## Working with Movie Data

Movielens tracks movies,review ratings and taglines. At the present time there are 25 million ratings that can be used for data analysis. This has been a very popular dataset to use predictive analytics techniques to predict how movies will perform. It is assumed that hire ratings will translate into box office and viewership numbers. It is possible to link this data back to the IMDb database, which contains more movie attributes such as cast and crew, studio and their own ratings. This can also add more attributes to evaluate with the predictive model. Finally there is themoviedborg (tmdb) dataset this adds boxoffice data as well as cast and crew data.

For this project we will use basic data from movielens and link to imdb data and tmdb as well. This will give you a fairly rich set of data to analyze.

## The Challenge!

You want to report on the status of english-speaking movies over the last 15 years. This includes trying to create a prediction model to predict top movies for next year. You have to produce a report of your findings and insights along with a description of your predictive model and how you validated it. Along with the report you need to produce an interactive dashboard that not only tells the story, present in your report, but also allows users to explore the data to discover insight for themselves.

You need to solve this Challenge in three Phases. Phase 1 will gather and report on data from verified sources. Phase 2 will provide a predictive model that can be used successfully to predict high ratings and / or revenue. Phase 2 will prepare a comprehensive interactive dashboard that presents all the findings produced in the two prior phases. A report will be produced as a final deliverable.

***Tools to be used in this project***

As a minimum you need to use the following tools or technologies. This can be augmented with whatever tools you desire. Be sure to mention what tools you used in the final report.

Oracle Autonomous Data Warehouse (ADW) data will be on instructors DB
Students can also load data on their own ADW instance if desired

SQL
Oracle Analytics
Microsoft Word
*R and RStudio are optional – For those that want to us RStudio Pro will have access to copy of the data on remote KU data server.*

## PHASE 1 – Data Discovery

At a minimum, from the available data, you need to provide movie information such as title, year released, director, leading cast members, average ratings, number of ratings given, and gross revenue. Descriptive Analytics dashboard pages that provide this information in a way non-technical viewers can understand and explore the data. All available data should be used as a basis of this work. You can exclude columns or tables if properly justified.

Questions to answer:

Using the data provide you need to answer the following questions (this is just a minimum!)

1. What are the average number of English speaking files released each year over the most recent 15 years?
2. Categorize #1 by genre.
3. What year had the largest number of films? By Genre? (Rank them?)
4. Can you report by gender of the Director? (Hint: OAC can enrich data by determining gender from name or maybe there is a column that will help)
5. What are the top Directors? Actor? Male or Female top Actors?
6. What are the top Rated and bottom rated films per year and overall?
7. Do the same by revenue.
8. Do Ratings directly correlate with Revenue?
9. Do tags add value to any analysis?
10. Do Actors add value to the analysis?
11. Do Directors add value to the analysis?
12. Are there any insights that can be made by using Production Company?
13. Can you determine profitability of the movies?

### Available Data

The data for this project can be found on my ADW server. Use the following information to setup a connection:

Server Name: KUCCADW2
Username: BA480DATA
Password: GoHawks#2175
Wallet File: Provided on Blackboard

This data is Read Only and you will not be able to make changes to the data. If you need to make changes or include your own data you need to do a data merge or transfer data to your ADW server. There will be a mini-lesson about that posted the week of April 27$^{th}$ ,if not sooner.

There is data from various sources. The primary data is Movielens data, followed by IMDb data. The remaining data sources should also be used to obtain additional demographic or metrics required for the model or reports.

It should be noted that simple single column inner joins may not provide proper joins with the data. Use of outer joins or multi-column joins may be necessary. Use the skills and process learned in the early lectures and assignments to help you guide thru the process. There will be some hints provided in these instructions, but they should not be considered the only things that can be done to gather the data properly to satisfy the requirements of this phase of the project.

