Université Nice Côte d'Azur
Enrico FORMENTI

Année 2020/21
M1 MIAGE
Maths pour Big Data

Examen sur projet

Table des matières

1 Le projet et sa structure

Le projet de cette année voudrais être représentatif d'un cas d'application réelle des notions vues en cours et il consistera donc de quatre parties :

(C) collecte des données et des informations ;

(E) étude et mise en œuvre des tests statistiques aptes à vérifier des hypothèses sur les informations collectées ;

(R) rapport sur le cas d'étude et discussion des résultats obtenus.

(P) présentation orale de l'ensemble du projet.

Chacune des ces parties sera détaillée dans la suite de ce document.

1.1 Collecte (C)

Pour la totalité des sujets proposés les données sont disponibles au téléchargement sur Internet mais il faudra parfois les nettoyer des données inutiles ou vérifier la consistance des valeurs. Pour cela il faudra donc écrire des petits scripts de pre-traitement. Pour cela il faudra utiliser les techniques vues pendant les TD1 et TD2. Il s'agira, évidemment, de choisir les plus adaptées.

La qualité des données qui en résultera sera fondamentale pour la fiabilité des conclusions que vous pourrez tirer par la suite.

1.2 Etude (E)

Chaque sujet prévoit de vérifier un certain nombre d'hypothèses en suivant les exercices-modèles que nous avons vu en cours. Les bases de données constituées au point précédent seront utilisées à la fois pour calculer les paramètres statistiques utiles à l'étude et aussi pour extraire l'échantillon sur lequel appliquer les tests d'hypothèses. Les modalités du découpage varient en fonction de la base de données et seront spécifiées dans le sujet.

1.3 Rapport (R)

La dernière partie du projet (mais pas la moins importante !) prévoit la production d'un petit rapport sous forme de page web (html) contenant :

— un descriptif du sujet et de ses finalités ;
— un descriptif des données utilisées ;
— la méthodologie suivie pour répondre aux questions ;
— les résultats obtenus et leur visualisation graphique autant que possible ;
— une courte présentation des membres du groupe et de leur rôle dans le projet.

1.4 Présentation orale (P)

La présentation orale devant les autres étudiants de la promo et l'enseignant visera à présenter (*a minima*) : le sujet, les méthodologies utilisées, les résultats obtenus. La durée de la présentation est fixée à 12 minutes (attention à ne pas dépasser !). Chaque membre du groupe devra présenter une partie. La présentation sera suivie d'une séance de questions (prof plus étudiants) d'une durée maximale de 5 minutes.

1.4.1 Durée de la présentation orale

La présentation orale aura une durée de 17 minutes dont 12 minutes de présentation (chaque membre du groupe devra y participer) et 5 minutes de questions.

2 Les groupes

Chaque groupe est composé de 2 ou 3 étudiants (4 pour certain sujets). Chaque étudiant est supposé participer au développement du projet et à la présentation orale. Les groupes de moins de 2 ou plus de 4 étudiants sont interdis. Les groupes sont ceux de la liste fournie par M. Renevier et ne peuvent être changés.

3 Choix du sujet

Chaque groupe choisit son sujet après une concertation entre ses membres. Il s'agira ensuite de communiquer à l'enseignant les informations suivantes :

— nom, prénom et adresse email de chaque membre ;
— liste des quatre sujets choisis par ordre de priorité (décroissante).

4 Attribution d'un sujet

Après avoir envoyé votre choix à l'enseignant vous allez recevoir un email de confirmation avec l'acceptation du groupe. A la même occasion il vous sera aussi attribué un sujet. La politique d'attribution des sujet sera "premier arrivé, premier servi". Tout groupe n'ayant pas encore choisi de sujet avant mardi 15 décembre 2020 à minuit se verra attribué un sujet d'office. Le sujet attribué (d'office ou pas) ne pourra en aucun cas être changé.

5 Ce qu'il faut rendre et quand

Il faudra rendre au professeur mail un fichier nommé `groupe-xxx.zip` où `xxx` est l'identifiant du groupe. Cet archive doit impérativement contenir quatre répertoires : `PRE`, `R`, `FIG` et `REP`. Ce répertoires doivent contenir, respectivement :

— les scripts ou programmes utilisés pour le traitement des données ou si la base de donnée a nécessité d'un pre-traitement ;
— les scripts R pour effectuer les statistiques et les vérifications d'hypothèses ;
— les scripts R pour générer les graphiques utilises dans le rapport ;

— le rapport final au format html avec les fichiers annexes (figures, tables, etc).

