TP2 (Analyse de données observationnelles)

ÉNONCÉ GÉNÉRAL DU TP2

Dans ce TP, vous ferez une analyse de sondages, données très utilisées dans toutes les sciences sociales. Les données utilisées dans ce TP sont à votre discrétion et il revient à vous de les trouver. Cela fait partie de l'apprentissage. Je demeure disponible pour vous conseiller bien sûr. L'objectif de ce TP est de vous familiariser à l'écriture d'un rapport de recherche classique avec données. Ce TP peut également vous mener à des pistes de réflexion pour votre travail de session.

Pour ce TP, je vous demande de décrire les données utilisées (méthodes de collecte, organisme ayant collecté ces données, etc.) et pour quelle(s) raison(s) ces dernières sont appropriées pour répondre à votre question de recherche. Je vous demande également de décrire les étapes par lesquelles vous êtes passé dans **votre codage** (nettoyage de données, uniformisation de variables, etc.).

Consignes et étapes:

DEUX OPITIONS SONT POSSIBLES:

La première option vous invite à croiser des données de sondage(s) (minimum 1 sondage) avec d'autres types de données observationnelles (par exemple des données économiques comme le PIB). Il faut expliquer pourquoi vous avez croisé les différentes bases de données. Exceptionnellement, si vous choisissez cette option, je vous permets d'utiliser d'autres types de données que des données de sondages, mais elles doivent venir de deux bases de données différentes.

La deuxième option vous invite à croiser des données de plusieurs sondages ensemble (minimum deux sondages). **Il est important d'expliquer pourquoi il était nécessaire de croiser les sondages choisis.** Cela peut être pour augmenter le nombre

d'observations dans le cas où l'échantillon serait trop faible (moins 100 observations) ou pour faire des études longitudinales (sur une certaine période de temps).

MENTIONNEZ L'OPTION CHOISIE DANS VOTRE TITRE. Exemple: *TP2: Étude de l'effet de la croissance économique (PIB) sur l'appui au parti du président américain sortant (Option 1).*

Étapes du TP

1: Écrire une courte introduction d'une demi-page à une page expliquant la question de recherche et les données utilisées pour répondre à cette dernière.

2: Écrire une section décrivant les données utilisées, les variables mobilisées et le croisement de variables. Vous pouvez appeler cette section *données et méthodes*. Expliquez clairement les transformations de variables effectuées (nettoyage, uniformisation). Je veux un maximum d'une page pour cette section. Je veux voir votre code de nettoyage de vos variables dans une annexe à la fin de votre document `Quarto`.

Exemple:

```
# 1 - Libraries ----

#install.packages("tidyverse")
suppressMessages(library(tidyverse))
#install.packages("tidyverse")
library(lubridate)
# #install.packages("haven")
library(haven)

# 2 - Data ----

Anes <- read_sav("_data/anes_timeseries_cdf_spss_20211118.sav")  # American Election Study

Usa_growth <- read_csv("_data/united-states-gdp-growth-rate.csv") # American GDP growth
```

```
New names:
Rows: 61 Columns: 4
-- Column specification
```

```
-------------------------------------------------------- Delimiter: "," dbl
(2): GDP Growth (%), Annual Change lgl (1): ...4 date (1): date
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
* `` -> `...4`
```

