UNIVERSITÉ DE NANTES

Open Data

La réussite scolaire

Par Samuel DEFOSSE et Simon THIVET

1. Introduction

Dans le cadre du module de données sur le Web, nous devions réaliser un exposé sur un thème choisi en s'appuyant sur l'Open data. La finalité de ce projet est d'exposer une réflexion scientifique sur ce thème avec les données disponibles.

Le choix du sujet est donc très important car, selon le thème choisi, il y a plus ou moins de jeux de données disponibles sur le web. Une réflexion approfondie doit être réalisée afin de sélectionner une problématique intéressante sur un thème entouré de nombreuses sources de données pertinentes.

Dans un premier temps, nous avions choisi un sujet sur l'évolution urbaine en fonction des zones inondables en France. Cependant, ce sujet nécessitait une mise en évidence trop complexe et les jeux de données disponibles n'étaient pas assez étoffés.

Dans un second temps, nous avons choisi le thème de l'éducation nationale, plus particulièrement la réussite scolaire, qui sera illustrée ici avec le baccalauréat, selon différents facteurs. Nous allons donc croiser plusieurs jeux de données en lien avec ce thème afin d'étoffer et de montrer une notre étude analytique.

Ce compte rendu met en exergue notre réflexion et notre processus d'analyse sur les différents jeux de données que nous avons sélectionné. Tout d'abord, nous détaillerons nos jeux de données, puis, dans un second temps, nous les étudierons. Enfin, dans un troisième temps, nous présenterons notre maquette qui permet d'exploiter et d'analyser les données.

UNIVERSITÉ DE NANTES

2. Nos jeux de données

A. Description

La première étape fut de trouver différents jeux de données sur le web en rapport avec notre sujet. Nous avons donc effectué nos recherches via le moteur de recherche Google en utilisant les mots clé suivants :

- taux de réussite
- budget
- diplôme
- participation
- baccalauréat
- origine sociale
- secteur privé et public

Ces différentes recherches nous ont permis de rassembler plusieurs jeux de données exploitables.

Nos jeux de données restent relativement simples et sont à l'échelle nationale et départementale. Les données que nous avons trouvé sont présentés sous différentes formes (Graphe, tableau, carte, cartographie) et sous différents formats (CSV, JSON, Excel)

B. Rappel sur l'Open Data

L'open data regroupe des données numériques accessibles par tous et dont l'usage est libre de droits. Elles peuvent provenir d'un domaine public ou privé (exemples : Entreprise privées, collectivité, service public, région ...). Ces données sont distribuées de manière très structurée ou non et doivent être exploitables via au moins une licence ouverte. La réutilisation de celles-ci ne sont soumises à aucune restriction technique, juridique ou financière. La totalité des données que nous avons sélectionnées respectent ces conditions.

C. Qualité des données

L'Open Data permet donc une forte accessibilité à un grand nombre de données de toutes sortes, cependant, il faut tout de même rester vigilant sur la cohérence et la pertinence de celles-ci. Il faut garder un oeil critique envers les données que nous utilisons. Toutes les données que nous avons utilisées étaient structurées, et la majorité étaient disponibles avec traitement automatisé intégré sur le site. Néanmoins, certains de nos jeux de données manquent de documentation, ce qui rend la compréhension des données difficiles. Nous pouvons ajouter aussi que la totalité de nos jeux sont statiques et ne sont pas mis à jour régulièrement.

