TP2 : Étude de l'effet de du revenu annuel familial (en dollars américains) sur le prix des propriétés (en dollars américains) (Option 1).

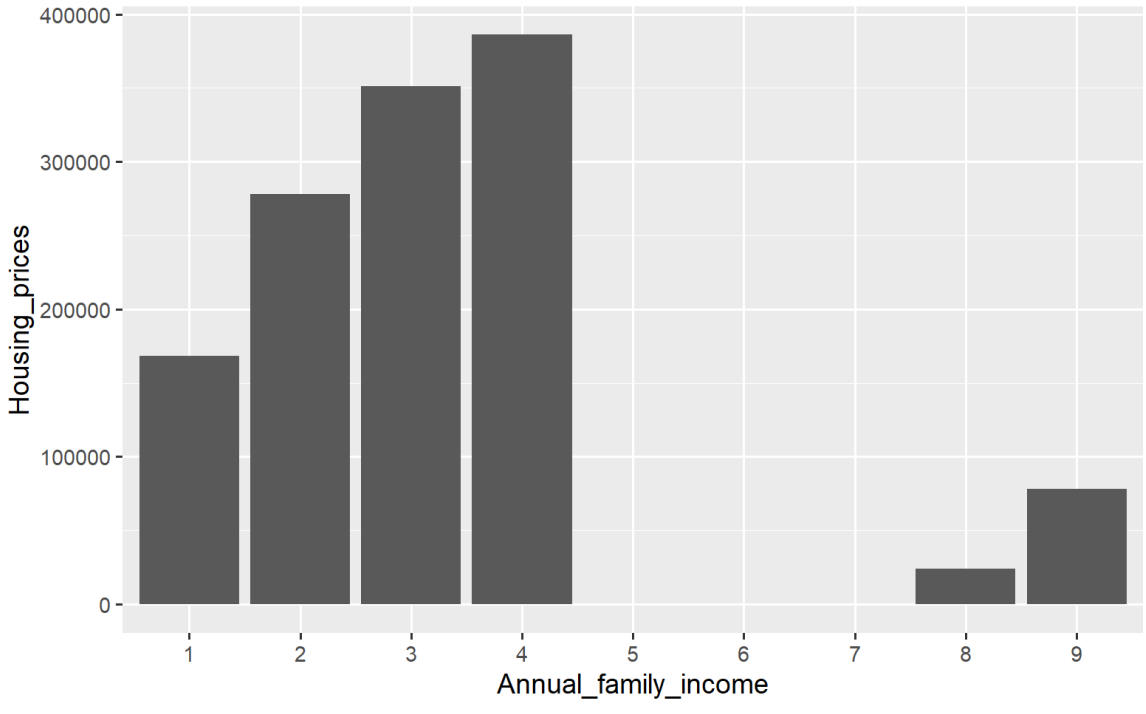
Introduction

Ma question de recherche cherchait à trouver s'il y avait un lien potentiel entre le revenu annuel d'une famille et le coût des propriétés. Ma question de recherche est la suivante : Est-ce qu'une augmentation du salaire annuel, exerce une influence sur l'augmentation du prix des propriété? Les variables inclues dans cette question de recherche sont : la variable indépendante (X) qui représente le salaire annuel d'une famille (pendant les 3 dernières années – 2019, 2020, 2021) et la variable dépendante (Y) est le prix d'une propriété sur le marché (pendant les 3 dernières années – 2019, 2020, 2021) Afin de mener cette courte recherche j'ai choisi l'option 1. Il fallu croiser 1 sondage et un d'autres types de données que des données de sondage. Afin de tenter de répondre à ma question de recherche, j'ai décidé de me concentrer sur le cas des États-Unis. Les revenus et les prix des propriétés ont grandement varié durant l'histoire américaine, cependant il a fallu choisir une période précise à étudier pour être capable de trouver la présence d'un lien ou non. C'est pour ça que j'ai choisi 3 des *Monmouth University National Poll* (de 2019 à 2021) pour mon sondage et pour les autres types de données j'ai choisi la banque de donnée de l'OCDE.

