Concours DATAVIZ : retour d'expérience en 1ère année de DUT STID

François-Xavier JOLLOIS, Florence MURI, Elisabeth OTTENWAELTER[1], Antoine ROLLAND[2], Sylvie VIGUIER[3]

TITLE

Dataviz Challenge

RÉSUMÉ

Les départements Statistique et Informatique Décisionnelle (STID) des IUT de France ont décidé de créer un challenge dédié aux étudiants de première année. Celui-ci a pour contenu la production d'une visualisation de données (dataviz) autour de données identiques pour tous, par groupes et dans le temps restreint d'une journée. Chaque groupe d'étudiants participant doit proposer une dataviz sur la problématique de leur choix. Nous détaillons ici cette expérience réussie et très enrichissante pour tous.

Mots-clés : *concours, data-visualisation, apprendre autrement, DUT, STID, statistique, décisionnel.*

ABSTRACT

The Departments of Statistics and Business Intelligence of French IUT (Technological University Institute) have decided to create a challenge dedicated to first-year students. This challenge is dedicated to the production of a data visualization (dataviz) with the same data used by all participants, formed in teams and in the short period of one day. Each team must propose a dataviz in connection with a specific problem they have chosen. We detail here this successful and very enriching experience for all.

Keywords: *challenge, dataviz, DUT, STID, statistics, business intelligence.*

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[4] et de la statistique. Dans ce cadre, il existe un enseignement dédié à la visualisation de données (ou data-visualisation, encore dit dataviz). Même si l'usage des graphiques en statistique, et particulièrement en statistique descriptive, est très ancien, l'approche théorique de la construction d'un graphique adapté à la présentation ou à l'exploration de données statistiques est un phénomène relativement nouveau.

[1]Département STID – IUT Paris Descartes, {prenom.nom}@parisdescartes.fr
[2]Département STID – IUT Lumière Lyon II, antoine.rolland@univ-lyon2.fr
[3]IUT de Perpignan, département STID à Carcassonne, viguier@univ-perp.fr

[4]D'après Wikipédia, « [Le décisionnel] désigne les moyens, les outils et les méthodes qui permettent de collecter, consolider, modéliser et restituer les données, matérielles ou immatérielles, d'une entreprise en vue d'offrir une aide à la décision et de permettre à un décideur d'avoir une vue d'ensemble de l'activité traitée. »

C'est à l'établissement d'un corpus théorique de la représentation des données que vise la data-visualisation. On consultera par exemple [1] pour une première approche historique de la data-visualisation. L'enseignement de data-visualisation en DUT STID se place au deuxième semestre de la première année, dans laquelle les enseignements intègrent, en amont, la statistique descriptive, la programmation, le reporting, la manipulation de données et la communication. C'est donc une suite logique dans le déroulement de l'année. Le contenu de ce cours intègre principalement des éléments de communication. Les compétences statistiques nécessaires à la réalisation de telles productions ont été vues avant, principalement dans le premier semestre. De même, les étudiants ont appris à programmer et à manipuler des données, ce qui leur donne la possibilité de jouer avec les données comme bon leur semble.

Afin de valoriser cet enseignement, les départements de Paris et de Carcassonne ont décidé en mai 2016 de réaliser un concours de data-visualisation, dédié à ces étudiants de première année. Ces deux concours ont été réalisés de manière disjointe, mais sur les mêmes données. Ce premier exercice a donné entière satisfaction, tant dans l'engagement des étudiants que dans les bénéfices retirés de ce concours. Suite à ces deux expériences réussies, l'ensemble des départements STID a souhaité élargir le concours aux départements intéressés en mai 2017 ; de fait, 10 départements (sur 12) ont pu se joindre au concours dans cette première édition nationale.

Le déroulement de ce concours, pour les deux éditions, est le même :

1. Les données sont présentées aux étudiants.
2. Les étudiants, répartis en groupes de 3 à 5 étudiants, ont la journée pour produire une data-visualisation :
   - ils sont libres de choisir la façon de les exploiter,
   - ils sont plus ou moins encadrés.
3. En fin de journée, ils présentent leur production devant un jury.

Nous présenterons dans la suite les objectifs attendus d'un tel événement, et nous détaillerons le déroulement du concours de ces deux années. Nous terminerons par un bilan et des perspectives pour les éditions futures.

2 Objectifs et attendus

Un tel exercice permet de viser des objectifs tant pour les étudiants, que pour les enseignants et même pour la formation STID en général. C'est ce que nous allons présenter dans la suite.

