Data Warehousing and Business Intelligence Project

By:
Mayank Gulaty
X15031705
National College of Ireland

Table of Contents

List of figures

Approach

I wanted to do a project based on music so I started searching for papers and I found 2 very good papers that inspired this project. One paper sees the evolution of music over 50 years from 1960 -2010 (Mauch, MacCallum, et al. 2015) and the other paper does the lyric based song sentiment classification (Xia, Wang, et al. 2008). In the second paper, problems in lyric sentiment classifications are discussed like many words like nouns and words in the lyrics are ambiguous. I wanted to do audio based sentiment analysis but due to lack of time and sources, I stuck with lyric based sentiment analysis. After deciding the topic, I started looking for APIs and got few APIs. Some were working, some were not and some had incomplete data. However, I managed to find some good sources at the end. After that I started studying about warehousing architectures and approaches and how to use APIs with R and started my project.

Introduction

This project implemented a working model of a data warehouse and showed its business intelligence capabilities. The paper will show the whole process of a data warehouse along with 3 case studies to show the business intelligence capabilities of data warehouse. The most popular

definition of the data warehouse is that it is a “subject oriented, integrated, non-volatile, time variant collection of data for management’s decision making” by Inmon told in his book Building the Data Warehouse in 1998. I will discuss my data sources and the architecture used and implementation of the project along with 3 case studies.
Case study 1 will evaluate how the emotions of joy and sadness in the lyrics of the songs changed over 10 years. Case study 2 will examine which genre’s lyrics were most positive and negative. Finally, case study 3 will see evolution of emotions of joy and sadness in songs’ lyrics in the various cities of United states.

Architecture and Implementation

There are many types of data warehouse architectures available but the main ones are Kimball and Inmon’s approach. Inmon approach is relational which is based on entity-relationship model, normalization and tables using joins.
Unlike Inmon’s top down approach, Kimball’s approach is bottom-up. This approach uses dimensional data modelling which means that it begins with tables rather entity relationship diagram. The tables are either dimension tables or fact tables.(Breslin 2004).
As, the time was limited for this project, I am going to use Kimball architecture because of its ease of use and implementation and not much planning is needed in the start as opposed to the Inmon approach where we have to design the models up front. Also, normalized data model is not required in Kimball. Kimball architecture is made up of independent dimensional data marts connected by a data bus architecture in a star schema and data consistency is achieved as the dimensions are conformed.

Technologies Used

Programming languages

- R for extracting data from API’s and writing it to excel files
- SQL for creating dimensions and fact tables

Database Management

- SQL server & SSIS for loading fact tables and dimensions
- Alchemy API for sentiment analysis

Additional Add-ons / Softwares

- Rvest package in R for web scraping
- SPSS for statistical analysis

- Tableau software for making graphs

Data Sources

Five data sources are being utilised in this data warehouse -

1. Project Whitburn (structured csv file) (See Appendix A)

2. List of songs scraped from Wikipedia page (unstructured HTML)

3. Song metadata from Musixmatch API (semi-structured XML)

4. Song Lyrics from Chartlyrics API (semi-structured XML)

5. Artist Country and City from echonest API (semi-structured XML)

Data Bus Architecture

Common Dimensions

Business Processes	Genre	Date	Sentiment	Geography
Emotional sentiments comparison	•	•	•	•
Sentiment Scores Comparison	•	•	•	•

Implementation and design process

This section explains the design process taken in this project. I chose Kimball's design process which is a 4 step process. (Kimball Group 2013)

1- Select the business process
2- Declare the grain
3- Identify the dimensions
4- Identify the facts

In my case the business process would analysis of the music lyrics over the years from 1990-2010.

Grain in project is the different sentiment scores. According to Kimball Group 2013, dimensions are sometimes soul of the data warehouse.Dimensions in my case are Genre, Sentiment, Year and geography.The fact table in my project is named FactSongs which contain the different sentiment scores and the foreign keys to all the dimensions which will all be linked to the dimensions via a data bus architecture in a star schema as shown below.

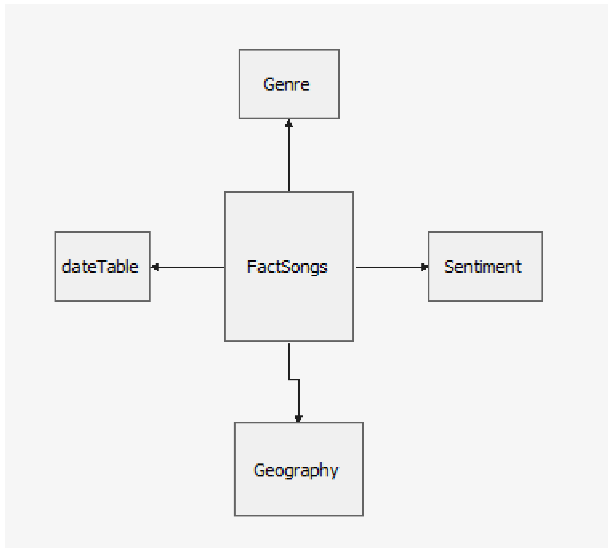


Figure 1 Star Schema of the Project

ETL (Extract Transform Load)

ETL process is the backbone of a data warehouse architecture. The role of ETL is to extract data from different sources, make sure that the quality of data is good and consistent so that these separate sources can be used together to answer a non trivial business intelligence query and to see a pattern to make strategic business decisions. The final form of data is very easy to use and user friendly so that even non coders can make the decisions based on the data.