Données et méthodes

En prenant les 3 bases de données du *Monmouth University National Poll* il a fallu extraire les données nécessaires pour répondre à ma question. J'ai nettoyé les 3 bases de données de manière individuelle en gardant que deux variables : le revenu annuel d'une famille (*QD9* représentant la variable dans la base de données) et l'année (« DATE »). La variable « DATE » est une colonne qui a dû être crée pour les besoins de la fusion des 2 bases de données qui apparaîtront après le nettoyage. 3 bases de données (IS19_clean, IS20_clean, IS21_clean) contenant la variable

Annuel_family_income (x) qui est la variable *QD9* renommée et DATE (y) ont été créé à partir du *Monmouth University National Poll*. Pour la banque de données de l'OCDE, j'ai sélectionné le coût des propriétés de 2019 à 2021, mais il a fallu se débarrasser de tous les autres pays présents pour ne seulement garder que la variable USA. Cela à créer la base de données OECD_Housing_clean avec la variable House_prices (x) autrefois nommée VALUE et DATE (y) qui était TIME. Avant de fusionner les données pour créer une base de données commune, il a fallu fusionner les 3 bases de données (IS19_clean, IS20_clean, IS21_clean) en une seule (Income_Surveys_clean). Puisque les 3 bases de données étaient en ordre d'années il a fallu les ajouter les unes aux autres sans compromettre les bases de données individuelles. Finalement, Income_Surveys_clean a pu être fusionnée avec OECD_Housing_clean, mais il a fallu choisir qui serait la variable indépendante et dépendante de cette nouvelle base de données combinée. Les deux bases de données ont été combinées sur la base commune de la variable DATE et c'est la Income_Surveys_clean qui a offert la variable indépendante sous la forme du Annuel_family_income (X) alors que la variable dépendante qui est House_prices (Y) est offerte dans OECD_Housing_clean. Cette fusion a permis de créer une base de données nommée Data_clean combinant les deux variables précédentes.



Description du croisement des variables de la base de données fusionnée

Le croisement des variables de la base de données fusionnée offre un résultat qui est complexe à analyser. Même si le diagramme, peut paraître comme l'option la moins intéressante pour analyser ce croissement, il reste le meilleur choix. La création d'un nuage de point donne un résultat complètement farfelu, tout comme la création d'une ligne qui passerait entre les deux variables. Cela est sûrement un résultat d'une programmation manquée sur R lors de la combinaison des variables qui n'étaient peut-être tout simplement pas faites pour être combinés, puisqu'elles ne permettent pas d'offrir de résultats intéressants à l'étude ou même une réponse quelconque. À ceci, s'ajoute le problème de la création du graphique qui laisse à désirer. À la fin, on se retrouve avec un diagramme à bandes qui offre une interprétation de la variable indépendante en tant que variable continue allant de 1 à 9. Cependant, les deux derniers chiffres doivent être omis puisqu'ils représentent des observations vides rendant le graphique encore plus complexe à déchiffrer lors de la recherche de réponses. Quant à la variable dépendante, censée être continue elle dépasse largement les limites qui étaient imposée par la base de données fusionnée une autre conséquence potentielle d'un codage défectueux.

Bibliographie

Monmouth University Polling Institute. 2019, « *Monmouth University National Poll, Number 187* », Monmouth University Polling Institute, https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/Q9NSNT&version=1.0

Monmouth University Polling Institute. 2020, « Monmouth University National Poll, Number 189 », Monmouth University Polling Institute https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/MJYHWX&version=1.0

Monmouth University Polling Institute. 2021, « *Monmouth University National Poll, Number 240* », Monmouth University Polling Institute, https://dataverse.unc.edu/dataset.xhtml?persistentId=doi:10.15139/S3/H74VUO&version=1.0

OECD. 2024, « *Housing prices (indicator)* », OECD, https://data.oecd.org/price/housing-prices.htm

