Report: Data Analysis and Processing

Part 1: Data Task

Notes:

- There are **2 Do-files** and **2 log files**. One Do file is for questions 1 to 8 which deal with data directly. The second Do-file is for regression and table questions from 9 to 13.
- There are dependencies that need to be installed in order to have publishable tables and outputs or in some cases run regressions with fixed effects to avoid heteroskedastic errors and bias. These are:
    - ssc install estout
    - ssc install outreg2
    - ssc install tabout
    - ssc install ietoolkit (very important)
    - ssc install reghdfe
    - ssc install ftools
- Do-files might take longer if entirely executed in one go because of the code which explores the entire dataset for clues and each variable analysis. Patience can help or keep an eye on clicking 'more' each time the run is stuck.
- Regarding regression questions, I avoided changing the nature of questions through extra data manipulation. In cases where results are not possible to obtain, I explain the situation in writing and how they can be addressed, but I didn't give out answers which are out of context of the questions asked. This is for question 12 and question 13.
- The focus was not to attempt to get the answer no matter what. The focus was to explain my thinking process and remain within the limit of doing this as a test and not as part of a full fledged research project. **Despite an electrical power problem suffered on the 20th and the 21st, I attempted to spend more time trying to find a conclusive statement on this issue, hence submitting after the deadline. I will humbly accept penalties.**

1. Data Import and Exploration

The code begins by importing the dataset "combined_data.dta" located at my local machine.

Whenever given data, I will try to analyze it first before I do anything else to gain insights about the data in order to deal with outliers and other irregularities in the dataset that could be potential sources of bias later in the process. I didn't do this by looking at summary statistics entirely, but by focusing on certain variables. This is why I started data exploration commands such as describe, codebook, and summarize to gain insights into the dataset's structure, variable descriptions, and summary statistics. Quality checks are performed using tabulate to identify missing values in some variables that I presume to be of interest and I generated histograms temporarily to visualize the distribution of certain variables.

From skimming through the test questions, I got a feeling that these variables involved in the data cleaning questions are of interest, thus, I wanted to learn more about them. The variables are:

- survey_complete

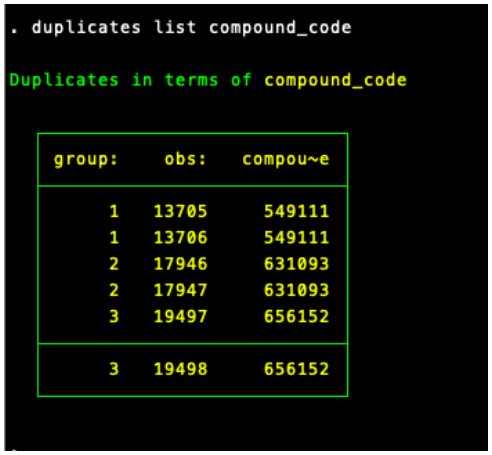
- number of visits post carto

Results indeed showed a few irregularities for survey_complete and number of visits post carto, which justified the next questions in the data task. However, by skimming other questions, I didn't identify other clues that could give me problems later, except that some polygons have not had any visit or had little to no variations, thus inferring that later on I need to work on this issue. The outputs of data pre-analysis are shared in a folder called **pre-analysis**.

2. Data Cleaning

In order to find duplicates, first I need to find a unique identifier variable for the dataset. It is obvious that it is 'compound_code' variable. I will use this variable to try to find duplicates and remove them.

Duplicates in the dataset based on the unique identifier variable "compound_code" are listed and reported.



```
. duplicates list compound_code

Duplicates in terms of compound_code

  group:    obs:   compou~e
      1   13705     549111
      1   13706     549111
      2   17946     631093
      2   17947     631093
      3   19497     656152
      3   19498     656152
```

Any duplicate observations are then dropped using duplicates drop compound_code, force. Note here I used the ***'force' command*** to try to remove all duplicates including the first instance of the duplicates.

To drop incomplete surveys, I had to maneuver some steps to convert the variable "survey_complete" which was in ***float*** format into a possible format for processing.