Back room: Preparing The data

1- Extracting
2- Cleaning
3- Conforming
4- Delivering

Extraction - In this project, structured, semi-structured and unstructured sources of data were used. Firstly, I web scraped the Billboard top 100 list from Wikipedia of the year 1990-2000. For this I used rvest package available in CRAN repository. (Wickhan, 2015). I extracted all of them in separate files and then merged to make it one. At this time, I had songs with their artists. I then extracted songs and artists using the loop and hit it on musixmatch api to get the genres. I got the song years from Project whitburn spreadsheet. It was hard getting this spreadsheet but somehow I managed to find one useful for my project. After that, I extracted the lyrics of songs from musixmatch API. Many lyrics were not found in musixmatch API, so I hit it on Chartlyrics API. And finally I got the artist city and country from the echonest API. (See Appendix A for code screenshots)

Cleaning – As the data was collected from so many sources, all the data was not clean like some lyrics were not found, some genres were not found and to move to the next stage, it should be cleaned. For this purpose, I used R programming language and Microsoft excel to clean the data. Cleaning involved editing of the data manually where only 1 field was missing and deletion of the rows where manual editing was not possible as many columns were missing.
Sentiment scores when extracted from the alchemy API to the data frame were in string format. So I had to transform it all to numeric value before loading it to the excel file.

Transformation – Some data transformation had to be done before loading like the sentiment scores extracted were in string format. So I converted that values to numeric values with the help of R.

Conforming - This step is required whenever two or more data sources are joined or merged. In my case Musixmatch ID binded my different data sources.

Delivering - In this part the dimension tables and the fact tables are loaded via SSIS (SQL Server Integration Services) and the cube is also updated. Below are the screenshots of my SSIS workflow, dimension tables and fact tables.

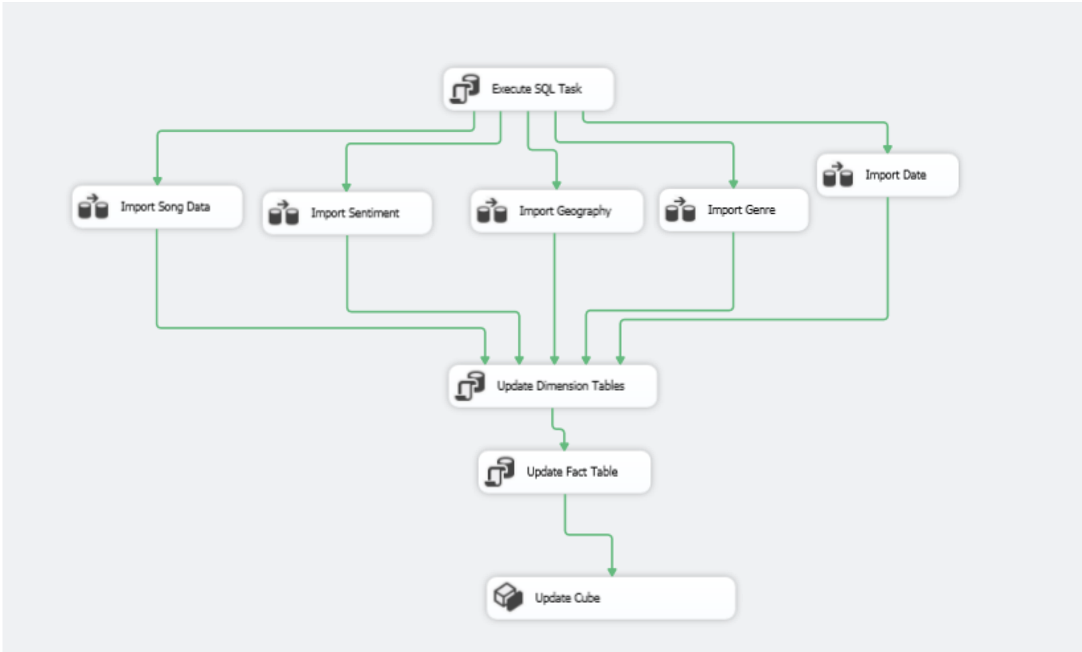


Figure 2 SSIS Workflow

	Sentiment_ID	Sentiment_Group_ID	Sentiment_Type
1	1	1	negative
2	2	1	negative
3	3	1	negative
4	4	2	positive
5	5	1	negative
6	6	1	negative
7	7	1	negative
8	8	1	negative
9	9	1	negative
10	10	2	positive
11	11	2	positive
12	12	2	positive

Figure 3 Sentiment Dimension

	GID	Artist_Country	Artist_City	Artist_Country_Group_ID	Artist_City_Group_ID
1	1	Canada	Kingston	4	83
2	2	United States	Los Angeles	23	89
3	3	Sweden	Gothenburg	21	60
4	4	United States	Nashville	23	101
5	5	United States	Perth Amboy	23	121
6	6	Sweden	Gothenburg	21	60
7	7	United States	Severn	23	149
8	8	United States	Rockport	23	134
9	9	United States	Miami	23	96
10	10	Puerto Rico	San Juan	20	142
11	11	United States	Harlem	23	67
12	12	United Kingdom	London	22	87

