IUT de Paris - Rives de Seine
Université Paris Cité

SciencesPo

CONSTRUCTIONS, VISUALISATIONS ET ANALYSES STATISTIQUES DE RESEAUX SOCIAUX ET ORGANISATIONNELS

EVA BERTRAND

BUT Science des Données Visualisation, Conception d'Outils Décisionnels - 2ème année - FI

2023 – 2024

IUT PARIS RIVES DE SEINE
Mme Claire STEFAN

FNSP – SCIENCES PO
Mr Emmanuel LAZEGA

REMERCIEMENT

Je voudrais tout d'abord remercier Monsieur Lazega, mon tuteur de stage à SciencesPo, pour m'avoir acceptée et accueillie pendant 12 semaines, durant lesquelles j'ai pu mettre en pratique mes compétences aux bénéfices de travaux qui ont été très intéressants et très enrichissants.

Je remercie Nathanael, d'avoir été mon binôme lors de cette période stage. Nous avons pu nous entraider quand il était nécessaire pour avancer dans nos missions.

Je souhaite aussi remercier Madame Stephan pour avoir été ma tutrice du côté IUT, lors de ce stage.

Un grand merci à tous les membres du laboratoire du CSO pour m'avoir accueillie.

Un remerciement à Madame Muri pour avoir lu et corrigé ce rapport de stage, que je l'espère, sera conforme aux attendus.

Table des matières

1. Introduction 4
1.1. Contexte du stage 4
1.2. Missions réalisées 4
1.3. Problématiques 5
2. Contexte 6
2.1. Présentation de l'entreprise 6
2.2. Présentation du service 6
2.3. Présentation de l'environnement 6
2.4. Organigramme 7
3.Mission 1 8
3.1. Description des objectifs et livrables 8
3.2. Cartographie de l'architecture des données 8
3.3. Tâches et compétences 9
3.4. Contribution 9
3.5. Réflexion sur les apports 10
4.Mission 2 11
4.1. Description des objectifs et livrables 11
4.2. Cartographie de l'architecture des données 11
4.3. Tâches et compétences 12
4.4. Contribution 13
4.5. Réflexion sur les apports 18
5.Conclusion générale 20
6.Bibliographie 21
6.1. Références 21
6.2. Lexique 21
6.3. Table des illustrations 21
7.Annexes 23
8. Démarche Portfolio 28

1. Introduction

1.1. Contexte du stage

Mon stage se déroule dans le cadre de ma formation en BUT Science des Données. En tant qu'étudiante de deuxième année en formation initiale, il m'était demandé de réaliser un stage d'une durée minimale de huit semaines dans une entreprise. L'objectif principal de ce stage est de me permettre de mettre en pratique les compétences et les connaissances acquises durant ma formation, tout en développant de nouvelles compétences.

En amont de mon postulat sur cette offre, j'ai communiqué mes candidatures à d'autres offres d'entreprises (BPCE, RATP, TF1), mais je n'ai eu que des retours négatifs. Puis j'ai commencé à postuler à des offres qui étaient transmises par ma professeur Madame Bonnot.

Ce qui m'a mené à postuler à l'offre actuelle c'est mon envie de découvrir la branche de la sociologie, et de voir comment ce que j'apprends au sein de ma formation peut être mis en pratique dans ce domaine.

La durée de mon stage s'étend sur une période de 11 semaines, du 8 avril 2024 au 21 juin 2024.

1.2. Missions réalisées

Afin d'assurer le bon déroulement du stage, j'ai d'abord assimilé les différentes notions de sociologie, sur différents sites, logiciels et scriptes R que j'aillais aborder lors de ce stage.

Durant cette période de stage, j'ai participé sur les deux missions présentées dans l'offre de monsieur Lazega :

Mission 1 : Etudes sur les données ARC

- Création des analyses multi-niveaux[1] entre les organisations et les individus,
- Aide à la documentation et à l'archivage sur le site de datasciencespo des données, scriptes et publications,
- Rédaction et conception de la documentation

Mission 2 : Etudes sur les données UPC

- Construction des visualisations des réseaux bipartites[2] et des analyses multi-niveaux entre des événements et individus,
- Analyses descriptives de variables concernant les événements et les individus.
- Rédaction et conception de la documentation

[1] Méthode statistique permettant d'étudier des données structurées en niveaux hiérarchiques
[2] Type de réseau où les nœuds sont divisés en deux groupes distincts, et les liens ne peuvent exister qu'entre les nœuds de groupes différents

1.3. Problématiques

Comme mon stage se déroulait sur une durée de 11 semaines, j'ai travaillé sur plusieurs tâches qui se passaient sur des missions différentes, ce qui fait que je n'avais pas de problématique principale qui pourrait recouvrir toutes les tâches effectuées.

Néanmoins lors de ce stage, j'ai rencontré des problématiques techniques :

La problématique technique la plus importante que j'ai rencontré est la compréhension de la sociologie et des réseaux sociaux, de leurs concepts même, des procédés d'analyses. J'ai dû aussi m'adapter à d'anciens scriptes, les comprendre et réussir à en tirer des conclusions. Au cours du stage, je me suis aussi familiarisée avec le logiciel UCINET[3], logiciel permettant de faire des analyses de données de réseaux sociaux.

Pour remédier à ces problèmes, je me suis beaucoup documentée et renseignée sur les principes et concepts de la sociologie. J'ai lu des livres [1] parlant des procédés de l'analyse de réseaux sociaux, des documents [2] et sites internet [3] renseignant sur les statistiques et analyses de réseaux sur Rstudio.

[3] Logiciel permettant l'analyse des données des réseaux sociaux, mesure de centralité, identification de sous-groupes.

2. Contexte

2.1. Présentation de l'entreprise

L'entreprise où j'effectue mon stage est l'Institut d'études Politiques de Paris, communément appelé Sciences Po. Sciences Po a été créée en 1872 par Emile Boutmy afin de répondre à la crise qui s'abat sur la France après la guerre de 1870. Le site se trouve au 1 place Saint Thomas à Paris. Sciences Po compte sous son enseigne un total 7 écoles : Dijon, le Havre, Menton, Nancy, Poitiers, Reims et Paris.

Les domaines étudiés dans cette école sont les sciences politiques, droit, l'histoire, la sociologie et l'économie.

2.2. Présentation du service

J'effectue mon stage au sein du Centre de Sociologie des Organisations (CSO). Il été fondé en 1964 par Michel Crozier, s'articulant autour du programme de recherche sur l'administration française. Il s'agit maintenant d'un laboratoire s'étendant sur la sociologie économique, les organisations, l'action publique, les professions, du travail, des mouvements sociaux et du droit.

Les recherches du laboratoire se positionnent autour de 5 axes :

- Droit, normes et régulations
- Travail, emploi et profession
- Gouvernance et organisations économiques
- Savoir, science et expertise
- Action publique et transformations de l'Etat

Le laboratoire est dirigé par Sophie Dubuisson-Quellier et Patrick Castel au poste de directeur adjoint.

2.3. Présentation de l'environnement

Lors de mon stage, mon tuteur est Emmanuel Lazega. Il est professeur des universités à Sciences Po, dont les thèmes de recherches sont la sociologie des organisations, la sociologie économique et de la sociologie des réseaux. Il contribue aussi au développement d'une sociologie néo-structurale qui articule notamment l'analyse organisationnelle et l'analyse de réseaux sociaux.

Durant mon stage, j'étais accompagné de mon camarade de classe Nathanael Slama.