2.1 Du point de vue des étudiants

Le premier objectif est bien évidemment de tester une partie des compétences acquises par les étudiants tout au long de leur première année de DUT, et spécialement les compétences

liées à la manipulation et la description statistique des données, mais aussi leur aptitude à choisir un angle de communication sur un sujet précis, et évidemment leurs compétences techniques de production de graphiques clairs et adaptés aux données et à la question posée. Ils ont ainsi l'occasion de mettre en œuvre les différents enseignements qu'ils ont reçus et ainsi mieux comprendre les interactions entre chacun. C'est une façon intéressante pour eux de voir que chaque brique ainsi acquise est une partie d'un tout.

Ensuite, ce concours est aussi l'occasion de confronter les étudiants, encore jeunes, à des problématiques métiers. Ils ont ainsi l'occasion de voir un aperçu du travail en entreprise, alors que la plupart n'ont effectué que le stage de découverte d'une semaine pendant le collège.

Un autre aspect intéressant de cet exercice est la mise en situation des étudiants autour d'un travail en équipe, qui nécessite de réelles interactions entre eux pour que le travail avance. Ils ont eu jusque là l'habitude de réaliser leurs travaux pratiques seuls ou en binôme. C'est un exercice compliqué, surtout sur un temps court et limité comme c'est le cas ici.

Enfin, l'expérience nouvelle pour les étudiants de travailler dans des conditions différentes de celles des enseignements plus classiques est très bien perçue. La nécessité de se prendre en charge collectivement pour la gestion du temps, pour la procédure d'organisation et le mode de production, l'ambiance de défi et de coopération avec un accompagnement bienveillant leur ont laissé un excellent souvenir.

2.2 Du point de vue des enseignants

L'exercice en question doit permettre un réel travail collaboratif à l'intérieur des départements. Afin de préparer au mieux les étudiants, et dans le cadre des enseignements de data-visualisation prévus dans le programme pédagogique national, les enseignants de statistique, d'informatique et de communication doivent se concerter et dialoguer.

Confrontés aux données en même temps que les étudiants, les enseignants peuvent conseiller, suggérer des pistes, faire des retours qualitatifs sur le travail des étudiants si ceux-ci sont demandeurs, mais ne sont pas là pour guider les étudiants vers une solution existante au préalable. Par ailleurs, et suivant leurs capacités, les enseignants peuvent aussi servir de personnes-ressource sur l'utilisation pratique des logiciels si besoin. Enfin, c'est l'occasion de discuter avec les étudiants différemment, et ainsi d'établir un lien avec ceux-ci autrement que dans le cadre classique d'un cours.

2.3 Du point de vue de la formation

Les 12 départements STID de France dialoguent et agissent ensemble à la promotion de la spécialité à travers l'association STID France[5]. Malgré un environnement général porteur et une réelle dynamique autour de l'économie de la donnée[6], le DUT STID souffre

[5]Site de STID France : `http://www.stid-france.fr`

[6]Voir, par exemple, le rapport franco-britannique sur l'économie de la donnée publié sur le site du ministère de l'économie et des finances le 13 juillet 2016 [2].

de manière générale d'un déficit de notoriété auprès des entreprises du domaine ainsi qu'auprès des bacheliers et futurs bacheliers. L'organisation d'un concours national est ainsi l'occasion de montrer d'une part aux entreprises les compétences de nos étudiants et d'autre part au grand public la vitalité de la formation. Cela nécessite néanmoins un travail particulier de communication, pour l'instant limité à l'environnement immédiat des départements concernés d'une part, et à la communauté statistique d'autre part à travers des retours d'expériences présentés lors des Journées de la Statistique 2017 à Avignon ou au CFIES[7] 2017 à Grenoble.

L'organisation d'une manifestation commune permet aussi à une communauté relativement petite et éclatée de développer un sentiment d'appartenance et ainsi de dresser des pistes de collaboration plus régulières, en parallèle des Journées Pédagogiques organisées tous les deux ans.

3 Déroulement

3.1 Édition 2016

L'idée d'organiser un concours de data-visualisation en première année de STID a germé début 2016. Cependant, les contraintes d'emplois du temps et de coordination entre les différents départements ont fait que seuls les départements de Paris et de Carcassonne ont pu mettre en place un concours interne à chaque département.

