/>) with brand value information of the year 2019; 2.Global 500 2018 (<https://brandfinance.com/knowledge-centre/reports/brand-finance-global-500-2018/>) with brand value information of the year 2018; 3.Global 500 2017 (<https://brandfinance.com/knowledge-centre/reports/brand-finance-global-500-2017/>) with brand value information of the year 2017; 4.Global 500 2016 (<https://brandfinance.com/knowledge-centre/reports/brand-finance-global-500-2016/>) with brand value information of the year 2016; and 5.Global 500 2015 (<https://brandfinance.com/knowledge-centre/reports/brand-finance-global-500-2015/>) with brand value information of the year 2015.

Each .pdf report constitutes a semi-structured source of data with a table that contains the name of the brands, country, position in the rank and the brand value from the current and the past year. However, relevant to this project are name, country and brand value of the current year of the report.

This dataset addresses the business requirements listed in Section 1 in the following ways: providing the list of 50 leading brands in 2019 and their brand value from last 5 years.

2.2 Source 2: Kaggle

Kaggle is a well known web community widely used for researchers, students, educators and data science professionals around the world for publishing and finding datasets and also exchanging acknowledgement.

The second source of data used for populating this datawarehouse contains extra information of geolocation that can be combined to the brands details. The data was retrieved from <https://www.kaggle.com/statchaitya/country-to-continent> in a comma separated values (.csv) file with 9 columns and 249 rows.

The downloaded file comprehends a structured source of data containing nine columns with geolocation information like: country, code_2, code_3, country_code, iso_3166_2, continent, sub_region, region_code, sub_region_code. However, relevant to this project are: country, continent and sub_region.



| countryContinent.csv (15.93 KB) | | | | | | | 6 of 9 columns | |
|---------------------------------|----------------------|----------------------|-----------------------------------------------------------------------------------|---------------------------------------------|------------------------------------------------------|-------------------------------------------------------------------------------------|----------------|--|
| | country | code_2 | # country_code | continent | sub_region | # region_code | | |
| | 249 unique values | 249 unique values |  | Africa 23% Americas 22% Other (4) 55% | Caribbean 11% Eastern Africa 8% Other (21) 81% |  | | |
| 1 | Afghanistan | AF | 4 | Asia | Southern Asia | 142 | | |
| 2 | Åland Islands | AX | 248 | Europe | Northern Europe | 150 | | |
| 3 | Albania | AL | 8 | Europe | Southern Europe | 150 | | |
| 4 | Algeria | DZ | 12 | Africa | Northern Africa | 2 | | |
| 5 | American Samoa | AS | 16 | Oceania | Polynesia | 9 | | |
| 6 | Andorra | AD | 20 | Europe | Southern Europe | 150 | | |
| 7 | Angola | AO | 24 | Africa | Middle Africa | 2 | | |
| 8 | Anguilla | AI | 660 | Americas | Caribbean | 19 | | |

Figure 2: Kaggle: Country to Continent

This dataset addresses the business requirements listed in Section 1 in the following ways: providing additional Geolocation information of brands like: country, sub region and continent.

2.3 Source 3: Twitter

Twitter is one of the most popular social network platforms. It is basically a microblog where users can post public short messages of 280 characters called as tweets. Users tend to use this social media to express thoughts. The high amount of twitter users around the world (more than 300 million active users according to Statista (2019)) makes this social media a very attractive source of data of comments and reviews for sentiment analysis.

Many people before buying a product or service nowadays use internet search engines to find reviews and make their mind. So, it is important for the companies to follow what people say about their products and services in the internet. Twitter is one of most common platforms used by customers for reviewing or complaining.



Figure 3: Tweets can express the satisfaction level of a customer

The third source of data used in this project is constituted by unstructured data provided by comments about the brands posted by twitter users. The tweets were collected during the Christmas week from 21st to 28th of december of 2019. In total, 1.3 Million tweets were captured using the twitter API during the specified dates. After removing retweets (replicated tweets) and tweets that were not in english, it was left a bit more than 441K tweets, that were later used for lexical sentiment analysis using the techniques

proposed by Årup Nielsen (2011), Hu & Liu (2004), Mohammad & Turney (2010).

The tweets were then grouped by date and brand and the average sentiment score was calculated, resulting in a dataset composed by each day, the list of brands, mean sentiment score and number of tweets.

This dataset addresses the business requirements listed in Section 1 in the following ways: providing tweets about the brands for future sentiment analysis and popularity calculation.

3 Related Work

One of the first works in the literature of twitter sentiment analysis was developed by Petrović et al. (2010) where 97 Million tweets posts were collected using the twitter API and then later categorized. Many studies were developed until 2013, but then twitter authorities decided to withdraw all public twitter datasets and no further works could be done with this data anymore.

There are many methods and technologies available for evaluation and opinion mining sentiment analysis. The works from Årup Nielsen (2011), Hu & Liu (2004) and Mohammad & Turney (2010) became popular techniques of lexicon sentiment analysis based in single words that are called by unigrams. One of the main reasons why those techniques were selected in this project is that their dictionary of lexicons was built and validated using crowdsourcing, reviews and twitter data, making an accurate sentiment analysis of posts on twitter.