Le projet est à rendre avant le **15 janvier 2021 à minuit**.

6 La notation

Chaque partie du projet sera notée sur 4 points et va donc constituer la note $C + E + R$ à laquelle s'ajoutera la note sur la présentation orale P (sur 4 points aussi) complétée par la note de participation Π (toujours sur 4 points). La note de participation sera attribuée individuellement à chaque étudiant sur la base de son attitude lors des présentations orales des autres groupes et de sa participation à la discussion générale à l'issue de chaque présentation orale. Il est donc indispensable d'être présents aux présentations orales (d'au moins à une grande partie).

Bon courage à tous !

7 Les sujets

Les sujets sont groupés dans les deux catégories suivantes. Chaque catégorie, voir chaque sujet, a ses propres spécificités.

7.1 Kaggle data sets

Ce groupe de sujets est tiré d'un site web qui met à disposition des utilisateur une énorme quantité de données, les plus disparates et vise devenir une référence pour la communauté de chercheurs et d'étudiants travaillant dans le machine learning et le big data plus en général. Les données sont librement téléchargeable à l'adresse fourni mais par contre il s'agit assez souvent de bases assez grandes ou qu'il faudra, le cas échéant, traiter avant de pouvoir les exploiter.

7.1.1 Consommation d'alcool et résultats scolaires

Nombre de groupes : 4
Taille d'un groupe : 2/3 personnes

La consommation d'alcool chez les jeunes étudiants est souvent objet de débat. Considérons les données disponibles dans cette base. Il s'agit d'un sondage mené auprès d'étudiants de deux lycées Portugais qui vise à vérifier s'il y a un lien entre consommation d'alcool et taux de réussite à l'école. D'autres informations sur la structure de la base se trouvent ici.

Les données sont séparées en deux fichiers selon la matière considérée (Mathématiques, Langue Portugaise). Il s'agira donc, dans un premier temps de fusionner les deux bases. Ensuite, par le biais d'une analyse de corrélation cherchez à établir quelles sont les variables qui sont corrélées et cherchez à comprendre/expliquer ces corrélations.

A présent, il faudra chercher à répondre à un certain nombre de questions en vous appuyant sur les données dont vous disposez :

1. est-ce que c'est vrai que la consommation d'alcool moyenne est plus élevée chez les étudiants dont les parents ont un niveau d'instruction modeste ?
2. est-ce que c'est vrai que le taux de réussite est supérieur chez les étudiants qui n'abusent pas d'alcool ?
3. est-ce que la proportion de garçons qui abusent d'alcool est significativement supérieure à celle des filles ?

Pour chaque question il faudra utiliser la base de donnée pour calculer les quantités statistiques concernées par la question (moyenne, variance, etc) et ensuite extraire un échantillon significatif E (la taille c'est à vous de la choisir selon ce qu'on a dit en cours, attention aussi aux modalités d'extraction !). Vous allez ensuite mettre en place des test d'hypothèses adaptés en fonction de la question qui est posée.

7.1.2 Gun violence data

Nombre de groupes : 4

Taille d'un groupe : 3 personnes

Il s'agira de mener une série de statistiques sur l'archive du même intitulé qui se trouve ici. Cet archive (au format CSV) contient des entrée relative aux faits divers causés par des armes à feu qui ont eu lieu aux États Unis entre le 1er janvier 2013 et le 31 mars 2018. Il s'agira de montrer l'évolution d'un certain nombre de variables pendant la période concernée, d'en calculer la moyenne et l'écart type (quand cela est possible). L'ensemble V des variables qui nous intéressent contient :

— nombre de blessés ;
— nombre de morts ;
— nombre de malfaiteurs ;
— age des malfaiteurs ;
— état dans lequel le fait divers a eu lieu.

Voici les questions qu'il faudrait étudier :

1. Est-ce qu'il y a des corrélations (linéaires) entre les variables de V ?
2. Prenez les données de la période entre le 1er janvier 2013 et le 31 décembre 2017 pour calculer les quantités statistiques (moyenne, écart type, etc) et considérez les données entre le 1 janvier 2018 et le 31 mars 2018 comme un échantillon. Est-ce que l'on peut dire que les valeurs moyens des variables dans V ont significativement changé par rapport au passé ?
3. Si l'on prend en compte aussi le mois de l'année dans lequel le fait divers a été commis, est-ce qu'il y a une corrélation forte entre le nombre de fait divers et le mois de l'année ? Quelles conclusions en tirez-vous ?