2.4. Organigramme

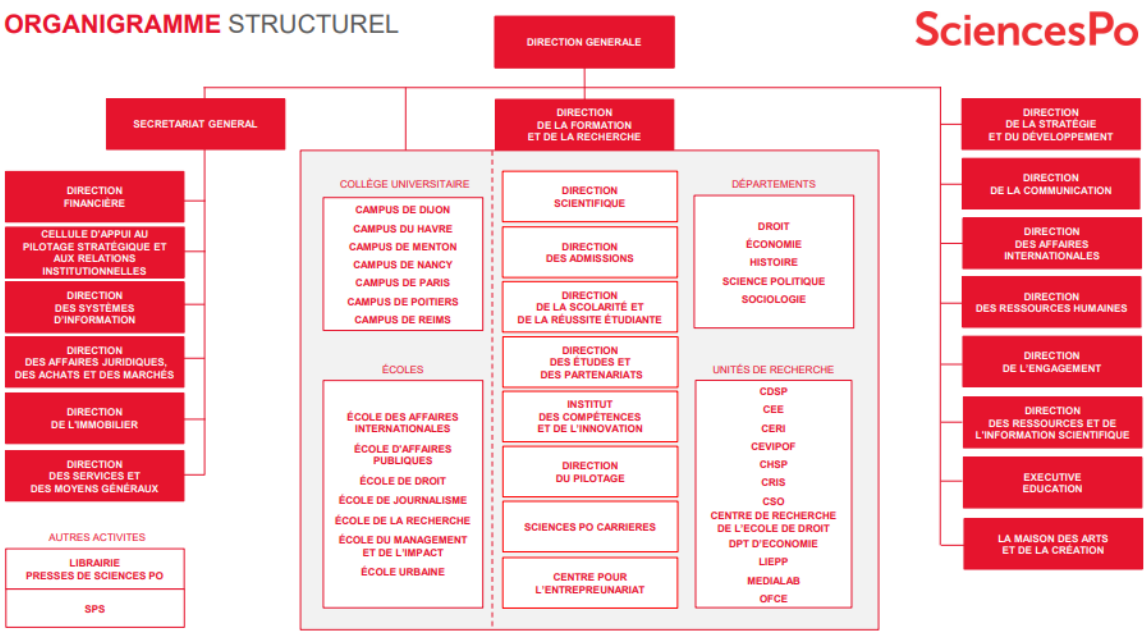


Figure 1 : Organigramme général de Sciences Po

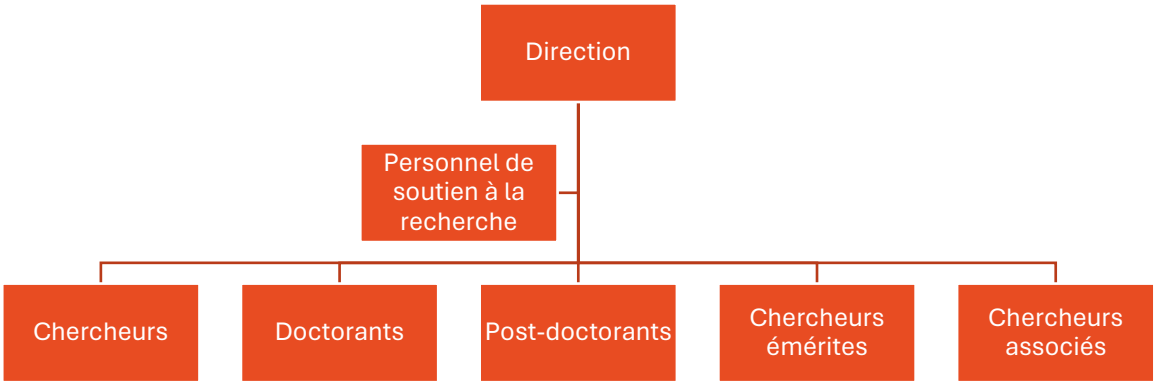


Figure 2 : Organigramme du CSO

3.Mission 1

3.1. Description des objectifs et livrables

J'ai travaillé peu de temps sur cette mission, mais parfois j'aide mon camarade quand il en a besoin. Néanmoins, quand je travaillais sur cette mission, je n'avais pas de calendrier prévisionnel. J'ai commencé par je nettoyage des bases et ensuite je devais procéder au traitement mais mon tuteur m'a confiée la seconde mission.

L'objectif général de la mission était de construire les visualisations multi-niveaux, d'écrire un script généraliste permettant de traiter automatiquement les bases de données du même format et de la même série de bases de données. De plus, Il fallait aider à la documentation et à la mise en ligne les bases de données, les scripts et les publications en liens avec l'enquête sur l'entrepôt de données de la recherche de Sciences Po[4].

Les livrables attendues à la fin de cette mission sont les différents scripts, un rapport qui explique les codes.

3.2. Cartographie de l'architecture des données

Ces données sont issues d'enquêtes sociologiques : les données ont été récoltées à la suite de réponses de questionnaires et d'entretiens. Elles portent sur les relations que le répondant a avec les autres membres de l'organisation. Ce qui fait que ce procédé mène à un fort taux de réponses. Les données sont stockées et archivées sur le site de datasciencespo.

Pour cette mission, j'ai surtout travaillé sur Rstudio et Excel. J'ai d'abord utilisé Excel pour visualiser et comprendre la structure des bases de données. Ensuite, j'ai utilisé Rstudio afin de réaliser le nettoyage et traitement des bases de données. De plus, j'ai été initié à un logiciel dédié à l'analyse de données de réseaux sociaux : Ucinet, afin de voir la structure des réseaux des chercheurs et la centralité des individus.

Dans cette mission, j'avais 14 bases de données qui correspondaient aux réseaux de chercheurs. Chaque base correspondait une question du questionnaire. Il y avait aussi 16 bases de données représentants les questions pour les réseaux des laboratoires.

Les bases de chercheurs ont le même de lignes que de colonnes : 129x129, et, les bases de laboratoires 133x133.

Toutes les bases étaient au même format : les valeurs dans la première colonne étaient les noms des entités répondantes et dans la première ligne, nous retrouvions les noms des entités

[4] data.sciencespo.fr

répondues. S'il y avait un zéro, le répondant n'avait pas répondu l'individu, dans le cas contraire il y avait un 1.

Avec ces bases de données, il fallait avant tout s'assurer dès l'importation sur Rstudio de l'anonymat des données, retirer le nom des laboratoires et des chercheurs.

3.3. Tâches et compétences

Dans cette mission, j'ai d'abord appris à utiliser le logiciel Ucinet, c'est un logiciel qui permet de mesurer la centralité d'un individu, de détecter des communautés au sein même d'un réseau de personnes et de visualiser les relations entre les différents nœuds (individus ou groupes) d'un réseau.

De plus, comme les bases des chercheurs étaient toutes au même format et contenaient des valeurs similaires, il était nécessaire de faire des fonctions. Ceci était dans le but de faciliter le nettoyage et le traitement, d'automatiser de façon à gagner du temps et de créer un programme générique.

3.4. Contribution

Etant donné que je ne suis pas restée longtemps sur cette mission, je n'ai fais que la première version des fonction de nettoyage et de traitement des bases des chercheurs et des laboratoires.

```
#CHARGEMENT DES DONNEES
  #La fonction ci-dessus permet quand l'on rentre le chemin de données de charger le chemin du fichier .xls
load_excel = function(file_path) {
  data = read_xls(file_path, sheet = 1)
  | #permet de supprimer la première colonne de la base et de garder toutes les autres.
  data = data %>% select(-1)
   #permet de renommer les colonnes restantes de la base de 1 à n (n = nb total de colonne)
  colnames(data) = 1:ncol(data)
  return(data)
}

# Chargement des données et nettoyage
rs1 = load_excel("C:/Users/gevab/OneDrive/Documents/BUT/Stage/Archivage ARC/Chercheurs répondants 128/V1RTR-1.xls")
rs2 = load_excel("C:/Users/gevab/OneDrive/Documents/BUT/Stage/Archivage ARC/Chercheurs répondants 128/V1RTR-2.xls")
rs3 = load_excel("C:/Users/gevab/OneDrive/Documents/BUT/Stage/Archivage ARC/Chercheurs répondants 128/V1RTR-3.xls")
rs4 = load_excel("C:/Users/gevab/OneDrive/Documents/BUT/Stage/Archivage ARC/Chercheurs répondants 128/V1RTR-4.xls")
rs5 = load_excel("C:/Users/gevab/OneDrive/Documents/BUT/Stage/Archivage ARC/Chercheurs répondants 128/V1RTR-5.xls")
rs6 = load_excel("C:/Users/gevab/OneDrive/Documents/BUT/Stage/Archivage ARC/Chercheurs répondants 128/V1RTR-6.xls")
rs7 = load_excel("C:/Users/gevab/OneDrive/Documents/BUT/Stage/Archivage ARC/Chercheurs répondants 128/V1RTR-7.xls")
rs8 = load_excel("C:/Users/gevab/OneDrive/Documents/BUT/Stage/Archivage ARC/Chercheurs répondants 128/V1RTR-8.xls")
rs9 = load_excel("C:/Users/gevab/OneDrive/Documents/BUT/Stage/Archivage ARC/Chercheurs répondants 128/V1RTR-9.xls")
rs10 = load_excel("C:/Users/gevab/OneDrive/Documents/BUT/Stage/Archivage ARC/Chercheurs répondants 128/V1RTR-10.xls")
rs11 = load_excel("C:/Users/gevab/OneDrive/Documents/BUT/Stage/Archivage ARC/Chercheurs répondants 128/V1RTR-11.xls")
rs12 = load_excel("C:/Users/gevab/OneDrive/Documents/BUT/Stage/Archivage ARC/Chercheurs répondants 128/V1RTR-12.xls")
rs13 = load_excel("C:/Users/gevab/OneDrive/Documents/BUT/Stage/Archivage ARC/Chercheurs répondants 128/V1RTR-13.xls")
rs14 = load_excel("C:/Users/gevab/OneDrive/Documents/BUT/Stage/Archivage ARC/Chercheurs répondants 128/V1RTR-14.xls")
```