3.1.1 Données

Les données utilisées concernaient les logements mis en location sur le site *AirBnB* le 2 septembre 2015, et mises à disposition par le site *InsideAirBnB* [3]. Un fichier était ainsi fourni, avec les informations suivantes :

- Résumé du logement : quartier, géolocalisation, type et prix du logement ;
- Détail du logement : type et nombre de lits, prix détaillés, présence ou non de divers appareils électroménagers ou autres équipements, acceptation ou non d'invités, idem pour les animaux, ...

3.1.2 Déroulement

Le concours s'est déroulé le 12 mai 2016 pour le département de Paris ; 9 équipes de 3 étudiants, tous volontaires, ont planché de 9h à 17h sur les données. A la fin de la journée ils ont présenté leurs travaux durant 5 minutes, devant un jury composé d'enseignants et de professionnels de la data-visualisation.

[7] CFIES : Colloque Francophone International sur l'Enseignement de la Statistique.

Pour le département de Carcassonne, les étudiants de première année ont eu une journée mobilisée, le 17 mai, pour plancher sur le même sujet. Ce travail était obligatoire, suivi par une évaluation de la production mais sans passage devant un jury.

3.1.3 Consignes

Après une brève présentation du site AirBnB et des données, il a été demandé aux étudiants de rendre en fin de journée (17h) une data-visualisation sur une problématique que chaque groupe devait définir. Ils ont dans un premier temps eu à explorer les données, pour comprendre chaque variable et les regrouper éventuellement par thème. Une fois cette première partie effectuée, ils avaient à choisir l'axe d'analyse (tarif, localisation, équipement, ...) et la façon de représenter les informations.

3.1.4 Encadrement

Le jeu de données a été proposé par François-Xavier Jollois (Paris). Il a été choisi pour la thématique mobilisante pour les étudiants et pour la simplicité des données attaquables sans manipulation préalable.

Les étudiants ont été accompagnés tout au long de la journée par deux « enseignants-référents », spécialistes en statistique et/ou en communication, qui ont pu répondre à leurs questions, éventuellement lever des points techniques bloquants, mais surtout les amener à améliorer le document produit en servant de vis-à-vis.

A Paris, le jury était composé d'enseignants du département, de la responsable de la communication de l'IUT, de deux dirigeants d'une entreprise spécialisée dans la data-visualisation en temps réel (IDCWare[8]), et d'un dirigeant d'une start-up hébergée à l'incubateur de l'IUT de Paris.

3.1.5 Résultats

Les étudiants ont réussi à produire des visualisations vraiment intéressantes, avec des choix de problématiques pertinents et assumés. Ils se sont pris totalement au jeu et ont grandement apprécié cette journée.

Pour Paris, en plus de lots plus communs, les 3 équipes gagnantes ont pu faire une présentation de leur data-visualisation, dans les locaux de IDCware, entreprise spécialisée dans la visualisation de données temps réel et présente dans le jury, durant un séminaire auquel ont participé d'autres entreprises du domaine et leurs clients.

A Carcassonne, l'ambiance collaborative et positive qui a été ressentie par les étudiants et l'équipe pédagogique a conforté la conviction qu'un tel évènement apportait une dimension toute nouvelle pour la formation, avec une affirmation de la conscience d'une appartenance à un groupe plus large que le seul département local. Il était évident qu'une telle expérience devait se développer.

[8]Site d'IDCWare : `http://www.idcware.com`

3.2 Édition 2017

3.2.1 Partenariat

Les départements STID et la SNCF, via le département *Innovation & Recherche* [4], ont décidé d'un partenariat pour l'édition 2017 du concours de data-visualisation. Ainsi, la SNCF a fourni les données (présentées ci-dessous). Des coachs, provenant des équipes statistiques et/ou données de la SNCF, sont venus dans presque tous les départements. Ceux-ci avaient pour mission d'aider les étudiants à comprendre les données et à mieux les appréhender.

3.2.2 Données

Le jeu de données représentait la série des meilleurs temps de parcours d'environ 40 trajets Origine → Destination depuis 1920. De plus, ces données ont été enrichies avec d'autres sources ouvertes pour ajouter les informations suivantes, pour chaque année et, le cas échéant, chaque origine et destination :

- population, taux de chômage, nombre d'emplois ;
- occupation des sols ;
- indices des prix à Paris ;
- fréquentations des principaux aéroports ;
- nombre d'immatriculations et livraison de carburant.

