devenez.fr
GROUPE FITEC

Date: Juillet 2020

Rapport Technique projet fil rouge

Prédiction de l'évolution du télétravail et impact de la crise sanitaire

Groupe 2:

Claude El tannoury

Mamadou Camara

Benoît Bouillaud

Awa Thiam

Responsable Formation:

Yassine El maataoui

Date:

9/07/2020

Table des matières

Introduction

Le télétravail a commencé aux États-Unis en 1950. Des questions se posent sur son évolution et sur l'efficacité des télétravailleurs. Différents paramètres sont à prendre en compte : pays, entreprises accordant un télétravail, sexe, âge, etc...

Par ailleurs, plus de 5,2 millions de salariés ont été en télétravail en France durant la crise sanitaire du covid-19. Nous souhaitons ainsi étudier l'effet de celle-ci sur les télétravailleurs en fonction de leurs différentes conditions de vie. L'identification des facteurs influençant le télétravail nous permettra d'élaborer de nouvelles stratégies et conditions plus adaptées à celui-ci.

I. Utilisation des outils collaboratifs

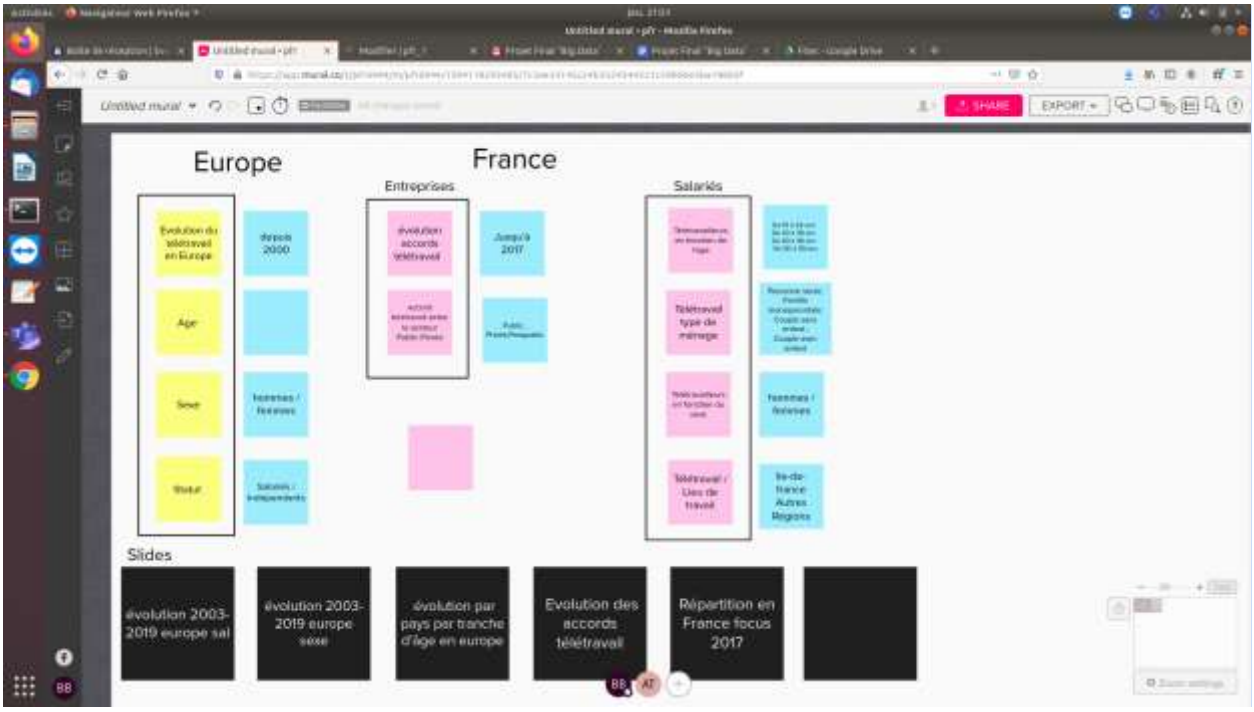
L'agilité définit une approche de gestion de projet aux antipodes des autres méthodes plus traditionnelles de type cycle en V. S'il est courant de raisonner davantage en mode "gestion de produit", ce dernier sera en l'espèce notre étude sur le télétravail. Les pratiques agiles ont démontré tout leur intérêt dans la réalisation de projets : une meilleure communication, une visibilité accrue

sur l'avancée du projet, mais aussi une souplesse et une réactivité accrues. C'est pour ces raisons que nous avons souhaité appliquer cet esprit agile en utilisant les outils collaboratifs suivants :

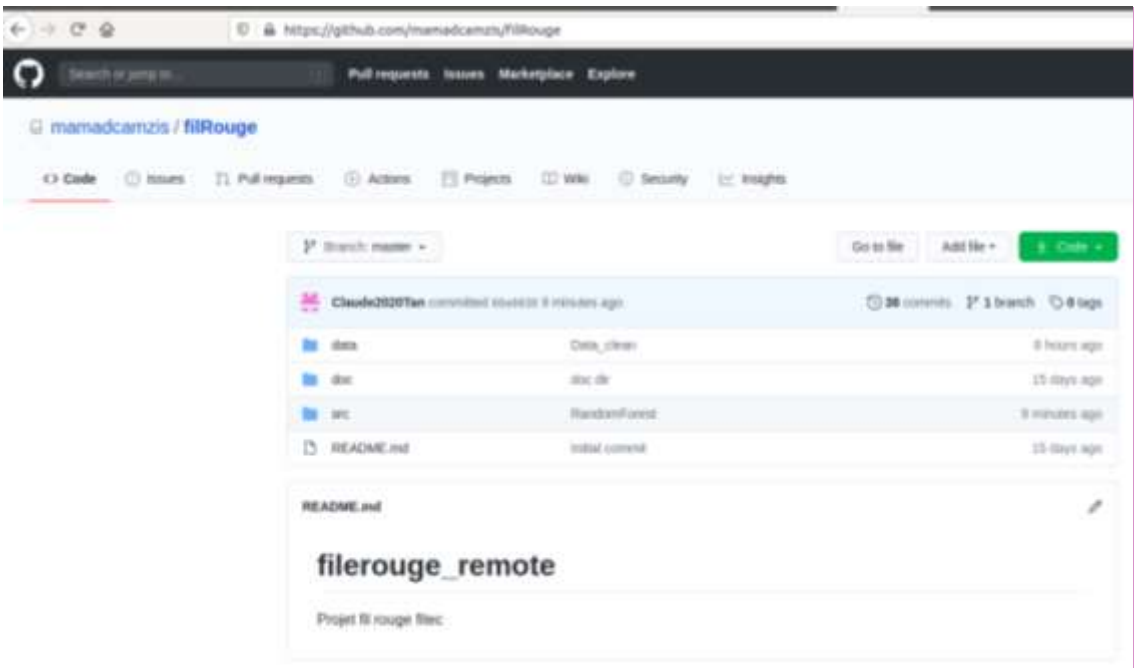
- **Google Collaboratory** : **Google Sheet** nous a permis de référencer les différents liens sources de données. Nous avons ainsi établi une catégorisation par thème / importance, ainsi qu'une répartition de ces liens entre les membres du groupe de travail. **Google Docs** nous a quant à lui permis de construire notre dossier de Projet final au fur et à mesure de l'avancement de la formation.

- **Trello**: Outils de gestion de projet incontournable, nous avons pu l'utiliser afin de créer un tableau digital nous permettant de répartir les tâches au sein de notre groupe et d'en actualiser le déroulement. Ci-dessous une copie d'écran :

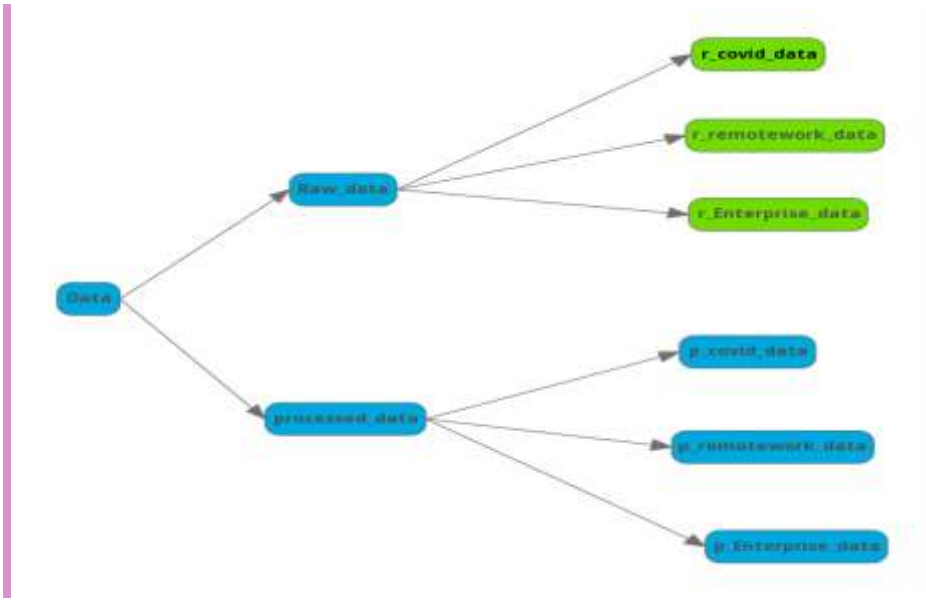
- **Mural**: Nous avons découvert cet outil assez simple, mais très utile lors de notre cours sur l'animation d'un atelier à l'aide de la méthode agile. Parmi les différents outils présentés par Patrick, celui-ci nous a paru intéressant et par ailleurs très intuitif. On a ainsi pu déposer nos idées rapidement sur des post-its, organiser notre pensée ainsi que notre travail, tout en partageant de manière interactive. Ci-dessous une copie d'écran :



