**IMDb Database Analytics Term Manuscript** Cesar Lopez, Michael Cheung, Brian Wonse, Tan Huynh

California State University, Los Angeles Undergraduates

CIS 4560 Dr. Jongwook Woo

emails: cesarlopez200@gmail.com, cheungmichael97@gmail.com,

brianwonse@gmail.com, moitinhdauptd@gmail.com

**Abstract**: What not better way to practice data analysis then practicing Big Data with the Hadoop File System. By repeatedly entering different HDFS and Hive statements, users will gain a better understanding of data processing and data analysis techniques. Users will be sorting through IMDB information that relates to films, television programs, home videos, etc.   
  
 **1. Introduction**

We will be using information gathered from IMDb to sort through the different media types of what is being watched from all around the world. IMDb is a great place for anyone to discover new things to watch, learn more about what they are watching, trivial information and even movie showing times. From this website, there is a huge amount of data that has been collected and given out to the public to access. [4]

**IMDb Weekly Usage in July 2017**

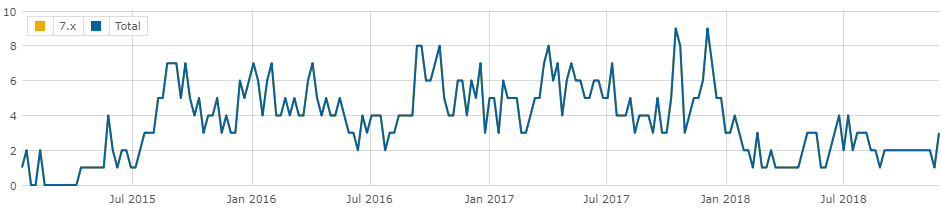


Figure 1.1. IMDb weekly usage in July 2017 in millions.

**1.1 Related Works**

This is not the first time that IMDb has been analyzed. There are other works and articles that have previously shown examples of IMDB database analysis. There’s no surprise to this because IMDb is a very popular site when it comes to a multimedia database. What makes us different in comparison is that the method to generate a visual form of the data is different and the analysis is based on our own educated inference while theirs is based off of their own. What’s more, is that some papers seek to automate predictions on viewing habits.

One of the standout databases is an IMDb exploratory data analysis, published in Russia on January 5th, 2015 by Ilya Ezepov. She downloaded and visualized data taken straight from IMDB through an R package called ggplot2.. The steps and the visuals are uniquely different in comparison to ours. What’s also different was the focus on which she built her predictive model. Her main focus was through movies while for us, we wanted to tackle movies and televised episodic media (shows). In fact, her goal is absolutely different in that she wants to find a correlation between major historic events and the impact it has on movie media. In her case, it’s more isolated as opposed to a broad and more generalized vision that we have for our goal. [1]

