Université Nice Côte d'Azur
Enrico FORMENTI

Année 2019/20
M1 Miage
TD Math pour Big Data

Examen sur projet

1 Le projet et structure

Le projet de cette année voudrais être représentatif d'un cas d'application réelle et consistera donc de trois parties :

(C) collecte des données et des informations ;

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

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

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

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

1.1 Collecte

Pour un certain nombre de sujets 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. La qualité des données qui en résultera sera fondamentale pour la fiabilité des conclusions que vous pourrez tirer par la suite.

1.2 Etude

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

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

La présentation orale devant les autres étudiants de la promo et l'enseignant visera a 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.

2 Les groupes

Chaque groupe est composé de 3 étudiants (4 pour certain sujets). Chaque étudiant est sensé participer au développement du projet et à la présentation orale. Les groupes de moins de 3 étudiants sont interdis. Pour déclarer un groupe il faut envoyer un mail à l'enseignant avec objet `[MATH2] groupe` et pour contenu les noms, prénoms et émail des membres. L'émail doit être celle déclarée sur le site `Piazza`. Tout étudiant ne faisant pas partie d'un groupe à 12h du vendredi 15 novembre 2019 sera affectée à un groupe d'office.

3 Les sujets

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

3.1 Kaggle data sets

Ce groupe de sujet est un peu différent des précédents car il ne nécessite pas de produire soit même les données à analyser. La base de donnée est 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.

3.1.1 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 ? Quels conclusions en tirez-vous ?

3.1.2 Évolution de la température aux US

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

Nous voudrions mener un étude sur l'évolution de la température de l'air aux US pendant une longue période. La base de donnée est téléchargeable ici dans la section intitulée "Meteorological". Pour procéder à l'étude il faudra télécharger tous les fichiers entre 1980 et 2019. Choisissez cinq états qui vous paraissent représentatifs des zones climatiques présentes. Vous allez développer vos statistiques exclusivement par rapport à ces états cibles. Les variables statistiques à considérer sont :

— température minimale ;
— température maximale ;
— température moyenne.

Les questions que vous devez analyser sont les suivantes :

1. Est-ce qu'il y a une corrélation (linéaire) entre ces trois variables ?

2. Prenez les données entre 1980 et 2014 pour calculer les quantités statistiques intéressantes (moyenne, écart type, etc) pour ces trois variables. Maintenant considérez les données de la période 2015-2019 comme étant l'échantillon. Est-ce que l'on peut affirmer que la moyenne des trois variables a significativement augmenté par rapport à la période 1980-2014 ?
3. Même question qu'au point précédent mais par rapport à l'écart type.
4. Sur la bases des données que vous avez étudié que peut-on conclure alors sur l'épineuse question du réchauffement climatique ?

3.1.3 Les taxi à Chicago (US)

Nombre de groupes : 2
Taille d'un groupe : 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" ?

3.1.4 Évolution de la composition de la population mondiale

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

Il est bien connu que la population mondiale compte désormais plus de sept milliards et demi d'individus et qu'elle est en augmentation constante. Par contre, on dit que la population européenne n'augment que très légèrement au fil des année et qu'elle de plus en plus âgée. Afin de vérifier ces deux affirmations, nous allons utiliser la base donnée qui se trouve ici. Il s'agira d'utiliser les données dans la plage 1960-2010 pour calculer moyenne et variance des variables aléatoires qu'il nous intéressent. Puis on utilisera les données de 2011 à 2017 comme échantillon pour la vérification des hypothèses. On utilisera la même technique d'exploitation des séries temporelles pour étudier la proportion de population âgée de plus de 65 ans par rapport au reste de la population. En prenant les données qui se trouvent ici, fournies par l'INSEE, est-ce que vous pouvez établir une corrélation entre le niveau de dépenses de santé et le vieillissement de la population ?

3.1.5 Les crypto-devises

Nombre de groupes : 3
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.

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.

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 ?
— 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 ?

3.2 Et maintenant venons sur des sujets plus français...

3.2.1 Evolution de la population française

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

Le site de l'INSEE nous met à disposition une série de données sur la population française. Sur ces pages, on trouve des données sur la répartition de la population par tranche d'age. Nous voudrions étudier le vieillissement de la population. Il faudra donc retravailler les données pour constituer une base contenant les séries temporelles de la proportion de la population ayant plus de 65 ans et celle ayant plus de 80 ans ainsi que leur répartition entre hommes et femmes par rapport à toutes les autres tranches d'ages confondues.

Le but c'est d'utiliser la plage 1970-2010 pour les paramètres statistiques qui nous intéressent et ensuite utiliser les donnée de 2011 à 2017 pour vérifier si la tendance est à la hausse comme on dit et pour quelle catégorie d'individus (hommes/femmes). Vous ferez la même étude aussi par rapport aux flux migratoires.

3.2.2 Evolution de la température en France

Nombre de groupes : 3
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.

3.2.3 Diffusion de la varicelle en France

Nombre de groupes : 2
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 ?

3.2.4 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.

Malheureusement les données concernant l'année 2019 ne seront mis en ligne qu'en octobre 2020, nous allons arrêter nous notre analyse au 2018. L'idée est de prendre les données entre 2005 et 2017 comme référence pour calculer les quantités statistiques intéressantes (moyenne, écart type, etc) et celles pour l'année 2018 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 2018 ?
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.

4 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 (celle donnée pour s'inscrire sur Piazza) de chaque membre ;
— liste des quatre sujets choisis par ordre de priorité (décroissante).

5 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 vendredi 22 novembre 2019 à 12h se verra attribué un sujet d'office. Le sujet attribué d'office ne pourra en aucun cas être changé.

6 Ce qu'il faut rendre et quand

Il faudra rendre au professeur (via Piazza) un fichier nommé `groupe-xxx.zip` où `xxx` est le numéro attribué à votre 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 décembre à minuit**.

7 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.

8 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 S (sur 4 points aussi) complétée par la note de participation P (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.

Bon courage à tous !