Figure 3 : Extrait du script, fonction pour nettoyer les bases des chercheurs

Néanmoins, ayant terminé les plus grandes parties de ma mission que je vais vous présenter prochainement, je reviens travailler avec mon camarade de stage sur cette première mission afin de l'aider à terminer.

3.5. Réflexion sur les apports

Etant données que je n'ai passé beaucoup de temps sur cette mission, je n'ai pas vraiment eu de problème. Toutefois, j'ai eu quelques difficultés lors de l'écriture du script. Je devais faire une fonction pour le traitement mais je ne suis pas très à l'aise avec l'écriture de fonctions et de boucles.

4.Mission 2

4.1. Description des objectifs et livrables

Pour cette mission, je n'avais de calendrier prévisionnel à proprement parler mais j'ai d'abord commencé par le nettoyage des bases de données, j'ai ensuite procédé au traitement des données et je suis passé à la visualisation. Après ces étapes, j'ai rédigé un rapport écrit.

L'objectif était de traiter 3 bases de données comportant des données d'événements et d'acteurs. Pour deux bases, il fallait faire des analyses statistiques et des visualisations sur certaines variables. Pour la dernière base, je devais reprendre des scripts des anciens stagiaires et poursuivre les visualisations de réseaux.

A la fin de cette mission, les livrables attendues étaient dans un dossier : des scripts R, les bases de données et un rapport expliquant les codes et les graphiques. Néanmoins, j'ai décidé dans le dossier attendu des créer un sous dossier contenant les graphiques sortant d'Excel,

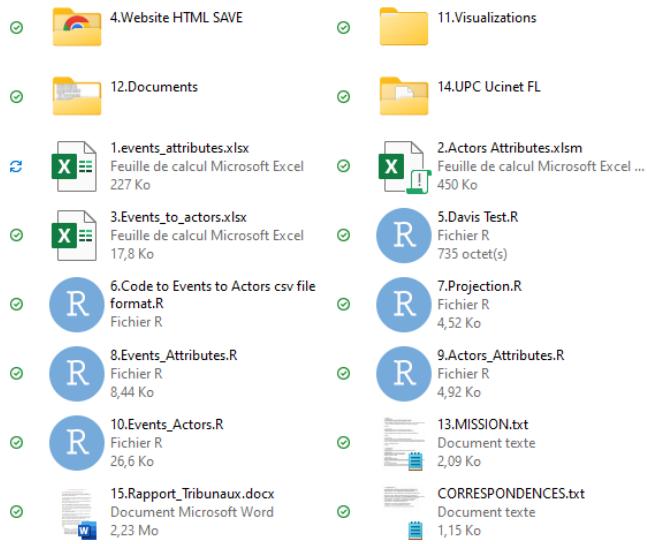


Figure 4 : Livrables à rendre

4.2. Cartographie de l'architecture des données

Le processus initial de récolte des données pour cette mission était des entretiens mais à cause de la COVID-19 ceux-ci ont été annulé. Toutefois en 2009, lors d'un des événements 38 personnes ont été interviewée sur leurs réseaux. Le reste des données ont été récolté en ligne via les sites des organisateurs des événements.

Pour cette mission, j'ai utilisé Excel et Rstudio. Excel m'a surtout servi à voir comment le bases étaient présentées et à apporter d'importantes modifications que je n'envisager pas de faire sur Rstudio. J'ai utilisé Rstudio afin d'écrire des scripts permettant d'importer, traiter et de créer des visualisations, qui par la suite, mènent à des analyses.

Lors de cette mission, j'ai travaillé sur 3 bases de données différentes, et, chacune d'entre elles devait être abordées différemment :

- La première base dont je me suis occupée comportais des informations sur des événements comme l'identifiant, le nom, la date, le lieu et les organisateurs. Cette base avait 4 feuilles Excel mais une seule devait être utilisée. Il y avait 216 observations et 17 variables. Pour cette base, il était question de créer des visualisations et faire des analyses sur différentes variables comme la répartition des événements ou la répartition des lieux.

- La seconde base que j'ai traitée contenait des informations « d'acteurs ». Il y avait 136 variables pour 2311 observations. Cette base contient l'identifiant, le nom, le pays, le genre, l'occupation lors des événements et la catégorisation professionnelle. Il y avait 2 feuilles mais il fallait en traiter qu'une seule. Pour cette base je devais créer des visualisations et analyses sur des variables comme le genre, la catégorie de métier ou la nationalité.

- La dernière base mettait en lien les événements et les acteurs. Il y a 3 colonnes et 216 observations. Deux colonnes donnaient les identifiants des événements et la troisième contenait la liste des « acteurs » ayant été présents à ces événements. Les attendus pour cette base étaient de donner les projections[5] des acteurs, créer le réseau bipartite entre les événements et les acteurs, et, de construire la visualisation multi-niveau des événements et acteurs.

Pour chaque base que j'ai traité les attendus étaient différents :

- Pour la première base de données, je devais vérifier si les informations par la stagiaire précédente étaient toujours correctes et apporter des modifications où celles-ci étaient nécessaires.

- Pour la seconde, j'ai dû compléter des informations, comme le genre, le pays d'appartenance et la catégorie, afin d'avoir un rendu plus complet lors des analyses et visualisations. De plus, j'ai dû retirer tous les doublons directement sur Excel, car en les retirant sur Rstudio, j'ai rencontré des problèmes par la suite dans le code.

- Pour la dernière base, j'ai dû modifier les listes d'acteurs par les nouveaux identifiants acteurs, ce qui fait suite aux modifications effectuées sur la base contenant les informations des acteurs.

4.3. Tâches et compétences

Pour cette mission, j'ai mis en pratique les compétences que j'ai acquis sur Rstudio, les différents procédés pour nettoyer, traiter les bases de données, et construire des visualisations qui répondent aux attentes. Avec cette mission, j'ai pris de nouveaux de procédés et de nouvelles

[5] Graphe qui transforme le réseau en un réseau monopartite en créant des liens entre les nœuds du même groupe

méthodes pour construire les visualisations. J'ai dû être plus rigoureuse et organisée dans mon travail.

4.4. Contribution

J'ai commencé à travailler sur la base contenant les informations des événements car il me semblait que celle-ci allait être la plus simple et la plus courte à traiter. Pour commencer j'ai d'abord ajouté à la base sur Excel deux nouvelle colonne qui contiennent les coordonnées GPS des lieux de rassemblement (latitude et longitude). Et j'ai aussi pour les différents organisateurs renommer certains qui sont les mêmes mais dont le nom était écrit différemment.

event_city	event_cnt	latitude	longitude
San Servolo	IT	45.416665	12.34986
	IT		
London	UK	51.5074456	-0.1277653
Brussels	BL	50.8465573	4.351697
Warsaw	PL	52.2337172	21.0714322
Brussels	BL	50.8465573	4.351697
Brussels	BL	50.8465573	4.351697
Brussels	BL	50.8465573	4.351697
Strasbough	FR	48.584614	7.7507127
Strasbough	FR	48.584615	7.7507128
Strasbough	FR	48.584616	7.7507129
Munich	DE	48.1371079	11.5753822
Munich	DE	48.1371080	11.5753823

Figure 5 : Extrait de la base de données des événements

Je suis ensuite passée sur Rstudio afin de commencer le nettoyage et le traitement de la base.