Figure 4 Geography Dimension

	Genre_ID	Genre_Group_ID	Genre
1	1	14	Rock
2	2	9	Pop
3	3	9	Pop
4	4	2	Christian & Gospel
5	5	14	Rock
6	6	9	Pop
7	7	12	R&B/Soul
8	8	14	Rock
9	9	10	Pop in Spanish
10	10	9	Pop
11	11	6	Hip Hop/Rap
12	12	12	R&B/Soul

Figure 5 Genre Dimension

	Date_ID	Year_Group_ID	Song_Year
1	1	2	1991
2	2	1	1990
3	3	5	1994
4	4	2	1991
5	5	1	1990
6	6	5	1994
7	7	11	2000
8	8	9	1998
9	9	3	1992
10	10	10	1999
11	11	6	1995
12	12	7	1996

Figure 6 Year Dimension

FId	Song_ID	Genre_ID	Geography_ID	Date_ID	Sentiment_ID	Sentiment_Score	Joy	Sadness
1	1	1	1	1	1	-0.3261540000	0.0035240000	0.0421980000
2	2	2	2	2	2	-0.1561480000	0.0028350000	0.0079360000
3	3	3	3	3	3	-0.0653763000	0.0069570000	0.0134790000
4	4	4	4	4	4	0.3583300000	0.7984700000	0.0096400000
5	5	5	5	5	5	-0.2647890000	0.0004770000	0.0056230000
6	6	6	6	6	6	-0.1438050000	0.0017320000	0.1264540000
7	7	7	7	7	7	-0.3767250000	0.0001710000	0.0161950000
8	8	8	8	8	8	-0.1700880000	0.0071790000	0.0200280000
9	9	9	9	9	9	-0.4089870000	0.0035010000	0.2198270000
10	10	10	10	10	10	0.0482900000	0.0136380000	0.0177000000
11	11	11	11	11	11	0.1377250000	0.0185140000	0.0054390000

Figure 7 Fact Table

Front room: Presenting The data

This is the part what end users and managers see. Querying is done in this area and the dimension tables can be accessed here via various means like OLAP cubes and interactive dashboards. I used pivot table and Tableau software for making the graphs for my case studies.

Limitations

Due to time limitations, I couldn't analyze a big data set and so I decided to go with top 100 songs of the year but after getting the genres of the songs and aggregating it, I saw that majority of the songs are from POP genre. So there was a skewness of the data towards the pop genre and United states country. Also, the sentiment analysis was done from one source and there can be a bit of difference when doing sentiment analysis from different sources. Though, in my project I am observing the evolution over years, I still tried to get the specific release date of the songs for a deeper analytics but could only get it for around 15 percent of my data as most of the dates were empty in the API and that's why I had to drop the idea of adding date dimension and only worked with years.

Future Work

I could get a very big dataset of the music and scale my project up to a bigger level to see the evolution in the past 50 years or more and I could get the release date of the songs by doing web scraping. It will take time but its possible. After the specific dates, I can get a deeper analytics over the specific months. I could also extract the BPM of the music for more relations.

Case Studies

Case Study 1

This case study observes how the joy and sadness in song lyrics changed over the years from 1990- 2000. As we can see from the graph that the sad songs were always more popular than the happy songs and the trend didn't change over the years. The songs with the lyrics of joy decreased as the years passed by.