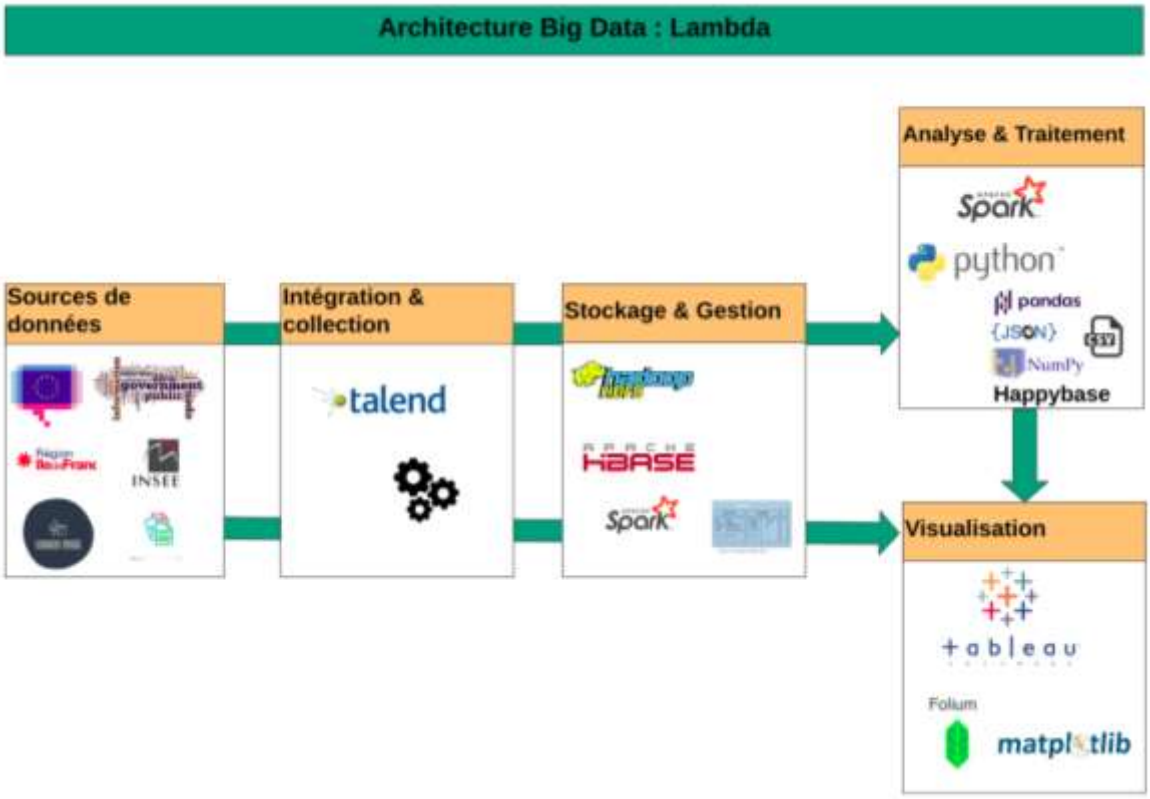
- **Github** : Certainement la plus grande plateforme d'hébergement de projets Git au monde, cet outil que certains membres avaient déjà rencontré lors de leurs expériences en milieu professionnel, nous a été pratique. Cela a fluidifié notre collaboration sur ce projet en optimisant l'accès à l'information et le versionning. Certains d'entre nous ont découvert cette plateforme dans le cadre de la formation, et ont donc pu bénéficier de la mise en application approfondie au cours du projet et du transfert de compétences entre stagiaires. Voici le lien github où nos données/Jupyter notebooks se trouvent : Projet_BIG_Data. Ci-dessous une copie d'écran :



- **Mindmup** : Application de cartographie mentale qui nous a permis d'échanger et d'organiser nos idées de manière interactive. A titre d'exemple, nous l'avons utilisé pour catégoriser nos données dans des folders sur Github (schéma ci-dessous):



II. Architecture générale de système logiciel



Architecture générale

Ce schéma représente l'architecture générale de notre système logiciel.

Les données sont récupérées depuis les différentes sources respectant le Règlement général sur la protection des données (RGPD), ensuite sont stockées sous **HDFS**, nettoyées et traitées via **Python** et ses bibliothèques (pandas, numpy, PySpark, csv, json...) et **Talend**.

Ensuite, nous avons utilisé Tableau pour la visualisation et l'analyse de nos données, pour la création de Dashboard et d'un storytelling de toutes nos données.

Après avoir terminé la visualisation et l'analyse de nos données, nous avons utilisé la librairie de **"Scikit Learn"** pour faire la prédiction sur nos données.

Cette Architecture que nous avons conçue est considérée "**Architecture Lambda"** comme les données sont stockées avant d'être utilisées.

En absence de serveur collaboratif nous permettant de stocker toutes les données dans un seul data Lake, nous avons donc partagé les données par thématiques entre nous tout en suivant l'architecture Big Data, et en procédant finalement à une phase de consolidation de toutes nos données via tableau.

III. Récupération des données & Web Scraping

La récupération des données s'est faite après une recherche des différentes sources sur plusieurs sites web et aussi dans les OpenData de l'Ile-de-France, de la France, de la Belgique, de l'Europe et des Etats-Unis.
Les sources de données conservées après une première analyse basée sur la pertinence par rapport à notre sujet et respectant la RGPD sont les suivantes :

- Covid USA : contient des données sur une étude menée aux Etats-Unis sur le télétravail dans le contexte de la crise Sanitaire (Covid-19)
- Evolution du télétravail en Europe en 2018 : nous avons des données sur l'évolution du télétravail dans les Pays de l'Europe en fonction de différents paramètres
- Covid France : ces données proviennent de l'OpenData de Corana-work et sur une enquête portant le télétravail pendant le confinement
- Télétravail en France 2017 : nous avons des données liées au télétravail en France en 2017 suivant l'âge, le sexe, la catégorie socio-professionnelle, les distances domicile-travail etc.
- Télétravail en Europe : ce sont les données sur l'évolution du télétravail de 1992 à 2019 selon plusieurs paramètres : la localisation, le sexe, l'âge, le statut professionnel, la fréquence du recours au télétravail
- Accord des entreprises sur le télétravail : ces données contiennent l'accord des entreprises sur le télétravail jusqu'en 2017 avec les dates des accords, le secteur etc.

Après cette phase de collecte, nous avons utilisé des outils tels que **Talend** pour télécharger des données depuis des sites web avec l'écriture de jobs, le **Scraping** avec Python.

Pour le scraping, nous avons utilisé la bibliothèque **BeautifulSoup** pour récupérer les données sur les sites.

Nous avons scrapé les données contenues dans les tableaux et les avons enregistré dans un fichier csv. Dans notre projet, nous avons récupéré des données dans des formes de tableaux différents où la structure des balises est complexe et le chargement des pages web est dynamique. Ces paramètres rendent le scraping difficile. Mais malgré cela, nous avons récupéré les données souhaitées.

Nous avons aussi scrapé les données de l'API 'open data soft' relatives au covid19, puis nous les avons sauvegardé dans un premier temps sous un fichier json en utilisant les librairies de python **request** et **json**. (le code src pour scraper ces données est sous le nom Scraping Data API).

Vous retrouverez l'intégralité des codes source dans le repo Github.

IV. Stockage des données brutes dans HDFS

Pour stocker les données brutes nous avons créé un **datalake** dans HDFS en mettent en place un cluster avec un **namenode**, trois **datanodes**.

Dans HDFS, nous avons la cohérence et la distribution, ce qui correspond à notre besoin dans le cadre de notre projet.

Les Étapes suivantes nous ont permis de mettre en place le cluster :

- Installation de HDFS

  Téléchargement HDFS

  ```
  wget http://apache.crihan.fr/dist/hadoop/common/hadoop-2.8.1/hadoop-2.8.1.tar.gz
  ```

  Dezipper

  ```
  tar xzf hadoop-2.8.1.tar.gz
  ```

- Configuration du port d'écoute dans core-site.xml

```
<configuration>
   <property>
      <name>fs.defaultFS</name>
      <value>hdfs://localhost:9000</value>
   </property>
</configuration>
```

- Création du namenode

  Répertoire de stockage des données du namenode

  ```
  Mkdir  /code/hdfs/datanode
  ```

  Configuration du namenode

  ```
  <configuration>
     <property>
        <name>dfs.name.dir</name>
        <value>/home/regis/code/hdfs/namenode/</value>
     </property>
  </configuration>
  ```

- Creation des trois datanodes

  Configuration datanode1

  ```
  <configuration>
     <property>
        <name>dfs.datanode.address</name>
        <value>0.0.0.0:50011</value>
     </property>
     <property>
        <name>dfs.datanode.http.address</name>
        <value>0.0.0.0:50076</value>
     </property>
     <property>
        <name>dfs.datanode.ipc.address</name>
        <value>0.0.0.0:50021</value>
     </property>
     <property>
        <name>dfs.data.dir</name>
        <value>/home/regis/code/hdfs/datanode1/</value>
     </property>
  </configuration>
  ```

- Envoie des données dans le cluster HDFS

Ici nous avons utilisé hdfs (librairie de python pour charger les données). Voir le code du script dans le repo Git suivant script-python .

Deux méthodes sont utilisées pour stocker les données sous HDFS.

Certains parmi nous a installé HDFS sur leur machine local, d'autres ont utilisé une image docker de HDFS.