The method proposed by Årup Nielsen (2011) (AFFIN lexicon) consists in giving to each word an score between a spectrum of negative sentiment (with values close to -5), to positive sentiment (values close to +5). The overall sentiment is calculated by the sum of the scores, a result more than 0 indicates a positive sentiment (good reviews), and the opposite a negative sentiment (bad reviews).

| | word | value |
|----|----------|-------|
| | <chr> | <dbl> |
| 1 | darkness | -1 |
| 2 | like | 2 |
| 3 | stop | -1 |
| 4 | tired | -2 |
| 5 | wasting | -2 |
| 6 | gift | 2 |
| 7 | free | 1 |
| 8 | dead | -3 |
| 9 | true | 2 |
| 10 | thanks | 2 |

Figure 4: Sample AFFIN lexicon summary

Different from the Årup Nielsen (2011) approach, Hu & Liu (2004) (BING lexicon) doesn't work with scores for each word but categorizing them into two classes: positive or negative. The overall sentiment is then given by calculating the difference between the two classes.

| | word | sentiment |
|----|----------|-----------|
| | <chr> | <chr> |
| 1 | darkness | negative |
| 2 | like | positive |
| 3 | dark | negative |
| 4 | tired | negative |
| 5 | wasting | negative |
| 6 | freaking | negative |
| 7 | free | positive |
| 8 | dead | negative |
| 9 | right | positive |
| 10 | aspire | positive |

Figure 5: Sample BING lexicon summary

The Mohammad & Turney (2010) (NRC lexicon) like Hu & Liu (2004) method works making a classification of each word but with a larger set of categories like: positive, negative, fear, joy, surprise, among others. Each word of the lexicon is characterized by a vector of yes/no columns indicating the categories that the given word is part of. The result of the sentiment analysis is given the by a vector containing for each column (representing the categories) the number of words classified in the respective category.

| | word | sentiment |
|----|----------|--------------|
| | <chr> | <chr> |
| 1 | resident | positive |
| 2 | delivery | anticipation |
| 3 | delivery | positive |
| 4 | darkness | anger |
| 5 | darkness | fear |
| 6 | darkness | negative |
| 7 | darkness | sadness |
| 8 | long | anticipation |
| 9 | time | anticipation |
| 10 | dark | sadness |

Figure 6: Sample NRC lexicon summary

Sentiment extracted from twitter can be an indicator of real life facts. Petrovic et al. (2010) performed a study to find correlation between twitter sentiment and public opinion measured from pool results. Over one billion of politic opinion tweets were extracted between 2008 and 2009. The dataset was analysed and correlated to facts and the experiment shows that more than 80% of correlation was achieved by the public opinion twitter sentiment analysis and the actual results of the elections.

With similar thinking of Petrovic et al. (2010) this datawarehouse is developed to analyse public opinion but instead of political election results, the brand value is the fact in analysis. Five years of brand value and public opinion on twitter data are analysed for future correlation tests. The geolocation of the companies is also an important factor for this project because it implies to the world visibility of the brand.

The dimensional modelling of the proposed datawarehouse follow the fundamental techniques based in Kimball & Ross (2013), one of the most popular books in the literature for developing a datawarehouse and business intelligence.

4 Data Model

The multidimensional model of the proposed datawarehouse project is composed by three dimension tables (DimBrand, DimLocation and DimDate) and one fact table (FactTable) following the structure of the star schema displayed in the figure bellow:

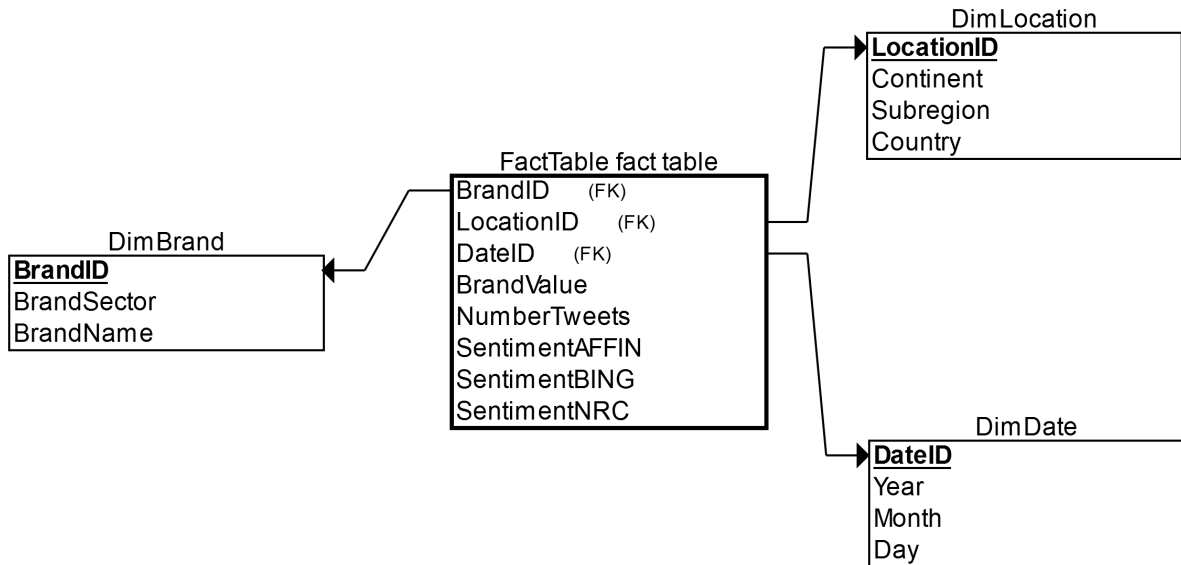


Figure 7: Star Schema Data Model

The first dimension modeled for this project was called DimBrand containing the relevant information of the brands in study. In total the fifty most valuable brands around the world were collected from the first source of data 2.1 Brand Finance. This dimension provides the industrial sector of the brand (BrandSector) and the name (BrandName) and can be referred by the primary key representing the brand identifier (BrandID). This is an hierarchical dimension composed by two levels of hierarchy sector (BrandSector) and Brand (BrandName).

The second dimension modeled for this project was called DimLocation and it contains the relevant information of the Location of the analysed brands in this study. The geolocation information used to populate this dimension was retrieved from the second source of data 2.2 Kaggle. This dimension provides for each country (Country), the sub-region (Subregion) and the continent (Continent). Each location can be referred by the primary key representing the location identifier (LocationID). This is also an hierarchical dimension composed by three levels of hierarchy Continent, Subregion and Country.