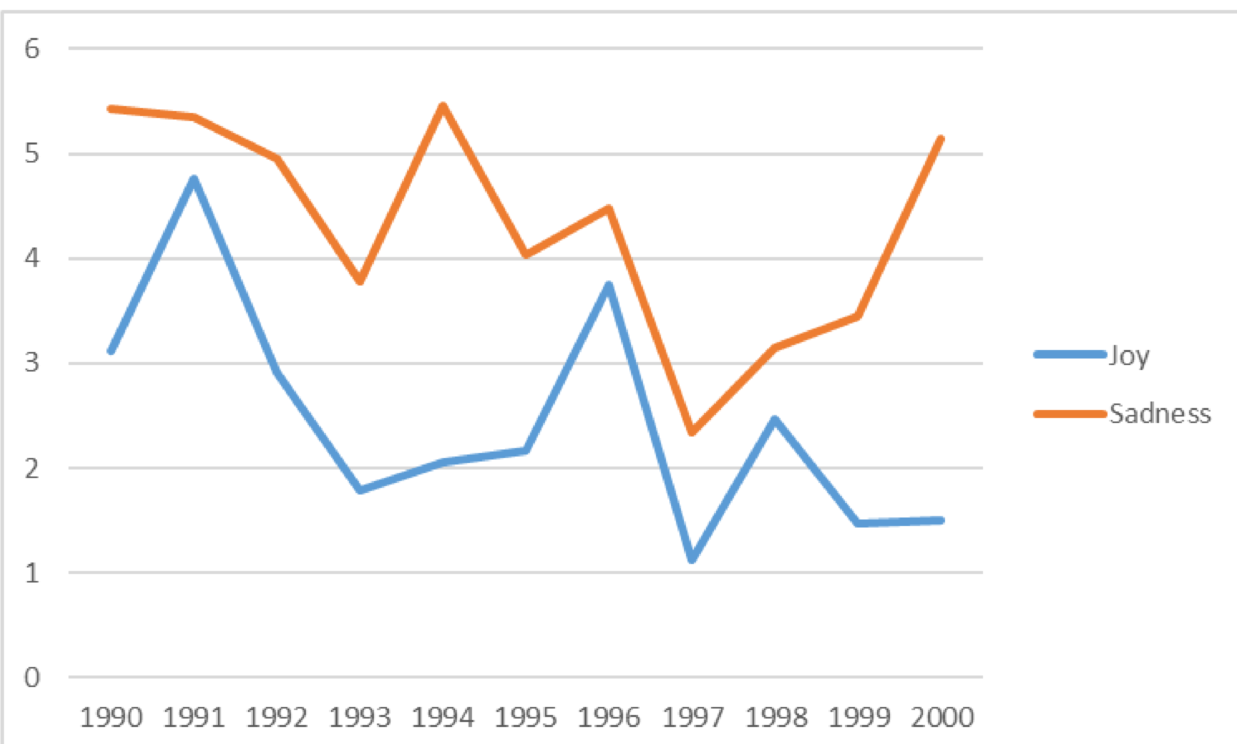
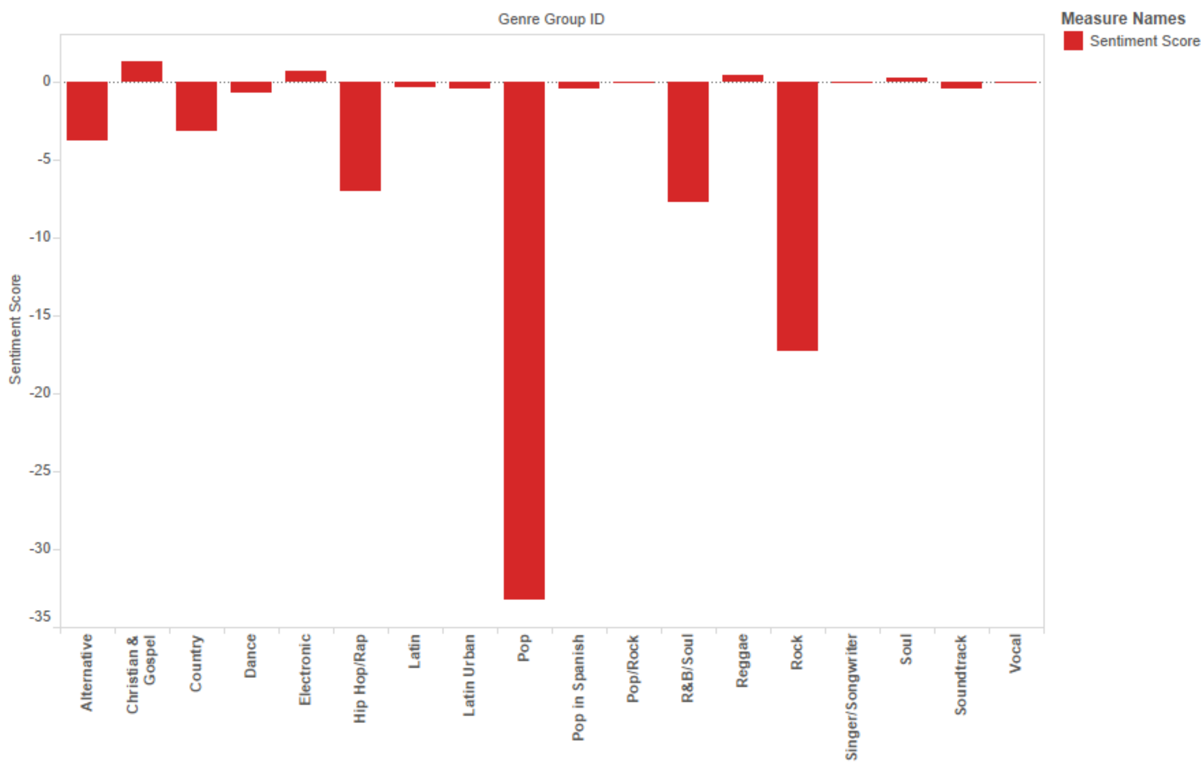