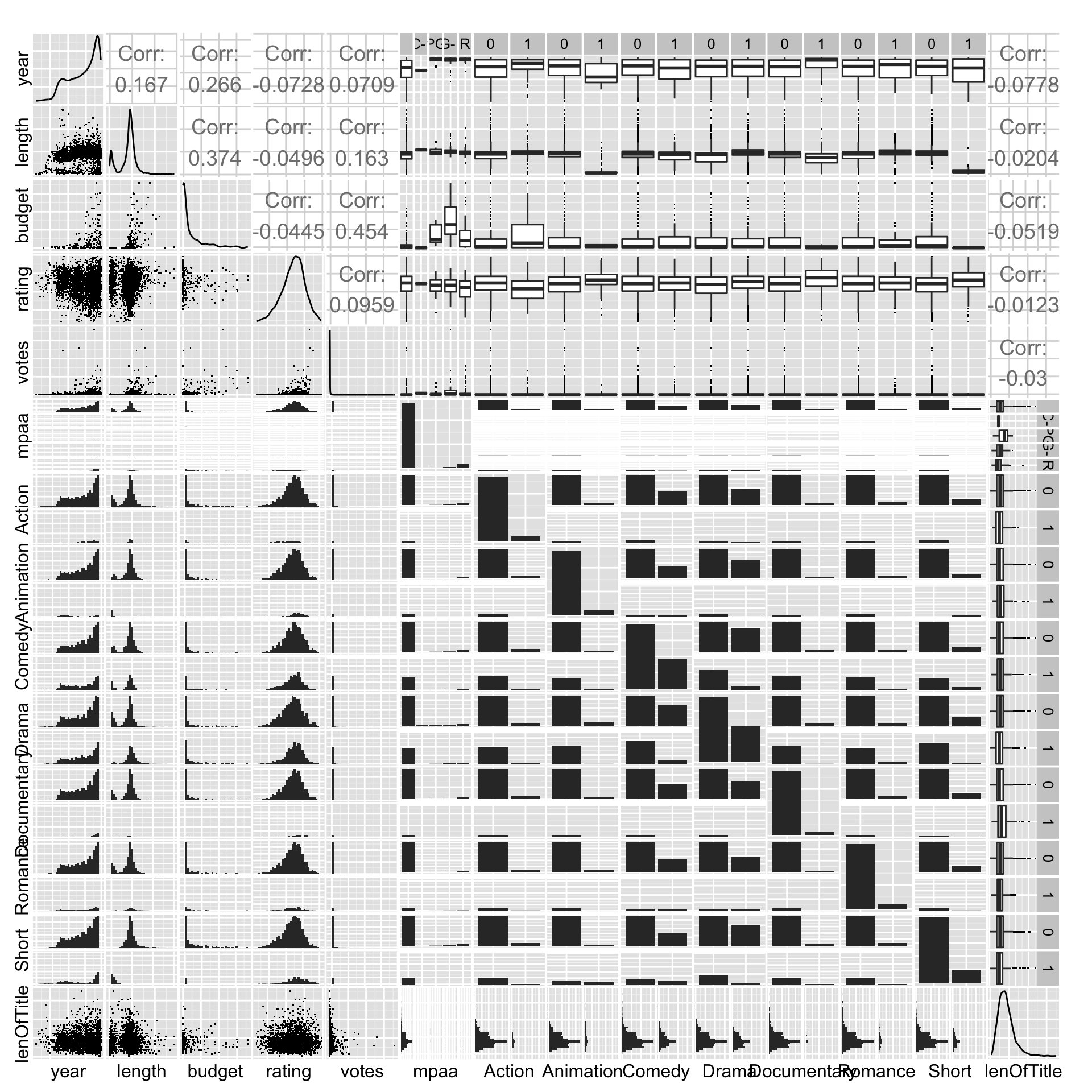


Figure 1.2. Ezepov’s full visualization from her data analysis paper. All charts included.

A more recent work on IMDB also utilized the ggplot2 visualization package. Work done by T.J Gray used two data sets for comparison and to build a multiple supervisory learning models. As opposed to a basic analysis, he tries to go for a full comparison through another database by Movie Lens. What he’s trying to compare are votes to ratings and their correlation between the two fields. Both databases are used to affirm or reject his hypothesis that votes are inversely related to ratings. [2]