J'ai d'abord installé tous les packages qui étaient nécessaire et j'ai importé la base. Puis j'ai retiré les colonnes qui ne me servirait pas dans mon code.

J'ai ensuite traité la base en créant deux nouvelles colonnes qui contiennent les années extraite de la variable contenant les dates des événements ou la période sur laquelle les événements ont eu lieu. Pour cela j'ai écrit une fonction qui selon les valeurs dans les cellules changes la méthode d'extraction. Je me suis en suite occupé de la variable des villes afin de les mettre au même format pour ne pas avoir des valeurs identiques mais écrite différemment.

Pour chaque visualisation, j'ai retraité les données selon les besoins que j'avais afin de créer les meilleures visualisations possibles

Après cette première base, je me suis occupée de celle des acteurs. Sur Excel, j'ai apporté d'importantes modifications, j'ai surtout apporté des informations qui étaient manquantes comme le genre de certaines personnes, le pays d'appartenance et la catégorisation professionnelle. J'ai aussi dû retirer tous les doublons, ce qui représente près des 300 individus supplémentaires.

En passant sur Rstudio, j'ai importé la base de données et j'ai retiré des colonnes afin d'assurer la notion d'anonymat qui est demandé, et, j'ai retiré des lignes qui était vides, ce qui fait que je suis passée d'une base de plus de 2000 lignes à une base de 883 lignes.

```
########## IMPORTATION ##########
actors_path = "C:/Users/gevab/OneDrive/Documents/BUT/Stage/UPC/"
aa = read_xlsx(file.path(actors_path,"2.Actors Attributes.xlsm"), sheet=1)
actatt = aa %>% select(-c(1:2, 5, (ncol(aa) - 1):ncol(aa))) #retire les colonnes 1, 2, 5 et les deux dernières
actatt = actatt[complete.cases(actatt[1]), ] #supprime les lignes où la valeur de la première colonne est vide
```

Figure 6 : Extrait de l'importation des données "Acteurs"

aa	2179 obs. of 136 variables
act_categ	883 obs. of 16 variables
act_cnt	42 obs. of 2 variables
act_gender	883 obs. of 4 variables
actatt	883 obs. of 131 variables

Figure 7 : Nettoyage de la base "Acteurs"

Je suis ensuite passée à l'étape des visualisations, comme précédemment j'ai appliqué le traitement sur les données selon mes besoins. J'ai créé les visualisations qui montrent les distributions de certaines variables comme le genre, le pays d'appartenance et la catégorisation professionnelle.

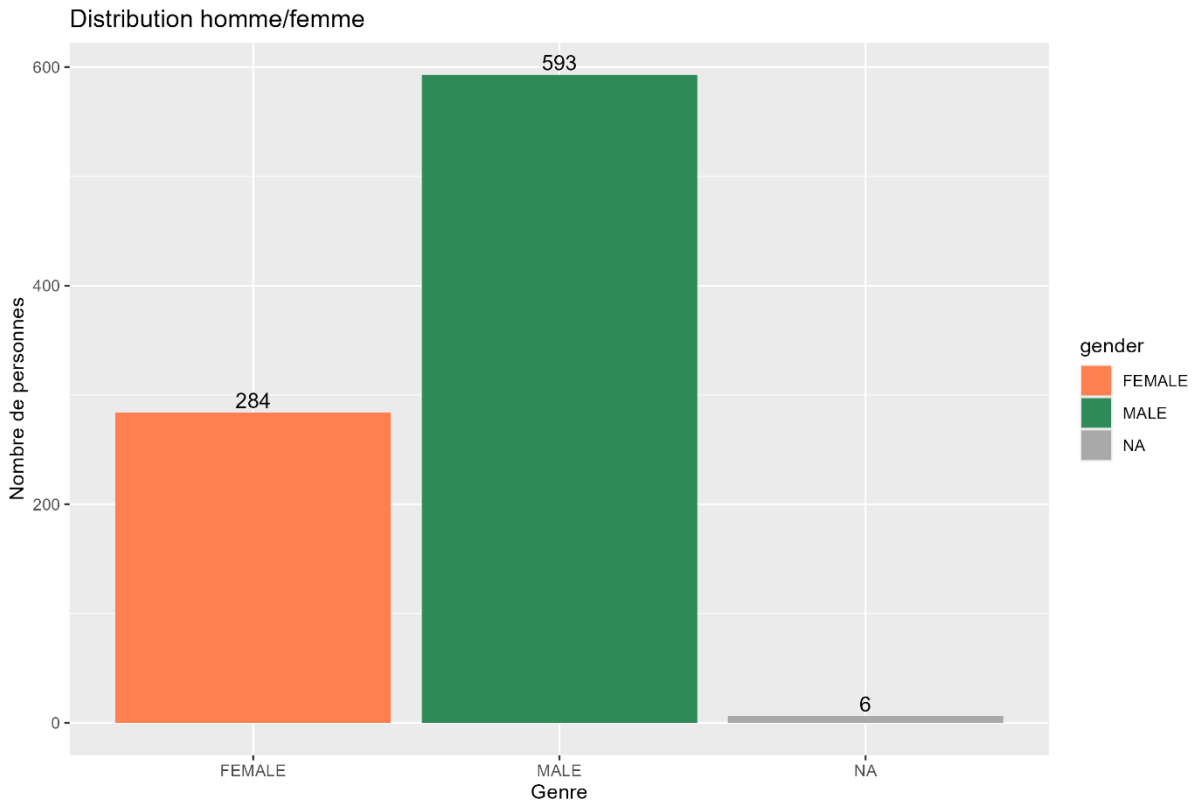


Figure 8 : Graphique représentant la distribution du genre des "Acteurs"

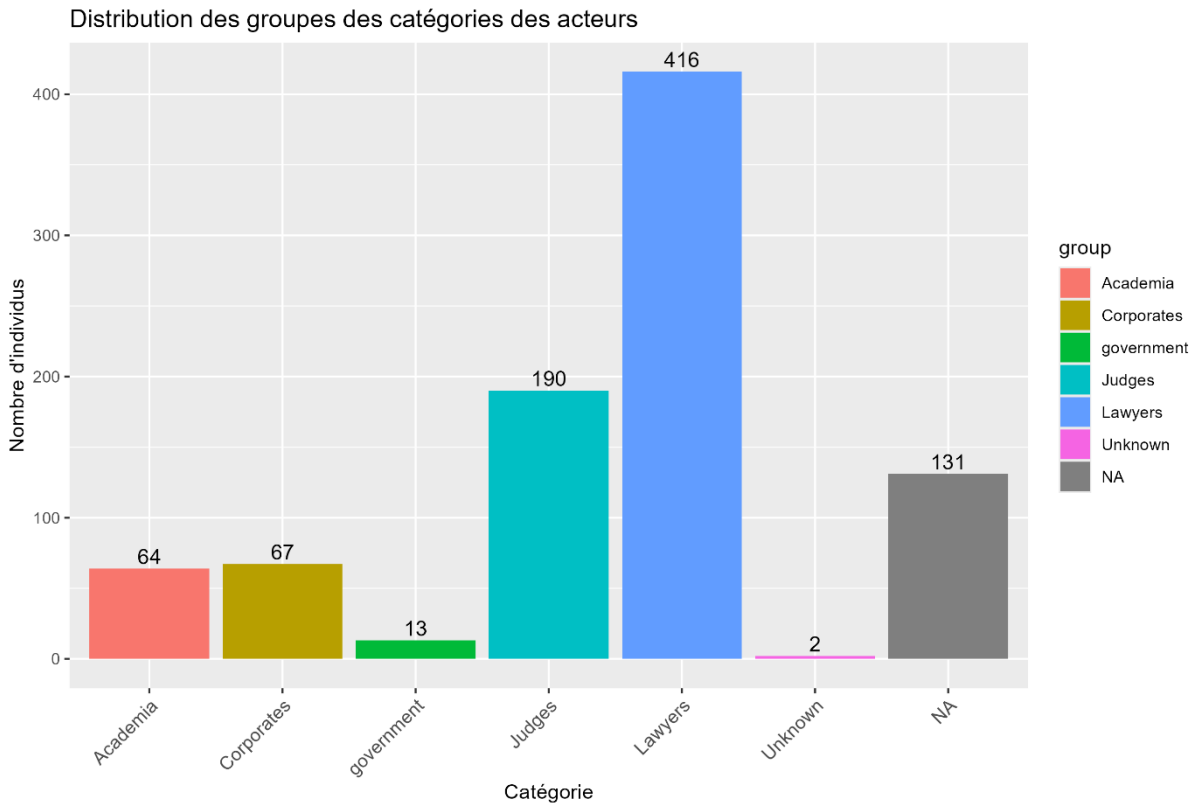


Figure 9 : Graphique représentant la distribution de la catégorie professionnelle des acteurs