### *MovieLens Data*

The data is comprised of 5 data files. The names and data subset contained in the files are provided below:

ML_MOVIES. -- This contains the movieid and title
ML_RATINGS – This contains the reviewerid, timestamp and rating for each movie.
ML_LINKS -- This provides the translation between movieid and imdb movie id

The following files can add additional "optional" attributes these joins will result in multiple rows for each movie, so be careful.

ML_TAGS -- This contains the tag for the movie that each reviewer provided
ML_Genomes-tags -- This file linked to genomes-scores provide the genres and score that are associated
ML_Genomes_scores -- This file contain the scores associated to the genomes

There is also a separate document provided by movielens contained in the project files.

Various joins can be made to evaluate the data. The following SQL provides the joins that can be used.

```
select * from ml_movies m – you should put in actual columns!
join ml_ratings r on (m.movieid = r.movieid) -- return more than one
row per reviewer (suggest using aggregate for AVG REVIEW)
join ml_links l on (m.movieid = l.movieid)
order by tile

-- There are other joins that you can do if you like
--join ml_tags t on (m.movieid = t.movieid and r.userid = t.userid)
-- This will return 1 or more tags for each movie
-- Join to genome
Join ml_genome_scores on (m.movieid = g.movieid and g.tagid = t.tagid)
Join ml_genome_tags on (g.tagid = t.tagid)
   Note: suggest aggregating these.. by most relevance maybe even top x
scores.
```

## *IMDB DATA*

IMDb data is fairly extensive, but we will be limited only a few datasets. These include the following tables that represent a subset of the available data.

IMDB_MOVIES – contains the basic information about the movies (only movies will be used here)

IMDB_RATING – this is the imdb avg rating. This can be compared with the ratings from movielens.

Joins to this data is simple using the imdb_id column between all the tables. Note that you can also link to the movie lens data using the imdb_id column in the ML_LINKS table.

## *TMDB Data*

TMDB_MOVIES – is data joined from several table to provide a table that has several import columns of data.

You will need to join all the table together to be able to have all the data necessary to finish this report. Not all tables are required for the analysis. You may find redundant data and unnecessary data as well. You need to find the appropriate select statements and joins to answer your questions.

I would suggest that you use SQL to create select statements that will get the data you need in a form necessary to do your reporting and analysis. You should find that one large select statement will work.

Once you have that you can export the data (using SQLDeveloper Desktop) to an excel or csv file that you can then upload to your ADW instance for further analysis. There will be 61 students working on this data and if you all work on the same db server you will have access issues and performance problems. This is not unlike the “Real World”. So do your initial discoveries on the source data and then determine what you need and pull it down so you can work with the data. You will find that while the base data is quite large you will not need all of the data to do the analysis.

Hint: In order to use the ML_RATINGS you will need to aggregate them and associate the results to the movie. There are only about 29K movies but 25 million ratings!

Doing this will reduce the actual row counts down to something that is very manageable.

## PHASE 2 – Predictive Analytics

During this phase you will evaluate which prediction methods to use to provide the required results. I suggest you try multiple methods and from the results choose the best method. Explain and document your process and results. Compare your predicted ratings and revenues with actual results to help determine the best predictive model.

You should have found during phase 1 that the dataset you are working will is not as big as the base data that you started with. You will need to spit the data up for training and testing data as well. This is another good reason to extract the data you need and bring it to your server for further analysis. You can then determine the best way to split your date (either by sampling or maybe by year). If you need ideas on how to do that I can produce a short video during the week of 27 April. Please let me know and if there are enough people who want this I will do the video.

Use as many attributes as you can in the models. If you can, provide an analysis of which attributes were significant and which were not.

Provide your final predictions and the results in report and visualization format.

## PHASE 3 – Report your Findings

In this phase I want you to prepare a set of Interactive Dashboard pages that provide your findings, insights and predictive model results.

Prepare a final report, using the report example provided, combining the results of Phase 1 and 2 and provide a summary and conclusions.

I will post a mini-lesson next week on how to do the report and hints on what I am looking for.