DimDate is the third dimension modeled for this datawarehouse and it contains the relevant information of the Dates of the brand related tweets where sentiment analysis was performed. The date information used to populate this dimension was generated automatically and then combined to the third source of data 2.3 Twitter. This dimension provides for each Date the day (Day), the month (Month) and the year (Year). Each date can be referred by the primary key representing the date identifier (DateID). This is also an hierarchical dimension composed by three levels of hierarchy Year, Month and Day.

The composition of the three dimensions above generates the fact table (FactTable). It contains the brand value (BrandValue), the number of tweets (NumberTweets), the AFFIN calculated sentiment analysis (SentimentAFFIN), the BING calculated sentiment analysis (SentimentBING) and the NRC calculated sentiment analysis (SentimentNRC).

Each brand from a specific location in a given year has a brand value, so the combination of the dimension brand, location and date can satisfy the first requirement of this data warehouse listed in Section 1. By combining the same dimensions the second requirement can also be satisfied listing the number of brands by sector and continent in 2019. The third requirement is satisfied by calculating the sum of brand value of brands of each country for each year available in the data warehouse.

5 Logical Data Map

The following logical data map describes how the relevant fields in each source of data presented in Section 2 map to the target data model fields during the ETL process.

Table 3: Logical Data Map describing all transformations, sources and destinations for all components of the data model illustrated in Figure 7

| Source | Column | Destination | Column | Type | Transformation |
|--------|----------------------------|-------------|-------------|-----------|----------------------------------------------------------------------------------------------------|
| 1 | Brand | DimBrand | BrandName | Dimension | Non latin alphabet characters removed (E.g. Apostrophe, Asterisk, At sign, Hyphen-minus, Low line) |
| 1 | Sector | DimBrand | BrandSector | Dimension | Non latin alphabet characters removed (E.g. Apostrophe, Asterisk, At sign, Hyphen-minus, Low line) |
| 1 | Brand Value | FactTable | BrandValue | Fact | Non numerical characters removed (E.g. \$, comma) |
| 2 | country | DimLocation | Country | Dimension | Non latin alphabet characters removed (E.g. Apostrophe, Asterisk, At sign, Hyphen-minus, Low line) |
| 2 | continent | DimLocation | Continent | Dimension | Non latin alphabet characters removed (E.g. Apostrophe, Asterisk, At sign, Hyphen-minus, Low line) |
| 2 | sub_region | DimLocation | Subregion | Dimension | Non latin alphabet characters removed (E.g. Apostrophe, Asterisk, At sign, Hyphen-minus, Low line) |
| 3 | created_at (YYYY-MM-DD) | DimDate | Day | Dimension | Remove the first 8 characters (I.e. YYYY-MM-) |
| 3 | created_at (YYYY-MM-DD) | DimDate | Month | Dimension | Remove the first 5 characters (I.e. YYYY-) AND remove the last 3 characters (I.e. -DD) |
| 3 | created_at (YYYY-MM-DD) | DimDate | YEAR | Dimension | Remove the last 6 characters (I.e. -MM-DD) |

Continued on next page

Table 3 – *Continued from previous page*

| Source | Column | Destination | Column | Type | Transformation |
|--------|------------|-------------|----------------|------|------------------------------------------------------------------------------------------------------|
| 3 | tweet_text | FactTable | SentimentAFFIN | Fact | Result of the performed AFFIN sentiment analysis (positive - negative) |
| 3 | tweet_text | FactTable | SentimentBING | Fact | Result of the performed BING sentiment analysis (positive classifications - negative classification) |
| 3 | tweet_text | FactTable | SentimentNRC | Fact | Result of the performed NRC sentiment analysis (positive classifications - negative classifications) |
| 3 | tweet_text | FactTable | NumberTweets | Fact | Count of tweets |

6 ETL Process

In this section the process of moving the data from the various sources into the data warehouse is detailed, as well as the transformations needed to conform the different types and formats of data.

The first step of the ETL process is creating a staging csv file with the brand information. The report with the 50 most valuable brands from 2019 from Source 2.1 Brand Finance is automatically downloaded from the url: https://brandfinance.com/images/upload/global_500_2019_locked_1.pdf. Then, using the R programming language tools the pdf file is scrapped to get the list of the 50 brands, their sector and country. The non Latin alphabet characters are removed from the brand, sector and country. The output produced by this step is a .csv file called brands.info.csv (brand_id, brand_name, country, sector). The produced file is automatically loaded to the table RAWBRAND-INFO with the same columns in the data warehouse using a data flow task in SSIS from flat file source to ole db destination.

The second phase of the ETL process is creating a staging csv file with the brand value from 2015 to 2019 of each one of the 50 brands. The reports from the last five years of brand value from Source 2.1 Brand Finance are automatically downloaded from the urls: https://brandfinance.com/images/upload/global_500_2015_for_print.pdf (2015), https://brandfinance.com/images/upload/global_500_2016_website.pdf (2016), https://brandfinance.com/images/upload/global_500_2017_locked_website.pdf (2017), https://brandfinance.com/images/upload/brand_finance_global_500_report_2018_locked_1.pdf (2018), https://brandfinance.com/images/upload/global_500_2019_locked_1.pdf (2019). Then a new dataframe is created for each brand from brands.info.csv, with the brand value of each year. The non numerical characters (i.e. \$, comma) are removed from the brand value. This process generates an output file called brand_values.csv (brand_id, brand_name, year, value). The produced file is automatically loaded to the table RAWBRANDVALUES with the same columns in the data warehouse using a data flow task in SSIS from flat file source to ole db destination.

The third part of the ETL process is downloading and transforming the geolocation information. The file from Source 2.2 Kaggle is manually downloaded (sitting behind a login) from the url: <https://www.kaggle.com/statchaitya/country-to-continent>. A new data frame is created with country, sub_region and continent and the non Latin alphabet characters are removed. This process generates an output file called locations.csv (country, sub_region, continent). The produced file is automatically loaded to the table RAWLOCATIONS with the same columns in the data warehouse using a data flow task in SSIS from flat file source to ole db destination.