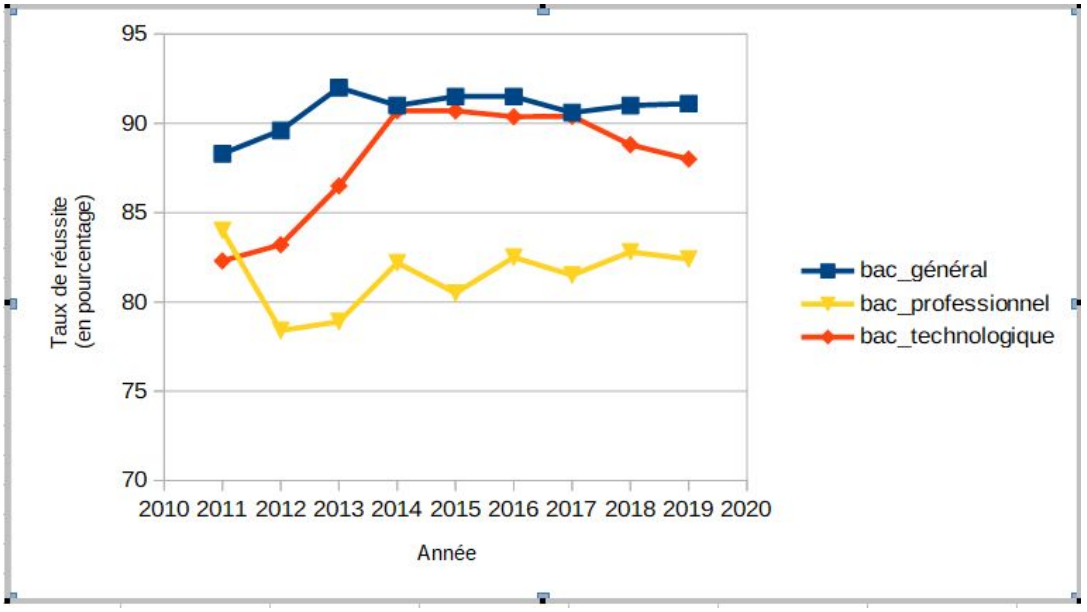
3.Analyse des données

Pour notre analyse de données sur la réussite scolaire, nous avons décidé de nous baser uniquement sur l'obtention du baccalauréat. Notre étude se porte à l'échelle nationale ainsi que régionale et sur plusieurs critères. Dans un premier temps nous allons analyser le taux de réussite par voie (général, technologique et professionnel), ensuite le taux de réussite des étudiants selon leurs origines sociales, et pour finir le taux de réussite selon le secteur des établissements (privé ou public).

A. Des disparités en fonction des voies

Pour cette étude nous avons travaillé sur des données à l'échelle nationale entre 2011 et 2019 en fonction des voies. Les données récupérées contiennent le taux de réussite en pourcentage par année et cela pour les 3 voies du baccalauréat : générale, technologique et professionnelle.

Nous avons croisé les 3 jeux de données afin d'obtenir le graphique suivant qui permet de comparer le taux de réussite de chaque voie, ainsi que son évolution, au fil des années. Nous avons aussi décidé de calculer la moyenne du taux d'admission sur cette période pour chaque voie.



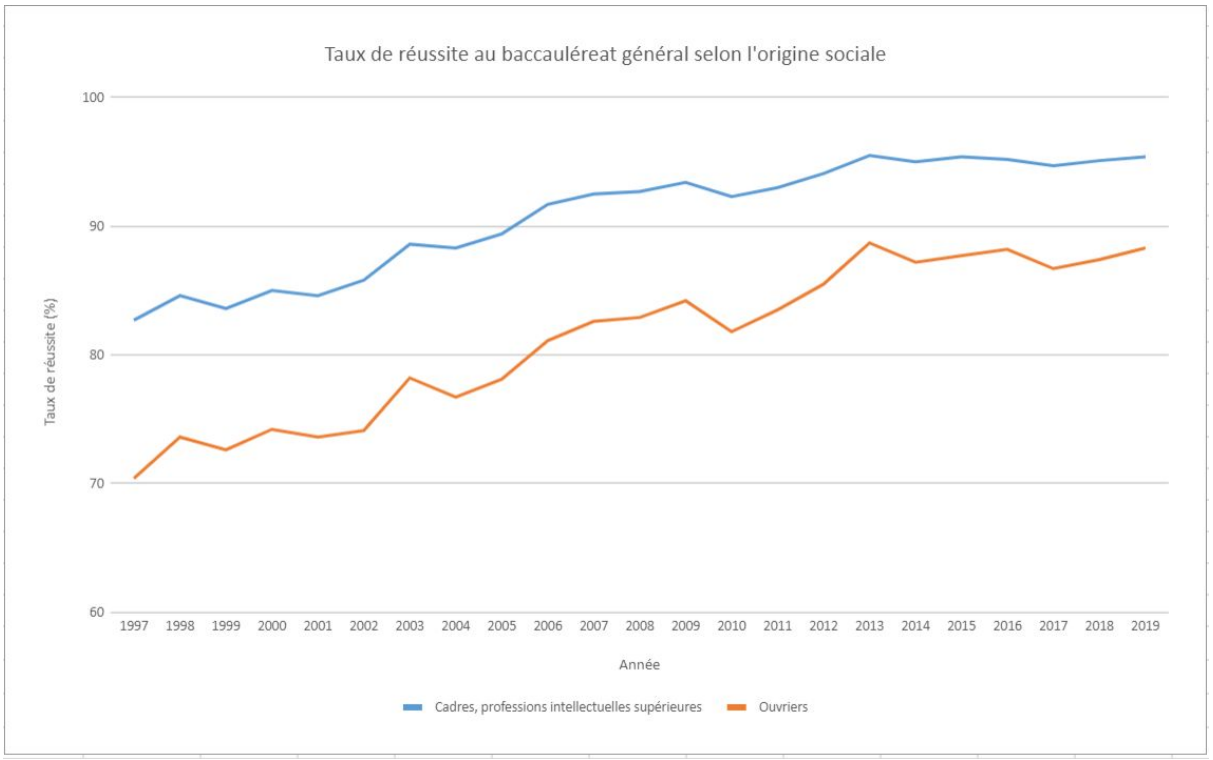
Voie	Moyenne taux d'admissions (%) de 2011 à 2019
Général	90,7
Technologique	87,8
Professionnel	81,5