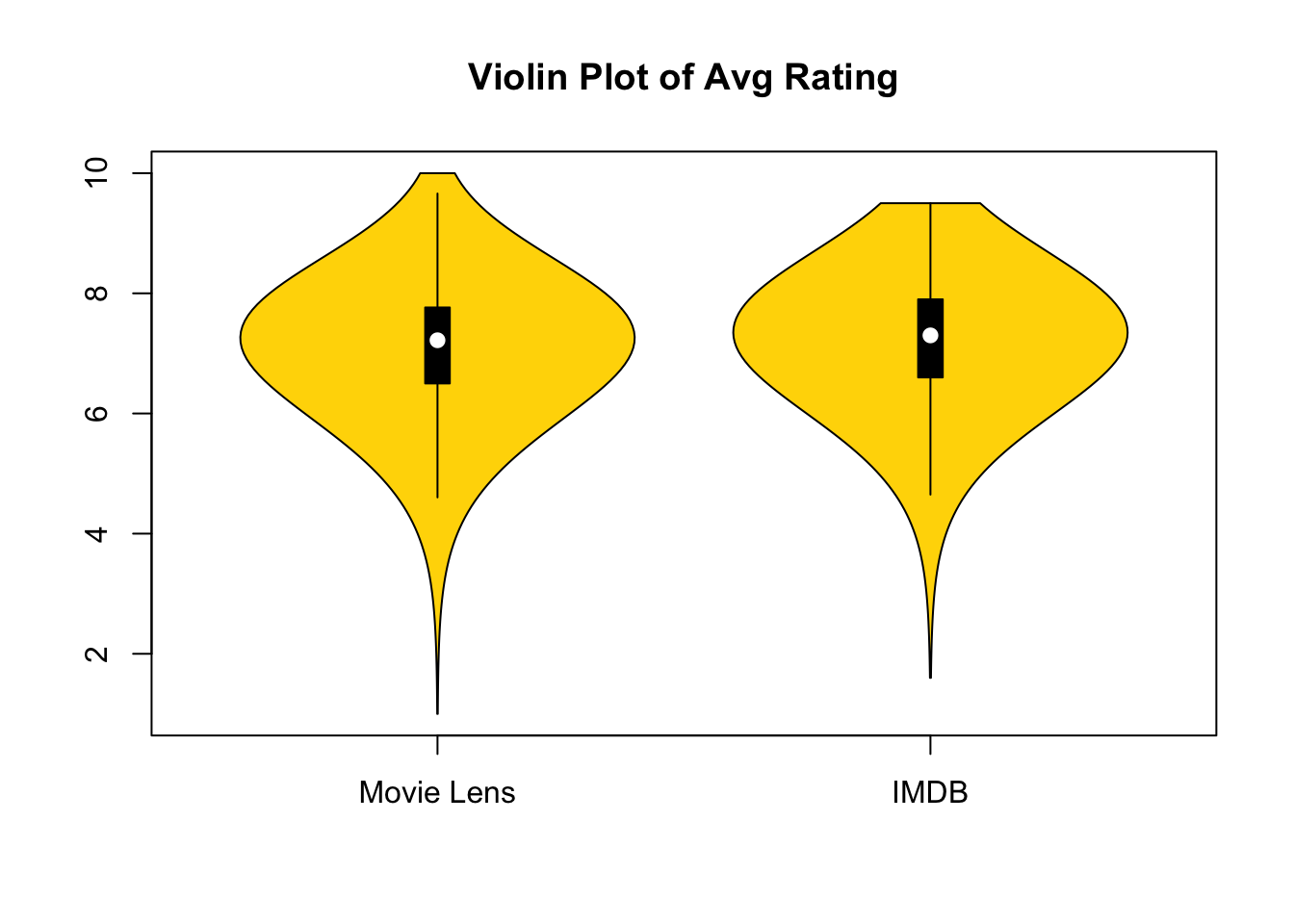


Figure 1.3.Gray’s main data visualization to compare the two databases in terms of ratings.

As mentioned before, our project is just mere analysis through visualization models. We plan to connect the dots of what we find behind the data. This is similar to the first work by Ezpov where she attempts to make a full on predictive model, however, it’s based off of her own selective fields that she believes would have more impact. (Historical era and movies.) Moreover, how she designed her experiment is different from ours in terms of how we executed the experiment with language usage and select visualization tools. What’s important to note is how her focus is only on movies while our focus is on all of IMDb which would also include various forms of media such as episodic shows.

**2. Background**

In this day and age, not only are pieces of technology advancing itself but also forms of entertainment. Media encompasses what forms of entertainment exists within the world of technology. Originally, media would affiliate itself with newspapers or books whether they may be of fact or fiction. Due to the rise of technology, that format has been subcategorized as a “print media”.

Nowadays, 3 forms of media exists, the aforementioned print media, software based virtual media (movies, shows, or multimedia), and Internet based platforms (Twitter, Facebook, YouTube). What we’ll be focusing on, are the statistics surrounding media primarily, movies and shows. The shows and movies had a large impact on the general populous through fantasies and propaganda that would shape the opinions of communities that had access to these forms of media.

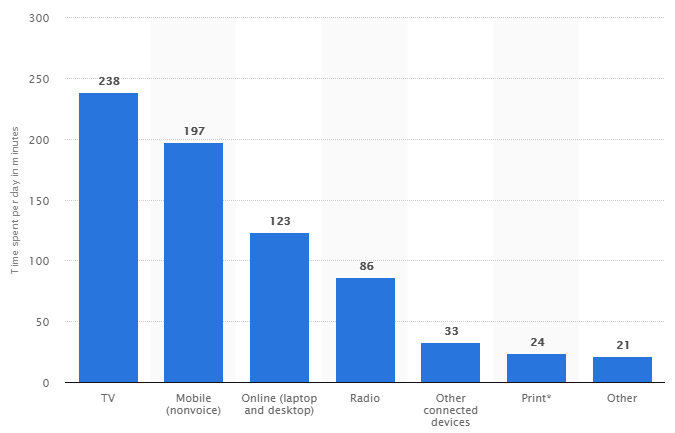


Figure 2.1. Average time spent with major media per day in the United States as of September 2017 (in minutes)