The fourth stage of the ETL process is using the twitter api to collect tweets related to the brands. For each brand a set of @name and #Hashtag was generated and then followed the steps from the tutorial for collecting tweets from <http://140dev.com/free-twitter-api-source-code-library/>. After a few days running, the table containing the tweets collected from the brands was exported to a csv file. The three sentiment analysis were performed for each tweet grouping them by date of tweet, brand and average sentiment scores calculated. The output generated is a csv file called

tweets_analysis.csv with date, brand_id, sentiment_affin, sentiment_bing, sentiment_nrc and tweets_count. The produced file is automatically loaded to the table RAWSENTIMENTS with the same columns in the data warehouse using a data flow task in SSIS from flat file source to ole db destination.

7 Application

In this section three BI Queries are created to satisfy each one of the three requirements listed in Section 1. Diverse features of the software Tableau 10.4.0 were used in order to create the graphics for more intuitive data visualization.

7.1 BI Query 1: What is the annual change of total brand value of leading brands in each country?

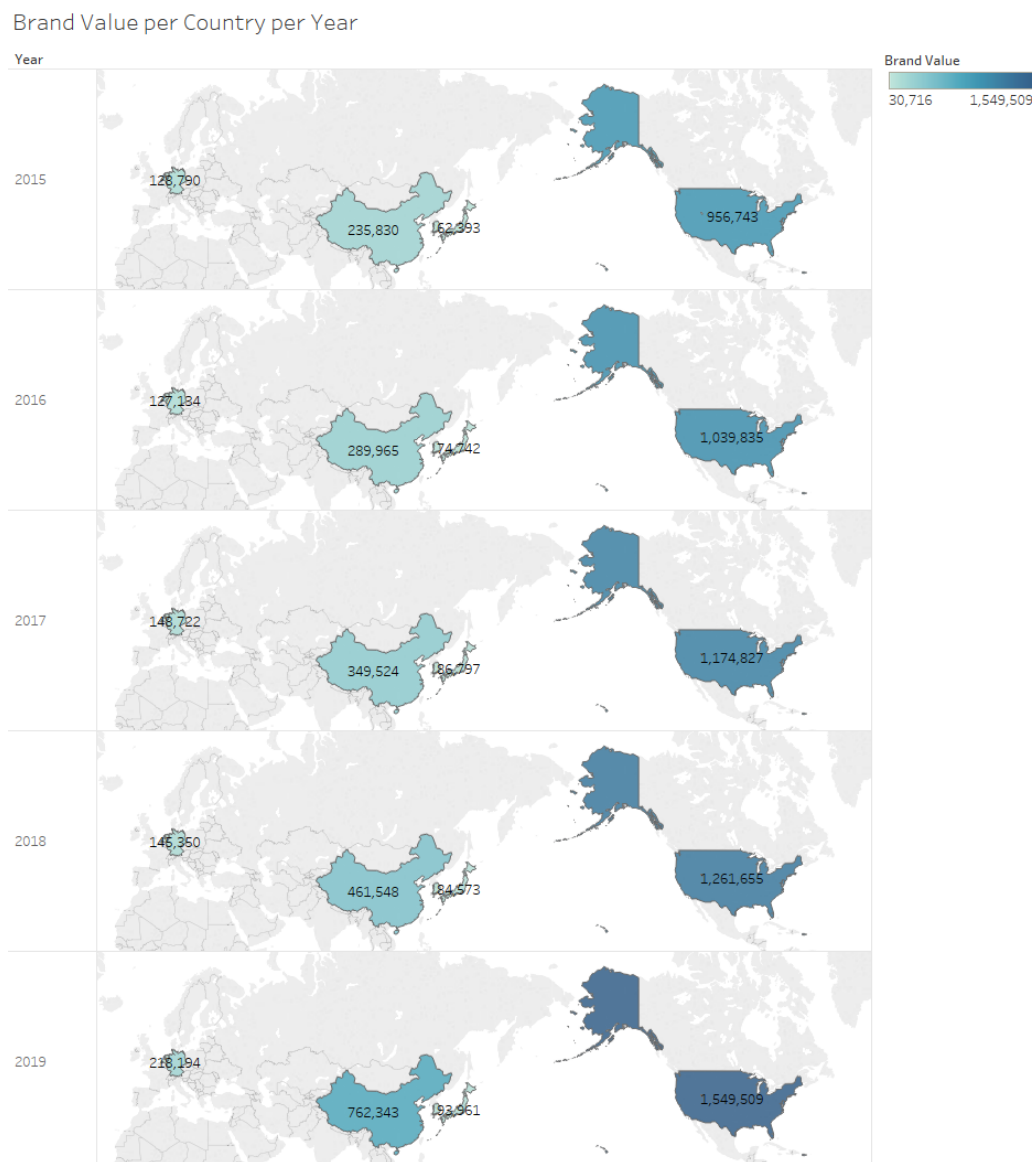


Figure 8: Results for BI Query 1

For this query, the contributing sources of data are: 2.1 Brand Finance with the list of brands and brand value; and 2.2 Kaggle providing information of the Geolocation.

As illustrated in Figure 8, the general findings shows that the USA is the country with highest total sum of brand value, and have always been during the last 5 years, followed by China and Germany. We can also identify that the total sum of brand value of the USA grew more than 60% between 2015 and 2019, China had the highest growth with 223% and Germany 59%.

7.2 BI Query 2: Which sectors have more occurrence of leading brands in each continent?

For this query, the contributing sources of data are: 2.1 Brand Finance with the list of brands and their industrial sector; and 2.2 Kaggle providing information of the Geolocation.