Figure 8 Case study 1: Emotion vs time

Case Study 2

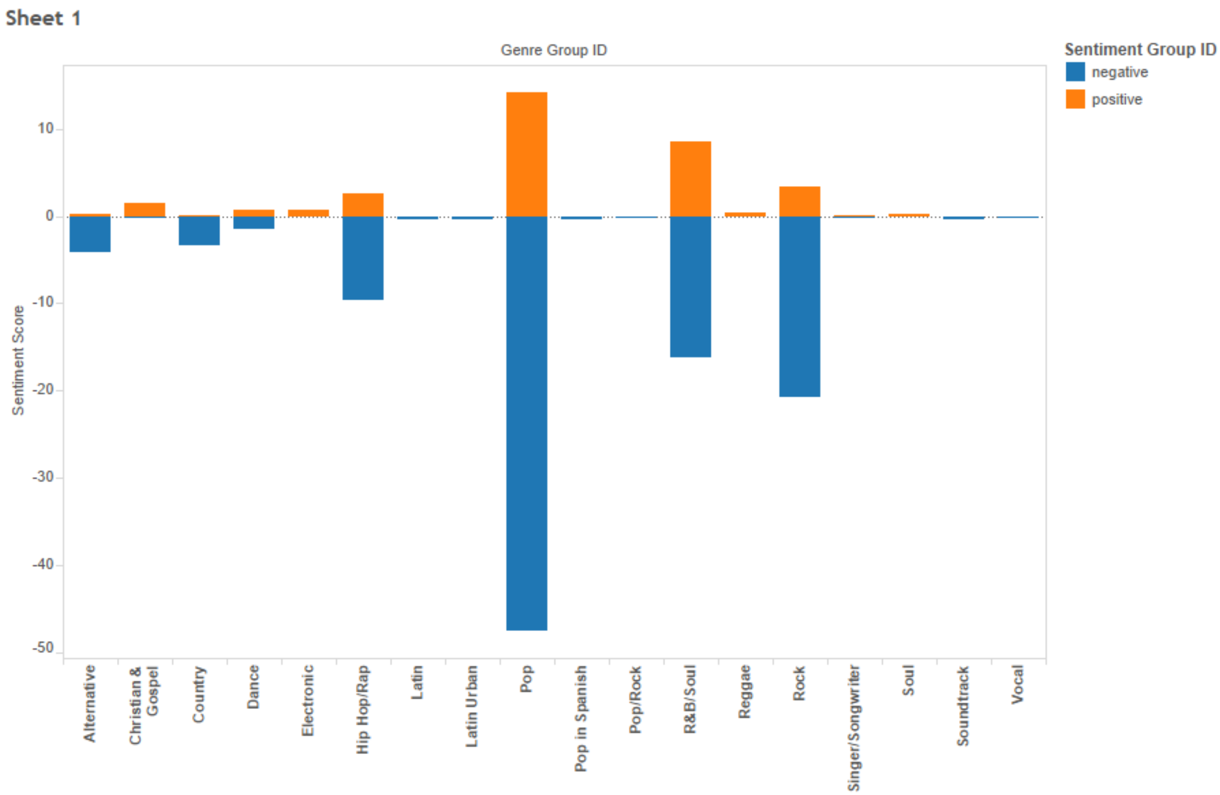
This case study will see the sentiment of the song lyrics whether positive or negative with respect to the specific genres. And as I mentioned, the data is skewed as the number of POP songs are more but still we can see a picture here that popular genres like POP, Hip/Hop, Rock were inclined towards negative. We can see that there are some genres like soul, reggae and Christian & gospel which were positive and its true as Christian & gospel is traditional music and mostly contains the positive lyrics.

Sheet 1



Sentiment Score for each Genre. Color shows details about Sentiment Score.

Figure 9 Case study 2.1



Sentiment Score for each Genre. Color shows details about Sentiment Group ID.

Figure 10 Case study 2.2

Case Study 3

Unlike case study 1, this case study will see the evolution in the various cities in a particular country i.e. United states. I chose United states because most of the artists were from United States so I could get a better picture. We can see that New York has the highest number of sad songs.

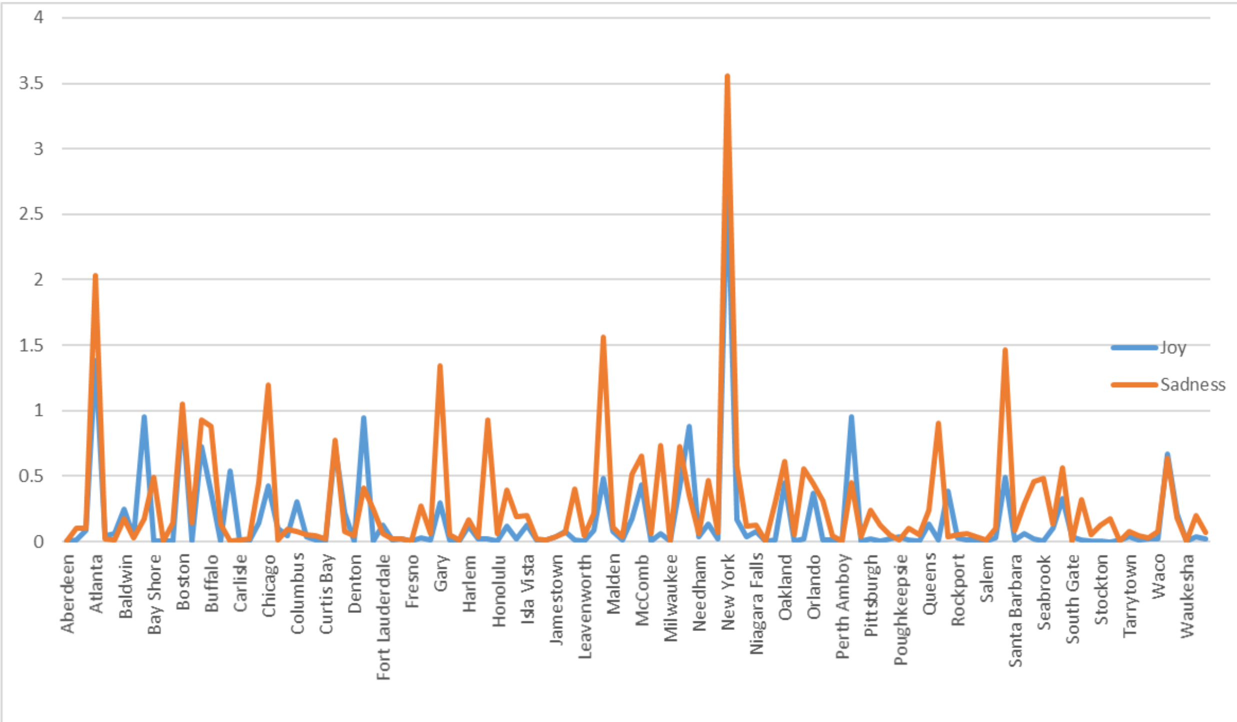


Figure 11 Case study 3.1

A much clearer view can be seen from the following bubble chart showing the same information as above. The darker the shade, the sad the song is.