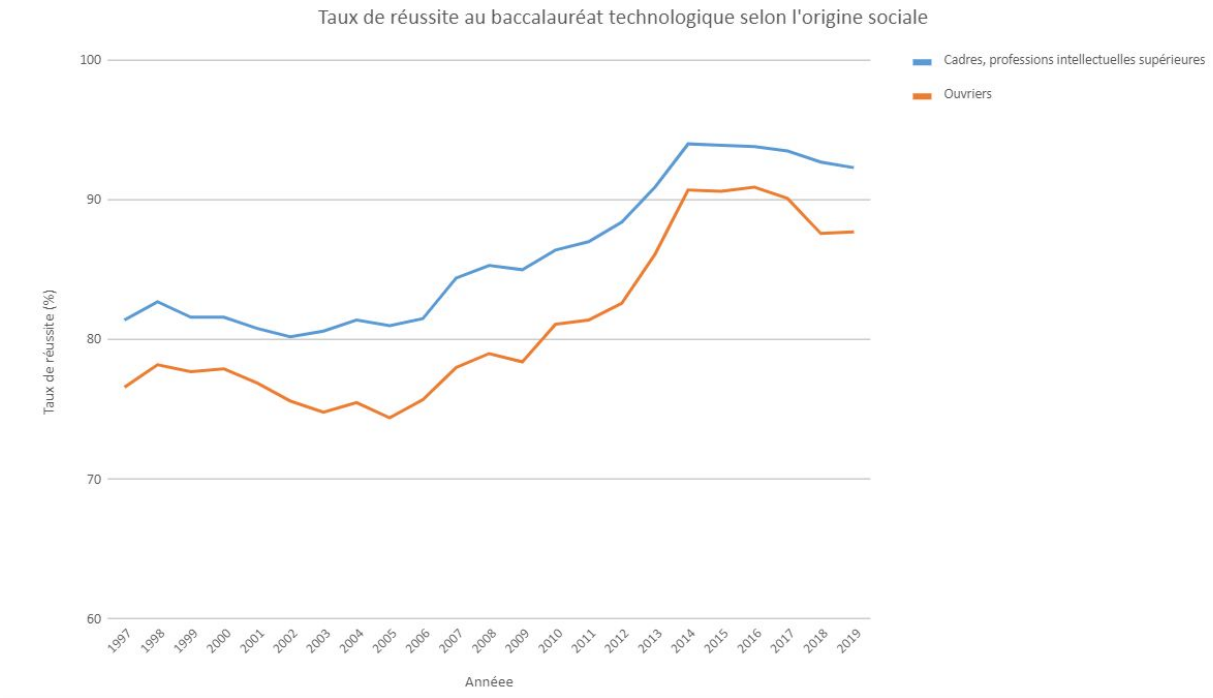
Nous constatons qu'à partir de 2014, le pourcentage de réussite des différentes séries varient relativement peu au fil des années. Nous remarquons aussi de grandes disparités : Le bac général obtient le meilleur taux, suivi du bac technologique et pour finir, le bac professionnel est nettement en dessous.
Ces disparités peuvent être expliquées par différents facteurs que nous allons traiter par la suite.