Pour continuer sur cette mission, je me suis confrontée à la dernière base de données. Pour cette étape de la mission j'ai d'abord dû remodifier les identifiants des acteurs dans la liste des participants par suite de la modification que j'ai effectué avec les doublons dans la seconde base.

	id_eve_xx	list_actors
1	1	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,2
2	2	4,5,7,10,11,13,15,17,18,19,22,25,27,28,29,30,31,32,33,34,35,36,
3	3	28,42,43,39,44,45,46,47,48,49,50,51,52,53,54,55
4	4	5,19,40,47,48,56,57
5	5	7,58,59,60,61,62,63
6	6	4,5,7,23,40,47,48,56,57,64,65,66,67,68,69
7	7	56
8	8	5,46,56,57,66,71,72,73,74,75
9	9	
10	10	5,7,12,19,25,42,47,57,76,77,78,79,80,81,82,83
11	11	44,48,70,84,85,86,87,88
12	12	44,48,52,70,85,86,87
13	13	42,70,89,90,91,92,93,94,95,96,97,98
14	14	5,28,47,48,70,76,95,99,100,101,102
15	15	5,28,47,48,51,95,99,100,101,102,103,104
16	16	7,28,47,48,70,99,100,101,103,105,106,107
17	17	2,9,25,56,71,72,76,96,108,109,110,111,112,113,114,115,116,117
18	18	28,36,44,56,70,72,76,96,114,124,125,126,127,128,129,130,131,1
19	19	5,7,31,36,44,46,56,67,72,76,79,124,125,126,127,128,139,140,14
20	20	44,47,56,70,72,76,125,127,140,143,154,155,156,173,174,175,17
21	21	5,6,7,13,14,15,16,18,22,23,26,29,39,158,159,160,161,162,163,16
22	22	1,2,3,7,9,19,22,25,26,27,29,30,124,169,184,185,186,187,188,189
23	23	1,2,4,5,7,12,16,19,22,29,30,33,34,39,40,73,76,79,124,187,188,18
24	24	2,3,4,5,7,9,12,13,15,17,20,22,25,27,29,30,36,39,40,59,60,61,73,
25	25	7,64,73,328
26	26	5,7,36,44,64,70,71,76,78,109,123,264,329,330,331,332,333,334
27	27	44,70,76,79,335,336
28	28	31
29	29	70,338,340,341
30	30	151

Figure 10 : Extrait de la base faisant lien entre les événements et les acteurs

Sur Rstudio, j'ai importé les trois bases afin de pour par la suite répondre aux attentes.

Toutefois, il est important de savoir que l'année précédente, Monsieur Lazega a eu une stagiaire qui avait travaillé dessus et dont j'ai récupéré les scripts pour continuer la mission. J'ai alors réutilisé ses scripts et j'ai continuer avec la visualisation des réseaux d'acteurs et d'événements, retravaillé les projections, visualisé le réseau bipartite et construit l'analyse multi-niveau.

Toutefois, il est important de savoir que j'ai reçu de l'aide de la part de David Schoch, qui m'a aiguillée sur le procédé à suivre pour la construction de l'analyse multi-niveaux.

```
#-- réseau des acteurs --#
#indentation des id
V(Actor_Graph)$id = Actors_Attributes$id_ind_xx[match(names(V(Actor_Graph)), Actors_Attributes$id_ind_xx)]

#mappage des couleurs
Actors_Attributes = Actors_Attributes %>%
  mutate(color = case_when(
    `Categorisation 1` == "judge" ~ "firebrick1",
    `Categorisation 1` %in% c("advocate", "barrister", "litigator", "solicitor", "lawyer", "attorney") ~ "gold",
    `Categorisation 1` %in% c("corporate", "counsel") ~ "turquoise1",
    `Categorisation 1` == "academia" ~ "violet",
    `Categorisation 1` == "government" ~ "olivedrab",
    `Categorisation 1` == "blank" ~ "royalblue",
    is.na(`Categorisation 1`) ~ "royalblue"
  ))
V(Actor_Graph)$color = Actors_Attributes$color[match(names(V(Actor_Graph)), Actors_Attributes$id_ind_xx)]
V(Actor_Graph)$categorisation = Actors_Attributes$`Categorisation 1`[match(names(V(Actor_Graph)), Actors_Attributes$id
V(Actor_Graph)$color = factor(V(Actor_Graph)$color,
                              levels = c("firebrick1", "gold", "turquoise1", "violet", "olivedrab", "royalblue"),
                              labels = c("judges", "lawyers", "corporates", "academia", "government", "Unknown"))

#-global visualizations-#
#Fruchterman Reingold
rep_act_global_fr = ggraph(Actor_Graph, layout = "fr") +
  geom_edge_link(edge_colour = "grey", width = 0.2) +
  geom_node_point(aes(fill = color), size = 3, shape = 24, color = "black") +
  labs(title = "Réseau acteurs", subtitle= "Fruchterman Reingold layout") +
  geom_node_text(aes(label = id), repel = TRUE) +
  scale_fill_manual(values = c("firebrick1", "gold", "turquoise1", "violet", "olivedrab", "royalblue"),
                    name = "Categorisation") +
  theme(legend.position = "bottom")
```