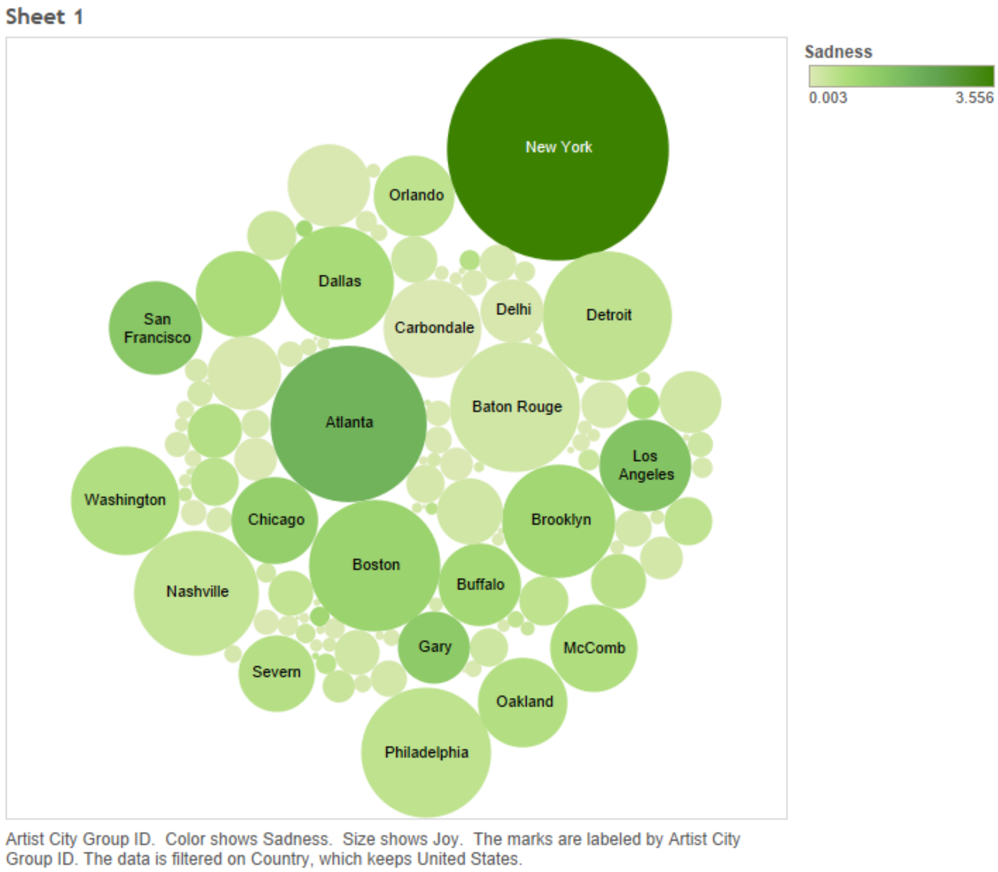


Figure 12 Case study 3.2 Bubble Chart

Statistics

I used SPSS to play around with some statistical methods and I did a pearson correlation test with my sentiment scores, joy and sadness scores. The output is shown below.

Correlations

		Sentiment_Score	Joy	Sadness
Sentiment_Score	Pearson Correlation	1	.439**	-.248**
	Sig. (2-tailed)		.000	.000
	N	629	629	629
Joy	Pearson Correlation	.439**	1	-.088*
	Sig. (2-tailed)	.000		.028
	N	629	629	629
Sadness	Pearson Correlation	-.248**	-.088*	1
	Sig. (2-tailed)	.000	.028	
	N	629	629	629

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 13 Correlation

We can see from the output that the sentiment score and joy are positively correlated and it's a fairly strong correlation of 0.439 which means higher the sentiment score higher the value of joy is. There is a negative correlation between Sentiment score and sadness though not very strong but not weak also. Then I ran a partial correlation test by controlling the sentiment score variable to see whether or not its really affecting the variables or not and from the following output we can see that the values changed significantly and thus we can conclude that sentiment score is affecting the other two variables. Again, the data in this project was less and I couldn't do much tests but with a bigger dataset and more measures we can run a good correlation and partial correlation to see which variable is affecting which.

Partial Corr

Correlations

Control Variables			Joy	Sadness
Sentiment_Score	Joy	Correlation	1.000	.024
		Significance (2-tailed)	.	.540
		df	0	626
	Sadness	Correlation	.024	1.000
		Significance (2-tailed)	.540	.
		df	626	0

Figure 14 Partial Correlation

Conclusion

This project saw how a data warehouse is implemented with all the steps from the data extraction to data staging and how the facts and dimensions are made. It was then used to answer 3 business intelligence queries.

List of References

Breslin, M., (2004). Data Warehousing Battle of the Giants : Comparing the Basics of the Kimball and Inmon Models. *Business Intelligence Journal*, pp.6–20.

