explore-data : Un outil d'enseignement de la statistique descriptive

François-Xavier JOLLOIS

Département STID - IUT Paris Descartes
143 avenue de Versailles, 75016 Paris
francois-xavier.jollois@parisdescartes.fr

Résumé. Après avec utilisé différents outils du marché pour enseigner la statistique descriptive, nous avons choisi cette année de développer en interne une application web dédiée à son enseignement. Nous avons pour cela utilisé le langage R, et particulièrement la librairie Shiny. Nous présentons ici les raisons de ce choix, ainsi que l'application en elle-même et le retour d'expérience après une première année d'utilisation.

Mots-clés. application web, statistique descriptive, R, shiny

Abstract. After using several market tools to teach descriptive statistics, we chose to develop a web application dedicated to teaching. For that, we used the R language and especially the shiny library. We present here the reasons for this choice, as well as the application in itself and the result after a first year of use.

Keywords. web application, descriptive statistics, R, shiny

1 Introduction

Le Diplôme Universitaire de Technologie (DUT) Statistique et Informatique Décisionnelle (STID) a pour but de former des étudiants tout juste diplômés du baccalauréat sur les domaines du décisionnel et de la statistique. La statistique descriptive y est enseigné dès le premier semestre.

Pour accompagner le cours en amphithéâtre ainsi que les travaux dirigés, les étudiants ont un ensemble de TP sur ordinateur leur permettant de pratiquer eux-mêmes ce qu'ils ont vu par ailleurs. Dans ce cadre, nous avons utilisé les années précédentes plusieurs outils du marché. Malheureusement, bien que ceux-ci soient prévus pour réaliser des calculs et graphiques statistiques, il nous est apparu que ce n'était pas pleinement satisfaisant pour différentes raisons.

2 Pourquoi un outil dédié ?

En premier lieu, nous détaillons ici deux raisons de ne pas utiliser un logiciel statistique. Pour la plupart d'entre eux, les informations fournies sont souvent trop complètes. Cet

afflux massif perturbe les étudiants, qui ont du mal à voir ce qui est essentiel et relié au cours de statistique descriptive. Par exemple, il y a des résultats de tests statistiques, avec des p-value. Or, ces notions ne sont réellement abordées qu'en deuxième année.

De plus, la terminologie utilisée dans chaque logiciel est, bien évidemment, différente du cours et pas toujours correcte. Plusieurs outils nécessitent, par exemple, de passer par la création d'un histogramme pour pouvoir réaliser un diagramme en barres. Cette confusion perturbe les étudiants, et reste parfois jusqu'à la fin du diplôme.

Ensuite, nous abordons les raisons de ne pas utiliser un langage statistique. Comme indiqué plus haut, les étudiants concernés sont en première année post-bac. Ceux-ci ont donc, en général, une très faible connaissance de l'outil informatique, et encore moins en programmation. Bien qu'il y ait un cours de programmation en parallèle, ils sont encore très fragiles sur le sujet durant ce premier semestre.

Nous avons aussi remarqué que la complexité, même relative, de la programmation est un frein important à la compréhension des étudiants. Ceux-ci sont particulièrement concentrés sur la partie codage et non sur l'analyse. Ainsi, ils n'en retirent pas les bénéfices escomptés.

Enfin, que ce soit pour un logiciel ou un langage de programmation, il est nécessaire pour eux d'avoir à disposition un ordinateur, sur lequel ils pourront installer l'outil pour un travail à la maison. Même si beaucoup d'entre eux sont équipés à la maison, ce n'est pas forcément le cas pour tous. De plus, certains outils peuvent s'avérer complexes à installer, voire impossible car nous ne disposions pas forcément de licences à domicile pour les étudiants.

Toutes ces raisons nous ont donc amenés à choisir de développer un outil en interne. Le choix d'une application web permet de répondre à la problématique de l'installation d'un outil, celle-ci étant disponible à partir de la plupart des ordinateurs (certains réseaux sécurisés bloquent l'application, problème en cours de résolution actuellement). Il n'y a pas le frein de la programmation pour les étudiants, puisque l'outil est tout en "clic-bouton". En outre, l'idée de la créer nous-même permet de répondre aux critiques sur le trop d'informations et la terminologie.