Figure 11 : Code pour la visualisation du réseau des acteurs

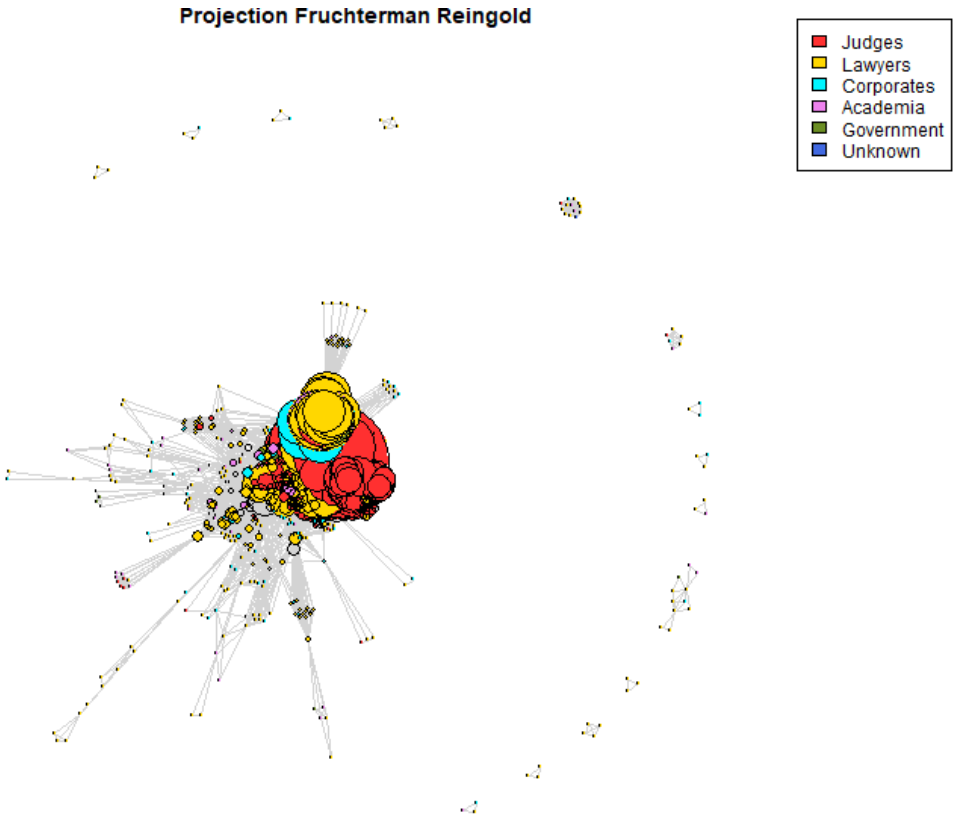


Figure 12 : Projection du réseau acteurs

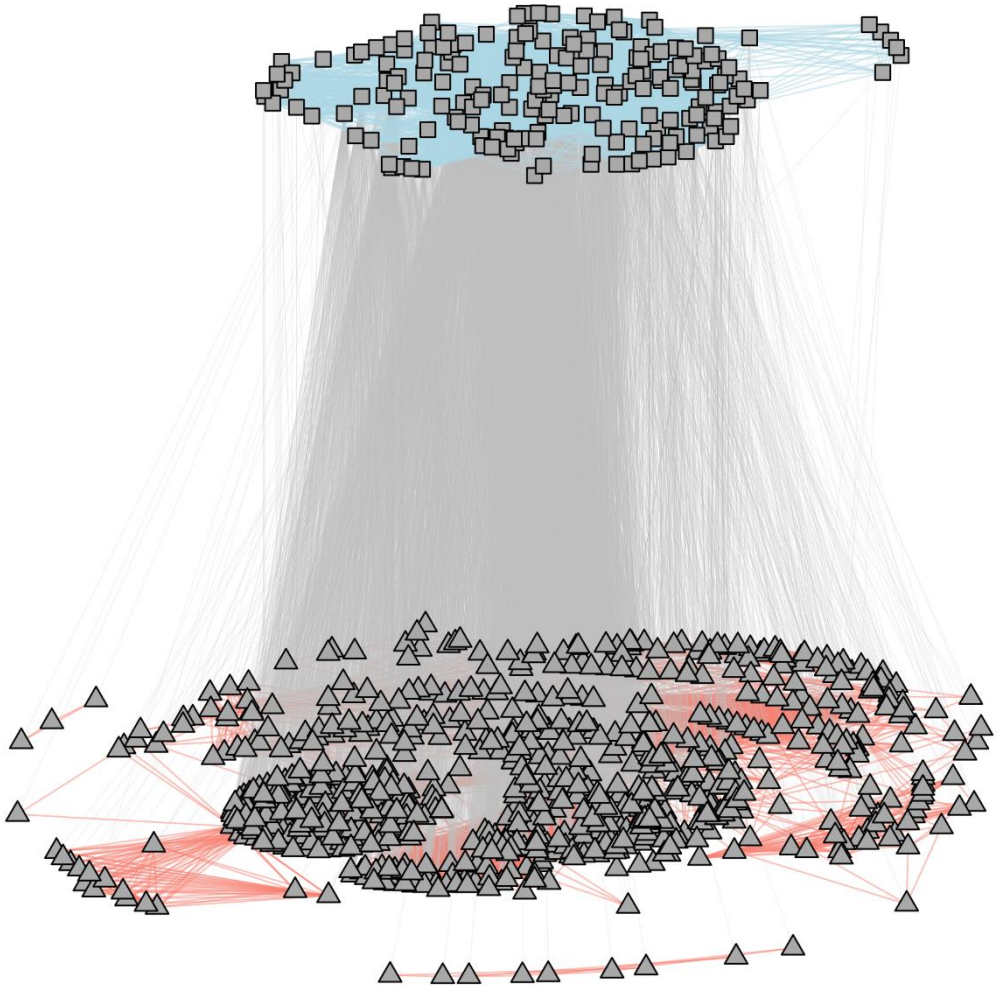


Figure 13 : Analyse Multi-niveau événements/acteurs

Etant donné que j'ai encore deux semaines de stage à compter de la remise de ce rapport de stage, la dernière figure doit encore être travaillée afin de données plus d'informations.

4.5. Réflexion sur les apports

Lors de cette mission, j'ai rencontré quelques difficultés pendant l'écriture des scripts et l'études des bases :

A plusieurs reprises j'ai dû modifier les bases de données, renseigner de nouvelles informations, retirer les doublons qui posaient des problèmes dans mes scripts R. Puis au vu de la taille des réseaux d'acteurs et d'événements, le temps d'exécution des codes devenait de plus en plus long et entrainait parfois des fermetures du logiciel R.

Au début de la mission, j'ai eu aussi quelques problèmes de compréhension des scripts de la stagiaire précédente et de réussir de me les approprier afin de continuer le travail attendu. Toutefois, ce fut un court moment et pendant ce temps je pouvais réaliser les autres tâches de cette mission.

5.Conclusion générale

Avant de commencer ce stage, j'avais quelques appréhensions. Je me demandais si j'allais être à la hauteur des attentes et si je pourrais réaliser les missions qui me seraient confiées. De plus, j'avais une certaine inquiétude concernant ce qui m'attendait dans le milieu professionnel et je me demandais si ce domaine allait réellement me plaire.

Mais au cours de ce stage, je me suis rendue compte que j'étais à la hauteur des attentes et que je pouvais accomplir les missions qui m'étaient confiées. De plus, je me suis très vite adaptée au rythme de travail et à l'environnement de l'institution. Tout cela m'a fait comprendre que je n'avais pas à avoir peur et que le monde professionnel n'était pas aussi effrayant que je l'avais imaginé.

En ce qui concerne le domaine de la sociologie, j'ai constaté que plus le temps passait, plus j'appréciais ce que je faisais. Et en approfondissant mes recherches pour avancer dans mon travail, plus mon intérêt pour ce domaine augmentait, alors que celui-ci m'était encore inconnu il y a quelques mois. J'aimerais désormais explorer les différentes possibilités qui pourraient se présenter dans ce domaine et me renseigner sur les démarches à suivre pour m'y engager davantage.

Au niveau technique, j'ai découvert de nouveaux procédés sur R et à force de me documenter, j'ai pu améliorer au fur et à mesure les scripts que j'ai produits. Et, j'ai aussi appris à me réorganiser quand il y avait des changements de dernières minutes à faire et à noter ce que je devais faire et ce que j'avais fait.

6.Bibliographie

6.1. Références

[1] E. Lazega, *Que sais-je ? - Réseaux sociaux et structures relationnelles* (Presses Universitaires de France, Paris, 2014)

[2] E.D. Kolaczyk, G. Csárdi, *Statistical Analysis of Network Data with R* (Springer, New York, 2014)

[3] D. Schoch, '*Basic Network Analysis in R*': https://mr.schochastics.net/

https://github.com/schochastics

6.2. Lexique

Analyse Multi-niveaux : Une méthode statistique permettant d'étudier des données structurées en niveaux hiérarchiques

Centralité : Mesures aidant à identifier les nœuds les plus influents ou critiques dans un réseau

Projection : consiste à transformer le réseau bipartite en un réseau monopartite en créant des liens entre les nœuds du même groupe (exemple : acteurs), basés sur leurs interactions avec les nœuds de l'autre groupe (exemple : événements).

Réseau bipartite : Un type de réseau où les nœuds sont divisés en deux groupes distincts, et les liens ne peuvent exister qu'entre les nœuds de groupes différents

Ucinet : Un logiciel permettant l'analyse des données des réseaux sociaux, mesure de centralité, identification de sous-groupes.