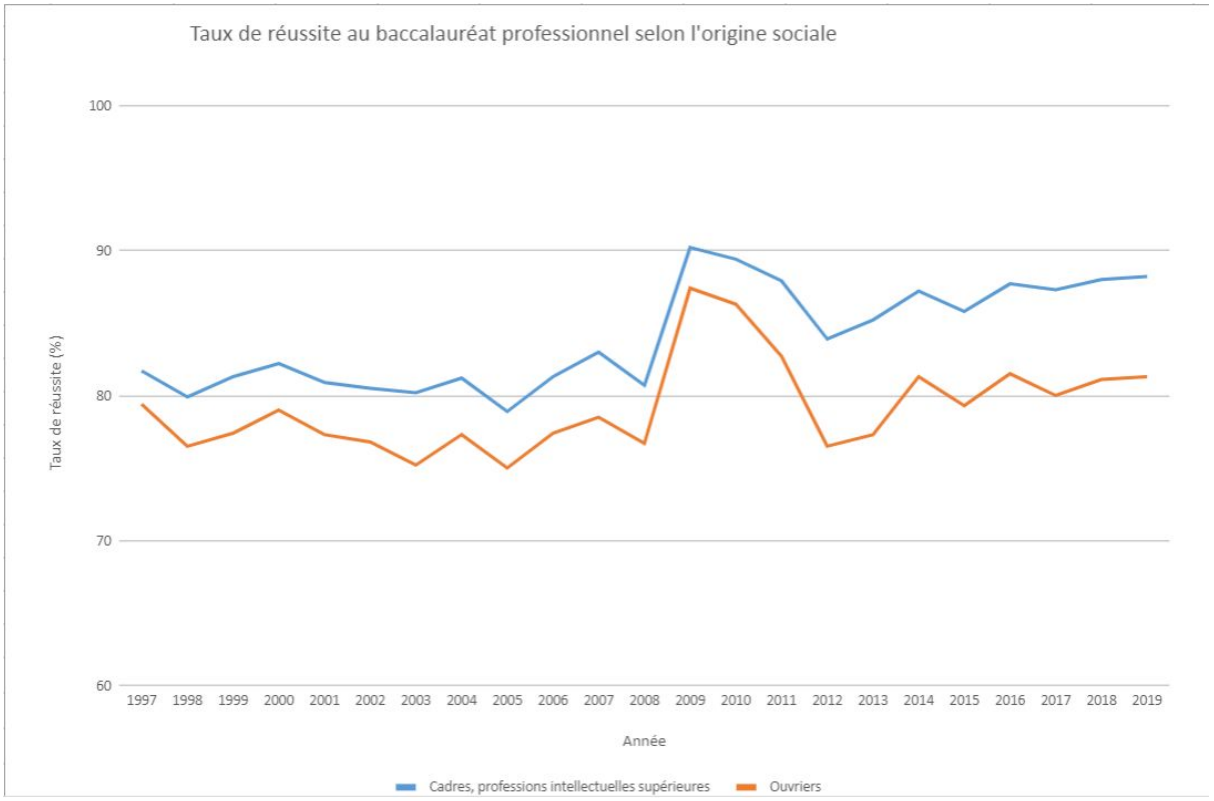
B. Le facteur social

Nous avons travaillé avec un jeu de données conséquent distribué au format CSV,JSON et Exce, qui est mis à jour chaque année (dernière mise à jour le 09/10/2020).

Nous avons décidé de comparer le taux de réussite au bac de la classe ouvrière et de la classe cadre (professions intellectuelles supérieures). Le choix de comparer uniquement ces deux classes sociales nous semble pertinent car, en analysant les données, nous avons remarqué que les données concernant ces classes reflètent bien la situation réelle. De plus, notre étude repose sur les données de 2011 à 2019 pour coller au contexte de notre précédente étude.

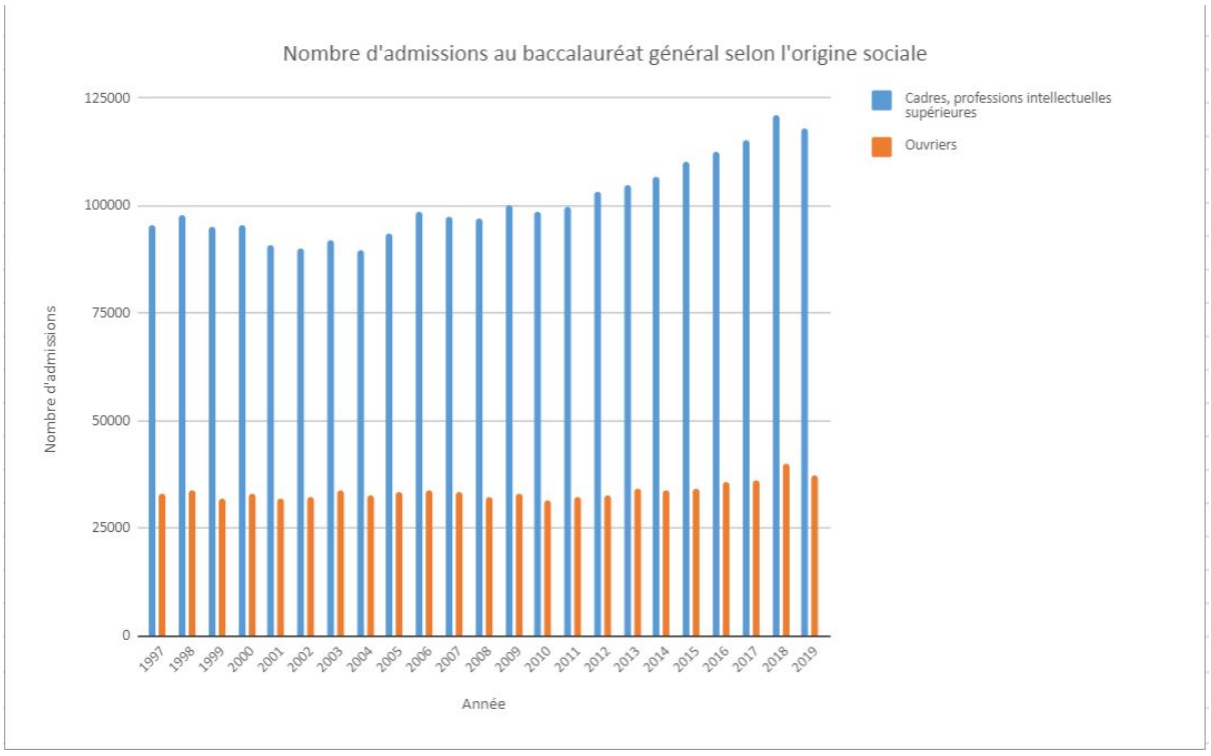

Taux de réussite au baccauléreat général selon l'origine sociale
100
90
80
70
60
Taux de réussite (%)
1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
Année
Cadres, professions intellectuelles supérieures
Ouvriers