3 Réalisation et détails

Pour réaliser cette application, nous avons utilisé le langage R [1], en utilisant la librairie Shiny [2], permettant la réalisation d'application web. L'intérêt d'une telle solution technique réside principalement dans la simplicité de programmation. Les aspects design et interface avec l'utilisateur sont eux gérés directement par l'outil. L'outil Shiny Server a permis le déploiement de celle-ci sur un serveur virtuel, hébergé sur une plateforme de serveurs virtuels, basée sur OpenNebula, dans les locaux de l'Université Paris Descartes[1]. Pour la réalisation des graphiques, nous avons choisi la librairie ggplot2 [3].

[1]Lien vers l'application déployée : `http://up5.fr/explore-data`

Afin que l'application puisse être utilisée et éventuellement améliorée, elle est développée sous la licence GPL v3 [4], et le code est disponible en ligne[2]. Elle se base donc principalement sur l'utilisation du clic-bouton, avec une interface épurée et simple pour ne pas perturber les étudiants. Elle est constituée de 5 onglets que nous détaillons ici.

Données L'utilisateur doit choisir le jeu de données qu'il souhaite analyser. Pour cela, il y d'une part des données déjà présentes (issues des librairies base et ggplot2). Il est aussi possible de charger des fichiers textes, en spécifiant quelques options (délimiteurs, noms des variables, séparateurs de décimales).

Variables Cet onglet présente chaque variable, en donnant son nom, son type et le nombre de valeurs distinctes. Il affiche les premières valeurs pour des variables numériques ou chaînes de caractères, et les modalités possibles pour des variables factor.

Sous-population Il est ici possible de faire une sélection d'un sous-ensemble d'individus, basée sur des critères logiques (par exemple, âge inférieur à une valeur spécifique).

Univarié Le but ici est de décrire soit une variable quantitative (toute variable numérique présente), soit une variable qualitative (toute variable sans restriction à l'heure actuelle). Il y a des informations de base (nombre de valeurs, nombre de valeurs manquantes, pourcentage de valeurs manquantes), les statistiques et graphiques usuels, en fonction de la nature de la variable à étudier.

Bivarié Le but ici est de décrire tous les croisements possibles entre deux variables. Seules les variables numériques sont logiquement considérées comme quantitatives, alors que toutes les variables peuvent être choisies comme une variable qualitative. En plus des informations de base (cf onglet Univarié), on dispose des calculs statistiques et graphiques classiques, défini par les types des variables à analyser.

4 Conclusion

Après cette première année d'utilisation, nous pouvons dire que l'expérience a été très fructueuse. Nous pouvons en retenir les points suivants.

En terme d'utilisation, nous n'avons rencontré aucun problème d'accès, ni de surcharge du serveur supportant l'application, ce qui était une crainte en début d'année. Le seul défaut de la situation actuelle est le blocage de l'application par certains pare-feux.

[2]Lien vers le projet : `https://github.com/fxjollois/explore-data`

Ensuite, les étudiants se sont rapidement approprié l'outil, sans éprouver de difficultés lors de son utilisation. Ils n'ont pas eu d'effort à faire pour faire le lien entre le cours et son application en TP. Les enseignants sont également satisfaits par l'outil.

Pour l'année prochaine, nous avons donc décidé de continuer d'utiliser cet outil, en ajoutant quelques fonctionnalités. Par exemple, il y aura la possibilité de créer une variable ordinale, basée sur une variable quantitative. Il devrait aussi y avoir le moyen de fusionner des modalités d'une variable qualitative. Toutes ces améliorations sont en cours de développement. Toutes propositions ou contributions au projet sont les bienvenues.

Bibliographie

[1] R Core Team (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. `https://www.R-project.org/`.
[2] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2017). shiny: Web Application Framework for R. `https://shiny.rstudio.com/`
[3] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. `http://ggplot2.org/`
[4] GNU GENERAL PUBLIC LICENSE, Version 3, 29 June 2007. Free Software Foundation, Inc. `http://fsf.org/`