**2.1 Modern Day Impact**

In the 21st century, media such as shows and movies play an extremely important role in people’s everyday lives. Media such as shows and movies has improved in quality and have become much more easily accessible in this day and age that many would make references in relation to popular media which is known as “pop culture references”. In a special case, video games are also included within this section of media. With high accessibility, there will come large viewerships, and with large viewerships, there will become more prevalent fandoms and presence. [3]

Various fandoms of pop culture may affect their own lifestyles accordingly based off of the media that they base their fandom over. With the outstanding evolution of technology that constantly progresses with time, not only has media improved and become more accessible, but the data behind viewership, profits, and ratings are accessible to the public.

**2.2 Ratings on Quality**

With advances in technology, quality of media has increased dramatically however, there will be some points within the media creation community where they could create something unfavorable to the public eye. [5] In order to ensure that the public is safe as well as to communicate to creators what would be favorable for the public is through certain statistical fields.

IMDB is an online database of information related to films, television programs, home videos and video games, and internet streams, including cast, production crew and personnel biographies, plot summaries, trivia, and fan reviews and ratings. It would go above and beyond to force every angle of a certain media to be revealed to create some form of quality control. With IMDb, we can find out what’s best, what’s not the best, and what other people are watching. What’s more important is the database that they have been growing over time.

**2.3 Relevance and How It’s Useful**

Databases based on the topic of media is massive in size. Due to the fact that media consumption is at a daily rate worldwide, and IMDb is known to be one of the top most used databases that’s always up to date, it’s best that we use this site as a focus for our analysis. From here we are able to sort through this data and focus on a specific part to analyze. We can utilize the data given to easily predict favorable media and other standout habits shown by the data regionally. What we are trying to do is to create a geomap and pie charts show shows counted mediatypes released in a specific years by region. So we can show the likes and dislikes that align with the times. Businesses can look at this and understand that certain events would affect what people would like during these situations and help them capitalize what kind of media they should be making.

**3. Preparation for the Exercise**

So why is did we choose this set of data? As we’ve said earlier, IMDb is one of the largest databases on media watched all around the world. The data it collects is updated frequently as new shows and movies come out. From a business standpoint, companies can use this data to figure out what’s popular in different regions. With this knowledge they can be able to come out with a new movie or show that has a high chance of selling out or doing good based on the research that is made.

Before we get into the main part of the lab, we need to do some prep. We’ll need to headover to IBM Bluemix and create an account there. With this account we are able to startup a time limited Analytics Engine configured with Spark and Hive. We will also need to create credentials so we will be able to access our Analytics Engine through SSH. Once created we can remotely login to our engine.

We can find hardware specs through different linux command. “Lscpu” gives us useful information such as number of cpu cores and each cores speed. Our engine has 4 cores with each core having a speed of about 2195 MHz.

**3.1 Downloading the Data**

Once the server is done being created, we will be able to use git bash and SSH into it. Once we’re in we will be downloading the data with the -wget command, from this link <https://datasets.imdbws.com/>(insert file name here). Once we finish downloading each individual file, we need to prepare it before we are able to analyze it. We must decompress and rename each file that was downloaded. Once we finish that, we have to send all the files into HDFS directory.

**3.2 Preparing and Moving Files**

After downloading the files we will need to prep them. Once files have been unzipped and renamed, we can move our downloaded files from our local directory into our HDFS directory.

**3.3 Connecting to the Database**

We need to open the Beeline client which will allow us to connect to HiveServer2. When connected, we are able to create Hive Tables in the HDFS directories.

**3.4 Creating Tables and Inserting Data**

6 tables will be created, each with columns based off data from IMDb’s website. Once the tables have been created, we will be able to insert the data into them for better organization. Now obviously 6 tables filled with this much data can be very tedious to sort through.

**3.5 Preparing Data**

We will be breaking down the 6 tables into 2 tables with less columns. With simplified tables, it is much easier to go through the data to get what we are looking for. The two tables that were just created will now be joined based on the columns that we found relevant for the purpose of this exercise. The table is stored in a private warehouse and we will need to create a new HDFS directory and store the table there for it to be downloaded into our local server.