Group Kimball, (2013). Kimball Dimensional Modeling Techniques. , pp.1–24.

Mauch, M., MacCallum, R.M., Levy, M. & Leroi, A.M., (2015). The evolution of popular music: USA 1960–2010. *Royal Society Open Science*, 2(5), p.150081. Available at: http://rsos.royalsocietypublishing.org/content/2/5/150081.abstract.

Xia, Y., Wang, L., Wong, K.F. & Xu, M., (2008). Sentiment vector space model for lyric-based song sentiment classification. *ACL-08: HLT - 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, (June), pp.133–136. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84859921031&partnerID=tZOtx3y1.

Wickhan, H. (2015) *Package 'rvest'* [Online]
Available at: https://cran.r-project.org/web/packages/rvest/rvest.pdf [Accessed 3 March 2016]

Appendix A: Screenshots of Code

```r
for(i in seq(1990,2000,1)){
urls <-paste0("https://en.wikipedia.org/wiki/Billboard_Year-End_Hot_100_singles_of_",i)
print(urls)
url <- urls
artist <- url %>%
  read_html() %>%
  html_nodes(xpath='//*[@id="mw-content-text"]/table[1]') %>%
  html_table()
artist <- artist[[1]]
Year<- i
bind<- cbind(artist,Year)
outputfile <- paste("Billboard","-",i,".xlsx")
write.xlsx(bind,outputfile,append = TRUE)
}
```

Figure 15 Wikipedia Scraping

```r
for (i in 1:1100)
{
  j=j+1;
  if(j<20)
  {
    tryCatch({
      b<- x[[i,3]]
      city = paste("http://developer.echonest.com/api/v4/artist/profile?",
                   "name=", b ,"&api_key=",
                   "0YUVFLK5MOXJ0XZTT",
                   "&bucket=","artist_location",
                   "&format=xml",sep="")
      xml= xmlParse(city)
      xmltop=xmlRoot(xml)
      a_city=xmlValue(xmltop[[2]][[2]][[1]])
      a_location=xmlValue(xmltop[[2]][[2]][[3]])
      a_country=xmlValue(xmltop[[2]][[2]][[4]])
      cityDF <- rbind(cityDF, a_city)
      locationDF <- rbind(locationDF, a_location)
      countryDF <- rbind(countryDF, a_country)
    },error = function(e){
```

Figure 16 Artist Location

```r
#sentiment text
for (i in 1:629) {
  ly<-x[i,5]
  url = paste("http://gateway-a.watsonplatform.net/calls/text/TextGetTextSentiment?","ap
  xml= xmlParse(url)
  xmltop=xmlRoot(xml)
  score=xmlValue(xmltop[[5]][[2]])
  type=xmlValue(xmltop[[5]][[3]])
  scoreDF <- rbind(scoreDF, score)
  typeDF <- rbind(typeDF, type)
}
```

Figure 17 Sentiment Analysis

```r
for(i in 1:1100)
{
  lVar <- x[[i,5]]
all = paste("http://api.musixmatch.com/ws/1.1/track.lyrics.get?",
            "track_id=", lVar,"&apikey=",
            "8808c0ad5181267c83b0941c79197c5f",
            "&format=xml",sep="")

xml <- xmlParse(all)
xml

statusCode<-sapply(getNodeSet(xml, "//status_code"), xmlValue)

statusCode<-as.numeric(statusCode)

if(statusCode == 404){
  lyrics<-"NA"
}else
  lyrics<-(sapply(getNodeSet(xml, "//lyrics_body"), xmlValue))

collect1 <- rbind(collect1, lyrics)
```

Figure 18 Grabbing Lyrics

```r
#emotional sentiment analysis
for (i in 1:629) {
  ly<-x[i,4]
  url = paste("http://gateway-a.watsonplatform.net/calls/text/TextGetEmotion?","apikey="
  xml= xmlParse(url)
  xmltop=xmlRoot(xml)
  anger=xmlValue(xmltop[[5]][[1]])
  disgust=xmlValue(xmltop[[5]][[2]])
  fear=xmlValue(xmltop[[5]][[3]])
  joy=xmlValue(xmltop[[5]][[4]])
  sadness=xmlValue(xmltop[[5]][[5]])
  angerDF <- rbind(angerDF, anger)
  disgustDF <- rbind(disgustDF, disgust)
  fearDF <- rbind(fearDF, fear)
  joyDF <- rbind(joyDF, joy)
  sadnessDF <- rbind(sadnessDF, sadness)
}
```