Installation hdfs en local

Utilisation d'image docker de HDFS

Les étapes suivantes nous ont permis de stocker les données sous HDFS:

Lancement du container

```
sudo docker run --hostname=quickstart.cloudera --privileged=true -t -i -v
/home/fitec/Documents/cloudera:/src --publish-all=true -p 8888:8888 -p 80:80 -p 7180:7180
cloudera/quickstart /usr/bin/docker-quickstart
```

Démarrage de Cloudera

```
cd /home/cloudera
./cloudera-manager – – express
```

Démarrage de cloudera , HDFS , Hive, Hbase, Yarn, zookeeper dans Cloudera (http://localhost:7180/)

- Stockage de données sur hdfs Via Python :
- Utilisation de librairie Pyspark,
- Création d'une application Pyspark
- Stockage des données sur HDFS,

Les données de Corona Work sont stockées sur HDFS via le code suivant 3_Depot_jsonfile_HDFS.ipynb.

Lien vers ce code est Code Depot HDFS Via Pyspark.

V. Nettoyage de données avec Python ou Talend

Placer la France dans un contexte européen nous a paru indispensable : étudier le télétravail en France est intéressant, le mettre en perspective en le comparant à nos voisins en Europe l'est d'autant plus. Différents liens obtenus notamment sur "data.europa.eu" nous ont permis d'obtenir des données au format CSV principalement.

La première étape après avoir récupéré le dataset sur l'évolution du télétravail en Europe a été de bien comprendre les informations s'y trouvant. Celles-ci énumèrent les personnes en emploi et travaillant à distance, en pourcentage du total de l'emploi, par sexe, âge et statut professionnel. 35 pays européens sont référencés dans l'étude. En revanche, tous n'offrent pas la même complétude en matière de données. Cela nous conduira plus tard, dans un souci de cohérence et d'exactitude, à privilégier les années à compter de 2003 dans notre travail de visualisation.

Quatre colonnes feront l'objet d'une attention toute particulière :

- Le Sexe :
    - T : total
    - M : males
    - F : females

Nous avons instinctivement pensé que le T correspondait au total des deux sexes et donc à la moyenne. Après vérification, nous avons pu confirmer cela sur certains pays. Toutefois, après avoir poussé ce contrôle un peu plus loin, nous nous sommes aperçus de nombreuses incohérences pour d'autres pays. En conséquence, nous avons décidé de sélectionner uniquement les colonnes Hommes et Femmes et de ne pas considérer la colonne Total.

- La Fréquence avec 3 options proposées :
    - NVR : never (jamais)
    - SMT : sometimes (quelquefois)
    - USU : usually (régulièrement)

Après analyse, nous avons constaté que les valeurs contenues dans la colonne "SMT" étaient les plus pertinentes et écartaient toute incohérence.

- L'Âge avec un grand nombre d'options disponibles.

Après avoir consulté la documentation du site europa.eu, nous avons pu déchiffrer chacune d'entre elles. Les âges sont ainsi principalement répartis sous forme d'intervalles. Ces derniers se superposant, il nous a fallu sélectionner 3 d'entre eux afin de couvrir l'ensemble de la population active de manière intelligente (15-24 / 25-49 / 50-64). Nous avons, dans d'autres cas et dans la création de visualisation, utilisé la valeur globale "Y_G15" qui représente les personnes de 15 ans et plus.

Statut : 9 statuts professionnels sont présentés :

- 'CFAM' = Travailleurs familiaux collaborant à l'entreprise familiale
- 'EMP' = Personnes occupées
- 'NCFAM' = Personnes occupées sauf travailleurs familiaux collaborant à l'entreprise familiale
- 'NRP' = Sans réponse
- 'NSAL' = Personnes occupées sauf salariés
- 'SAL' = salarié
- 'SELF' = Travailleurs indépendants
- 'SELF_NS' = Travailleurs indépendants sans salariés
- 'SELF_S' = Travailleurs indépendants avec salariés (employeurs)

Si, pour notre travail de projection via le machine learning, nous avons utilisé le maximum de données et donc l'ensemble des différents statuts professionnels, nous nous sommes plutôt concentrés sur le statut de salariés pour tout ce qui est visualisation. En effet, le statut de salariés étant encore le plus important dans notre société actuelle (25,5 millions en France au 4ème trimestre 2019), nous trouvions cela plus pertinent.

Ce traitement a été fait sous Python. Le nettoyage ainsi que le réagencement des données a principalement été obtenu en utilisant les bibliothèques incontournables que sont Pandas ou encore Numpy.
Lien vers le notebook correspondant : PFR_europe_lien_3.ipynb

Voici une description des colonnes utilisées dans les données sur le télétravail en France en 2017 et sur les accords des entreprises :

- Distance_domicile_travail: distance entre le domicile et le lieu de travail
- Pourcentage: pourcentage des personnes
- Entreprise: noms des entreprises
- Secteur: secteur d'activité des entreprises

- Date_accord :date de signature des accords
- Ensemble: ensemble des salariés
- Cadres: les cadres
- Employes: les salariés
- Ouvriers: les ouvriers
- Lieu_travail : lieu de travail des salariés

L'évolution du télétravail dans les différents pays et particulièrement en France pose des questions : quel est l'impact du covid 19 sur le télétravail, les télétravailleurs ? Quelles sont les conditions maximisant l'efficacité d'un télétravailleur? Les données récupérées depuis l'API répondent à ses différentes questions.
La structure des données était compliquée à parser. Cette complexité réside dans le fait que ces données correspondent à des ensembles de sous dictionnaires à l'intérieur de dictionnaires. En gros, il y avait plusieurs couches de dictionnaires à l'intérieur d'un seul dictionnaire. Le parsing des données était difficile et malgré la complexité de la structure des données, nous avons pu réorganiser ces dernières et les sauvegarder dans un fichier csv.

Ces données correspondent à un questionnaire réalisé durant le confinement. Les questions posées sont:

- Vous êtes : (homme, femme)
- Quel âge avez-vous ?
- Votre employeur est une : (Petite, moyenne, entreprise), association,
- Actuellement vous êtes : statut professionnel
- Dans votre métier vous êtes amené.e à : travail d'équipe, cadre manager, seul
- combien de temps mettez-vous à vous rendre sur votre lieu de travail ?
- Avant le confinement, à quelle fréquence aviez-vous recours au télétravail ?
- Où êtes-vous confiné.e ?
- Quelle est la surface de votre lieu de confinement en m² ?
- Avez-vous accès à un espace extérieur ? Combien de fois êtes-vous sorti.e au cours de la dernière semaine ?
- Combien de personnes au total vivent actuellement dans votre lieu de confinement ?
- Parmi vous, combien sont dans leur résidence principale ? Parmi ces personnes, combien sont des enfants de moins de 12 ans ?

- Est-ce que vous diriez que le confinement améliore vos relations avec vos co-confinés ?
- Pouvez-vous vous isoler dans une pièce du logement pour travailler sans être dérangé.e en cas de besoin ?
- Votre équipement en bureautique est-il satisfaisant
- Les consignes de télétravail communiquées par votre employeur vous paraissent-elles efficaces ?
- Depuis le confinement, la fréquence des interactions avec vos collègues a-t-elle changée ?
- si le travail devient pour toujours ???
- Depuis le confinement, comment jugeriez-vous votre concentration dans le travail ?
- Quelles sont les causes importantes de distraction?
- Depuis le confinement, votre temps de travail effectif a-t-il changé ?
- Avez-vous réaménagé vos horaires de travail habituels ?
- Pensez-vous que le télétravail confiné ait un impact sur l'efficacité du travail de votre équipe/projet ?
- Avez-vous progressé sur l'utilisation des outils collaboratifs (visioconférence, messagerie, documents partagés…) :
- Lorsque la situation sera revenue à la normale, aimeriez-vous continuer à télétravailler ?
- Selon vous, quels sont les conséquences positives du télétravail en mode "confiné" ?
- Selon vous, quels sont les conséquences négatives du télétravail en mode "confiné" ?
- Comment évaluez-vous vos conditions de travail AVANT le confinement ?
- Comment évaluez-vous vos conditions de travail APRES le début du confinement ?
- Depuis le confinement, vous faites de l'activité physique :
- Depuis le confinement, vous mangez :
- Depuis le confinement, vous buvez de l'alcool
- Depuis le confinement, vous prenez soin de votre apparence :
- Depuis le confinement, vous êtes en contact avec votre famille et vos amis :
- Depuis le confinement, vous suivez l'actualité :
- Diriez-vous que vous avez des inquiétudes sur la pérennité de votre emploi à cause du Covid-19 ?
- Est-ce que vous avez des contrariétés particulières liées à la propagation du Covid 19 et ses conséquences, si oui lesquelles ?

- Combien de personnes proches avec des symptômes du Covid 19 connaissez-vous ? Personnes malades ?
- Comment évaluez-vous votre bien être AVANT le début du confinement ?
- Comment évaluez-vous votre bien être APRES le début du confinement ?
- Quelles sont vos nouvelles priorités ?
- Région
- geom point : long, latitude

La majorité des réponses à ces questions est :

- soit une évaluation entre 1 et 5 (correspondant respectivement au moins fort et au plus fort)
- soit des données catégoriques
- Et quelques-unes correspond à des entiers

Les librairies utilisées pour analyser et nettoyer ces données sont : json, pandas, matplotlib, seaborn, Sickit Learrn.