Comparison of Amount of Leading Brands per Sectors per Continent

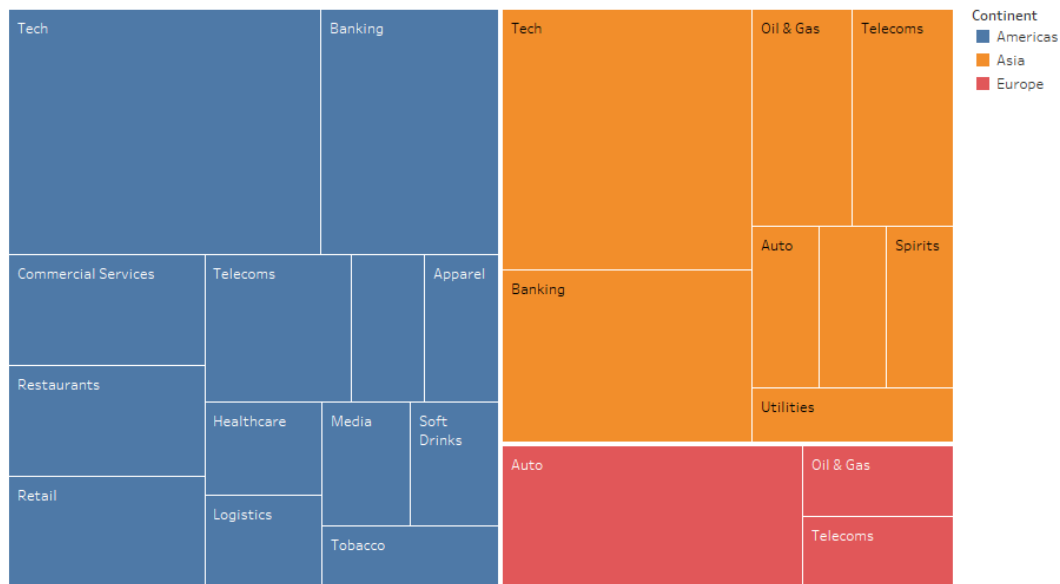


Figure 9: Results for BI Query 2

The general findings in Figure 9 shows that the two leading sectors are the same in Asia and the Americas, those are Tech and Banking. The European leading sector is the Automobile and the Americas is the continent with more diversified sectors with leading brands.

7.3 BI Query 3: Which brands have the most active engagement on Twitter?

For this query, the contributing sources of data are: 2.1 Brand Finance with the list of brands; and 2.3 Twitter providing tweets for future sentiment analysis and popularity calculation.

As illustrated in Figure 10, the general findings shows that Amazon, China Construction Bank, Youtube, Google and Apple are the top five most popular brands on twitter

between 21st to 28th of December, although, the brand with highest sentiment analysis in this period was Disney. Its also very impressive how popular was the brand Amazon with more than 13.5 thousand relevant tweets collected during the period.

Sentiment of Trending Brands

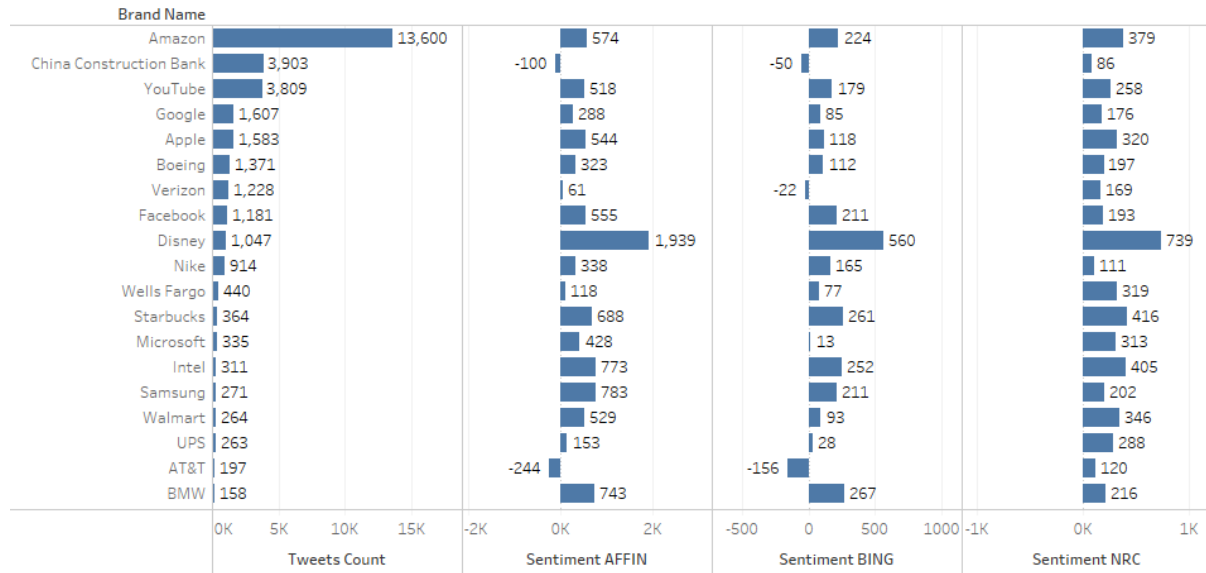


Figure 10: Results for BI Query 3

7.4 Discussion

The insights gathered from the first BI Query indicate that the brand value growth between the years of 2015 and 2019 of the USA and Germany was quite similar, around 60%. The economic ascension of China through the years can be visualized by the growth of leading brands in the country making the brand value rise in more than 200%. It is a very good indicator of China's worth for world brand investors. This way China can overcome USA in cumulative brand value of leading brands in less than 10 years.

The second BI Query exemplifies the competition of the Americas (Mostly USA) and Asia (Mostly China) in Tech, Banking and even Telecoms sectors. Although the Americas are still the Continent with more leading brands in 2019. The second BI Query also evidenced the leading sector in Europe, the Automobile sector is the one with more leading brands in Europe.

