ÉCOLE POLYTECHNIQUE DE MONTRÉAL

DÉPARTEMENT DE MATHÉMATIQUES ET DE GÉNIE INDUSTRIEL

TRAVAIL DE SESSION 1/2

PAR

FRANCIS FORGET : 1790076

SAMUEL CHAPLEAU : 1798907

VINCENT LABONTÉ : 1793452

TRAVAIL PRÉSENTÉ À SÉBASTIEN LE DIGABEL
DANS LE CADRE DU COURS MTH2302D
PROBABILITÉS ET STATISTIQUE

7 OCTOBRE 2016

Contexte général des données

Dans le cadre du travail de session du cours MTH2302D Probabilités et statistiques, nous avons choisi d'analyser des données relatives au domaine cinématographique. Selon nous, il est facile de constater la croissance, au fil du temps, de la popularité de la cinématographie, qui se traduit, aujourd'hui, de plus en plus à travers l'émergence de superproductions. Évidemment, derrière cette évolution se cachent des données massives ayant probablement des corrélations intéressantes. De plus, la réputation de certains sites Web cultes de compilations de critiques de films ainsi que de base de données comme *IMDb* et *Metacritic* nous incite grandement à considérer l'information résultante qu'ils recèlent pour pouvoir, conséquemment, y proposer diverses conclusions. C'est donc à l'intérieur de ce contexte d'actualité que nous avons décidé d'entreprendre notre cueillette de données.

Le travail qui suit comportera, d'abord, une section sur la provenance des données qui expliquera plus en profondeur les techniques et les lieux virtuels de collecte de l'information nécessaire à l'analyse. Par la suite, il y aura un second segment qui décrira la forme des données sélectionnées. Finalement, la dernière partie du travail se composera de questions ouvertes portant sur l'observation, l'interdépendance et la pertinence des résultats précédemment détaillés.

Provenance des données

Pour ce qui est de la provenance de nos données, nous avons utilisé deux interfaces de programmation applicative (API) totalement publiques. Ces deux interfaces sont disponibles aux liens suivant : www.themoviedb.org et www.omdbapi.com.

De façon plus détaillée, nous avons exécuté des scripts Python afin d'aller collecter toutes les informations nécessaires. Les scripts s'occupent de parcourir tous les films des deux bases de données afin de filtrer les films valides et de ne prendre que les variables intéressantes à notre étude. Par la suite, d'autres scripts s'occupent de faire la conversion des fichiers JSON, préalablement enregistrés, vers des fichiers CSV. Ceux-ci peuvent ensuite être ouverts à l'aide d'Excel. Les fichiers sources sont disponibles à l'adresse suivante : www.github.com/vincentlabonte/MTH2302D.

Description de la forme des données

Après avoir employé les techniques présentées dans la section précédente, nous avons réussi à collecter une grande quantité de données pour nous permettre d'analyser un bon échantillon. En effet, nous avons obtenu des informations sur exactement 3278 films, rassemblant beaucoup de films sortis de 2000 à 2016. Pour chacun de ces films, nous avons emmagasiné 10 types de données différentes. La liste qui suit contient la description de chacune de ces variables collectées.

Titre : Représente le titre du film. C'est une variable de type discrète à caractère descriptif seulement servant à distinguer les films dans le fichier Excel.

Date de sortie : Représente la date de sortie du film. C'est une variable de type discrète caractérisée par le jour, le mois et l'année de sortie du film sous la forme AAAA-MM-JJ.

Budget : Représente le budget utilisé pour le film en dollars américains. C'est une variable de type discrète, mais qui sera employé comme une variable continue due à sa proportion.

Durée : Représente la durée, le temps complet, du film en minutes. C'est une variable continue qui a été discrétisée sous la forme de minutes.

Classement : Représente le référencement cinématographique pour noter la pertinence du film pour certains publics. C'est une variable discrète prenant des valeurs telles que *G*, *PG*, *PG-13*, *R* et *NC-17*.

Genre : Représente le genre cinématographique du film. C'est une variable discrète prenant des valeurs telles que *Comedy*, *Crime*, *Drama*, *Fantasy*, *Romance* et *Sci-Fi*.

Réalisateur : Représente le réalisateur du film. C'est une variable discrète sous la forme du prénom et nom ou du surnom du réalisateur.

Récompenses : Représente le nombre de récompenses du film. C'est une variable discrète comprenant les victoires aux Oscars et aux Golden Globes ainsi que toute autre victoire ou nomination reçue.

Metascore : Représente la valeur du Metascore attribuée au film sur le site www.metacritic.com. C'est une variable continue, puisque c'est une variable calculée à

partir de valeurs attribuées par des critiques, mais qui a été discrétisée sous une échelle de 0 à 100.

Cote IMDb : Représente la valeur de la cote IMDb attribuée au film sur le site www.imdb.com. C'est une variable continue, puisque c'est une variable calculée à partir de valeurs attribuées par les utilisateurs du site qui a été discrétisée à la décimale près.

Questions ouvertes

À la suite de la cueillette de données, plusieurs questions nous sont venues en tête. La liste qui suit comporte celles qui nous paraissent les plus susceptibles d'avoir une corrélation réaliste ou qui nous semblent être tout simplement les plus intéressantes à analyser.

- Existe-t-il une dépendance entre la popularité du film et son budget ?
- Existe-t-il un lien entre le réalisateur d'un film et la popularité de celui-ci ?
- Existe-t-il une corrélation entre le genre d'un film et sa durée ?
- Existe-t-il une relation entre le réalisateur d'un film et son obtention de récompenses ?
- Existe-t-il un lien entre l'obtention de récompenses et le budget attribué à un film ?
- Est-il possible de conclure que la popularité d'un film joue un rôle sur les récompenses qu'un film peut obtenir ?

N. B. La popularité peut être définie en fonction du Metascore ou en fonction des cotes attribuées par les utilisateurs du site IMDb.