**3.6 Preparing File for PowerBi**

In order to be able to load our excel data into PowerBi, we must format our data into a flat table. Each column should follow the same data type and there should be a header for each column. All data should also be formatted as a table in excel. The file we be delimited by tabs so that the data is organized into tabs. Then we have to format all the cells then save the file as csv.

**4. Result and Analysis**

Now that file is prepped , we are ready to use PowerBi. We can download it officially on the Microsoft website and install it on our computer or use the online version by logging in into our account. And we can use some query codes to analyze the data.

**4.1 Result using PowerBi**

After opening PowerBi, we go to My Workspace, and click on Create button, and choose Report. A small window will be popped up, we upload the dataset from there and hit the Create button again.

**4.2 Geo-spatial visualization**

If we would like to show the different types of media released after 1970, we need to: select the Map icon, then under Location pick Region filed, and add startYear in the Legend and also add mediaType under Size. We can filter the results based on what we have set up, so the map will show up the result as the picture below:

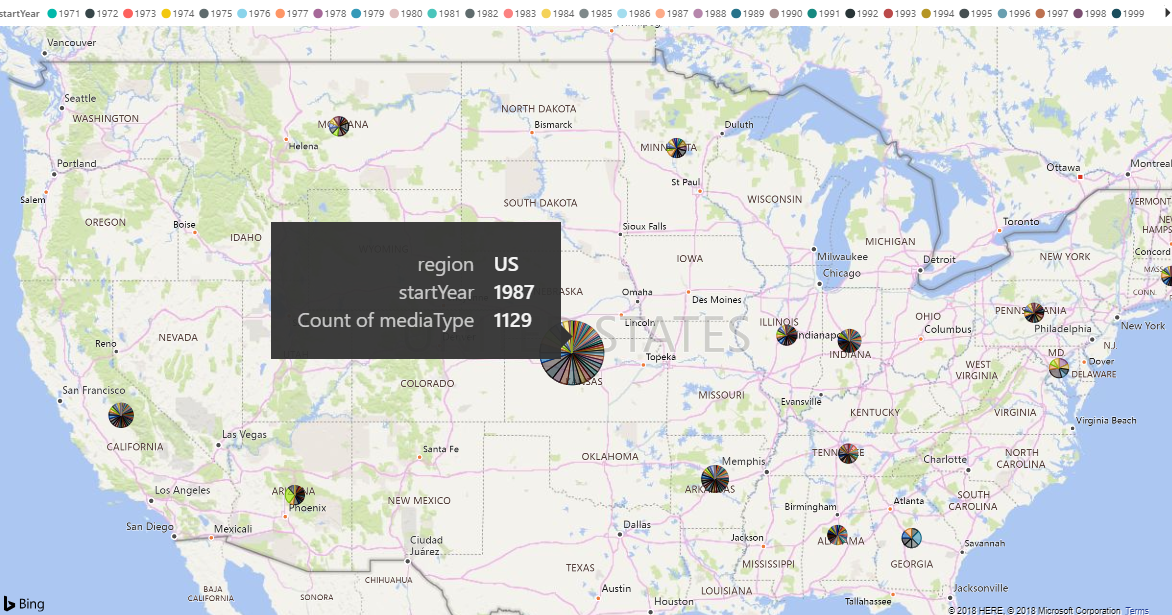


Figure 4.2 Showing movies, tvEpisode,and tvSeries released after 1970 in PowerBi.

**4.3 Analysis**

In order to analyze our data for some particular purposes, we could use some query codes. Some examples are included in figures below.