'Incomplete surveys indicated by the variable "survey_complete" as "No" are dropped using drop if survey_complete == "No". Comments in my code on this part explain the necessity of understanding variable types before operations like converting to strings and applying labels for better data management.

3. Data Merging

The CSV file "nearest_property.csv" is read into Stata using the 'insheet' command and it is saved as a Stata dataset.

Before merging, common variables or the unique identifier "compound_code" type are checked in both datasets to ensure compatibility. The "compound_code" variable in the main dataset is converted to a long

format to match the format in "nearest_property," followed by merging the datasets based on "compound_code."

```
.
. * Merge the datasets based on compound_code

. merge 1:1 compound_code using `temp1', keep(master match)

    Result                           # of obs.
    -----------------------------------------
    not matched                        21,761
        from master                    21,761  (_merge==1)
        from using                          0  (_merge==2)

    matched                            21,679  (_merge==3)
    -----------------------------------------


.
. * Save the merged dataset
```

On this question, I did a little demo to try to find what is the common variable to merge based on, but it is unnecessary for this exercise. I think it is important when one has a large dataset and I demonstrated how I can do it in the code. I found ‘compound code’ to be the variable of interest when merging.

4. Missing Value Imputation

Here, it was challenging to know exactly which variables are being talked about in this question. This is a repeating pattern in the entire data task where direct variable names aren’t mentioned, but instead labels. I took steps to first identify which variable corresponds to 'number of visits post carto' using code. This is to demonstrate how it can be done just in case there is less time and a very large dataset. However, it is still nice to mention direct variable names to save time.

I found out that it is represented by ***‘nb_visit_post_carto’.***

Missing values in the variable "nb_visit_post_carto" are imputed using the polygon average approach. The polygon average for "nb_visit_post_carto" is calculated and missing values are replaced with the polygon average. However, after this process there are still missing values. I am not surprised because when I did the data exploration work on question one, I expected this kind of behavior. The remaining missing values of nb_visit_post_carto were not replaced because the corresponding polygon average is missing. This could be because those polygons only have missing values for nb_visit_post_carto.

```
.
.    * Calculate the polygon average for nb_visit_post_carto

. egen polygon_avg = mean(nb_visit_post_carto), by(a7)
(21761 missing values generated)


.
.    * Impute missing values of nb_visit_post_carto using the polygon averag
> e

. replace nb_visit_post_carto = polygon_avg if mi(nb_visit_post_carto)
(6,935 real changes made)
```

If not all missing values are imputed, I chose to use ***a global average*** that is calculated and used to impute remaining missing values.

5. Variable Transformation

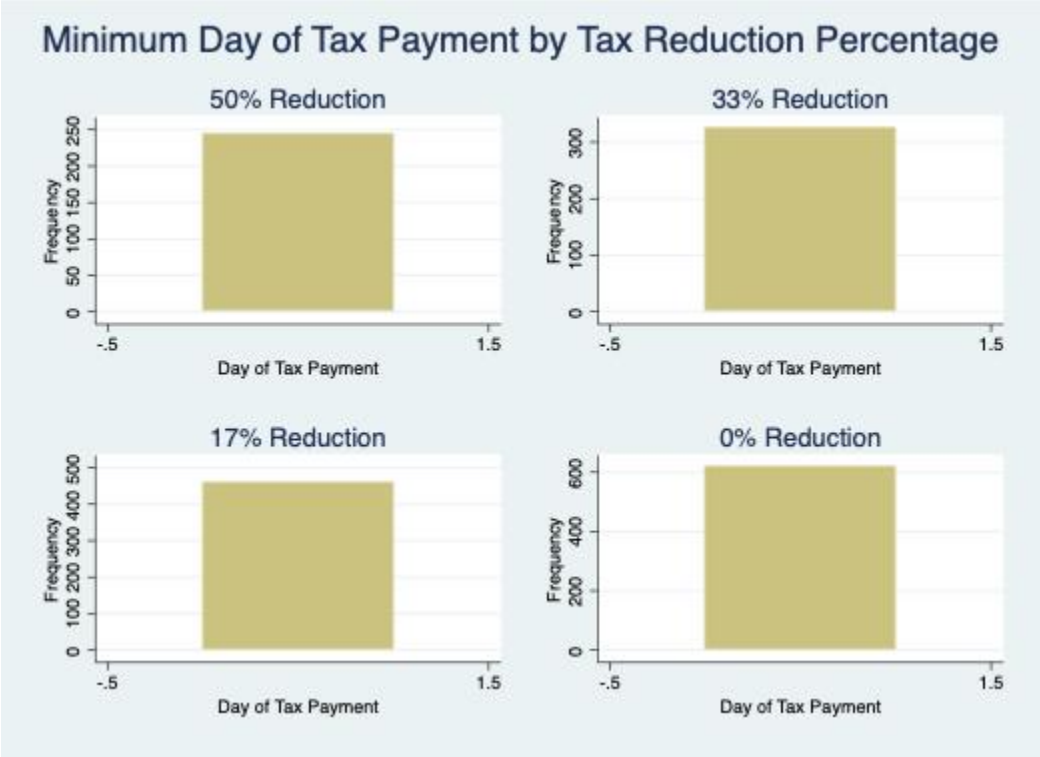
This is a very straightforward question. The string variable "rate" was converted to a numeric variable to facilitate further computations. Additionally, the log version of the numeric variable was generated for potential statistical analyses and modeling.

