<https://docs.google.com/spreadsheets/d/1h_0Vtv8hCXWnpNeBQMtpEfWaflLCCqJr/edit#gid=214178044>

1. Agradecemos el PDF compartido. Da por resuelto el tema de nombres de variables
   * Sin embargo, necesitamos las definiciones de los campos y sus valores aún.
2. Si podemos generar un entendimiento sobre cómo generan los ID de unidades creo que pudiéramos aclarar parte/todos los siguientes temas:
   * Importante para datos de 2021 incluir 2019 y 2020, e incluir ID de individuo cómo se detalla en la solicitud.
   * No sabemos qué ocurre en casos particulares en donde el ID de unidad bajo el cual está una persona puede cambiar año tras año (ex: si en 2019 rinde planilla como individuo, y en 2020 lo hace su pareja).
   * Hemos identificado IDs duplicados dentro del mismo año contributivo para ambos 2019 y 2020.

* La explicación provista de personas cuyo cónyuge muere explica la mayoría de los casos. Sin embargo, unos pocos casos no tienen cambios de estado sino en otros campos.
* Además, algunas unidades pasan de Individuo a Casado (contrario a la explicación en el e-mail sobre la muerte de cónyuge) - pero aquí es posible que no estemos identificando correctamente la primera y segunda ocurrencia de un duplicado - cómo hacemos eso?
  + En algunos casos, la información categórica de un mismo ID en el 2019 es distinta a la que se provee para el 2020.
* Generalmente es info de estado personal o trabajo, pero alrededor de 1% de los IDs en ambos años cambian de sexo entre 2019 y 2020. La mayoría son de F a M o M a F (hay casos en los cuales van de no tener info a tenerla, pero son pocos).

1. Work credit income is always 0 in 2020; sometimes work credit income > total income.
2. EITC inconsistente
   * Dando por entendido que el ‘flag’ no se ajusta si el monto es ajustado luego de verificación, quisiéramos confirmar entonces que para determinar los montos de crédito por trabajo, sería mejor ignorar el ‘flag’.
   * Además quisiéramos aclarar si hubo algún cambio en cómo se lleva este récord entre 2019 y 2020; en 2020 no hay casos con ‘No’ y monto > $0, y la cantidad de casos con ‘Yes’ y monto de $0 son muchísimo menos.
   * Si el crédito por trabajo reduce la responsabilidad contributiva, se incluye el total, o solo la porción que el contribuyente recibe en depósito?
     1. En reintegro esta el neto de todos los creditos y demas
     2. Tenemos reintegro?
   * Quisiéramos confirmar si casos con ‘Yes’ y montos de $0 son casos de personas que al parecer creían que eran elegibles, pero al Hacienda verificar se determinó que no lo eran. Similarmente con casos con ‘No’ y monto mayor a $0 (ie: si creían que no eran elegibles, pero Hacienda determinó que sí)
3. Data entry
   * Dando por entendido que algunos valores sospechosos surgen de posibles errores por parte de contribuyentes, buscamos aclarar qué columnas están propensas a esto. Por lo entendido, la mayoría de las columnas se afectan (info demográfica de contribuyente, pareja, y dependientes, cantidades de ingreso y pensión)

* Buscamos confirmar si los montos de crédito por trabajo, y responsabilidad contributiva se afectan.
  + Además, quisiéramos aclarar unidades con ‘Yes’ en el campo de pensión, teniendo un monto positivo, pero sin información de dependientes.

1. Draft lista resumida

* Nombres y definiciones de campos
  + El archivo no contiene los nombres de cada campo, por lo que se deben inferir según los valores que estos contienen.
  + El archivo no está acompañado de un anejo que detalle las definiciones de los campos y de los valores que estos contienen.
* Desconocemos cómo se generan los IDs únicos por unidad contributiva
  + No sabemos qué ocurre en casos particulares en donde el ID de unidad bajo el cual está una persona puede cambiar año tras año (ex: si en 2019 rinde planilla como individuo, y en 2020 lo hace su pareja).
  + Hemos identificado IDs duplicados dentro del mismo año contributivo para ambos 2019 y 2020.
  + En algunos casos, la información categórica de un mismo ID en el 2019 es distinta a la que se provee para el 2020.
* Campos con valores sospechosos
  + Años de nacimiento (de quienes rinden planillas y sus cónyuges) incluyen los primeros diez años de los 1900s y los pasados 10 años (2012-2022).
  + Gran número de unidades con ‘Yes’ en el campo de sobre crédito por trabajo, pero con un monto de ‘$0’.
  + El conteo de unidades con ‘Yes’ en el campo de pensión parece bajo. El monto total de pensión parece alto relativo a dicho conteo.