Bouzamondo_TP2

2024-02-05

Données et méthodes

#Libraries

```
library(tidyverse)
```

```
Warning: le package 'tidyverse' a été compilé avec la version R 4.3.2

Warning: le package 'ggplot2' a été compilé avec la version R 4.3.2

Warning: le package 'tidyr' a été compilé avec la version R 4.3.2

Warning: le package 'readr' a été compilé avec la version R 4.3.2

Warning: le package 'purrr' a été compilé avec la version R 4.3.2

Warning: le package 'dplyr' a été compilé avec la version R 4.3.2

Warning: le package 'stringr' a été compilé avec la version R 4.3.2

Warning: le package 'forcats' a été compilé avec la version R 4.3.2

Warning: le package 'lubridate' a été compilé avec la version R 4.3.2
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(lubridate)
library(haven)
```

```
Warning: le package 'haven' a été compilé avec la version R 4.3.2
```

```
library(reshape2)
```

```
Warning: le package 'reshape2' a été compilé avec la version R 4.3.2


Attachement du package : 'reshape2'

L'objet suivant est masqué depuis 'package:tidyr':

    smiths
```

```
library(reshape)
```

```
Warning: le package 'reshape' a été compilé avec la version R 4.3.2


Attachement du package : 'reshape'

Les objets suivants sont masqués depuis 'package:reshape2':

    colsplit, melt, recast
```

```
L'objet suivant est masqué depuis 'package:lubridate':

    stamp

L'objet suivant est masqué depuis 'package:dplyr':

    rename

Les objets suivants sont masqués depuis 'package:tidyr':

    expand, smiths
```

```
library(dplyr)
library(plyr)
```

```
Warning: le package 'plyr' a été compilé avec la version R 4.3.2

------------------------------------------------------------------------------
You have loaded plyr after dplyr - this is likely to cause problems.
If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
library(plyr); library(dplyr)
------------------------------------------------------------------------------

Attachement du package : 'plyr'

Les objets suivants sont masqués depuis 'package:reshape':

    rename, round_any

Les objets suivants sont masqués depuis 'package:dplyr':

    arrange, count, desc, failwith, id, mutate, rename, summarise,
    summarize

L'objet suivant est masqué depuis 'package:purrr':

    compact
```

```
library(ggplot2)
```

#2 - Data

```
Income_Survey_2019 <- read_por("/Users/15150135/Documents/Analyse Big Data/fas_1001/_tp/_t

Income_Survey_2020 <- read_por("/Users/15150135/Documents/Analyse Big Data/fas_1001/_tp/_t

Income_Survey_2021 <- haven::read_por("/Users/15150135/Documents/Analyse Big Data/fas_1001

OECD_Housing <- read.csv("/Users/15150135/Documents/Analyse Big Data/fas_1001/_tp/_tp2/OEC
```

Annexe

#3 - Nettoyage de variables

#3.1 - Income_Survey_2019

```
IS19_clean <- Income_Survey_2019 |>
select(QD9, DATE) |>
dplyr::rename(Annual_family_income = QD9) |>
dplyr::mutate(DATE = as.character(DATE))
```

```
IS19_clean$DATE <- c("2019")
```

#3.2 - Income_Survey_2020

```
IS20_clean <- Income_Survey_2020 |>
select(QD9) |>
dplyr::rename(Annual_family_income = QD9)
```

```
IS20_clean$DATE <- c("2020")
```

#3.3 - Income_Survey_2021

```
IS21_clean <- Income_Survey_2021 |>
select(QD9) |>
dplyr::rename(Annual_family_income = QD9)
```

```
IS21_clean$DATE <- c("2021")
```

#3.4 - OECD_Housing

```r
OECD_Housing_clean <- OECD_Housing |>
select(LOCATION, Value, TIME) |>
dplyr::rename(DATE = TIME) |>
dplyr::rename(House_prices = Value) |>
subset(LOCATION == "USA")
```

```r
OECD_Housing_clean$LOCATION <- NULL
```

#3.5 Fusion des données

```r
Income_Surveys_clean <- bind_rows(IS19_clean, IS20_clean, IS21_clean)
```

```r
Data_clean <- merge(x = Income_Surveys_clean$Annual_family_income, y = OECD_Housing_clean$
```

```r
#setnames(Data_clean, "x", "Annual_family_income")
#setnames(Data_clean, "y", "Housing_prices")
```

Graphique

```r
ggplot(Data_clean, aes(x=x, y=y)) +
  geom_bar(stat = "identity") +
  scale_x_discrete(limits = c("1", "2", "3", "4", "5", "6", "7", "8", "9")) +
  labs(x = "Annual_family_income", y ="Housing_prices")
```