6.3. Table des illustrations

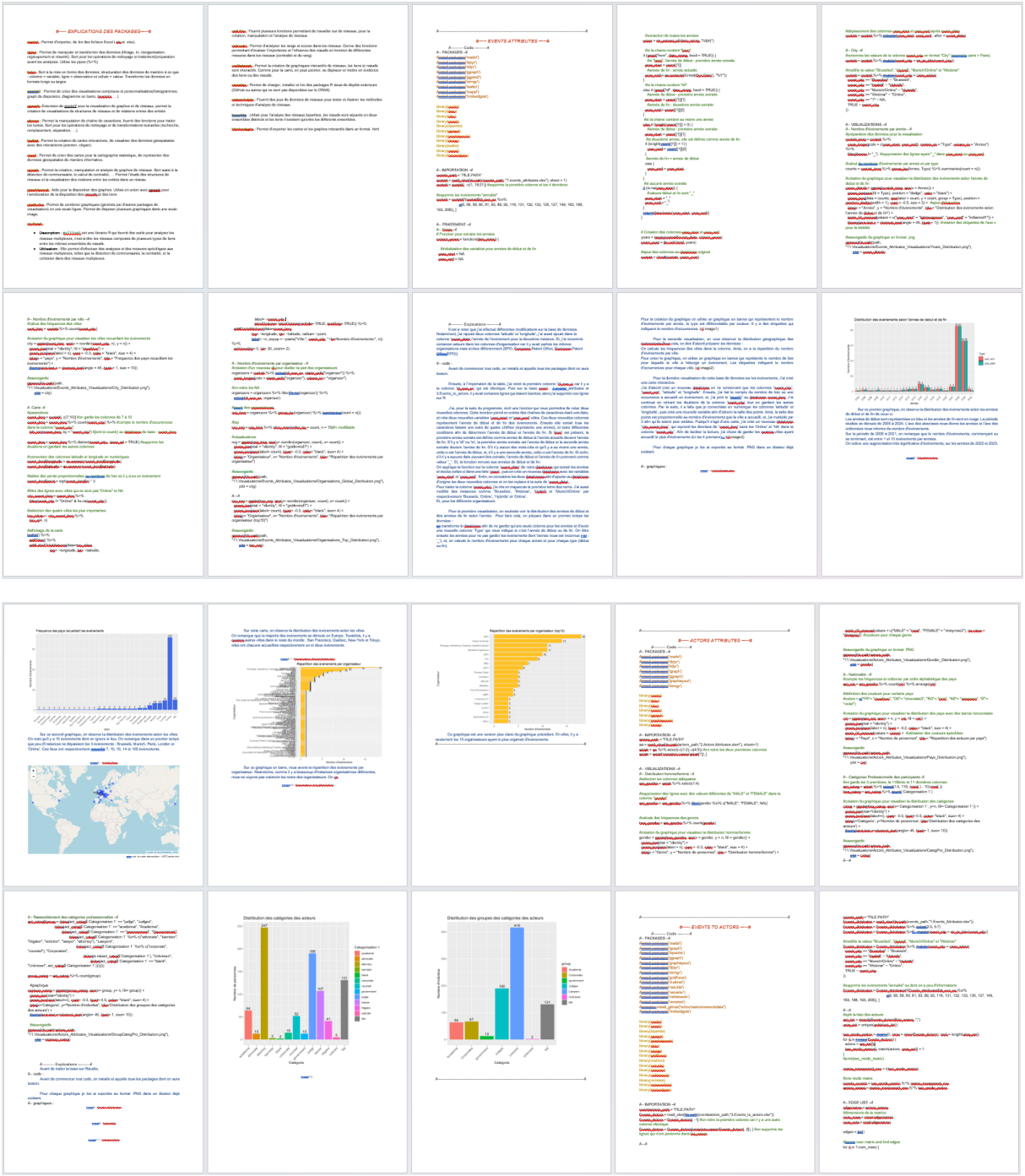
Figure 1 : Organigramme général de Sciences Po 7
Figure 2 : Organigramme du CSO 7
Figure 3 : Livrables à rendre 11
Figure 4 : Extrait de la base de données des événements 13
Figure 5 : Extrait de l'importation des données "Acteurs" 14
Figure 6 : Nettoyage de la base "Acteurs" 14
Figure 7 : Graphique représentant la distribution du genre des "Acteurs" 15
Figure 8 : Graphique représentant la distribution de la catégorie professionnelle des acteurs 15
Figure 9 : Extrait de la base faisant lien entre les événements et les acteurs 16
Figure 10 : Code pour la visualisation du réseau des acteurs 17
Figure 11 : Projection du réseau acteurs 17
Figure 12 : Analyse Multi-niveau événements/acteurs 18

Figure 13 : Rapport sur la mission UPC 24
Figure 14 : Extrait du code de la base événement, fonction de traitement 25
Figure 15 : Extrait du code pour la création des projections 26
Figure 16 : Code pour construire l'analyse multi-niveaux 27

7.Annexes

Annexe 1 : Rapport sur le projet UPC

Il est important de noter que lors de l'écriture de ce rapport de stage, l'écriture du rapport sur les données UPC n'est pas encore abouti.



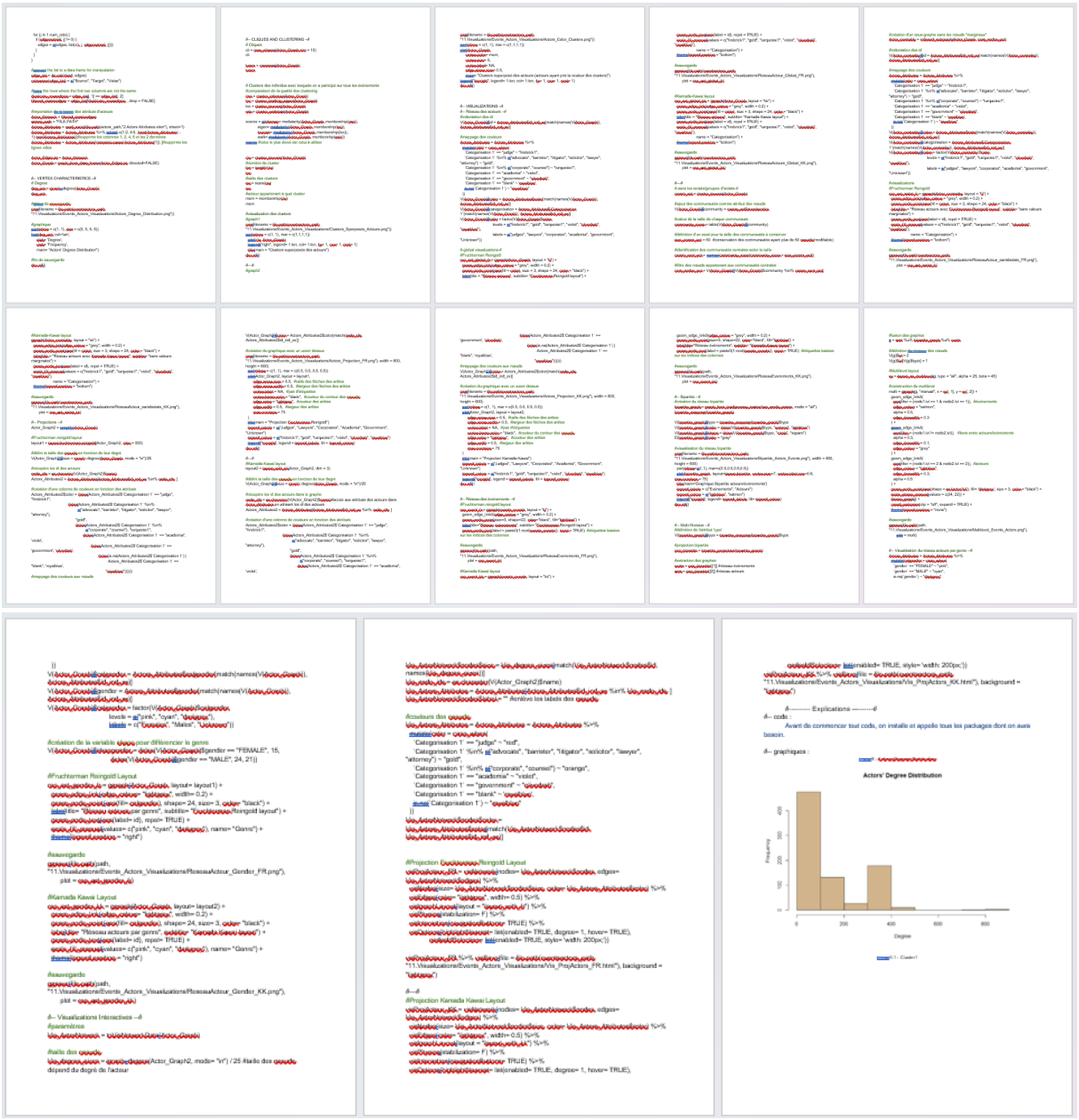


Figure 14 : Rapport sur la mission UPC

Annexe 2 : Extrait des scripts pour la mission UPC

```
########## TRAITEMENT ##########
#-- Years --#
  #Fonction pour extraire les années
extract_years = function(date_string) {

    #initialisation des variables pour années de début et de fin
  year_start = NA
  year_end = NA

    #extraction de toutes les années
  years = str_extract_all(date_string, "\\d{4}")

    #si la chaine contient "from"
  if (grepl("from", date_string, fixed = TRUE)) {
      #si "from", l'année de début : première année extraite
    year_start = years[[1]]
      #année de fin : année actuelle
    year_end = as.numeric(format(Sys.Date(), "%Y"))
  }
    #si la chaine contient "till"
  else if (grepl("till", date_string, fixed = TRUE)) {
      #année de début : première année extraite
    year_start = years[[1]][1]
      #année de fin : deuxième année extraite
    year_end = years[[1]][2]
  }
   #si la chaine contient au moins une année
  else if (length(years[[1]]) > 0) {
      #année de début : première année extraite
    year_start = years[[1]][1]
     #si deuxième année, elle est définie comme année de fin
    if (length(years[[1]]) > 1) {
      year_end = years[[1]][2]
    }
     #année de fin = année de début
    else {
      year_end = year_start
    }
  }
   #si aucune année extraite
  if (is.na(year_start)) {
      #valeurs debut et fin sont "_"
    year_start = "_"
    year_end = "_"
  }
  return(data.frame(year_start, year_end))
}
```