6. Dummy Variable Creation

Again this was a very straightforward question. Dummy variables representing the percentage of tax rate reduction were created using the "reduction_pct" variable. These dummy variables, named pct50, pct67, pct83, and pct100, were appropriately labeled to indicate the corresponding tax reduction percentages.

7. Data Visualization

Histograms were plotted for the subset of observations that paid taxes, categorized by tax reduction percentages. These histograms provided insights into the distribution of tax payment days across different reduction levels. The individual histograms were then combined into a single publication-suitable graph with clear labels and titles.



All graphs for this question are found in the results folder and they are labeled starting with ***'Q7'.***

8. Statistical Analysis

The average taxes paid amount was calculated for observations in polygons where over 5% of compounds paid taxes. This analysis provided a quantitative understanding of tax payment patterns within the dataset.

```
.   * Calculate the average taxes paid amount

. summarize taxes_paid_amt, meanonly

.
.   * Display the average taxes paid amount rounded to three decimal places

. di "The average taxes paid amount for all observations in polygons where o
> ver 5% of the compounds paid taxes is " round(r(mean), 0.001)
The average taxes paid amount for all observations in polygons where over 5%
>  of the compounds paid taxes is 2610.623
```

So, the average taxes paid amount for all observations in polygons where over 5% of the compounds paid taxes is **2610.623**

Note: I assume this is where the data intensive cleaning and manipulation work gets done, therefore I saved the last data file from here as the final dataset for this entire exercise. The final dataset is called '***merged_data.dta***'

9. Balance table and tests.

This question is a bit interesting and tricky. I will prioritize simplicity to save time. I assume that a balance table, in the context of experimental or observational studies, is used to assess whether the treatment groups (in this case, different levels of tax reductions) are balanced with respect to key characteristics or covariates. I assume that if this was real research, the goal would be to ensure that any differences in outcomes between treatment groups can be attributed to the treatment itself rather than pre-existing differences in characteristics.

Despite this, again, the variable specifying the treatments is not mentioned by its direct name, which prompted me to spend significant time trying to check if the data doesn't have other variables corresponding to treatments directly. I quickly remembered that in the final dataset I had, there I generated some dummy variables corresponding to the amount of tax reduced.

I choose the following characteristics:

1. **Property characteristics:** dist_city_center, dist_commune_buildings, dist_public_schools, dist_roads, roof, and walls.
2. **Owner characteristics:** age_prop, sex_prop
3. **Treatment group indicators:** pct50, pct67, pct83, pct100 (control)

Note: For variables walls and roof which are categorical, it is necessary to convert them into binary variables so that their analysis in the balance table can have some meaning. With the '***iebaltab***' command, this step is necessary, but can be a starting step for other subsequent steps. If I ignore it, I will have to not

treat the results for walls and roof quality as statistically meaningful in the balance table. I commented out this step in the code, but for a lengthy more deep analysis, it is necessary.