The third BI Query shows that two most commented brands on twitter are: 1st Amazon (USA) and 2nd China Construction Bank (China). Amazon is also the most valuable brand in 2019, what show us a relation between brand value and twitter popularity. The more people talking about a brand the higher is its brand value. It indicates that it is important for a business/company/brand to analyse public opinion to add more brand value to the brand.

8 Conclusion and Future Work

Although, the gathered tweets were more than 1.3 million, the application was only able to use a bit more than 400 thousand tweets. The reasons for discarding those tweets could be because they were retweets (replicated tweets), or they were in a different language than English, an emoticon or just urls/links.

To overcome the limits of lexicon based unigrams for sentiment analysis, a suggestion of future work could be using also machine learning techniques for text processing. More powerful techniques for data analysis allied to this data warehouse can bring more insights about the facts that can variate the brand value. Another suggestion is to use more data sources in order to collect other features that can be of good addition to the project.

The proposed data warehouse can be used as a business application for analysis of world leading brands. Correlations between features like Geolocation, brand value, sector and public opinion can be tested and analysed. This project can be very useful for future analysis of marketing strategies of leading brands and their effects to the brand value variation.

References

- Hu, M. & Liu, B. (2004), Mining and summarizing customer reviews, *in* ‘Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining’, KDD ’04, Association for Computing Machinery, New York, NY, USA, p. 168–177.
URL: <https://doi.org/10.1145/1014052.1014073>
- Kimball, R. & Ross, M. (2013), *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*, 3rd edn, Wiley Publishing.
- Mohammad, S. M. & Turney, P. D. (2010), Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon, *in* ‘Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text’, CAAGET ’10, Association for Computational Linguistics, USA, p. 26–34.
- Petrovic, S., Osborne, M. & Lavrenko, V. (2010), The edinburgh twitter corpus, *in* ‘Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media’, Association for Computational Linguistics, pp. 25–26.
- Petrović, S., Osborne, M. & Lavrenko, V. (2010), The edinburgh twitter corpus, pp. 25–26.
- Statista (2019), ‘Twitter: number of active users 2010-2019’.
URL: <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>
- Årup Nielsen, F. (2011), ‘A new anew: Evaluation of a word list for sentiment analysis in microblogs’.

Appendix

twitter_ETL.R

```
library(reshape)
library(textcat)
library(cld2)
library(cld3)
library(tidyverse)
library(tidytext)
library(textdata)

#1367172
#loading tweets and cleaning
tweets <- read.csv("tweets.csv", quote = "", sep = "\t", row.names = NULL)
tweets$tweet_text <- as.character(tweets$tweet_text)

#removing retweets
tweets$RT <- startsWith(tweets$tweet_text, "RT")
tweets <- tweets[!tweets$RT, ]

#creating the hashtags variables
brands_names <- c(1:50)
comment <- "brands_names<-c('Amazon',
'Apple',
'Google',
'Microsoft',
'Samsung',
'AT&T',
'Facebook',
'ICBC',
'Verizon',
'China_Construction_Bank',
'Walmart',
'Huawei',
'Mercedes-Benz',
'Ping_An',
'China_Mobile',
'Agricultural_Bank_of_China',
'Toyota',
'State_Grid',
'Bank_of_China',
'WeChat',
'Tencent_QQ',
'Home_Depot',
'Taobao',
'T_Deutsche_Telekom',
'Disney',
'Shell',
'Volkswagen',
'NTT_Group',
```

```

#####'BMW',
#####'Wells_Fargo',
#####'Starbucks',
#####'YouTube',
#####'PetroChina',
#####'Bank_of_America',
#####'Tmall',
#####'Citi',
#####'Chase',
#####'Coca-Cola',
#####'Marlboro',
#####'IBM',
#####'Nike',
#####'Boeing',
#####'McDonalds',
#####'UnitedHealthcare',
#####'Moutai',
#####'Deloitte',
#####'Porsche',
#####'UPS',
#####'Sinopec',
#####'Intel')"
```

```

brands_hashtags <- c('#amazon|@amazon',
                      '#apple|@apple',
                      '#google|@google',
                      '#microsoft|@microsoft',
                      '#samsung|@samsung',
                      '#att|@att',
                      '#facebook|@facebook',
                      '#icbc|@icbc',
                      '#verizon|@verizon',
                      '#ChinaConstructionBank|China|Construction|Bank',
                      '#walmart|@walmart',
                      '#huawei|@huawei',
                      '#mercedesbenz|@mercedesbenz',
                      '#pingan|@pingan_group',
                      '#chinamobile',
                      '#agriculturalbankofchina|@agriculturalbankofchina',
                      '#toyota|@toyota',
                      '#stategrid|@stategrid',
                      '#bankofchina|@bankofchina',
                      '#wechat|@wechat',
                      '#tencent|@TencentGlobal',
                      '#homedepot|@homedepot',
                      '#taobao|@taobaotaobao',
                      '#telekom|@telekom_group',
                      '#disney|@disney',
                      '#shell|@shell',
                      '#volkswagen|@volkswagen',
                      '#nttgroup|@GlobalNTT',
```

```

      '#bmw|@bmw',
      '#wellsfargo|@WellsFargo',
      '#starbucks|@starbucks',
      '#youtube|@youtube',
      '#petrochina|@chinapetro',
      '#bankofamerica|@bankofamerica',
      '#tmall|@tmall',
      '#citi|@citi',
      '#chase|@chase',
      '#cocacola|@cocacola',
      '#marlboro|@marlboro',
      '#ibm|@ibm',
      '#nike|@nike',
      '#boeing|@BoeingSpace',
      '#mcdonalds|@mcdonalds',
      '#unitedhealthcare|@UHC',
      '#moutai|@MoutaiGlobal',
      '#deloitte|@deloitte',
      '#porsche|@porsche',
      '#ups|@ups',
      '#sinopec|@sinopec|@SinopecNews',
      '#intel|@intel')
names(brands_hashtags) <- brands_names
number_brands = length(brands_hashtags)

sentiment_analysis <- function(brand_name, hashtag) {

  data <- filter(tweets, grepl(hashtag, tweets$tweet_text))

  if (nrow(data) == 0 ){
    return(NULL)
  }

  #Get only tweets in english
  data <- data %>% mutate(textcat = textcat(x = tweet_text),
                        cld2 = cld2::detect_language(text = tweet_text, pl
                        cld3 = cld3::detect_language(text = tweet_text)) %
    select(tweet_text, textcat, cld2, cld3, created_at) %>%
    filter(cld2 == "en" & cld3 == "en")
  if (nrow(data) == 0 ){
    return(NULL)
  }
  data$textcat <- textcat(x = data$tweet_text)
  data$cld2 <- cld2::detect_language(text = data$tweet_text, plain_text = FA
  data$cld3 <- cld3::detect_language(text = data$tweet_text)
  data <- data[data$cld2 == "en" & data$cld3 == "en", ]

  #Format date
  f <- "%Y-%m-%d_%H:%M:%S"
  data$created_at <- as.Date(data$created_at, format=f)