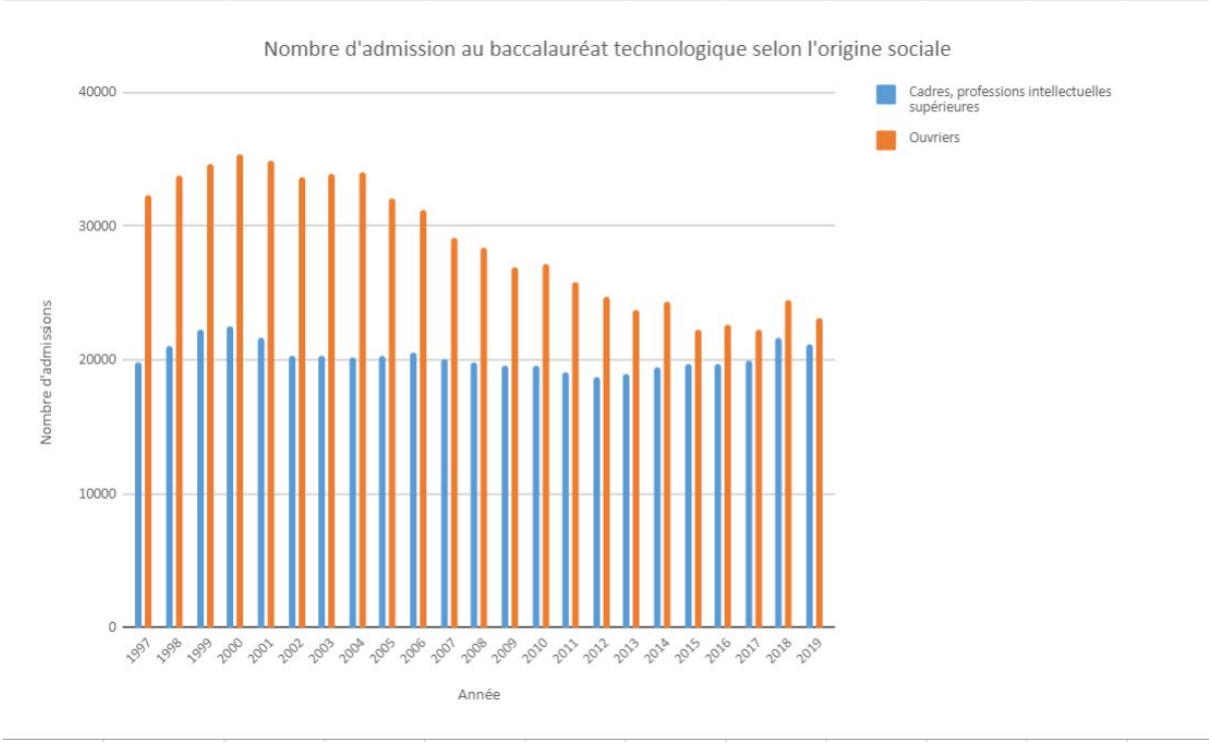

Taux de réussite au baccalauréat technologique selon l'origine sociale
100
90
80
70
60
Taux de réussite (%)
1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
Annéee
Cadres, professions intellectuelles supérieures
Ouvriers

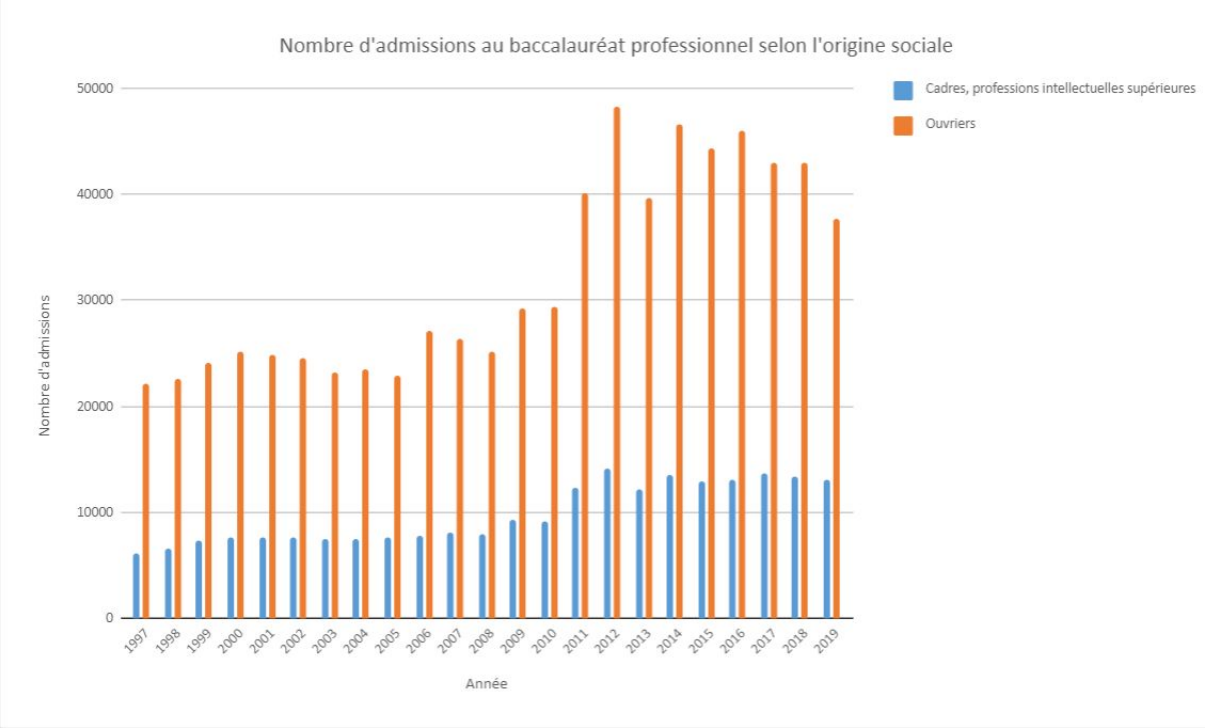


La première analyse que nous pouvons effectuer à travers ce croisement de données est que quelque soit la voie du baccalauréat, les candidats venant de la classe profession intellectuelle ont un meilleur taux d'admission que ceux venant de la classe ouvrière.On constate donc que le milieu dans lequel évolue le candidat est un des facteurs de réussite au baccalauréat et cela pour n'importe quelle voie.