7.1.3 Les taxi à Chicago (US)

Nombre de groupes : 4
Taille d'un groupe : 3/4 personnes

Dans ce projet il s'agira d'étudier un certain nombre de questions sur les courses de taxi de la ville de Chicago aux US. La base de donnée se trouve ici [1] Dans cette base sont mémorisés (entre autres) :

— les prix (P)
— la durée (D)
— la longueur (L) du parcours

des courses effectuées par les chauffeurs de taxi de la ville de Chicago depuis 2013 jusqu'à aujourd'hui. Vous devez vous intéresser aux questions suivantes après avoir étudié la distribution des valeurs P, D et L :

1. Utilisez les données des années entre 2013 et 2018 pour calculer la moyenne et la variance de P, D, L et les données connus pour 2019 comme échantillon pour tester si l'on peut affirmer que la moyenne de ces variables change sensiblement cette année ou pas (on prendra $\alpha = 5\%$).
2. Ne connaissant pas la formule pour la constructions des tarifs des courses, proposez une étude pour établir si le prix des courses est liée et de quelle manière à la longueur du trajet ou au temps de parcours.
3. Divisez la journée en plage horaires de 4h. Est-ce que vous pouvez via des tests d'hypothèses vérifier si l'une de ces plages peut être considérée comme "plage de pointe" ?

7.1.4 Les crypto-devises

Nombre de groupes : 4
Taille d'un groupe : 3 personnes

Le marché des crypto-devises s'est révélé assez volatile et les professionnels (et non) font profusion de conseils sur les cours. Nous voudrions aussi faire quelques considérations en exploitants les données qui nous sont mises à disposition ici. L'idée est d'utiliser les données des cours de 2016 et 2017 pour calculer moyenne et variance d'un certain nombre de variables comme :

— cours minimal du Bitcoin (moyen) ;
— cours maximal du Bitcoin (moyen) ;
— valeur des transactions (moyen) ;
— proportions des transactions en Bitcoin par rapport aux autres crypto-devises.

et utiliser les données de 2018 comme échantillon. Essayez de répondre aux questions suivantes :

— Est-ce que la moyenne de ces variables a considérablement changé en 2018 par rapport à la période 2016-2018 ?
— Est-ce que l'on obtient les mêmes conclusions si l'on considère une fenêtre de temps plus grande pour calculer les moyennes ?

1. Attention ! La base de donnée a une taille importante, il faudra donc prévoir sur votre disque dur plus de 120Go libres si vous voulez télécharger toute la base.

— Utilisez un test de corrélation pour vérifier si les autres crypto-devises sont ou pas corrélées au Bitcoin.
— Est-ce que la distribution du cours sur l'année 2018 du Bitcoin suit une distribution statistique que vous connaissez ?

7.1.5 Quelques considérations sur Netflix

Nombre de groupes : 4
Taille d'un groupe : 3 personnes

Netflix est probablement le système de vidéo à la demande le plus répandu au monde. Nous avons trouvé ici une base de donnée historique sur les programmes qui ont été proposés jusqu'en 2019. L'idée est séparer les données en deux parties H et A. La partie H (historique) contiendra toutes les données dès le début jusqu'à 2018 et A les données du 2019 [2]. Vous allez utiliser H pour calculer les quantités statistiques remarquables utiles à vos fins (moyenne, écart type, etc) et A sera utilise comme échantillon pour vos test statistiques.

Vous chercherez à répondre aux questions suivantes :

1. est-ce que c'est vrai que le temps moyen entre la sortie du film et son ajout au carnet Netflix s'est raccourci ?
2. à votre avis quelle est la cause de la tendance que vous remarquez au point précédent ? Proposez une solution pour pouvoir donner une réponse plus réaliste au point 1) et refaites les calculs en ce sens.
3. est-ce que c'est vrai que la proportion de show télévisés qui sont mis au programme considérablement augmenté par rapport aux autres programmes (films, séries) ?
4. même question que la précédente mais en prenant en compte la durée en minutes.s

7.2 Et maintenant venons sur des sujets plus français…

Dans cette section nous trouvons des bases de données de provenance gouvernementale ou institutionnelle.