```

```

data$created_at[1:round(nrow(data)/2)] <-
  data$created_at[1:round(nrow(data)/2)] - 1

noTweets <- table(data$created_at)
text_df <- data_frame(date = data$created_at, text = data$tweet_text)
text_df <- text_df %>% unnest_tokens(word, text)
afin <- get_sentiments("afinn")
bing <- get_sentiments("bing")
nrc <- get_sentiments("nrc")

afinSent <- text_df %>%
  inner_join(afin) %>%
  group_by(index = date) %>%
  summarise(sentiment = sum(value))

bingSent <- text_df %>%
  inner_join(bing) %>%
  count(index = date, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = (if (!exists("positive")) 0 else positive) - (if (!ex

nrcSent <- text_df %>%
  inner_join(nrc) %>%
  count(index = date, sentiment) %>%
  spread(sentiment, n, fill = 0) %>%
  mutate(sentiment = (if (!exists("positive")) 0 else positive) - (if (!ex

for(i in 1:length(noTweets)) {
  #print(i)
  #i=1
  if (!is.na(afinSent$sentiment[i])) {
    afinSent$sentiment[i] <- (afinSent$sentiment[i] / noTweets[[i]]) * 100
  }
  if (!is.na(bingSent$sentiment[i])) {
    bingSent$sentiment[i] <- (bingSent$sentiment[i] / noTweets[[i]]) * 100
  }
  nrcSent[i, -1] <- (nrcSent[i, -1] / noTweets[[i]]) * 100
}

tweets_count <- data.frame(noTweets)
names(tweets_count) <- c("date", "count")
tweets_count$date <- as.character(tweets_count$date)

sentimentTweets <- data.frame(afinSent)
sentimentTweets <- merge(sentimentTweets, bingSent[, c("index", "sentiment")])
sentimentTweets <- merge(sentimentTweets, nrcSent[, c("index", "sentiment")])
names(sentimentTweets) <- c("date", "afin", "bing", "nrc")
sentimentTweets$date <- as.character(sentimentTweets$date)

```

```

    sentimentTweets <- merge(sentimentTweets, tweets_count[, c("date", "count")

# tweetData <- melt(sentimentTweets, id="date")
return(cbind(sentimentTweets, "brand_id"=brand_name));
}

tweets_sentiments <- NULL
w<- NULL
for (w in 1:number_brands){
  #print(w)
  #print(brands_hashtags[w])
  print(names(brands_hashtags[w]))
  tweetData <- sentiment_analysis(names(brands_hashtags[w]), brands_hashtags
  if (!is.null(tweets_sentiments)) {
    if (!is.null(tweetData)) {
      tweets_sentiments <- rbind(tweets_sentiments, tweetData)
    }
  } else {
    tweets_sentiments <- tweetData
  }
}

tweets_sentiments$date <- format(as.Date(tweets_sentiments$date, format="%Y-
write.csv(tweets_sentiments, file = "tweets_analysis.csv", row.names = FALSE

```

brands_ETL.R

```

library(rJava)
library(tabulizer)
library(dplyr)
library(reshape)
library(textcat)
library(cld2)
library(cld3)
library(tidyverse)
library(tidytext)
library(textdata)

report2019url <- 'http://brandfinance.com/images/upload/brand_finance_global
extractedTables2019 <- extract_tables(report2019url)
table <- do.call(rbind, extractedTables2019[7])
leadingBrands <- rbind(table[4:53,1:9], table[4:53,10:18])

colnames(leadingBrands) <- c(RANK_2018 = "RANK_2018", RANK_2017 = "RANK_2017",
                             BRAND_NAME = "BRAND_NAME", COUNTRY = "COUNTRY",
                             BRAND_VALUE_2018 = "BRAND_VALUE_2018", PERCENT_CHANGE =
                             BRAND_VALUE_2017 = "BRAND_VALUE_2017", BRAND_RATING_201
                             BRAND_RATING_2017 = "BRAND_RATING_2017")