Ces sources de données sont disponibles à des niveaux géographiques et des périmètres temporels qui peuvent être différents. Ainsi, quand une information est absente, le fichier a été complété à partir de la dernière donnée disponible.

3.2.3 Déroulement

Le jeudi 18 mai 2017 était la journée choisie pour que la quasi-totalité des départements STID participe à ce challenge. Plus de 300 étudiants répartis sur toute la France, regroupés en équipes de 3 à 5, ont donc planché sur le sujet de 9h à 17h, après une brève présentation des données vers 8h30. Dans chaque site, les groupes ont présenté ensuite leur production devant un jury, propre à chaque site, composé d'enseignants et du coach SNCF présent. Chaque jury a choisi un travail devant concourir pour la finale nationale.

Les neuf data-visualisations sélectionnées (deux départements étaient sur le même site) ont ainsi été notées une seconde fois par un jury composé d'enseignants de chaque département participant au challenge. Celui-ci a dégagé un groupe de 3 productions (venant de Niort, Vannes et Paris), qui ont participé à une finale nationale, où les travaux sont évalués par SNCF. A l'issue du dernier jury, le groupe de l'IUT de **Niort** a été désigné vainqueur.

3.2.4 Encadrement

Les données ont été choisies par la SNCF avec la même exigence de simplicité pour permettre aux étudiants de se concentrer sur la problématique sans perdre de temps sur la manipulation des données.

Les étudiants ont été accompagnés tout au long de la journée par plusieurs personnes (dans la majorité des sites) :

- un « coach » de la SNCF, membre de services « statistiques » et/ou « données » de la SNCF, venant du pôle national ou des pôles régionaux. Ces « coachs » ont présenté le jeu de données, indiqué des pistes de réflexion sur une approche métier, mais aussi ont pu guider les étudiants vers telle ou telle représentation suivant les sollicitations ;
- des enseignants spécialistes des statistiques descriptives et/ou de communication qui ont pu accompagner les étudiants d'un point de vue méthodologique et technique, là aussi en répondant aux sollicitations. Le projet est bien le projet des étudiants ; cependant, il est de peu d'intérêt de laisser une équipe tourner en rond si on peut l'aider à avancer.

La composition des jurys au sein des départements était variable suivant les départements, mais incluait a minima le référent SNCF et les enseignants ayant suivi la journée. Le jury final était composé uniquement de personnes du service statistique de la SNCF.

3.2.5 Consignes

Les consignes ont été divulguées en même temps que les données. Chaque groupe devait rendre, à 17h, une data-visualisation portant sur une problématique définie par le groupe. Il s'agissait donc dans un premier temps d'explorer les données pour faire ressortir une idée intéressante à présenter (par exemple l'évolution du temps de trajet depuis 1920 par région de destination, ou l'impact de l'arrivée du TGV sur l'évolution du temps de trajet...). Dans un deuxième temps, le groupe a dû trouver la meilleure représentation visuelle pour montrer le résultat souhaité. Enfin, tant le logiciel utilisé que la forme de la restitution étaient libres. Chaque groupe était néanmoins invité à fournir un léger texte accompagnant la data-visualisation afin d'expliquer la démarche et la question posée pour les personnes absentes de la restitution orale.

3.2.6 Résultats

Les data-visualisations proposées par les étudiants ont été de qualités inégales, et dans l'ensemble relativement décevantes. Nous pouvons pointer quelques points d'attention sur ces travaux.

Difficulté à choisir un angle de vue. La plupart des étudiants ont été décontenancés par les données et ont eu du mal à dégager un sujet particulier à mettre en avant. La

quantité de données (nombre de variables) et l'absence de lien évident entre toutes ces données a par exemple conduit beaucoup de groupes à proposer une explication exogène à la SNCF de la baisse du temps de transport (par le taux de chômage ou autre par exemple), confondant ainsi corrélation et causalité. Au final, les meilleures data-visualisations sont celles qui se sont concentrées sur les données SNCF pures, mettant en avant simplement la baisse du temps de transport. Les représentants de la SNCF ont cependant noté des problématiques intéressantes choisies par les étudiants, telles que les circonstances du développement du réseau ferré, le lien entre l'amélioration des transports en commun et le développement des villes, l'impact de la prise en compte de l'environnement sur les trajets ferroviaires, ...

