De-identification of Data:

Given the high sensitivity of the data – survey on corruption perceptions – and the potential risk of retaliation, our team went to great lengths to deidentify the data. Personal identifiers were either removed or aggregated to preserve the anonymity of individual respondents.[1] At the same time, an important consideration was allowing for institutional analysis of the data, incorporating identifiers such as ministries. Our approach sought to balance these competing concerns.

Note that in our context, we had information on both the survey respondents and the population surveyed, through an administrative data on personnel. This allowed us to engage in two verification processes for anonymity:

1) If respondents are individually identifiable within bins of the survey itself, i.e. are the bins large enough that individual respondents cannot be singled out.
2) If respondents are individually identifiable within bins of the administrative data, i.e. can individual respondents be identified within the personnel data, using personal identifiers.

The personal identifiers in the data include:

1. State.
2. Ministry.
3. Gender.
4. Education level.
5. Hierarchical position (leadership).

The first step was to aggregate states (27) into regions (5). Ministries were left as is, to allow for institutional analysis, as well as gender. Education levels were aggregated to 2 levels: higher education and non-higher education. Due to the importance of having the perspective of public sector administrative leaders, we decided to retain this personal identifier. However, given their limited number, for all respondents who were in leadership position, we removed any other personal identifier.

To assess the risk of identification, we used the **sdcMicro** package to assess the global risk to identifying respondents. For verification process 1, to assess whether individual respondents can be identified within the survey, we first generated bins for:

1. Ministry.

[1] The de-identification process is encoded in the **protect_confidentiliaty.R** file.

2. Region.
3. Gender.
4. Education bin.

Based on these bins, we calculated the risk of identifying individuals. That is, we first created a count F_{survey} all individuals with a given set of characteristics (e.g. Ministry of Education, Region North, gender female and higher education) and then estimated the risk factor of any given individual to be identified in that bin: 1/ F_{survey}. We then calculate the global risk for all individuals in our survey sample as an average of all risk factors. Our de-identification strategy gives us a global risk of 1.53%, far below the recommended 5% for de-identification purposes.

For verification process 2, we assessed whether individual survey respondents could be linked to personnel data. To do so, we first constructed analogous bins in the survey and personnel data. For instance, the same bin outlined above (e.g. Ministry of Education, Region North, gender female and higher education) was constructed for both the survey and personnel data. Then, we constructed an F_{survey} and $F_{personnel}$ by summing the total number of respondents and civil servants respectively. The risk factor is then $F_{survey}/F_{personnel}$. The average of risk factor for all bins is 4.7%, which is below the recommended 5% as well.