The provided table (balance.csv) compares the means of various property and property owner characteristics across different treatment groups. It shows the mean values for each variable, along with the standard errors (in parentheses), for four different treatment groups, denoted as (1), (2), (3), and (4) which denotes the treatments groups pct50, pct67, pct83, and pct100 respectively. The table also includes pairwise t-test results for the differences in means between each pair of treatment groups.

The table was produced using pairwise t-tests to compare the means of each variable across the treatment groups. Regarding the validity of the experiment, I infer the following from the results:

1. For most variables, the differences in means between treatment groups are not statistically significant at the 1%, 5%, or 10% levels. This suggests that the treatment groups are reasonably balanced in terms of these characteristics, supporting the validity of the experiment.
2. There are a few exceptions where statistically significant differences exist:
    a. dist_commune_buildings: The mean difference between groups (1) and (2) is significant at the 10% level.
    b. walls: The mean differences between groups (1) and (2), as well as groups (3) and (4), are significant at the 10% and 1% levels, respectively.

These significant differences in wall quality and distance to commune buildings across some treatment groups may raise concerns about the balance and validity of the experiment. Further investigation or adjustments (e.g., controlling for these variables in the analysis) might be necessary to account for these imbalances. However, this is a sample test only for this question and does not imply that this argument holds for the entire dataset or when other variables are considered.

Overall, the balance table suggests that the treatment groups are reasonably balanced for most variables. The pairwise t-tests were used to produce this balance table and assess the statistical significance of the differences between treatment groups.

10. Regression

Results:

	(1)	(2)	(3)
VARIABLES	All Properties	Constant Bonus	Proportional Bonus
pct50	0.00396	0.00617	0.033
	(0.0408)	(0.0564)	(0.060)
pct67	0.0550	0.00530	0.115**
	(0.0376)	(0.0510)	(0.056)
pct83	0.0459	0.0317	0.074
	(0.0339)	(0.0459)	(0.050)
Constant	0.620***	0.664***	0.570***
	(0.0220)	(0.0301)	(0.032)
Observations	1,260	655	594
R-squared	0.003	0.001	0.008
Standard errors in parentheses			
*** p<0.01, ** p<0.05, * p<0.1			

11. Regression

Understanding the Problem with this question

The problem arises when running a fixed effects regression with polygon fixed effects. The variables pct50, pct67, and pct83 do not vary within some polygons, leading to perfect multicollinearity. This means that Stata cannot separately identify the effects of these variables and the polygon fixed effects, causing these variables to be dropped from the regression.

Potential Solutions

There are several potential solutions that I can try to this issue:

1. **Excluding the problematic polygons:** This solution involves removing the polygons that are causing the problem from the dataset. However, this may not be ideal if these polygons are important for the analysis.
2. **Changing the level of analysis:** Instead of analyzing the data at the property level, the data could be aggregated to the polygon level. This would involve creating new variables that represent the average or total values for each polygon.
3. **Including time-varying controls:** If there are other variables that vary over time and could be affecting both the tax reductions and the outcome, these could be included in the model to help isolate the effect of the tax reductions.

Instead of going through these complex steps, I recalled the existence of a package named '***ftools***' which has the command to regress '***reghdfe***'. This command performs a regression with multiple levels of fixed effects, absorbing the a7 variable (polygon fixed effects). In my code, the command is used instead of a normal reg command to include fixed effects, and the vce(robust) option is used to compute robust standard errors. The absorb(a7) option tells Stata to include fixed effects for the a7 variable.

1. To include polygon fixed effects controls for any unobserved, time-invariant characteristics of each polygon, this meant that the coefficients on the variables of interest (pct50, pct67, pct83) now represent the within-polygon effect of these variables on visit_post_carto. In other words, they measure the effect of changes in these variables within the same polygon over time.
2. The vce(robust) option is used to compute robust standard errors, which are robust to heteroskedasticity and certain types of specification errors. This is often a good choice as it makes fewer assumptions about the error term and can provide more reliable standard errors in the presence of heteroskedasticity or other violations of the classical linear regression model assumptions.