Les étapes à suivre pour l'analyse de ces données sont :
1-Transformer le dictionnaire en un dataFrame lisible ne contenant que les données qui peuvent nous intéresser
2- Supprimer les colonnes doublées (Vous êtes), les colonnes ne contenant pas assez de valeurs (région, latitude, longitude) et les colonnes inutiles...
3- Remplacer les Nan par les modes si c'est une évaluation ou catégories, et par médiane ou moyenne si c'est des integers.
Dans notre cas, nous avons remplacé par les variables integers et médianes car la distribution était asymétrique et les données catégoriques (non numériques) par mode.
4- Transformer les string en 0, 1 , 2 etc.. à l'aide de Label encoder et les ajouter au dataFrame. Cette étape est indispensable pour l'analyse et la prédiction dans les étapes suivantes de notre étude.

Le Code correspondent au nettoyage et à l'analyse des résultats de ce questionnaire est fait dans un notebook python accessible sous ce lien Nettoyage des données

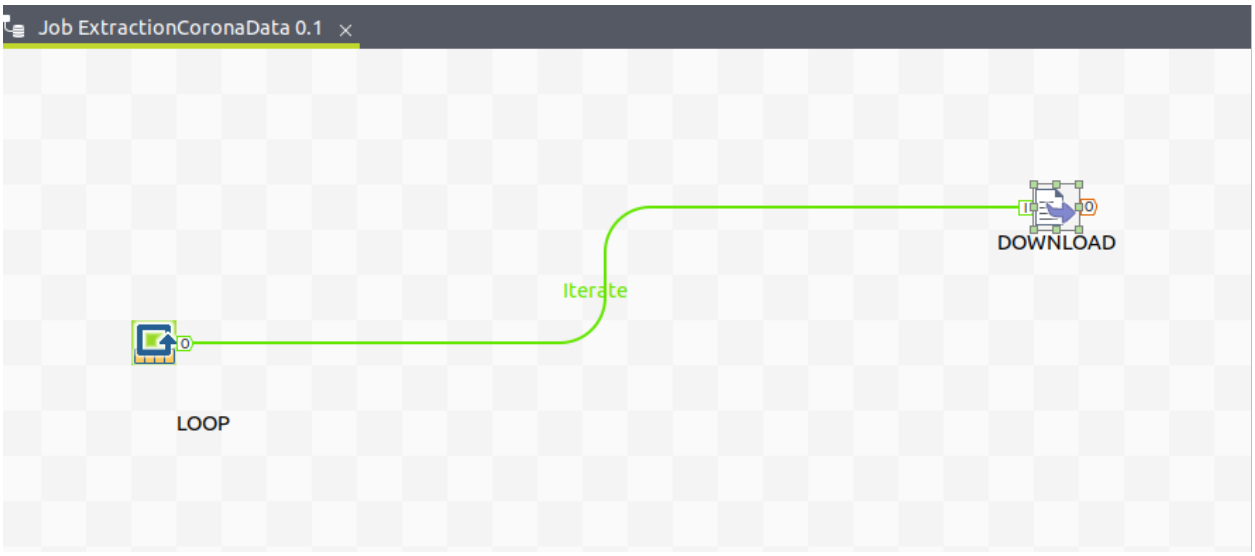
Pendant la phase de recherche de données liées au télétravail et à la pandémie du COVID-19, nous avons trouvé des données à ce sujet dans le cadre des Etats-Unis.
Ces données ont été obtenues grâce à un questionnaire qui a été soumis aux citoyens américains. Il fallait donc répondre à la question suivante: “Avez-vous commencé à faire du télétravail durant les 4 dernières semaines”.
Il y a six possibilités de réponses aux choix:

- “Je continue de me déplacer pour aller travailler”
- “J’ai été récemment licencié ou mis à pied”
- “Habituellement je me déplace à mon lieu de travail, maintenant je fais du télétravail”
- “Habituellement je fais du télétravail et je continue à le faire”
- “Habituellement je fais du télétravail et maintenant je me déplace à mon lieu de travail”
- “Aucune de ces réponses / Je travaille sans salaire

Partant de ça, nous avons importé les données en utilisant talend. Pour ce faire, nous avons utilisé deux composantes et un lien `iterate` entre les deux composantes

- *tForeach* pour faire une boucle sur tous les fichiers
- *tFileFetch* pour télécharger un fichier via un lien

Voici la configuration du Job talend qui nous a permis de télécharger le fichier

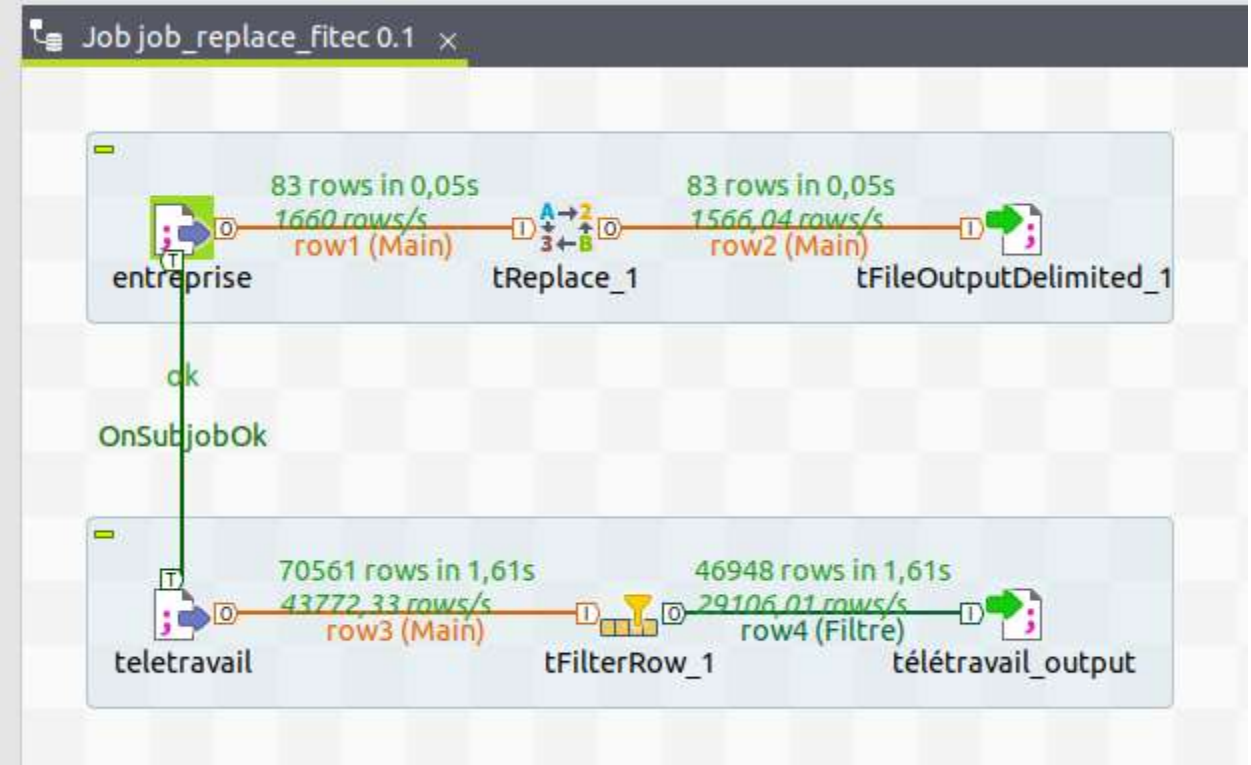


Ensuite le notebook analyse remote usa nous a permis de nettoyer les données et de faire des analyses sous python, pandas, seaborn, ainsi que matplotlib. Nous avons finalement stocké le dataset nettoyé sur Hdfs.

Nous avons aussi utilisé talend pour nettoyer les données. Le Job suivant nous a permis de le faire avec ses composantes suivants :

- tFileInputDelimited: données d'entrées (fichiers)
- tReplace: recherche et remplace les valeurs des colonnes d'entrée pour nettoyer le fichier
- tFilterRow : filtre des lignes d'entrée en définissant une ou plusieurs conditions sur les colonnes sélectionnées.
- tFileOutputDelimited: enregistre les données nettoyées dans un fichier

Voici le job talend qui nous a permis de faire cette tâche:



VI. Visualisation des données

A. Intégration des données sous Tableau

Un premier fichier CSV, finalement abandonné dans notre analyse par la suite par manque de pertinence, a dans un premier temps mis en exergue certains problèmes rencontrés avec Tableau. Nous constations avoir un fichier dépassant la limite de 1000 lignes sous **Tableau Public**. Des "split" avaient donc été faits afin d'alimenter la source de données sous Tableau, pour ensuite appliquer des "merge" dans le but d'avoir la totalité de nos données.

Ces contraintes nous ont fait basculer sur **Tableau Software** afin d'avoir toute la latitude nécessaire à notre projet. En effet, ce dernier n'a aucune limite en nombre de données et est par ailleurs pourvu de fonctionnalités plus avancées. La nouvelle contrainte fut le temps : la version d'essai étant limité à 14 jours. Nous aurons par la suite réussi à obtenir une licence étudiant en faisant la démarche en ligne grâce à notre attestation Fitec.

Une fois les contraintes logicielles outrepassées, nous pouvions aborder le traitement d'un fichier beaucoup plus volumineux (70,000 lignes) retraçant l'évolution du télétravail en Europe depuis le début des années 2000. Pré-nettoyé sous Python, ce dernier demanda un moment afin de bien l'appréhender et comprendre chacune des informations s'y trouvant. Ayant été formé sur ce logiciel par Sébastien Lamour dans le cadre de la formation, nous décidons de poursuivre sur notre lancée en attaquant notre travail de visualisation. Nous avons pu constater une nouvelle fois la puissance de Tableau et l'étendue des fonctionnalités proposées (interactivité entre les pages, génération rapide de graphiques…).

