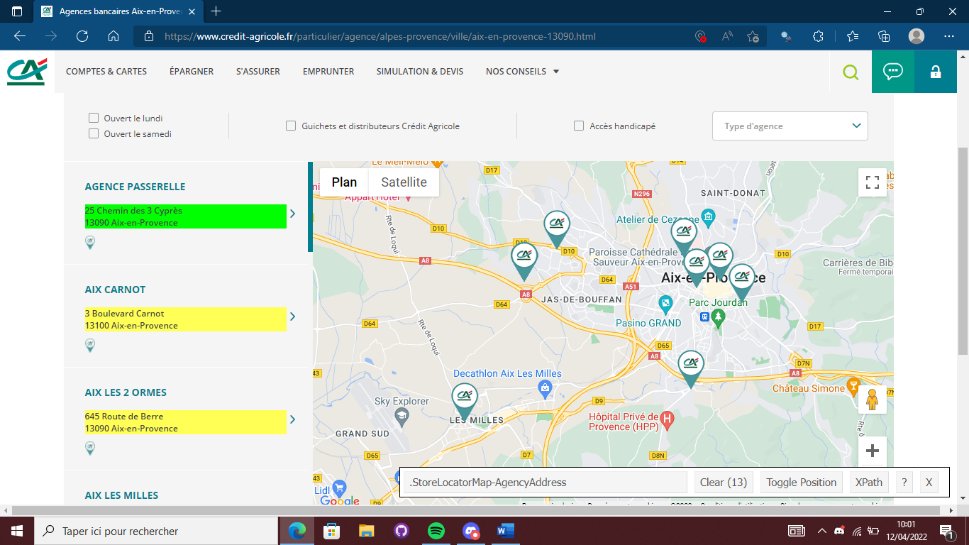
III) Réalisation

1) Récupération des données

Le web scraping consiste à récolter les données sur des sites web. Nous créons deux jeux de données avec cette méthode : le premier *bdd\_coordonnees\_banques2022* est sur les banques, c'est-à-dire le nom de la banque, son type, sa longitude et latitude et l'adresse, le deuxième *bdd\_social\_ze2020* contient les données socio-économiques pour chaque zone d'emploi de la France métropolitaine de 2020.

1. Banques

Pour récupérer les données des banques, nous avons utilisé six sites pour avoir les adresses, cinq sites des banques : Crédit Agricole, Société Générale, Crédit Mutuel, Crédit Mutuel de Bretagne et Banque Populaire. Pour la BNP Paribas, nous n'avons pas pu récupérer les données sur leur site donc nous avons utilisé un autre site où les données étaient disponibles.

Nous avons réalisé le web scraping sur *R* avec le package *rvest*. Ci-dessous, c’est un exemple pour sélectionner les adresses du Crédit Agricole :

En vert, nous avons la case sélectionnée, en jaune les cases qui sont sélectionnées à partir de la case verte car elles sont du même type. Nous pouvons aussi refuser des cases qui apparaitront en rouge. « .StoreLocatorMap-AgencyAddress » est le nom pour récupérer les informations sélectionnées en vert et jaune.

Le site du Crédit Agricole présente les régions puis dans chaque région les villes par ordre alphabétique où contiennent au moins une agence, les villes avec des arrondissements sont détaillées (Paris 01 par exemple). L’url est sous la forme suivante :

**https://www.credit-agricole.fr/particulier/agence/**region**/ville/**ville**-**code postal**.html.** Nous avons dû utiliser un fichier annexe contenant les codes postaux associés aux villes de France. Sur le site certaines villes ont plusieurs codes postaux ce qui entrainent des doublons d’adresses.

Les sites de la Société Générale et de la Banque Populaire étaient faits de la même façon. Les urls étaient de la forme :

**https://agences.**nom de la banque**.fr/banque-assurance/agences-**departement**-**numero du département. Nous avons fait un web scraping pour récupérer la liste de chaque département ainsi que leur numéro. A partir de ce lien, nous pouvons récupérer toutes les adresses d’un département.

Le site du Crédit Mutuel est composé d’une liste de département puis des villes contenant une agence dans un département sélectionné. Dans une troisième page imbriquée nous avons une liste d’adresses d’agences de la ville choisie ainsi que d’autres aux alentours. Ça ne pose pas de problèmes en revanche cela entraine une quantité très importante de doublons d’adresses.

Le Crédit Mutuel de Bretagne est composé de plusieurs pages imbriquées aussi, les noms des départements puis le nom des villes avec leur code postal et enfin sur une troisième page l’adresse exacte de l’agence. Nous avons dû formater l’url pour accéder à la page des villes et à celle de l’adresse.

Pour la BNP Paribas, le site utilisé lui aussi divise les agences en département. Si le nombre d’agences est important, la page est divisée en plusieurs pages. L’url est sous la forme suivante :

**https://www.moneyvox.fr/pratique/agences/bnp-paribas/**numero du département**/** s’il y a plusieurs pages il faut rajouter un « numéro de la page**/** ».

Après avoir réalisé ce web scraping nous avons fait un premier tri des adresses car nous avions des doublons à cause du format du site ou alors que l’agence était partagée en différente parties (par exemple : l’agence à Montélimar 23 rue Raymond Daujat qui a une partie pour les entreprises et apparait donc deux fois dans les adresses récupérées).

A partir des adresses, nous avons utilisé le package *BanR* qui permet de convertir une adresse française en longitude et latitude. Pour chaque banque il a fallu modifier des adresses manuellement car elles ne donnaient pas de résultats avec la fonction du package. Nous réalisons un deuxième tri des adresses avec les couples longitudes, latitudes car certaines adresses changeaient seulement de code postal mais désignées la même agence. Les longitudes et latitudes étaient donc identiques. Nous obtenons après cela 14704 agences différentes.