7.2.1 Evolution de la température en France

Nombre de groupes : 2
Taille d'un groupe : 3 personnes

On voudrais étudier l'évolution de la température atmosphérique sur le territoire français. Le problème est que les données historiques de Météo France sont payantes (200000€/an). On se contentera donc d'utiliser les quelques données publiques qu'on trouve ici pour arriver à nos fins. Choisissez cinq villes et considérez les variables statistiques (par rapport aux villes choisies bien sûr) :

— température minimale 12h
— température maximale

Découpez les données en deux parties : avant 2019 et courant 2019. Vous utiliserez le premier jeu de données pour calculer les quantités statistiques intéressantes (moyenne, écart type, etc) et celles du deuxième seront utilisées comme échantillon. Essayez de répondre aux questions suivantes :

1. Est-ce qu'il y a eu un changement significatif de la moyenne de la température par rapport au passé ?
2. Répétez les tests précédents en découpant les données autrement : avant 2015 et courant 2019. Que remarquez-vous ?
3. Essayez de comprendre à présent si le réchauffement (ou pas) est présent aussi bien en zone rurale que dans les grands villes. Pour avoir le classement des villes en zones rurales on utilisera les données sur les Zones de Revitalisation Rurale fournies ici.

7.2.2 Diffusion de la varicelle en France

Nombre de groupes : 2

2. Ici quand nous parlons d'année, nous voulons dire l'année d'ajout du programme (film, série ou spectacle télé) au carnet de Netflix.

Taille d'un groupe : 3 personnes

On parle d'épidémie d'une certaine maladie lorsque le seuil défini par les organismes de surveillance gouvernementaux est dépassé pendant deux semaines consécutives. Ce « nombre de cas attendus » ou « seuil » correspond au nombre de cas observés par rapport à la même époque les années précédentes, ce n'est donc pas un palier fixe.

L'Institut de recherche pour la valorisation des données de santé (IRSAN) analyse des centaines de milliers de données médicales issues de l'activité d'environ 1000 médecins qui envoient leurs données. Dans la base donnée que vous pouvez trouver ici les services de surveillance de la santé ont mémorisé les données sur diffusion des cas de varicelle. Il vous est demandé de :

— Chercher à détecter s'il y a eu épidémie ou pas dans les cinq dernières années.
— Si l'on prend comme échantillon les données pour l'année 2019 et comme base pour calculer (moyenne, écart type, etc) le cinq année qui vont du 2014 au 2018 (extrêmes inclus). Est-ce que l'on peut affirmer que le nombre de cas moyens est sensiblement différent par rapport aux cinq ans précédents ?
— Par un étude de la corrélation, essayez de voir si vous pouvez détecter où l'épidémie est localisée.
— Est-ce que vous pouvez établir un lien directe entre la proportion de cas et population de la région ?

7.2.3 Et si on parlait de sécurité routière ?

Nombre de groupes : 2
Taille d'un groupe : 3 personnes

Le journaux et la télévision nous rappellent sans cesse les problématiques liées à la sécurité routière. Cette année (ainsi que la passée) nous assistons à une hausse du nombre de morts sur la route après des années de baisse constante. Le problème c'est comment et où intervenir pour invertir la tendance actuelle. Vous allez à chercher quelques réponse partielle en analysant les données qui se trouvent ici et qui nous sont mises à disposition gratuitement par l'ONISIR (Observatoire National Interministériel de la Sécurité Routière). Vous pouvez trouver une notice détaillée sur comment exploiter ces données ici. Ce serait bien de la lire attentivement avant de commencer le projet.

L'idée est de prendre les données entre 2005 et 2018 comme référence pour calculer les quantités statistiques intéressantes (moyenne, écart type, etc) et celles pour l'année 2019 comme échantillon. Vous allez chercher à répondre aux questions suivantes :

1. Est-ce que c'est vrai que la proportion d'accidents mortels (impliquant au moins un décès) est augmentée en 2019 ?
2. Même question que la précédent par rapport aux blessés hospitalisés.
3. Si l'on voulez intervenir par une manœuvre soit au niveau législatif soit au niveau de contraste local (contrôles policiers, etc) on pourrait chercher à comprendre si la catégorie de véhicule est significative (vous considérerez seulement trois catégories : deux roues, voitures, poids lourds). Mettez en place un test qui pourrait répondre à cette question.
4. Toujours en allant dans le sens de la question précédente cherchez à comprendre si l'autoroute est plus dangereuse qu'une route nationale ou une route départementale.