```



```

leadingBrands[,1] <- as.numeric(leadingBrands[,1])
leadingBrands[,2] <- as.numeric(leadingBrands[,2])
leadingBrands[,5] <- as.numeric(gsub(',', '', leadingBrands[,5]))
leadingBrands[,6] <- as.numeric(gsub('%', '', leadingBrands[,6]))
leadingBrands[,7] <- as.numeric(gsub(',', '', leadingBrands[,7]))
write.csv(leadingBrands, file = "C:\\Users\\MOLAP\\Desktop\\datasets\\brand_

report2018url <- 'http://brandfinance.com/images/upload/brand_finance_global
extractedTables2018 <- extract_tables(report2018url)
table <- do.call(rbind, extractedTables2018[7])
leadingBrands <- rbind(table[4:53,1:9], table[4:53,10:18])

colnames(leadingBrands) <- c(RANK_2018 = "RANK_2018", RANK_2017 = "RANK_2017",
                             BRAND_NAME = "BRAND_NAME", COUNTRY = "COUNTRY",
                             BRAND_VALUE_2018 = "BRAND_VALUE_2018", PERCENT_C
                             BRAND_VALUE_2017 = "BRAND_VALUE_2017", BRAND_RAT
                             BRAND_RATING_2017 = "BRAND_RATING_2017")

leadingBrands[,1] <- as.numeric(leadingBrands[,1])
leadingBrands[,2] <- as.numeric(leadingBrands[,2])
leadingBrands[,5] <- as.numeric(gsub(',', '', leadingBrands[,5]))
leadingBrands[,6] <- as.numeric(gsub('%', '', leadingBrands[,6]))
leadingBrands[,7] <- as.numeric(gsub(',', '', leadingBrands[,7]))
write.csv(leadingBrands, file = "C:\\Users\\MOLAP\\Desktop\\datasets\\brand_

report2017url <- 'http://brandfinance.com/images/upload/brand_finance_global
extractedTables2017 <- extract_tables(report2017url)
table <- do.call(rbind, extractedTables2017[7])
leadingBrands <- rbind(table[4:53,1:9], table[4:53,10:18])

colnames(leadingBrands) <- c(RANK_2018 = "RANK_2018", RANK_2017 = "RANK_2017",
                             BRAND_NAME = "BRAND_NAME", COUNTRY = "COUNTRY",
                             BRAND_VALUE_2018 = "BRAND_VALUE_2018", PERCENT_C
                             BRAND_VALUE_2017 = "BRAND_VALUE_2017", BRAND_RAT
                             BRAND_RATING_2017 = "BRAND_RATING_2017")

leadingBrands[,1] <- as.numeric(leadingBrands[,1])
leadingBrands[,2] <- as.numeric(leadingBrands[,2])
leadingBrands[,5] <- as.numeric(gsub(',', '', leadingBrands[,5]))
leadingBrands[,6] <- as.numeric(gsub('%', '', leadingBrands[,6]))
leadingBrands[,7] <- as.numeric(gsub(',', '', leadingBrands[,7]))
write.csv(leadingBrands, file = "C:\\Users\\MOLAP\\Desktop\\datasets\\brand_

report2016url <- 'http://brandfinance.com/images/upload/brand_finance_global
extractedTables2016 <- extract_tables(report2016url)
table <- do.call(rbind, extractedTables2016[7])
leadingBrands <- rbind(table[4:53,1:9], table[4:53,10:18])

```

```

colnames(leadingBrands) <- c(RANK_2018 = "RANK_2018", RANK_2016 = "RANK_2016",
                             BRAND_NAME = "BRAND_NAME", COUNTRY = "COUNTRY",
                             BRAND_VALUE_2018 = "BRAND_VALUE_2018", PERCENT_C
                             BRAND_VALUE_2016 = "BRAND_VALUE_2016", BRAND_RAT
                             BRAND_RATING_2016 = "BRAND_RATING_2016")

leadingBrands[,1] <- as.numeric(leadingBrands[,1])
leadingBrands[,2] <- as.numeric(leadingBrands[,2])
leadingBrands[,5] <- as.numeric(gsub(',', '', leadingBrands[,5]))
leadingBrands[,6] <- as.numeric(gsub('%', '', leadingBrands[,6]))
leadingBrands[,7] <- as.numeric(gsub(',', '', leadingBrands[,7]))
write.csv(leadingBrands, file = "C:\\Users\\MOLAP\\Desktop\\datasets\\brand_

report2015url <- 'http://brandfinance.com/images/upload/brand_finance_global
extractedTables2015 <- extract_tables(report2015url)
table <- do.call(rbind, extractedTables2015[7])
leadingBrands <- rbind(table[4:53,1:9], table[4:53,10:18])

colnames(leadingBrands) <- c(RANK_2018 = "RANK_2018", RANK_2015 = "RANK_2015",
                             BRAND_NAME = "BRAND_NAME", COUNTRY = "COUNTRY",
                             BRAND_VALUE_2018 = "BRAND_VALUE_2018", PERCENT_C
                             BRAND_VALUE_2015 = "BRAND_VALUE_2015", BRAND_RAT
                             BRAND_RATING_2015 = "BRAND_RATING_2015")

leadingBrands[,1] <- as.numeric(leadingBrands[,1])
leadingBrands[,2] <- as.numeric(leadingBrands[,2])
leadingBrands[,5] <- as.numeric(gsub(',', '', leadingBrands[,5]))
leadingBrands[,6] <- as.numeric(gsub('%', '', leadingBrands[,6]))
leadingBrands[,7] <- as.numeric(gsub(',', '', leadingBrands[,7]))
write.csv(leadingBrands, file = "C:\\Users\\MOLAP\\Desktop\\datasets\\brand_

```

location_ETL.R

```

library(rJava)
library(tabulizer)
library(dplyr)
library(reshape)
library(textcat)
library(cld2)
library(cld3)
library(tidyverse)
library(tidytext)
library(textdata)

locations <- read.csv("C:\\Users\\MOLAP\\Desktop\\datasets\\countrycontinent

locations <- locations[,c(1,3,7)]

```

```
locations[,1] <- as.character(gsub("'", "", locations[,1]))
locations[,1] <- as.character(gsub("%'", "", locations[,1]))
locations[,1] <- as.character(gsub("'", "", locations[,1]))
locations[,2] <- as.character(gsub("'", "", locations[,2]))
locations[,2] <- as.character(gsub("%'", "", locations[,2]))
locations[,2] <- as.character(gsub("'", "", locations[,2]))
locations[,3] <- as.character(gsub("'", "", locations[,3]))
locations[,3] <- as.character(gsub("%'", "", locations[,3]))
locations[,3] <- as.character(gsub("'", "", locations[,3]))

write.csv(locations, file = "locations.csv", row.names = F)
```