12. Regression

The regression output indicates that all the variables (pct50, pct67, pct83, visited, visits) have been omitted from the model. This typically happens when there is perfect multicollinearity, meaning that one variable can be perfectly predicted by a combination of other variables.

```
. regress taxes_paid $xlist

      Source |       SS           df       MS      Number of obs   =     1,209
-------------+----------------------------------   F(5, 1203)      =         .
       Model |           0         5           0   Prob > F        =         .
    Residual |           0     1,203           0   R-squared       =         .
-------------+----------------------------------   Adj R-squared   =         .
       Total |           0     1,208           0   Root MSE        =         0

------------------------------------------------------------------------------
  taxes_paid |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       pct50 |           0  (omitted)
       pct67 |           0  (omitted)
       pct83 |           0  (omitted)
     visited |           0  (omitted)
      visits |           0  (omitted)
       _cons |           1          .        .       .            .           .
------------------------------------------------------------------------------
```

```
. * Run regression for all properties to attempt to fix multicollinearity

. reghdfe taxes_paid $xlist, vce(robust)
(MWFE estimator converged in 1 iterations)
warning: dependent variable taxes_paid is likely perfectly explained by the fixed effects
> (tol = 1.0e-09)
warning: missing F statistic; dropped variables due to collinearity or too few clusters

HDFE Linear regression                            Number of obs   =      1,209
Absorbing 1 HDFE group                            F(   5,   1203) =          .
                                                  Prob > F        =          .
                                                  Root MSE        =     0.0000

------------------------------------------------------------------------------
             |               Robust
  taxes_paid |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       pct50 |           0  (omitted)
       pct67 |           0  (omitted)
       pct83 |           0  (omitted)
     visited |           0  (omitted)
      visits |           0  (omitted)
       _cons |           1          .        .       .            .           .
------------------------------------------------------------------------------
```

In this case, it seems like multicollinearity can distort the coefficient estimates and inflate the standard errors, which makes it difficult to determine the individual impact of each predictor on the response variable.

To address multicollinearity, there are multiple approaches that I could take:

1. Remove some of the highly correlated predictors: If two or more predictors are highly correlated, I can consider removing one or more of them. The choice of which one to remove should be based on further understanding of the data and research objectives.
2. Combine the correlated predictors: If the predictors that are highly correlated represent similar underlying phenomena, I can consider combining them into a single predictor.
3. Using regularization methods: I can use techniques like Ridge Regression or Lasso can help in handling multicollinearity.

It is important to understand the assumptions of the regression model and check them using appropriate diagnostic plots and tests. If the assumptions are violated, the results of the regression may not be valid. It's also crucial to interpret the results in the context of the data and the research question. Therefore, in order to avoid going out of context, I rest the case here and declare that I cannot reach conclusive results that allow me to make any analysis using the variables given to be in the regression without doing further data manipulation that can be out of context of the asked question.

13. Regression

It does not make sense to apply the natural logarithm (ln) transformation to a dummy variable '***taxes paid***', hence no export of any results.

A dummy variable is a binary variable that takes on the values 0 or 1. The natural logarithm of 0 is undefined, and the natural logarithm of 1 is 0. Therefore, applying the ln transformation to a dummy variable would result in undefined or zero values, which would not be meaningful or useful in a regression analysis.

In general, the ln transformation is used on continuous, positive variables to help linearize relationships, stabilize variance, or normalize the distribution of the variable. It's not appropriate or useful for binary variables. I avoided going into making a probit or a logistic regression or even attempting to use other methods such as including interaction terms.