D'autres fichiers relatifs à l'évolution des accords de télétravail en France ou au télétravail durant la période du Covid-19 dans notre pays seront également traités sous Tableau Software afin de produire la majorité des visualisations. Vous trouvez dans le lien ci-dessous Tableau Pojet

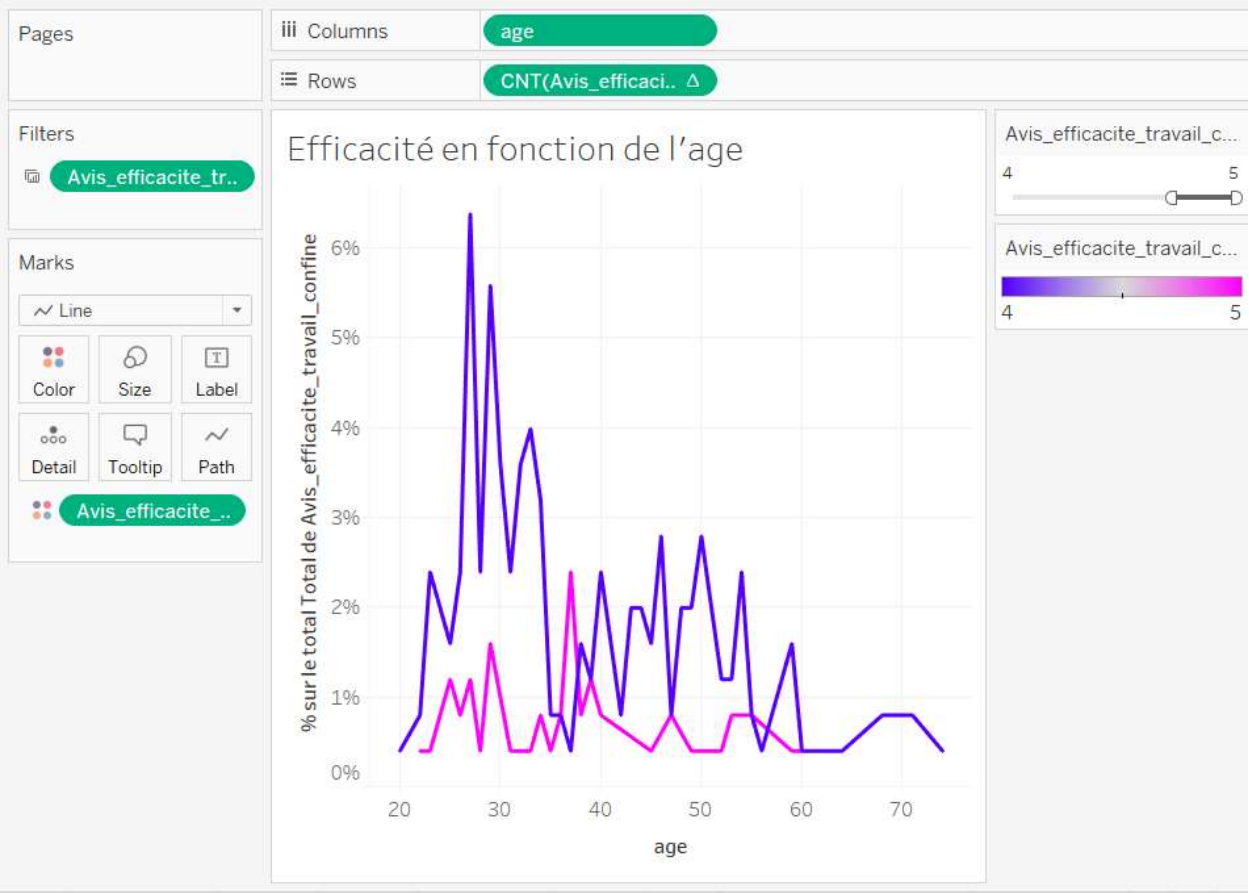
B. Création de feuille de calcul

Après l'ajout des sources de données et quelques travaux préliminaires sur les colonnes comme la conversion des types, le pivot, les alias, les champs calculées etc, nous avons créé différentes feuilles suivants les analyses faites au préalable.

Chaque feuille sous tableau correspond à un graphe. Pour créer un graphe, on définit les mesures (contiennent des valeurs numériques, quantitatives que vous pouvez mesurer) et les dimensions (contiennent des valeurs qualitatives par exemple noms, dates ou données géographiques). Ensuite, on peut personnaliser le graphe selon nos besoins comme la création de filtres ,de groupes, de sets, de légendes, de paramètres…

Dans le context de notre projet, nous avons créé des feuilles avec des filtres, des groupes, des sets, des légendes…Nous avons aussi personnalisé les graphes, les dashboards en respectant les règles de visualisation et de UI (User Interface).

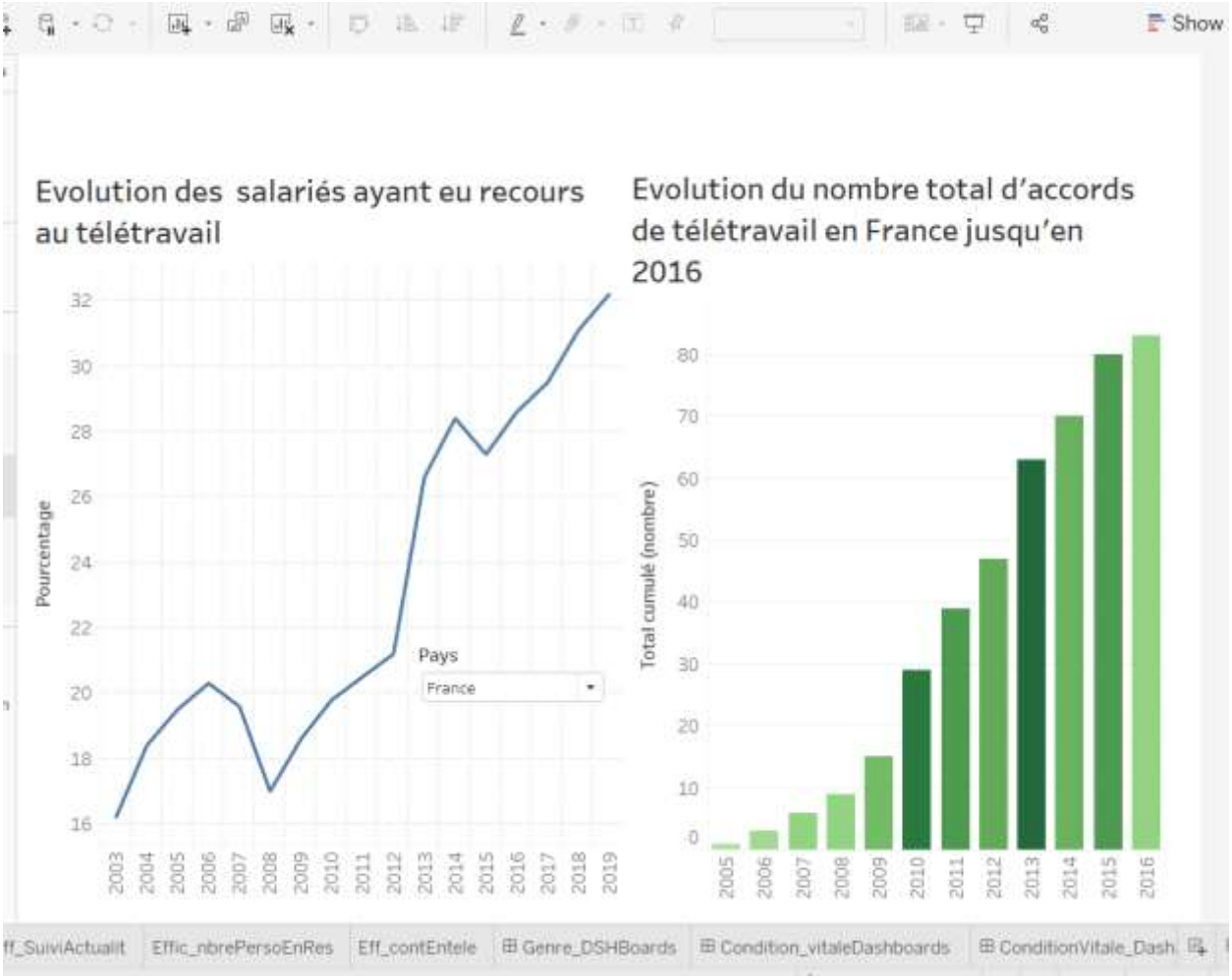
La figure suivante nous montre un exemple de feuille sous Tableau Software :



C. Création de Dashboard

Le dashboard permet la visualisation de données brutes en les rendant plus accessibles et compréhensibles. Pour cela, il fait appel à différentes représentations visuelles et différents types de hiérarchisation de la donnée.