Data-visualisation et statistiques descriptives. Beaucoup de groupes, une fois leur problématique établie, se sont contentés de produire des statistiques descriptives uni ou bivariées, présentées de manière très classique par des diagrammes en bâtons, diagrammes circulaires ou nuages de points lissés, sans chercher à faire preuve de recherche sur la meilleure manière de mettre en avant les phénomènes marquants dans les données. Les data-visualisations primées au niveau national se distinguent justement par l'accent mis sur l'adéquation de la représentation graphique à la question posée.

Choix du logiciel. Même si le choix du logiciel était libre, la qualité des documents produits grâce à Tableau ©[9], en particulier grâce au côté interactif, est largement meilleure que ce qui a été produit simplement avec Excel. Cependant, utiliser le logiciel Tableau n'est une condition ni nécessaire ni suffisante pour être finaliste. L'utilisation d'un logiciel performant peut certes améliorer une bonne data-visualisation, mais ne viendra combler ni une absence de réflexion sur la problématique ni un mauvais traitement des données.

4 Conclusion et futur

Reprenons les objectifs listés en début d'article.

Tester les compétences que les étudiants ont acquises tout au long de leur première année de DUT. Cet objectif a été atteint, mais de manière générale les enseignants ont été plutôt déçus des travaux proposés par les étudiants. Le passage du format de deux départements volontaires à l'ensemble des départements, et pour la plupart d'entre eux à l'ensemble des étudiants, s'est accompagné d'une baisse de la qualité des data-visualisations produites. On ne peut cependant mettre de côté les difficultés liées à la compréhension du sujet et à l'établissement d'une problématique. Une édition supplémentaire est donc nécessaire avant de statuer sur l'influence du volontariat sur la qualité des résultats.

[9]Site du logiciel Tableau © : `https://www.tableau.com`, avec une vidéo de démonstration à cette adresse `https://www.tableau.com/fr-fr#hero-video`

Confronter les étudiants à des problématiques métiers. Ce point a été atteint, sans que cela soit un réel succès comme noté précédemment. La masse des données proposées a été préjudiciable à une analyse des étudiants en terme « métier » et a favorisé, par la diversité des indicateurs socio-économiques mis à disposition, une vue plus externe, plutôt qu'une identification aux services de la SNCF.

Mise en situation des étudiants autour d'un travail en équipe. Ce point est indéniablement une réussite, par la contrainte systématique dans les départements de réunir les étudiants par 3, 4 ou 5.

Concertation des enseignants de différentes disciplines. Ce point reste à approfondir. L'organisation de cette première édition nationale a reposé sur la désignation d'un enseignant référent par département. Celui-ci n'a généralement pas souhaité ou réussi à mobiliser d'autres collègues sur le sujet. La connaissance relativement tardive du sujet n'a pas non plus facilité un travail préparatoire en équipe enseignante. Pour les départements où l'équipe enseignante s'est mobilisée en nombre, ce fut un moment très porteur par les échanges des points de vue quant aux manières de porter des appréciations aux étudiants.

Interagir avec les étudiants de manière différente que dans le cadre d'un cours. De facto, cet objectif a été atteint pour les enseignants référents.

Communication de la formation STID. Ce point est globalement décevant, de par l'absence de plan média concerté. La production d'un communiqué de presse à disposition des départements en amont de l'événement, ainsi que la désignation d'un *community manager* pour STID France apte à animer la journée sur les réseaux sociaux est impérative pour les prochaines éditions.

Travail en commun aux départements STID en France. Ce fut un premier pas. La coordination des emplois du temps de 10 départements a nécessité une anticipation de plusieurs mois. La définition conjointe du sujet et le processus de jury est encore à améliorer. Il appartient cependant maintenant aux enseignants de data-visualisation des différents départements de rebondir sur ce concours pour échanger sur leurs pratiques. Enfin, ce travail en commun sur des lieux très distants a pu se faire grâce à la volonté d'une équipe inter-départements très engagée et ouverte aux nouvelles formes d'enseignement. Il a permis aux étudiants participants de ressentir à leur niveau cette appartenance au réseau STID. Au-delà du concours, ils se sentent plus concernés par tout ce qui touche à cette formation, se renseignant plus volontiers sur les pratiques des différents départements. On peut penser que cela initiera d'autres échanges, ainsi qu'une plus grande mobilité des étudiants.