1. Draft tabla detallada

| **Tema** | **Problema** | **Ejemplos de impacto** | **Solución** |
| --- | --- | --- | --- |
| 1 (1) | El archivo no contiene los nombres de cada campo. Los nombres de campos se deben inferir según los valores que estos contienen. | Puede llevar a interpretar campos como algo distinto a lo que realmente representan (ex: no es del todo claro cómo se diferencian los cuatro campos de ingreso entre sí) | Proveer nombres de campos en el archivo según el órden que aparecen en cada fila. |
| 2 (1) | El archivo no está acompañado de un anejo que detalle las definiciones de los campos (todos) y de los valores que estos contienen (en casos de campos categóricos). | El campo de SSN incluye como posibles valores a 'SSN', 'PSSN', 'ITIN' - carecemos de certeza de los significados de estos valores. | Proveer un diccionario que detalle las definiciones de cada campo y sus valores (en casos de campos categóricos). |
| 3 (meet) | El campo binario (Yes/No) que indica si la unidad recibió crédito por trabajo no siempre parece consistente con el campo del monto de crédito por trabajo. Por ejemplo, hay casos marcados como ‘Yes’ en el primer campo, pero con un valor de ‘$0’ en el segundo (y viceversa). | Las sumas de crédito por trabajo (ya sea por año o municipio) difieren al sumar por estos dos métodos:  i) sin tomar en cuenta el campo binario  ii) tomándolo en cuenta, e ignorando los montos de aquellas unidades con ‘No’ en el primer campo. | Aclarar la definición del campo según se indica en el tema #2, y luego determinar si efectivamente hay inconsistencias. |
| 4 (2) | Desconocemos cómo se generan los IDs únicos por unidad contributiva. Por ello, no sabemos qué ocurre en casos particulares en donde el ID de unidad bajo el cual está una persona puede cambiar año tras año. | Tomemos personas A y B que rinden planillas como individuos en 2019. Si en 2020 se casan y la persona A rinde su planilla conjunto con B, entendemos que se preservará solo el ID de la persona A, y no el de la persona B.  Tomemos persona A que rinde planilla como casado conjunto con B en el 2019, y la persona C que lo hace como individuo. Si luego en 2020 las personas A y B se divorcian, y C se rinde planilla como casado con la persona A, entendemos que se preservará solo el ID original de la persona C. Esto ya que la persona B tendría un ID nuevo como individuo, y el ID de la persona A pasaría a ser el de la persona C.  En fin, esto imposibilita análisis propios tipo panel time-series (ya sea a nivel de individuos o a nivel de unidades).  Ver además los temas #5 y #6. | Detallar el proceso de generación de IDs de unidades para habilitar al equipo de datos para manejar casos como los mencionados en el ejemplo tal que se pueda dar seguimiento a toda unidad e individuo a través del tiempo (siempre y cuando llene planilla).  Entendemos que la solución posiblemente requiera añadir un campo de ID único por persona, tal que sea posible dar seguimiento a una persona sin importar si esta cambia estado personal y/o cónyuge año tras año. Este campo se solicitó en la petición de información pero, debido al formato de los datos, parece ser el único que no fue incluido. |
| 5 (2) | Relacionado al tema #4, hemos identificado IDs duplicados dentro del mismo año contributivo - 73 para 2019 y 102 para 2020.  En todos los casos, cada ID que se duplica dentro de un mismo año aparece un total de dos veces. Los campos de estado personal e información del cónyuge tienden a ser distintos para cada par de duplicados, pero hay otros campos que una que otra vez también difieren entre pares.  En 2019 (2020), hay uno (dos) duplicado(s) para el cual la información en todos los campos es idéntica. | Esto imposibilita el escoger una observación (o fila) para cada par de IDs duplicados dentro de un mismo año, ya que desconocemos cuál de las filas es precisa. | Aclarar cómo se genera el campo según se indica en el tema #4, y luego determinar si efectivamente hay inconsistencias. |
| 6 (meet, maybe) | Para los 980,724 IDs que están bajo ambos años contributivos del 2019 y 2020, el siguiente listado desglosa la cantidad de veces que algunos campos categóricos cambiaron para el mismo ID de un año para otro:  SSN = 3;  fecha\_nacimiento = 0;  estado\_personal = 35,417 (3.6%);  sexo = 16,799 (1.71%);  NAICS = 148,964 (15%);  cat\_empleado = 107,607 (11%);  fecha\_nacimiento\_pareja = 24,074 (2.4%);  sexo\_pareja = 33,274 (3.4%);  Siendo estos campos categóricos generalmente fijos año tras año para cada persona, quisiéramos confirmar que la información representa la realidad de las unidades, y que los IDs efectivamente representan la misma unidad contributiva independientemente del año. | Entendemos como posibles causas/efectos:  i) IDs que surgen en ambos años representan unidades distintas, lo cual sería obstáculo para análisis tipo panel a través del tiempo  ii) IDs que surgen en ambos años tienen una correción de información en alguno de los dos años relativo al otro - en este caso no habría inconsistencias  iii) IDs realmente reportan cambios en sus planillas entre 2019 y 2020 - en este caso no habría inconsistencias | Aclarar cómo se genera el campo según se indica en el tema #4, y luego determinar si efectivamente hay inconsistencias. |
| 7 | Para ambos quienes rinden planillas y cónyuges, hay fechas de nacimiento entre 2018 y 2022, e incluso en los primeros diez años de 1900s. Entendemos que las observaciones de mayor edad son sospechosas, y las de menor simplemente imposibles.  Conteo aprox. de observaciones por fechas de nacimiento para quienes rinde planilla:  entre 1901 y 1922: 1300  entre 2005 y 2022: 250  entre 2010 y 2022: 177 | Esta información introduce ruido a cualquier análisis que tomen en cuenta la edad de quien rinde la planilla en la unidad. | Limpieza de los campos de año de nacimiento, ambos para quien rinde planilla y para cónyuges. |
| 8 | El conteo de ‘Yes’ para el indicador sobre pensión parece muy bajo respecto a la cantidad de observaciones en ambos años. El monto, por otro lado, parece muy alto relativo a la cantidad de observaciones.  En 2019 (2020), 193 (224) tienen un ‘Yes’ y 1,365,449 (1,329,737) un ‘No’.  En 2019 (2020), la suma del monto de pensión iguala $3,690,384 ($4,110,748).  En 2019 (2020), approx. 100 (125) observaciones con un ‘Yes’ tienen un monto mayor de $10,000. | Esta información introduce ruido a cualquier análisis que tomen en cuenta temas de pensión. | Limpieza de ambos campos de pensión. |