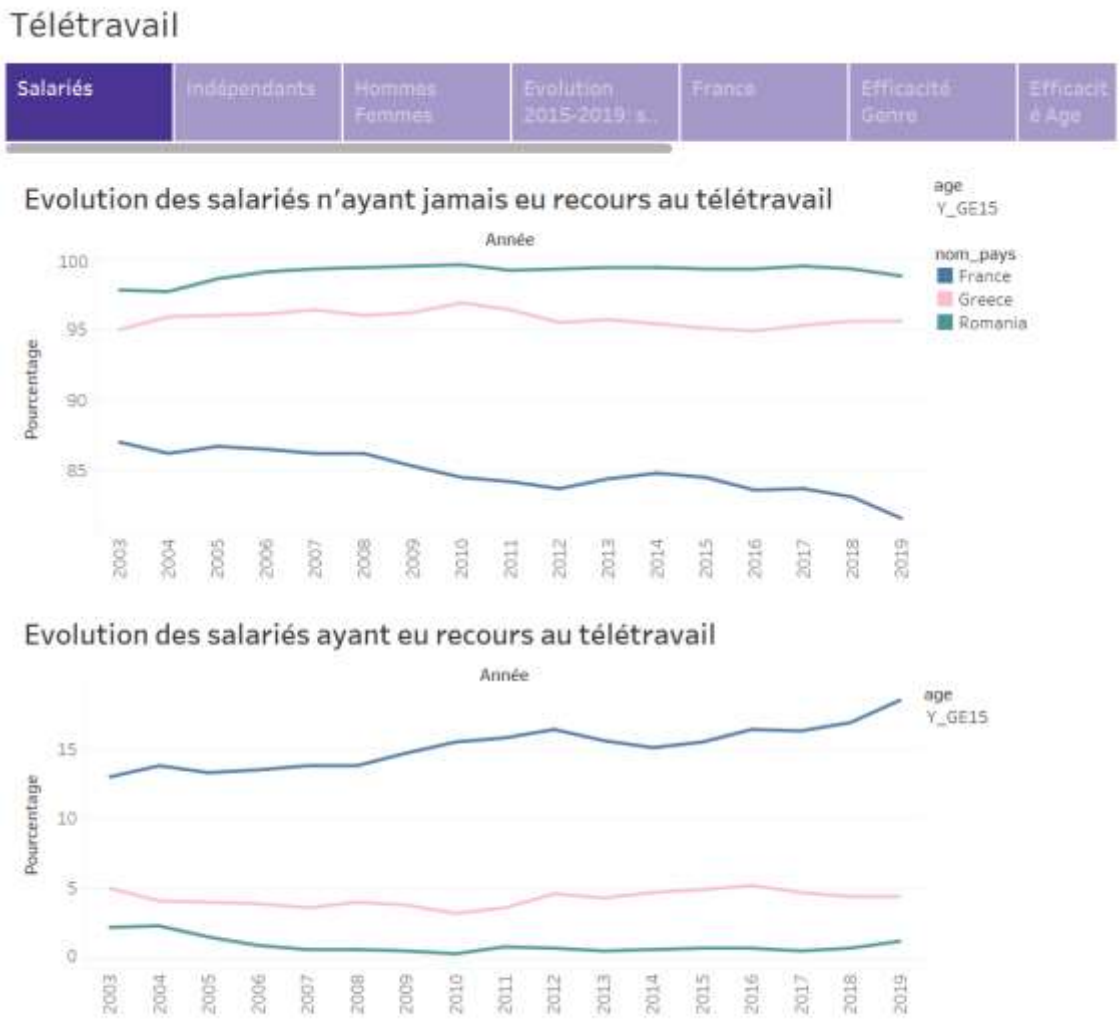
Les feuilles (graphes) obtenues seront regroupées par dashboard dans Tableau. Nous avons organisé les différents dashboards par thématique. Après, on applique des actions (filter, navigate to) à ceux-ci pour filtrer et naviguer entre les différentes feuilles ou dashboards pour interagir avec l'utilisateur.



VII. **Création de storytelling**

Storytelling ou l'art de représenter une histoire est le concept utilisé par Tableau pour mieux structurer et raconter une histoire sur les données.

Nous avons des données sur le télétravail au cours des années et en période de confinement. Dans tableau, nous avons regroupé les dashboards obtenus à partir de ces données dans une même histoire qui nous permet de visualiser et d'analyser les données.



VIII. Machine Learning et prédiction

Le notebook ml régresssion contient notre modélisation et les prédictions que nous avons pu faire sur les données via du Machine learning.

Nous avons créé un modèle de régression qui permet de prédire le pourcentage de télétravailleurs par rapport à l'âge, le sexe, statut du travailleur, la fréquence.

Nous avons obtenu un très bon score (R2 score) de l'ordre de 97% sur le training set et 96% sur le test set (cf ml régression)

Figure: Comparaison des y_train et y_train prédit

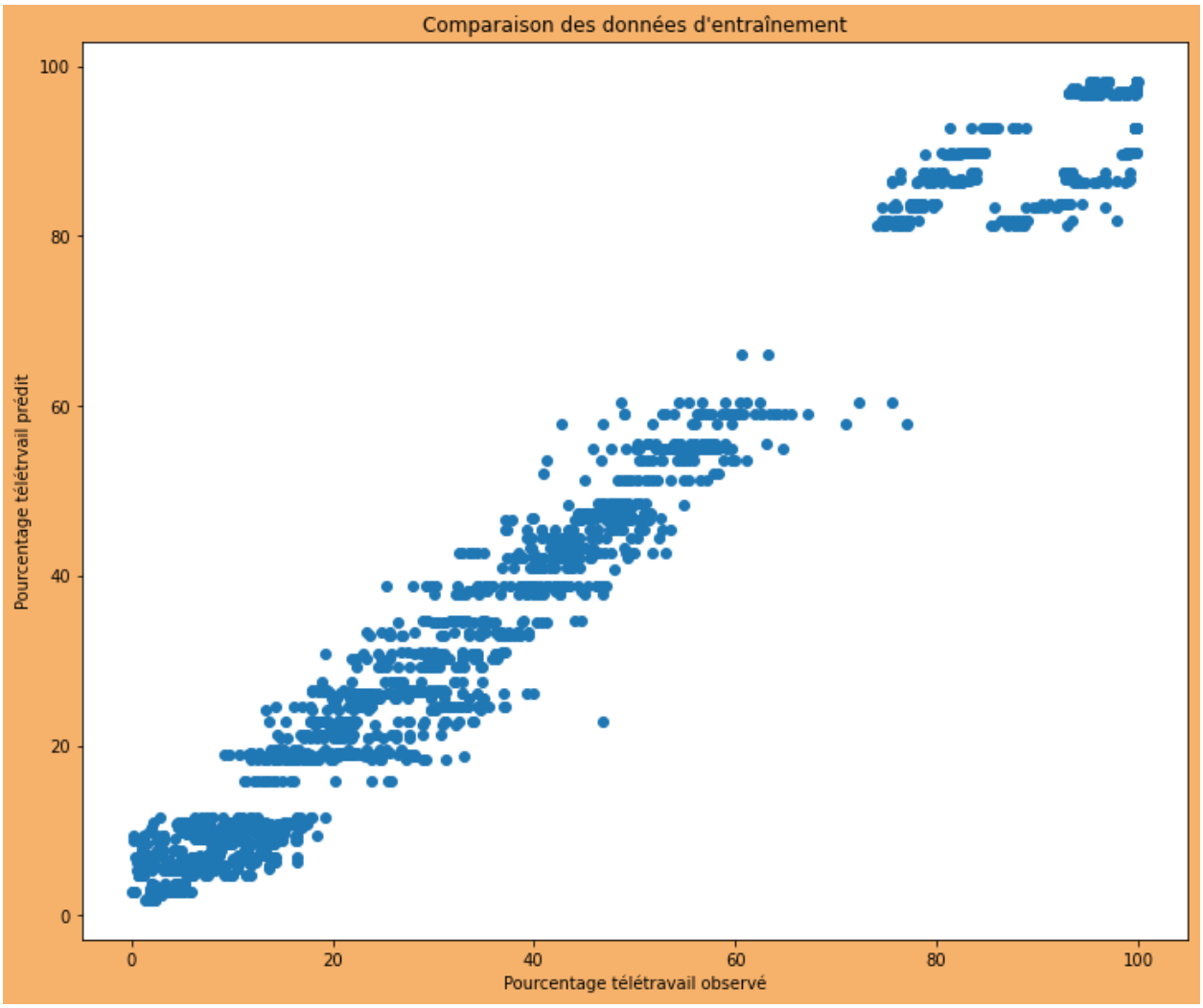
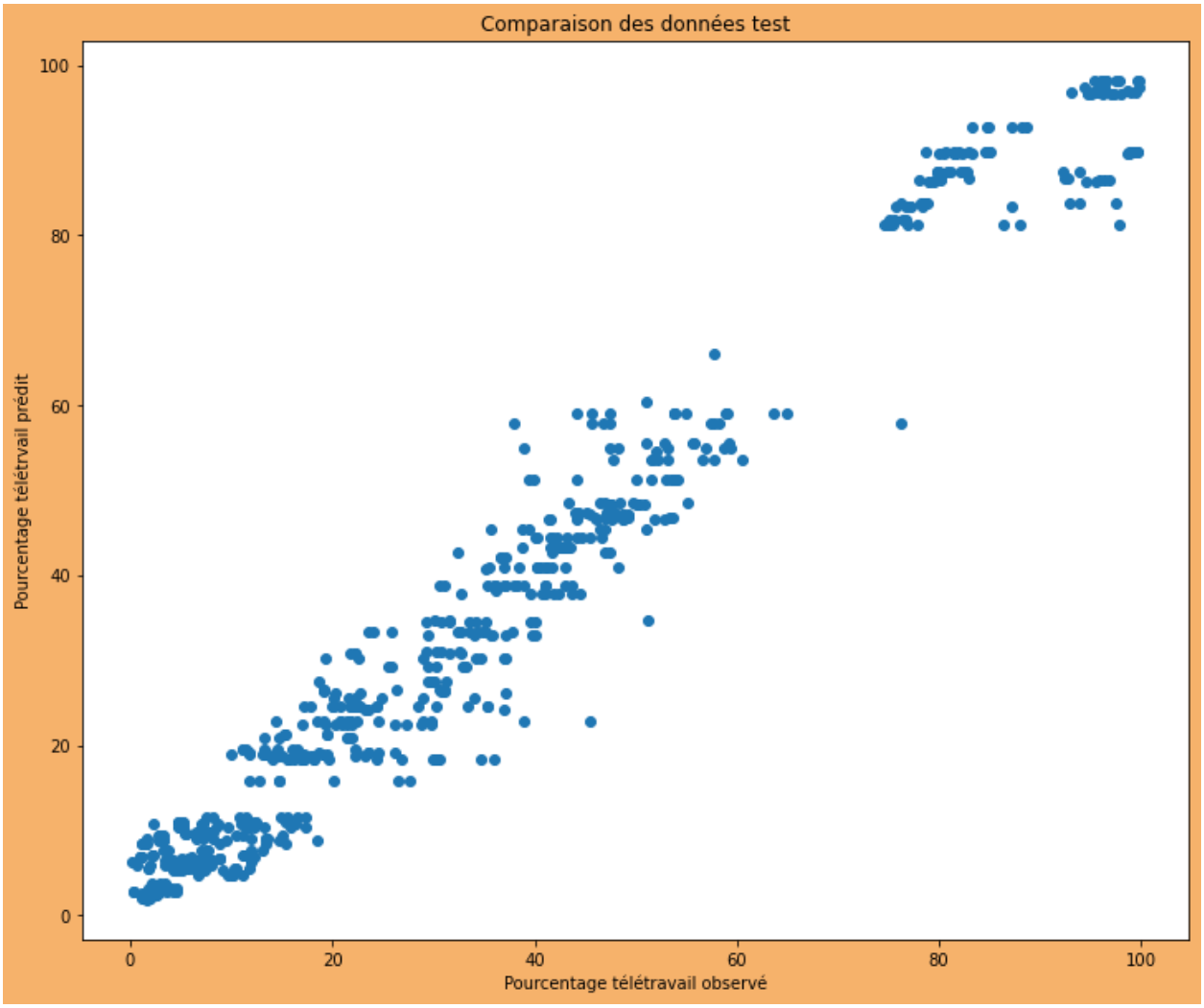