Pour pousser notre analyse plus loin et voir si cela impactait sur notre première analyse sur le taux de réussite par voie tout milieu confondu, nous avons décidé d'exploiter aussi le nombre d'admis et de comparer les chiffres via les graphiques en colonne suivants.


Nombre d'admissions au baccalauréat général selon l'origine sociale
Cadres, professions intellectuelles supérieures
Ouvriers
Nombre d'admissions
125000
100000
75000
50000
25000
0
1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
Année


Nombre d'admission au baccalauréat technologique selon l'origine sociale
Cadres, professions intellectuelles supérieures
Ouvriers
Nombre d'admissions
40000
30000
20000
10000
0
1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
Année



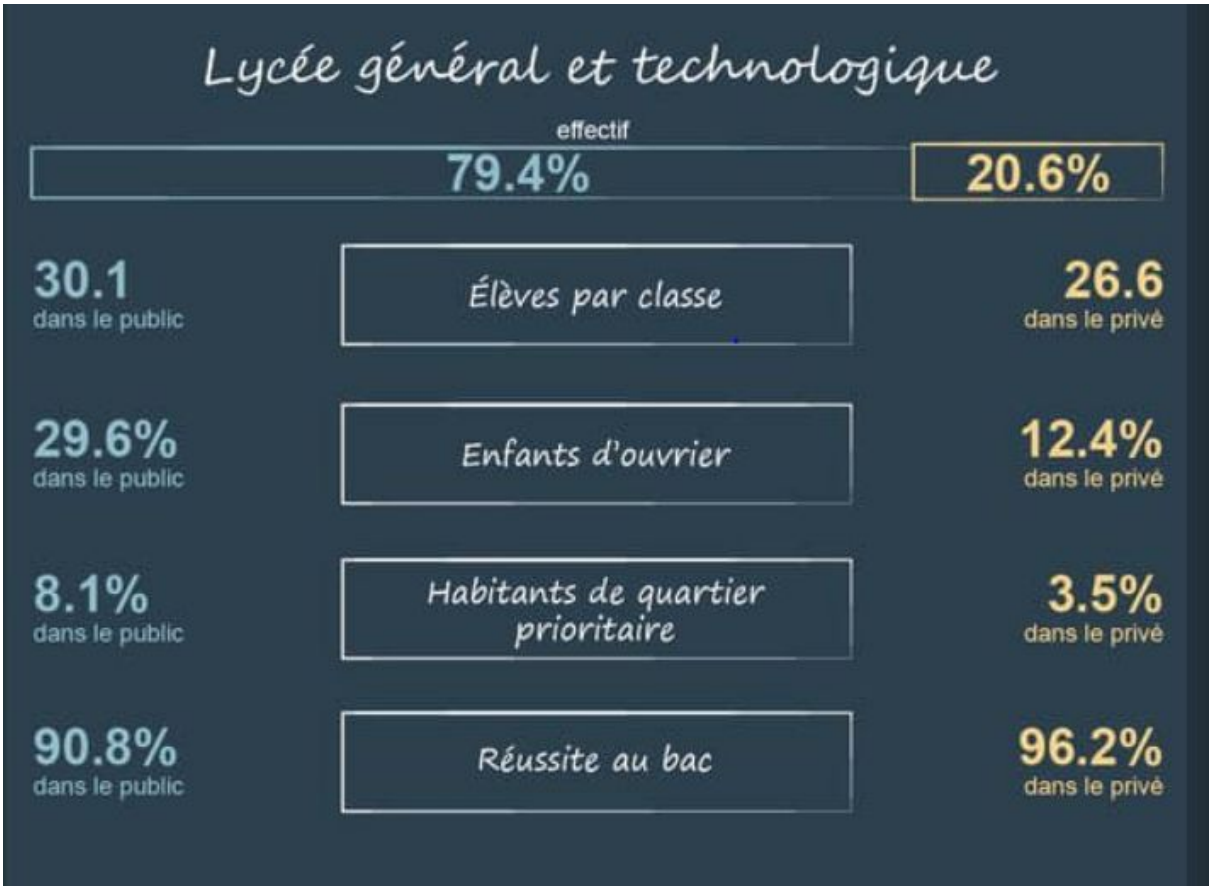
On observe en premier lieu que pour un baccalauréat général il y a beaucoup plus de candidats venant de la classe profession intellectuelle et inversement pour le baccalauréat pro, tandis que pour le baccalauréat technologique la répartition est plus équitable avec une légère plus grande proportion de candidats venant de la classe ouvrière.
Cela corrèle avec notre analyse sur la disparité en fonction des voies, en effet le baccalauréat général comprend un plus grand nombre de candidats venant de la classe profession intellectuelle avec un taux de réussite supérieure à ceux venant de la classe ouvrière ce qui explique le fort taux de réussite tout milieu confondu. A l'inverse pour le bac professionnel, il y a largement plus d'étudiants venant de la classe ouvrière avec un taux de réussite inférieur à ceux venant de la classe profession intellectuelle, ce qui concorde avec le fait que le taux de réussite au baccalauréat est le plus faible des trois voies tout milieu confondu.