Part 2: <u>French Part</u>

<u>Question 2, 1:</u>

L'Organisation d'Études Économiques au Kasaï (ODEKA) conduit une R.C.T dans la ville de Kananga, pour laquelle elle est responsable de la collecte de données Baseline, du déploiement de l'intervention et de la collecte de données Endline. La randomisation des groupes (traitement ou contrôle) est faite au niveau des quartiers. L'intervention a déjà commencé dans certains quartiers, cependant, dans le quartier de Tchiniambi, la collecte de données Baseline s'éternise ... alors que notre partenaire local (le ministère des infrastructures, travaux publics et reconstruction, aménagement du territoire, urbanisme et habitat,

affaires foncières et relation avec le parlement) s'impatiente ... Nous devons lancer l'intervention au plus vite! Espoir, un des enquêteurs affectés au quartier Tshishimbi, a finalement pu prendre rendez-vous avec le dernier ménage que nous avons besoin de sonder pour clôturer la collecte Baseline dans ce quartier. Cependant, ce rendez-vous interviendra après la randomisation, de telle sorte que les autres ménages du quartier auront déjà été mis au courant par les enquêteurs de leur groupe d'intervention (traitement ou contrôle).

Ce que J'en pense:

Importance de la collecte des données *Baseline* dans ce cas
La collecte des données *Baseline* est une étape très importante dans la conduite d'un essai contrôlé randomisé (ECR). Elle fournit un aperçu de la situation actuelle avant la mise en œuvre de l'intervention, permettant une comparaison des résultats entre les groupes de traitement et de contrôle. La collecte des données *Baseline* aide à établir un point de référence par rapport auquel l'impact de l'intervention peut être mesuré.

Dans le cas de l'ECR de l'ODEKA à Kananga, la collecte des données *Baseline* dans le quartier de Tshinsambi est nécessaire pour établir les conditions initiales et les caractéristiques des ménages avant le début de l'intervention. Ces données serviront de point de référence pour évaluer l'efficacité de l'intervention à l'avenir.

La situation décrite est problématique
La situation décrite soulève des préoccupations potentielles quant au risque de compromettre l'intégrité de l' ECR en raison du retard de la collecte des données *Baseline* dans le quartier de Tshinsambi. Plus précisément, si le dernier ménage est interrogé après la randomisation et le déploiement de l'intervention dans d'autres ménages du même quartier, il existe un risque de contamination ou de réponses biaisées de la part de ce ménage où ils pourraient essayer d'influencer le groupe dans lequel ils veulent être ou donner d'autres réponses incorrectes. Cette situation pourrait se produire si le ménage prend connaissance du statut de traitement ou de contrôle de ses voisins, influençant potentiellement ses réponses lors de l'enquête *Baseline*.

Comment aborder la situation ?
À mon avis, pour maintenir la rigueur et la validité de l'ECR, il est crucial de s'assurer que toute la collecte des données *Baseline* au sein d'un quartier soit terminée avant que la randomisation et le déploiement de l'intervention n'aient lieu dans ce quartier. Toute dérogation à ce protocole pourrait introduire un biais et compromettre la capacité de l'étude à tirer des conclusions causales précises.

Pour aborder cette situation tout en préservant l'intégrité de l'ECR, voici quelques options potentielles que je propose :

1. Retarder la randomisation et le déploiement de l'intervention pour l'ensemble du quartier de Tshinsambi jusqu'à ce que l'enquête *Baseline* du dernier ménage soit terminée. Cela garantit qu'aucun ménage n'est au courant de son statut de traitement ou de contrôle ou de celui de ses voisins avant d'avoir complété l'enquête *Baseline*.

2. Si le report de l'intervention n'est pas possible, envisager de réaffecter le dernier ménage à un autre quartier où la collecte des données *Baseline* est encore en cours et où aucune intervention n'a débuté. Cela préviendrait toute contamination au sein de Tshinsambi.

3. En dernier recours, exclure le dernier ménage de l'échantillon de l'étude pour Tshinsambi. Bien que cela puisse légèrement réduire la puissance statistique ou entraîner des erreurs statistiques, cela évite de compromettre la validité de l'ECR en prévenant tout biais potentiel dans les réponses *Baseline* des autres ménages.

Quelle que soit l'option choisie, une documentation claire et une transparence sur la dérogation au protocole initial sont essentielles. De plus, les chercheurs devraient examiner attentivement les caractéristiques du dernier ménage pour évaluer si son exclusion pourrait introduire un biais systématique dans l'échantillon de l'étude.