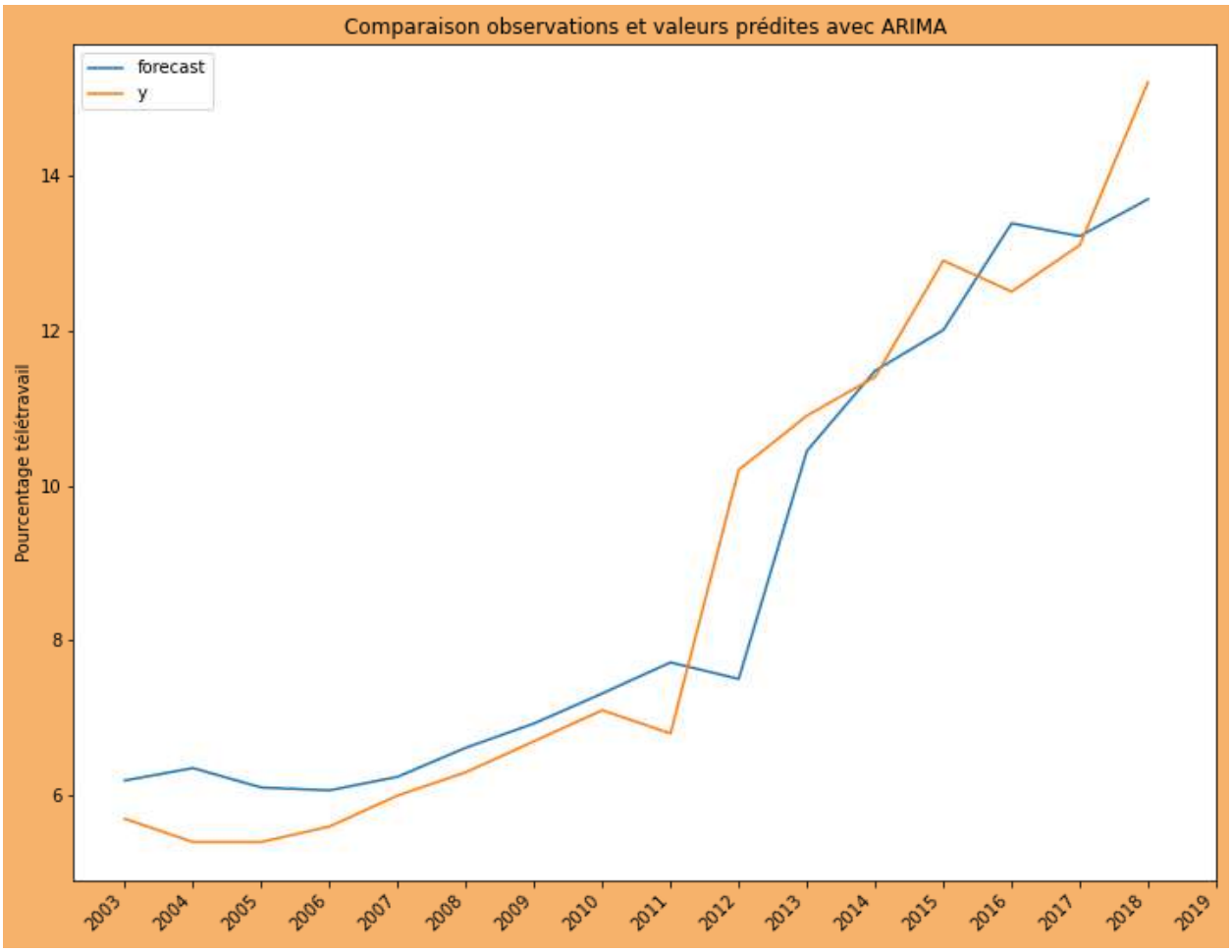
Figure: Comparaison des y_test et y_test prédit



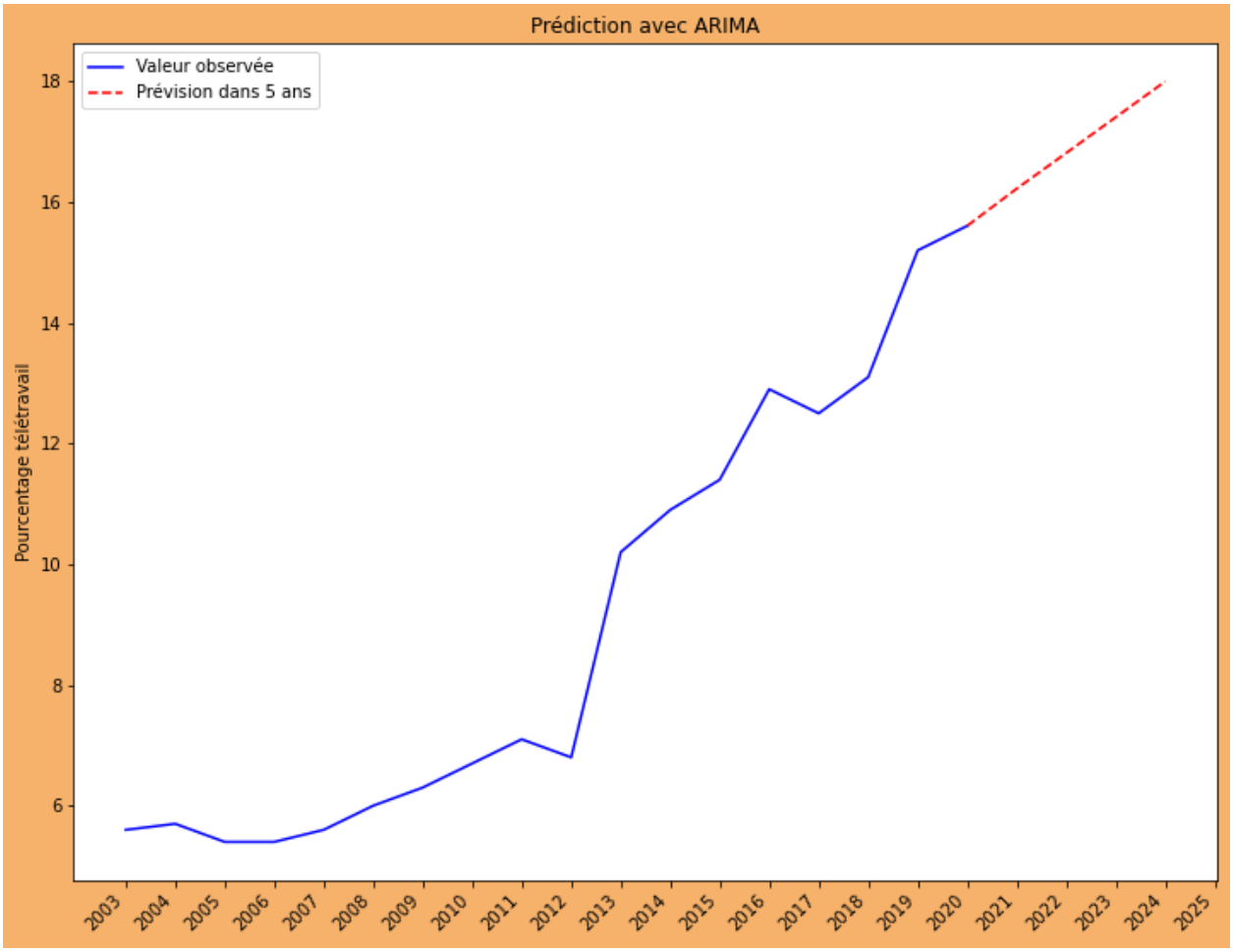
Ensuite, pour prédire le pourcentage de télétravailleurs dans les années à venir nous avons créé un modèle ARIMA. Pour créer une série temporelle, nous avons restreint notre dataset sur la population âgée de 25 à 49 ans (intervalle le plus significatif), nous nous sommes focalisés sur les salariés de 2003 à 2019 (fiabilité des donnée télétravail à partir de 2003)

Nous avons ainsi obtenu une erreur quadratique (mean squared error) de 2.526. Ceci nous a permis de faire une projection du pourcentage de télétravail dans les cinq années à venir (cf le notebook ml régression).