On observe donc une grande disparité dans la réussite scolaire selon l'origine sociale depuis un certain temps, et sans réelle évolution puisque les candidats venant de classe ouvrière ne réduisent jamais l'écart en taux de réussite avec les candidats venant de la classe professionnelle intellectuelle.

C. Le facteur du secteur de l'établissement (privé/public)

Nous trouvions intéressant de comparer le taux de réussite au bac entre le secteur public et privé, cependant nous n'avons pas trouvé de jeux de données exploitables à l'échelle nationale. Ces jeux de données sont trop conséquents et complexes et recensaient les résultats lycée par lycée. Il n'est pas pertinent de comparer uniquement les résultats du baccalauréat de seulement 2 lycées (un public et un privé).

Néanmoins, la tendance confirme le fait que les lycées du secteur privé possèdent un taux de réussite au bac supérieur à ceux du secteur public. De plus, on voit que les enfants venant d'un milieu plus aisé sont beaucoup plus représentés que les autres.
Voici une capture montrant bien la situation :



UNIVERSITÉ DE NANTES

4.La Maquette

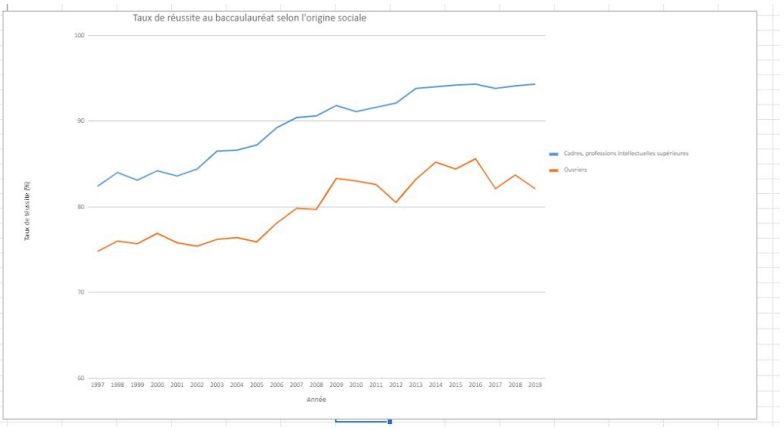
Pour la maquette, nous avons décidé de simplifier l'utilisation du jeu de données sur le taux de réussite au baccalauréat selon l'origine sociale via un fichier excel. En effet, le jeu de données étant plutôt conséquent, il est difficile de récupérer les informations recherchées par l'utilisateur. Nous avons donc décidé de mettre en place un fichier excel qui permet selon la feuille choisie de pouvoir afficher les taux d'admissions pour une classe sociale et de les comparer avec ceux d'une autre classe sociale.
Nous avons aussi mis en place une feuille qui réorganise le jeu de données.

Données réorganisées :

Taux réussite bac										
	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Agriculteurs exploitants	80,9	82,9	82,9	85	84,5	84	85,7	86	86	87,6
Artisans, commerçants, chefs d'entreprise	77	78,9	78,7	80	79	79,1	81,1	80,8	81,1	83,3
Autres personnes sans activité professionnelle	69,9	71	70,3	71,8	69,9	69,9	72,3	70,3	70,9	73,8
Cadres, professions intellectuelles supérieures	82,4	84	83,1	84,2	83,6	84,4	86,5	86,6	87,2	89,2
Cadres, professions intellectuelles supérieures : professeurs et assimilés	85,1	86,3	85,7	86,7	86,3	86,6	88,4	88,6	88,9	90,7
Employés	76,5	78,2	77,7	78,8	78,2	77,8	79,1	78,6	79,2	81
Ensemble	77,3	78,9	78,3	79,5	78,6	78,6	80,1	79,7	79,9	82,1
Indéterminés	67,4	69,2	68,6	70,1	69,2	69	70,7	70,4	68,9	72,9
Ouvriers	74,8	76	75,7	76,9	75,8	75,4	76,2	76,4	75,9	78,1
Professions intermédiaires	78,3	80	79,5	80,9	80,1	80,1	82	81,6	82,2	84,2
Professions intermédiaires : instituteurs et assimilés	81,7	83,2	82,4	84,7	84,3	84	86,5	85,9	86,9	89,1
Retraités	72,1	73,6	73,8	74,8	73,8	73,2	73,6	74	73,4	76,2
Taux réussite bac général										
	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
Agriculteurs exploitants	78,7	82,6	82,7	84,4	83,8	84,3	87,8	87,2	88,2	90,3
Artisans, commerçants, chefs d'entreprise	74,6	77,5	77,1	79	78,3	79,5	82,9	82	83,9	86,5
Autres personnes sans activité professionnelle	66,4	69,3	68,7	70,5	69,2	71	75,9	72	74,9	78
Cadres, professions intellectuelles supérieures	82,7	84,6	83,6	85	84,6	85,8	88,6	88,3	89,4	91,7
Cadres, professions intellectuelles supérieures : professeurs et assimilés	85,7	87,1	86,6	87,5	87,2	87,9	90,2	90,1	90,7	93,1
Employés	73,9	76,9	76,2	77,3	77,1	77,9	81,2	79,8	81,8	84,3
Ensemble	76,6	79,2	78,4	79,9	79,4	80,3	83,7	82,5	84,1	86,6
Indéterminés	61,9	64,3	63,5	66,5	68,2	68,3	73,7	70,5	71,9	74,8
Ouvriers	70,4	73,6	72,6	74,2	73,6	74,1	78,2	76,7	78,1	81,1
Professions intermédiaires	77,4	79,8	78,9	80,7	80,2	81	84,6	83,5	85,3	87,5
Professions intermédiaires : instituteurs et assimilés	81,9	83,5	82,4	84,9	84,8	85,5	88,6	88	89,2	90,9
Retraités	69,8	73	73,5	75	73,9	75,1	79,7	78,3	80,7	82,3