Références

[1] Michael Friendly, "A Brief History of Data Visualization", dans *Handbook of Data Visualization*, Springer, 2008 (DOI 10.1007/978-3-540-33037-0_2), p. 19

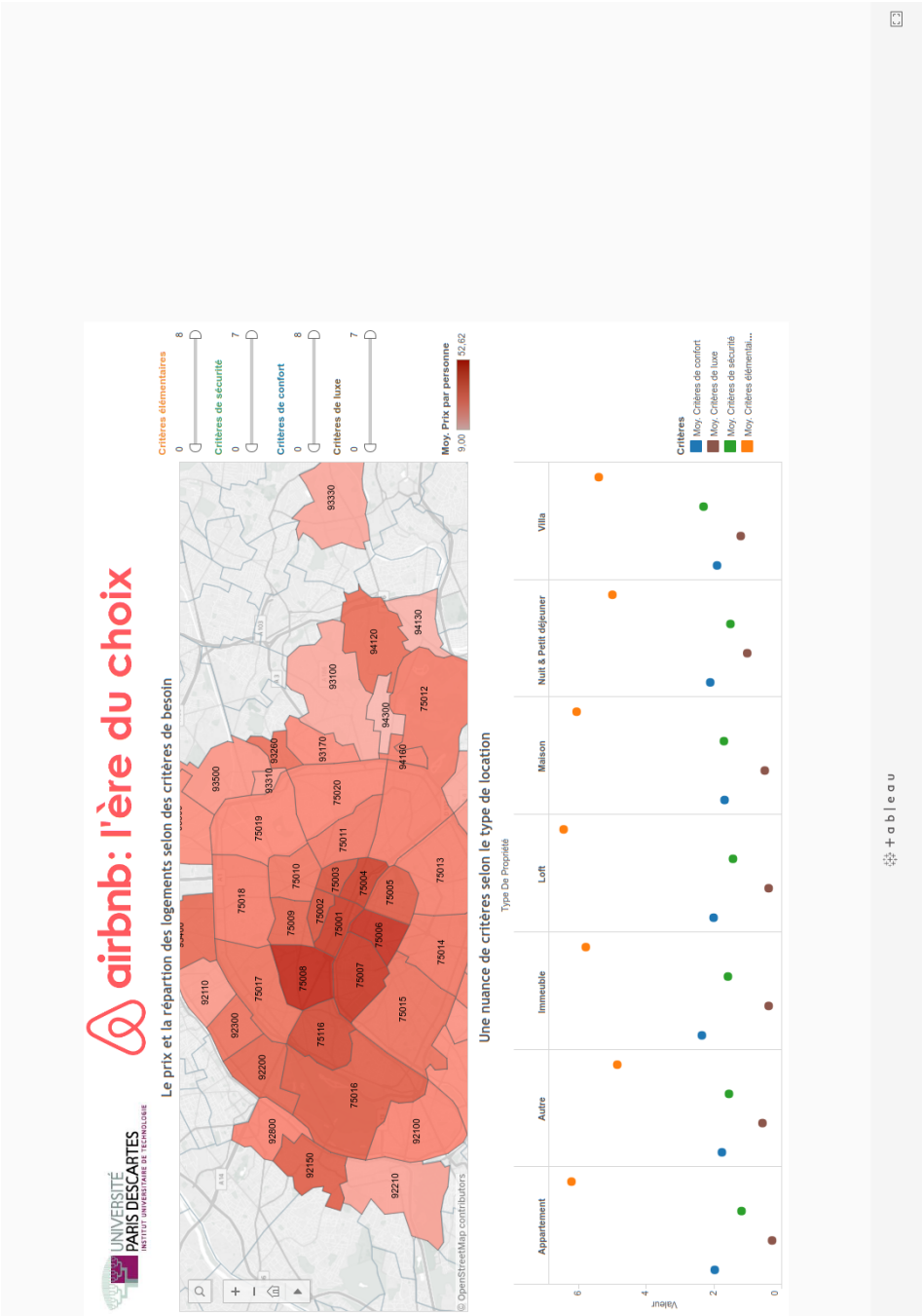
[2] Nigel Shadboldt et Henri Verdier, *La révolution de la donnée au service de la croissance. Innovation, Infrastructure, Compétences et « Pouvoir d'agir » à l'ère numérique*, Ministère de l'économie et des finances, juillet 2016. http://www.anrt.asso.fr/fr/la-revolution-de-la-donnee-au-service-de-la-croissance-19635

[3] Inside AirBnB, http://insideairbnb.com/.

[4] SNCF, Innovation & Recherche, http://www.sncf.com/fr/innovation-recherche

Annexe

Lauréat 2016



Concours Dataviz

Lauréat 2017