Comparaison séries temporelles et valeurs prédites



Prévision 5 ans à venir à partir de 2020



Nous avons aussi essayé d'appliquer des algorithmes de machine learning sur les données correspondantes au télétravail dans le cadre du covid 19 en France afin de pouvoir classifier nos données.

En premier lieu, nous avons fait une segmentation des données via un algorithme de clustering (K Means). Après plusieurs essais, notre clustering n'a pa permis de sortir une nette séparation des groupes (clusters).

Nous avons ainsi créé un modèle de classification (Random Forest) qui permet de prédire l'efficacité des télétravailleurs. Nous avons obtenu un taux de justesse (accuracy score) de 0.76 sur les données d'entraînement et 0.7 sur les données test. Par contrainte de temps, nous n'avons pas optimisé le modèle .

Pour plus de détails sur la modélisation et prédiction: voir le notebook suivant:

ml classification .

Conclusion

En résumé, ce projet professionnel fut enrichissant à plusieurs niveaux. Il nous a donné l'opportunité de mettre en pratique l'enseignement reçu lors de cette formation du point de vue du savoir-faire, comme du savoir-être. En effet, les compétences techniques et humaines furent tout autant demandées dans le cadre de ce projet.

Du point de vue technique:

En tant que stagiaire Analyst Big Data, ce projet a été pour nous l'occasion de revêtir plusieurs casquettes dans le domaine, et ce de manière plus ou moins avancée : en passant par le rôle de Data Scientist, Data Analyst, Data Engineer ou encore Architecte Big Data. Ce fut très instructif à plusieurs égards.

La récolte, le traitement, l'analyse et la visualisation des données nous ont permis d'une part, de mettre en pratique des connaissances théoriques acquises. Que ce soit sur l'écosystème Hadoop, Python, Tableau Software avec les composants comme HDFS, PySpark, Pandas, Tableau Desktop, Sklearn etc. Et d'autres part, d'analyser l'évolution du télétravail en fonction des pays, de l'âge, du statut professionnel, dans le cadre du covid -19. Cette étude a ainsi pu apporter des éléments de réponses quant à la question très importante de savoir dans quelles conditions un télétravailleur est le plus efficace.

Pour réaliser ce travail durant le temps imparti qui fut très court (inférieur à deux mois et ce, conjointement à la formation suivie), nous avons pu mettre en œuvre cet esprit agile récemment appris : nous sommes auto-organisés, nous avons utilisé des outils collaboratifs, avons planifié des réunions en visioconférence de manière régulière afin de suivre l'avancée de notre projet.

Nous avons réparti le travail en deux grandes parties : la première relative à l'architecture logiciel et base de données, et la seconde sur l'analyse des données. Chacune de ces deux parties ayant été divisées en un ensemble de petites itérations ou tâches. Lors de notre travail de visualisation sous

Tableau, nous avons pu fonctionner sous forme d'itération afin d'intégrer chacune de nos réalisations à notre projet global.

Ce fut un plaisir de travailler ensemble sur ce projet intéressant qui est aujourd'hui au cœur de l'actualité.

Appendix:

- liens vers nos réalisations techniques:

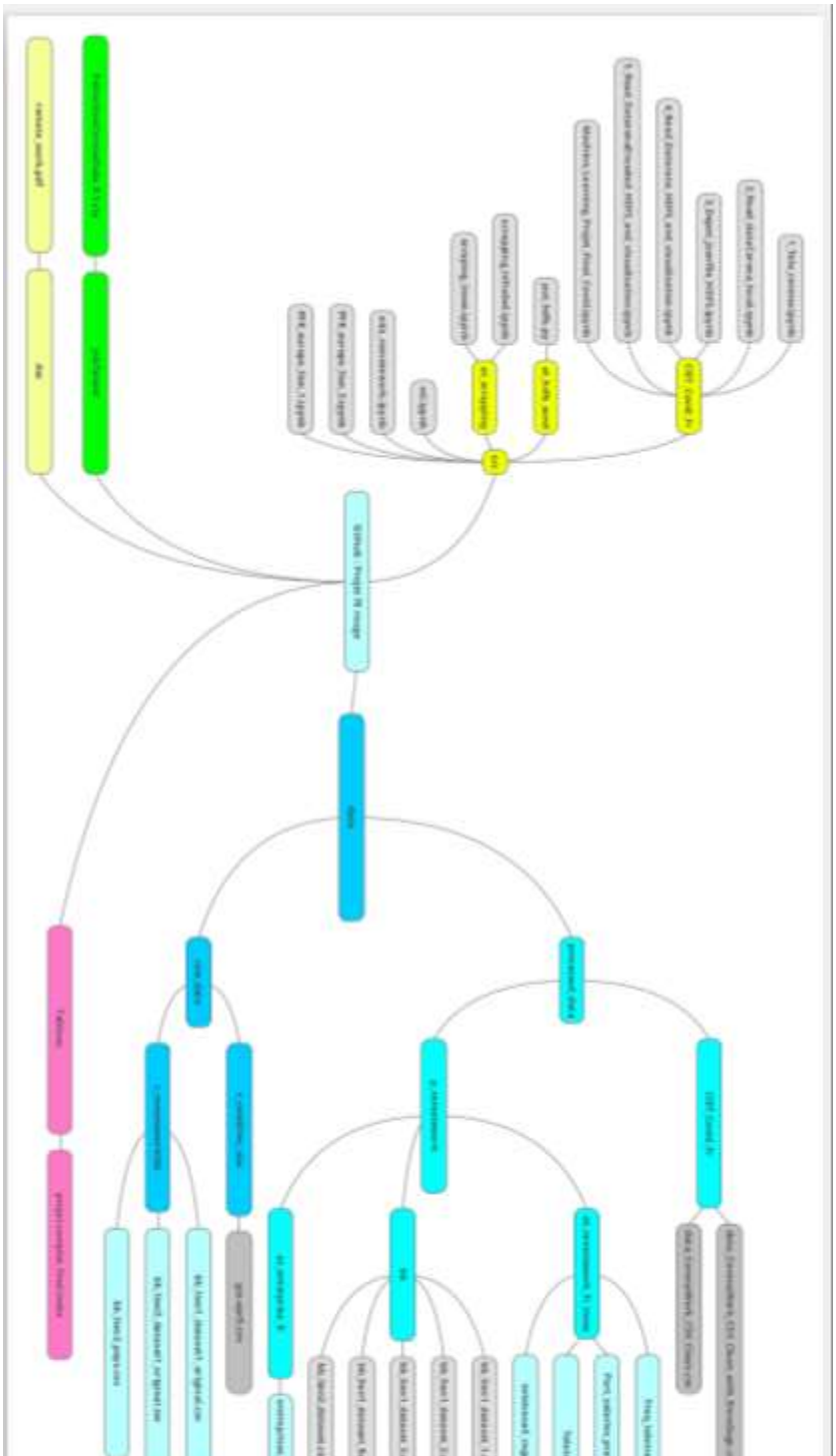
Vous trouverez toutes nos codes et toutes nos réalisations sur le lien github :
https://github.com/mamadcamzis/filRouge

L'architecture consiste en 4 répertoires principale :

1. **data** : données brut sous le nom raw_data; les data_clean sous le noms processed data
2. **src** : contient tous les codes d'analyse
3. **doc** pour quelques documentations
4. **Tableau:** pour nos Dashboards et notre telling story
5. **Jobtalend:** pour les codes Jobtalend.

L'architecture sur Git est ci-dessous détaillée:

Date: Juillet 2020



Vous pouvez trouver le pdf correspondant à cette architecture sous Architecture des répertoires de nos données

- Contraintes rencontrées

Nombreuses sont les contraintes rencontrées durant notre projet final. Parmi celles-ci, des problématiques liées au sujet et aux données, ou encore relatives aux serveurs avec la question de savoir où créer un seul Data Lake pour stocker nos données...

A titre d'exemple:

Plusieurs sujets ont suscité notre intérêt au début de la formation.

Nous avons donc fait trois propositions:

- La première était de traiter le sujet relatif au télétravail
- La deuxième concernait l'Impact de la météo et des monuments historiques sur l'attractivité touristique en France (voir la fiche projet ci-dessous)

- La troisième était relative à la variation des prix immobiliers d'entreprises en contexte de crise en île-de-France (voir la fiche projet ci-dessous)

Ces 3 sujets ont été très intéressants pour nous. Mais en faisant la bibliographie, nous avons remarqué que le second sujet avait été abordé par d'autres personnes qui ont même réussi à créer une application qui s'appellent "Monuments-guide-de-voyage" répondant précisément à nos objectifs.

Pour ce qui est du 3ème sujet, nous avons rencontré des problèmes pour trouver les prix immobiliers des entreprises. Le manque de donnée ne nous a pas permi d'avancer sur ce sujet. En conséquence, nous avons pris la décision de nous concentrer sur le sujet de l'évolution du télétravail en fonction de différents paramètres et dans le cadre du Covid.
Ce sujet d'actualité est pris très au sérieux par de nombreuses sociétés, et peuvent être l'objet de divers débat: les entreprises se doivent-elles encore aujourd'hui et à l'avenir d'avoir une grande superficie de bureaux, ou bien peuvent-elles envisager de nouvelles perspectives et stratégies immobilières, quel type de travailleur est-il privilégié pour faire du télétravail...

Date: Juillet 2020