Figure 15 : Extrait du code de la base événement, fonction de traitement

```r
#-----------------#
#-- Projections --#
Actor_Graph2 = simplify(Actor_Graph)

#Fruchterman.reingold layout
layout1 = layout.fruchterman.reingold(Actor_Graph2, niter = 500)

#défini la taille des noeuds en fonction de leur degré
V(Actor_Graph2)$size = igraph::degree(Actor_Graph, mode = "in")/25

#récupère les id des acteurs
node_ids = as.character(V(Actor_Graph2)$name)
Actors_Attributes2 = Actors_Attributes[Actors_Attributes$id_ind_xx %in% node_ids, ]

#création d'une colonne de couleurs en fonction des attributs
Actors_Attributes2$color = ifelse(Actors_Attributes2$`Categorisation 1` == "judge", "firebrick1",
                                  ifelse(Actors_Attributes2$`Categorisation 1` %in%
                                           c("advocate", "barrister", "litigator", "solicitor", "lawyer", "attorney"),
                                         "gold",
                                         ifelse(Actors_Attributes2$`Categorisation 1` %in%
                                                  c("corporate", "counsel"), "turquoise1",
                                                ifelse(Actors_Attributes2$`Categorisation 1` == "academia", 'violet',
                                                       ifelse(Actors_Attributes2$`Categorisation 1` == 'government', 'olivedrab',
                                                              ifelse(is.na(Actors_Attributes2$`Categorisation 1`) |
                                                                       Actors_Attributes2$`Categorisation 1` == "blank", 'royalblue',
                                                                     "royalblue"))))))
#mappage des couleurs aux nœuds
V(Actor_Graph2)$color = Actors_Attributes2$color[match(node_ids, Actors_Attributes2$id_ind_xx)]

#création du graphique avec un zoom dessus
png(filename = file.path(eventsactors_path, "11.Visualizations/Events_Actors_Visualizations/Actors_Projection_FR.png"), width = 800, height = 600)
  par(mfrow = c(1, 1), mar = c(0.5, 0.5, 0.5, 0.5))
  plot(Actor_Graph2, layout = layout1,
       edge.arrow.size = 0.5,  #taille des flèches des arêtes
       edge.arrow.width = 0.5,  #largeur des flèches des arêtes
       vertex.label = NA,  #pas d'étiquettes
       vertex.frame.color = "black",  #couleur du contour des noeuds
       edge.color = "lightgrey",  #couleur des arêtes
       edge.width = 0.5,  #largeur des arêtes
       max.overlaps = 75
  )
  title(main = "Projection Fruchterman Reingold")
  legend_labels = c("Judges", "Lawyers", "Corporates", "Academia", "Government", "Unknown")
  legend_colors = c("firebrick1", "gold", "turquoise1", "violet", "olivedrab", "royalblue")
  legend("topright", legend = legend_labels, fill = legend_colors)
dev.off()

```

Figure 16 : Extrait du code pour la création des projections

```r
#-- MULTINIVEAUX --#
  #définition de l'attribut 'type'
v(bipartite_graph)$type = bipartite_mapping(bipartite_graph)$type

  #projection bipartite
proj_bipartite = bipartite_projection(bipartite_graph)

  #extraction des graphes (projections)
evnts = proj_bipartite[[1]] #réseau événements
acts = proj_bipartite[[2]] #réseau acteurs

  #fusion des graphes
g = acts %u% bipartite_graph %u% evnts

  #définition du niveaux des nœuds
v(g)$lvl = 2
v(g)$lvl[v(g)$type] = 1

  #Multilevel layout
xy = layout_as_multilevel(g, type = "all", alpha = 25, beta = 45)

  #construction du multilevel
multi = ggraph(g, "manual", x = xy[, 1], y = xy[, 2]) +
  geom_edge_link0(
    aes(filter = (node1.lvl == 1 & node2.lvl == 1)), #événements
    edge_colour = "salmon",
    alpha = 0.5,
    edge_linewidth = 0.3
  ) +
  geom_edge_link0(
    aes(filter = (node1.lvl != node2.lvl)),  #liens entre événements et acteurs
    alpha = 0.3,
    edge_linewidth = 0.1,
    edge_colour = "grey"
  ) +
  geom_edge_link0(
    aes(filter = (node1.lvl == 2 & node2.lvl == 2)),  #acteurs
    edge_colour = "lightblue",
    edge_linewidth = 0.3,
    alpha = 0.5
  ) +
  geom_node_point(aes(shape = as.factor(lvl)), fill= 'darkgrey', size = 3, color = "black") +
  scale_shape_manual(values = c(24, 22)) +
  theme_graph() +
  coord_cartesian(clip = "off", expand = TRUE) +
  theme(legend.position = "none")

#sauvegarde
ggsave(file.path(eventsactors_path, "11.Visualizations/Events_Actors_Visualizations/Multilevel_Events_Actors.png"),
       plot = multi)
```

Figure 17 : Code pour construire l'analyse multi-niveaux

8. Démarche Portfolio

Ce stage m'a aidée à renforcer certains apprentissages critiques et composantes essentielles présents dans les compétences du BUT Science des Données :

- Pour la première compétence « Traiter des données à des fins décisionnelles », je comprends mieux la nécessité de tester, de corriger et documenter son programme, j'ai su intervenir dans les étapes du cycle de vie de la donnée et écrire un programme structuré et documenté.

Pendant la période de stage, j'ai à plusieurs reprises modifié mon code afin de corriger des erreurs ou pour améliorer les visualisations. De plus j'ai rédigé des commentaires explicatifs pour chaque étape des codes que j'ai écrit. La documentation des codes est nécessaire afin que d'autres personnes puissent les comprendre et les réutiliser.

- Dans la deuxième compétence « Analyser statistiquement des données », je comprends mieux le fait de prendre en compte le contexte d'étude et mettre en évidence les tendances et informations principale. J'ai aussi su identifier et mettre en œuvre les techniques adaptées aux attentes du client et je comprends mieux grâce au stage l'intérêts des analyses multivariées afin de synthétiser et résumer l'information résumé par plusieurs variables.

Lors du stage, j'ai construit des projections à partir de données bipartites et j'ai ajouté une variable en plus afin de mettre en évidences des informations qui peuvent être pertinentes lors de l'analyse. De plus, comme cité précédemment, j'ai adapté mes codes et les techniques de nettoyage, de traitement et de visualisation afin de répondre le plus aux attentes.

- Pour la troisième compétence « Valoriser une production dans un contexte professionnel », j'ai dû m'exprimer correctement que ce soit en français ou en anglais, à l'oral ou à l'écrit. J'ai aussi mieux compris la nécessité de choisir des indicateurs pertinents pour communiquer sur les résultats.

Lors de ce stage, j'ai rédigé un rapport explicatif des codes et des graphiques sortants, et celui-ci devait être écrit dans un français correct afin d'améliorer la compréhension de mes codes pour des étudiants qui l'utiliseront. Et, lors de la visualisation, j'ai dû faire des choix pour les variables à utiliser.

- Dans la quatrième compétence « Développer un outil décisionnel », j'ai renforcé ma mise en œuvre de structuration des données adaptée à leurs caractéristiques, mes réalisations de solutions de visualisation spécifiques aux données métier et j'ai mieux perçu les enjeux de l'automatisation et de l'interopérabilité d'un ensemble de tâches.

Par exemple, j'ai adapté mes scripts et mes visualisations selon les caractéristiques des données que j'utilisais. Et, au moment de me première mission, j'ai compris qu'il était nécessaire d'automatiser mon code comme le même procédé était à chaque fois appliqué ces les bases.