Figure 19 Emotional Sentiment Analysis

1992	29	a	1992_029	21	17	9	1	3		Prince	Prince & The N.P.G.		Diamonds And Pearls	4:14	
1992	30	a	1992_030	18	9	4	1	3		Madonna	Madonna	Erotica	Erotica	5:10	
1992	31	a	1992_031	25	22	10	3	4		Billy Ray Cyrus	Cyrus, Billy Ray	Some Gave .	Achy Breaky Heart	3:23	
1992	32	a	1992_032	21	18	8	2	4		Joe Public	Joe Public		Live And Learn	3:56	
1992	33	a	1992_033	22	19	8	1	4		Celine Dion	Dion, Celine	Celine Dion	If You Asked Me To	3:52	
1992	34	a	1992_034	20	14	7	1	4		Shakespear's Sister	Shakespear's Sister	Hormonally	Stay	3:40	
1992	35	a	1992_035	33	22	10	2	5		Ce Ce Peniston	Peniston, Ce Ce	Finally	Finally	4:20	
1992	36	a	1992_036	21	17	7	2	5		Sophie B. Hawkins	Hawkins, Sophie B.		Damn I Wish I Was Your Lover	5:19	
1992	37	a	1992_037	37	30	12	1	5		Jon Secada	Secada, Jon		Just Another Day	4:11	
1992	38	a	1992_038	20	16	7	1	5		Hi-Five	Hi-Five		She's Playing Hard To Get	4:32	
1992	39	a	1992_039	20	16	7	1	5		Mariah Carey	Carey, Mariah		Make It Happen	4:05	
1992	40	a	1992_040	20	17	4	1	5		M.C. Hammer	M.C. Hammer		2 Legit 2 Quit	7:55	
1992	41	a	1992_041	22	17	9	2	6		TLC	TLC		Ain't 2 Proud To Beg	4:09	
1992	42	a	1992_042	20	19	8	2	6		Nirvana	Nirvana	Nevermind -	Smells Like Teen Spirit	4:59	
1992	43	a	1992_043	25	20	7	1	6		Tevin Campbell	Campbell, Tevin	T.E.V.I.N.	Tell Me What You Want Me To Do	4:09	
1992	44	a	1992_044	24	17	7	1	6		En Vogue	En Vogue		Giving Him Something He Can Feel	3:50	
1992	45	a	1992_045	26	20	6	1	6		Tom Cochrane	Cochrane, Tom	Mad Mad W	Life Is A Highway	4:17	
1992	46	a	1992_046	22	17	5	1	6		Arrested Development	Arrested Development		Tennessee	4:35	
1992	47	a	1992_047	26	16	5	1	6		K.W.S.	K.W.S.		Please Don't Go	4:08	
1992	48	a	1992_048	24	16	5	1	6		Mint Condition	Mint Condition		Breakin' My Heart (Pretty Brown Ey	4:44	
1992	49	a	1992_049	20	11	4	1	6		Michael Jackson	Jackson, Michael		In The Closet	4:48	
1992	50	a	1992_050	23	18	4	1	6		Technotronic	Technot Featuring Ya Kid K		Move This	3:33	
1992	51	a	1992_051	23	21	10	2	7		Bobby Brown	Brown, Bobby	Bobby	Good Enough	3:52	
1992	52	a	1992_052	27	21	10	1	7		TLC	TLC		What About Your Friends	4:00	
1992	53	a	1992_053	31	23	9	1	7		Mary J. Blige	Blige, Mary J.	What's The (	Real Love	4:25	
1992	54	a	1992_054	20	14	4	1	7		Genesis	Genesis	We Can't Da	I Can't Dance	3:53	
1992	55	a	1992_055	20	12	2	1	7		M.C. Hammer	M.C. Hammer		Addams Groove	3:54	
1992	56	a	1992_056	20	12	3	2	8		Firehouse	Firehouse		When I Look Into Your Eyes	4:01	
1992	57	a	1992_057	23	18	4	1	8		Arrested Development	Arrested Developme	3 Years, 5 M	People Everyday	3:28	
1992	58	a	1992_058	20	17	3	1	8		Amy Grant	Grant, Amy	Heart In Motio	Good For Me	3:59	CS
1992	59	a	1992_059	20	16	1	1	8		En Vogue	En Vogue		Free Your Mind	4:06	
1992	60	a	1992_060	20	11	4	4	9		Ugly Kid Joe	Ugly Kid Joe	As Ugly As I	Everything About You	4:00	
1992	61	a	1992_061	20	14	5	2	9		Cover Girls, The	Cover Girls, The	Here It Is	Wishing On A Star	4:35	
1992	62	a	1992_062	20	16	2	1	9		Richard Marx	Marx, Richard		Hazard	5:08	
1992	63	a	1992_063	20	14	2	1	9		Celine Dion	Dion, Ce & Peabo Br	Beauty and t	Beauty And The Beast	3:59	
1992	64	a	1992_064	22	18	1	1	9		Elton John	John, Elton		The One	5:42	
1992	65	a	1992_065	20	15	1	1	9		U2	U2	Achtung Bal	Mysterious Ways	4:00	
1992	66	a	1992_066	20	18	3	3	10		Luther Vandross	Vandros & Janet Jackson with BB		The Best Things In Life Are Free	4:32	
1992	67	a	1992_067	20	14	1	1	10		U2	U2	Achtung Bal	One	4:32	
1992	68	a	1992_068	20	12	1	1	10		George Michael	Michael, George		Too Funky	3:40	
1992	69	a	1992_069	20	14	0	2	11		Michael Bolton	Bolton, Michael	Timeless (Th	To Love Somebody	4:00	
1992	70	a	1992_070	18	12	0	2	11		KLF, The	KLF, Th Featuring Ta	The White R	Justified And Ancient	3:38	
1992	71	a	1992_071	28	21	0	1	11		Jodeci	Jodeci		Come & Talk To Me	4:07	
1992	72	a	1992_072	20	15	0	1	11		Bryan Adams	Adams, Bryan	Waking Up 1	Do I Have To Say The Words?	4:17	
1992	73	a	1992_073	20	13	0	3	12		Karyn White	White, Karyn		The Way I Feel About You	4:18	

Figure 20 Project Whitburn csv file screenshot