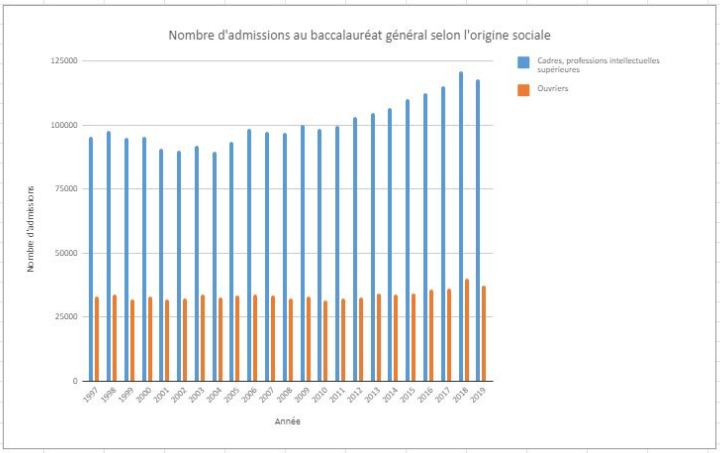
Exemple affichage graphique taux de réussite :

Année	Cadres, professions inte	Ouvriers
1997	82,4	74,8
1998	84	76
1999	83,1	75,7
2000	84,2	76,9
2001	83,6	75,8
2002	84,4	75,4
2003	86,5	76,2
2004	86,6	76,4
2005	87,2	75,9
2006	89,2	78,1
2007	90,4	79,8
2008	90,6	79,7
2009	91,8	83,3
2010	91,1	83
2011	91,6	82,6
2012	92,1	80,5
2013	93,8	83,2
2014	94	85,2
2015	94,2	84,4
2016	94,3	85,6
2017	93,8	82,1
2018	94,1	83,7
2019	94,3	82,1



Exemple affichage graphique nombre d'admissions :

Année	Cadres, professio	Ouvriers
1997	95290	33038
1998	97746	33672
1999	94905	31987
2000	95288	32880
2001	90697	31723
2002	90109	32118
2003	91819	33764
2004	89485	32783
2005	93490	33432
2006	98452	33898
2007	97521	33259
2008	97120	32222
2009	99980	33123
2010	98561	31466
2011	99611	32051
2012	103143	32633
2013	104859	34269
2014	106620	33708
2015	110045	34069
2016	112280	35666
2017	115305	36286
2018	121044	40055
2019	117877	37304



Cette maquette permet d'exploiter plus facilement les données du fichier, tout en étant intuitif pour l'utilisateur et nous a facilité l'exploitation des données pour notre analyse.
Le seul problème ici est que les données ne pourront pas se mettre à jour automatiquement, il faudra ajouter manuellement les derniers résultats pour pouvoir actualiser les données du fichier.

5.Conclusion

Notre recherche de données sur le taux de réussite scolaire nous a permis de mettre en application les principes de l'Open Data. Cela nous a obligé à avoir un regard critique sur les données trouvées et à juger si elles étaient exploitables ou non.

Le thème scolaire est vaste et contient de nombreux jeux de données et qui sont régulièrement mis à jour, il n'a donc pas été difficile d'en trouver. Le plus difficile a été de réussir à exploiter ces données étant données la volumétrie conséquente de certains des jeux de données.

Cette analyse nous a permis de conclure qu'il y a des disparités dans la réussite scolaire depuis un certain temps, et que malgré les actions de l'éducation nationale cela ne tend pas à disparaître. Il sera intéressant d'étudier les résultats du baccalauréat de la session 2021 avec la nouvelle réforme et voir si cela a un impact sur le taux de réussite.

UNIVERSITÉ DE NANTES

6.Webographie

https://www.insee.fr/fr/statistiques
https://www.data.gouv.fr
https://www.bfmtv.com/societe/education/enseignement-public-et-prive-le-comparatif-en-chiffres_AN-201711270046.html