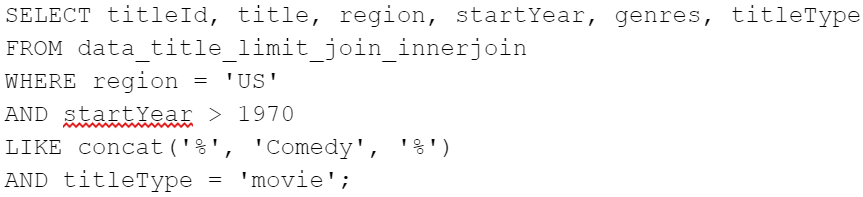


Figure 4.3a Count US Comedy movies after 1970

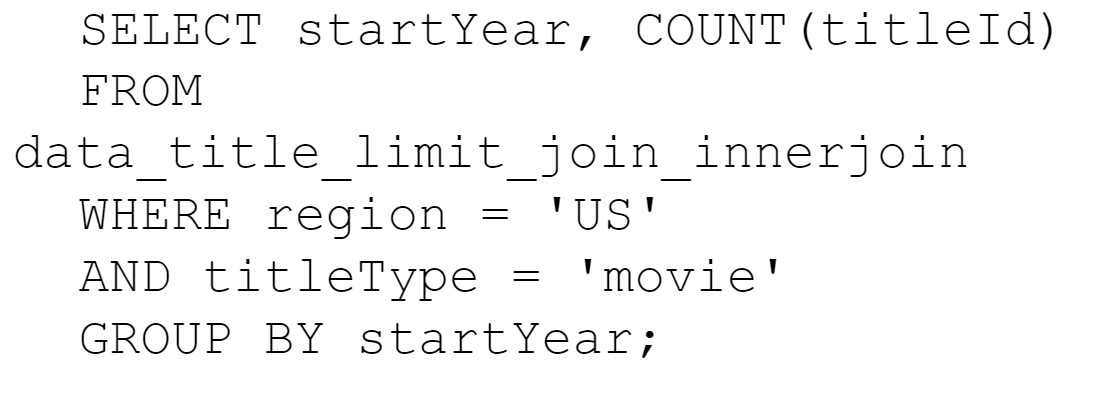


Figure 4.3b Count US movies that released each year

For each year we can count the total number of media released. We can limit our rows by limiting years. As most movies are released in the US we limit our research to titles whos region is US. This also helps us eliminate counting duplicate titles for our analysis.We then can further limit our counted results by specifying a genre and title type in the WHERE clause.

This counted number of media can be compared with other factors. Ilya Ezepov used a similar analysis to find correlations between major historic events and the impact it has on movie media. If we we to compare the number of media released each year with other historic factors we can find trends that might explain some correlations between the two. Could more sport related content be released in years that hold large events such as the Olympics?

**5. Problems Encountered**

Throughout the development of the lab, we encountered 3 main errors that took a lot of time to solve. Most of these errors were small syntactic mistakes.

The first major error I encountered was specifying the hdfs location of a table. I was not specifying the correct URI and was encountering a write permission error.

The second mistake I encountered took over 2 days to try to figure out. Joining tables proved to be more complicated than previously imagined. At first I was using LEFT OUTER JOIN to join all the contents of the first left handed tables with the contents of the right table where both table’s id matched. This was producing a cartison like product with many missing rows. Many matched ID’s were missing and there were some rows that were being duplicated. My solution to this issue was to using an INNER JOIN and the DISTINCT keyword with the SELECT statement.

The next issue dealt with finding the correct storage location when securely copying one file from one directory to another. Specifying the URI was causing issues because I was running the scp command in the shell connected to our local server. I should have been entering the scp command with a shell that was connected to my local machine. After switching environments my URI problem was solved.

**6. Conclusion**

The exercises in this project ease us into practicing Big Data with the Hadoop File System. In our project we collect, manipulate, and prepare IMDB data to be analysed. IMDB is a huge database with lots and lots of information collected on various types of media from around the world. With this exercise we planned on being able to organize all of this to see the media released after 1970 that is being watched in different regions.

We begin by creating an IBM’s Analytics Engine which allows us to connect to it’s server to download and organize the data we collected. Our data set is larger than 2GB and can be considered Big Data. Each file downloaded has over 1,048,576 lines which makes manually sorting through the raw data tedious. We were able to create tables and organize the data into the tables for easier analyzation. With services like PowerBi, we were able to use the data that we have collected so it can be displayed into a Geo-Spatial Visualization.

Even though this was a fairly easy exercise to get into Big Data, we did encounter problems. Luckily we didn’t bump into too many. Once they were resolved we continue to power on. From this exercise we were able to create a map and pinpoint within every region the release of movies and tv shows.

**References**  
[1] E. Ilya, “IMDb exploratory data analysis” Amazon AWS public analytics, January 5, 2015.

[2] G. T.J, “Movie Lens & IMDB Data Analysis” Amazon AWS public analytics, April 28, 2018.

[3] F. Nikki, “Do You IMDB?” LA Weekly, August 5, 2004.

[4] C, Jerri, “The Internet Movie Database” Lifewire, November 9, 2018.

[5] S. Aaron, A. Anderson, “Social Media Use in 2018” Pew Internet, March 1, 2018.