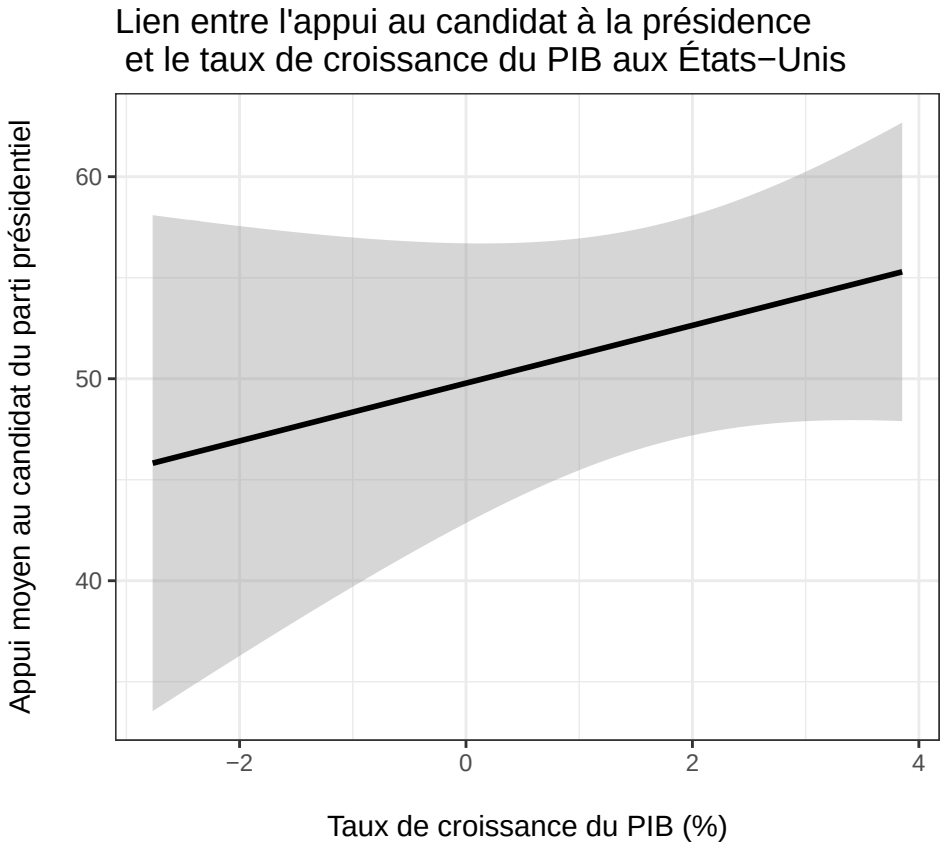
```r
# 3 - Nettoyage de variables ----

# 3.1 - US Election data ----

Anes_clean <-  Anes |>
  select(VCF0004, VCF0424, VCF0426) |>
  ### Nettoyage + transformation de variables ###
  rename(year    = VCF0004,
         sup_dem = VCF0424,
         sup_rep = VCF0426) |>
  filter(year >= 1992) |>
  na.omit() |>
  group_by(year) |>
  mutate(sup_dem  = mean(sup_dem, na.rm = T),
         sup_rep  = mean(sup_rep, na.rm = T),
         sup_pres = case_when(year %in% c(1992:2000,2012:2020) ~ sup_dem,
                              year %in% c(2004:2008) ~ sup_rep)) |>
  distinct()

# 3.2 - US growth data ----

Usa_growth_clean <- Usa_growth |>
  ### Nettoyage +  transformation de variables ###
  select(date, "GDP Growth (%)") |>
  rename(growth = "GDP Growth (%)") |>
  mutate(year = year(date)) |>
  select(-date) |>
  filter(year %in% c(1992,1996,2004,2008,2012,2016,2020))

# 4 - Fusion des données ----

Data_clean <- left_join(Anes_clean, Usa_growth_clean) |> # Jonction des données
  na.omit()
```

```
Joining with `by = join_by(year)`
```

3: Ensuite, écrivez une section qui décrit les résultats de votre analyse. Un graphique montrant le croisement de vos variables devrait être dans cette section. Vous pouvez intégrer votre graphique avec un fichier `.png` ou `.jpg` dans votre `Quarto` ou directement à l'intérieur de votre fichier `Quarto` en spécifiant ces arguments dans les paramètres de votre bloc de code `R`: `r, eval=TRUE, echo=FALSE`.

Exemple:

```
`geom_smooth()` using formula = 'y ~ x'
```



Le type de graphique utilisé est à votre discrétion et/ou selon l'analyse mobilisée. **Vous n'avez pas besoin de me montrer le code du graphique**, je n'évalue pas ce dernier. J'évalue simplement le croisement fait entre vos variables représenté par le graphique. Votre graphique doit clairement montrer le croisement des variables dans le cas de **l'option 1**.

Dans le cas de **l'option 2**, vous devez spécifier clairement la provenance de toutes les variables utilisées. Par exemple, dans mon graphique plus haut, l'appui au candidat du parti présidentiel provient de l' *American National Election Study*. Vous devez spécifier le nombre d'observations dans chacune des bases de données utilisées et dans la base de données fusionnée.

Je veux maximum une page pour cette section.

4: Remettez vos fiches sur `GitHub` en suivant les procédures habituelles.

> ⚠ **ATTENTION**
>
> **RAPPEL DE LA DATE DE REMISE: 5 février avant minuit**.

Évaluation (sur 5 points):

- **Justification des bases de données utilisées (cas d'étude, pourquoi ces bases de données aident-elles à répondre à la question de recherche): 1.5 point**.
- **Explication des transformations de variables effectuées (nettoyage, uniformisation, etc.): 2 points**
- **Description du croisement des variables de la base de données fusionnée (explication du lien entre les variables, analyse des résultats de l'étude): 1.5 point**.
- **BONUS - Création d'une échelle de mesure: 0.5 point**.