En résumé, à mon avis, il est crucial de prioriser l'intégrité des processus de randomisation et de collecte des données *Baseline*, même si cela nécessite de retarder les interventions ou de prendre des décisions difficiles quant à l'exclusion de certains ménages de l'analyse.

Question 2, 2:

Dans le cadre d'un projet de recherche mené à Kananga, nous souhaitons, entre autres, mesurer la perception des citoyens concernant la stabilité du gouvernement provincial. Les PIs (Principal Investigators) vous font parvenir le sondage Baseline « prêt à être déployé » pour le jour suivant. Alors qu'il est déjà tard, vous vous apercevez qu'une des questions n'a pas été traduite en français: "Do you think current governance is stable enough or too unstable?"

Ce que je ferais:

Voici une situation d'une simple erreur d'omission de traduction. Je veux clarifier que cela peut avoir des conséquences potentielles si ce n'est pas pris en charge correctement.

Dans cette situation, où une question clé de l'enquête n'a pas été correctement traduite en français avant le déploiement prévu de l'enquête de référence, je prendrais ou recommanderais les mesures suivantes :

1. Informer immédiatement les chercheurs principaux (PIs) de cette omission et du manque de traduction française pour la question. Je pourrais aussi expliquer l'importance d'avoir toutes les questions disponibles dans la langue locale (français dans ce cas) pour assurer une collecte de données précise et cohérente.
2. Demander un court report du déploiement de l'enquête de référence, ne serait-ce que d'un ou trois jours, pour permettre une traduction et un examen appropriés de la question. Ce délai est crucial pour préserver l'intégrité du processus de collecte des données.
3. Contacter un traducteur professionnel maîtrisant l'anglais et le français pour traduire avec précision la question d'enquête manquante. S'assurer que la traduction est culturellement adaptée au public cible à Kananga.

4. Examiner attentivement la question traduite avec les PIs et les autres membres de l'équipe concernés pour vérifier son exactitude, sa clarté et son alignement avec le ton et le style globaux de l'enquête.
5. Si le temps le permet, envisager de réaliser un petit test pilote de la question traduite auprès de quelques répondants de la population cible. Cela peut permettre d'identifier et de résoudre toute ambiguïté ou incompréhension potentielle avant le déploiement complet.
6. Une fois la traduction validée, mettre à jour le document d'enquête de référence avec la nouvelle question en français, en confirmant que tous les autres éléments de l'enquête sont exacts et prêts pour le déploiement.
7. Notifier les PIs et tous les membres de l'équipe concernés au sujet du document d'enquête mis à jour. Insister sur l'importance d'utiliser la version révisée pour la collecte de données afin de maintenir la cohérence et la validité.
8. Surveiller le déploiement de l'enquête de référence pour s'assurer que la question mise à jour est correctement utilisée et que les répondants la comprennent et y répondent de manière appropriée.

À l'ère actuelle, où le partage d'informations est très rapide, je tirerais également parti des avantages de la technologie moderne pour résoudre ce problème plus rapidement. Je peux communiquer rapidement le problème, suggérer des solutions et interrompre le déploiement pendant quelques heures ou quelques jours pour permettre une enquête complète.

Assurer une traduction précise des questions d'enquête, en particulier celles liées à des sujets sensibles ou spécifiques au contexte comme la stabilité de la gouvernance, est crucial pour recueillir des données fiables et significatives.

Une question traduite à la hâte ou de manière inappropriée pourrait entraîner des erreurs de mesure, compromettant la validité des résultats de la recherche. C'est pourquoi je préférerais prendre un court délai plutôt que de risquer de collecter des données erronées qui pourraient compromettre l'ensemble de l'étude.

En prenant ces mesures rapidement et en accordant la priorité à l'exactitude de la traduction et de l'examen, je peux efficacement résoudre le problème de traduction manquante sans compromettre la qualité et la validité du processus de collecte de données pour ce projet de recherche.

Ended: Mar 22, 2024 12:00 PM