OPSO79-1-UCSH2021

Tipos de datos cuantitativos e introducción a dplyr en R. Bloque teórico.

27/08/2021

Introducción

Introducción

Cuando analizamos datos cuantitativos, representaciones numéricas del mundo social, nos encontraremos con cosas muy diferentes.

Revisaremos los distintos tipos de datos cuantitativos que nos encontraremos según distintos criterios (orden, nivel de medida y nivel de agregación)

Definiremos también cuestiones básicas como que es una variable, que es una observación, que implica que un dato esté ordenado y que procesos garantizan que un dato sea de calidad.

Según el tipo de datos que tengamos dependerá el tipo de análisis a realizar (o incluso acciones previas como ordenar el dato).

Producción de un orden

Los datos rectangulares (tablas, *data frame*) y de calidad son una producción que demanda **tiempo**, **trabajo** y **dinero**.

Sin importar el cómo son producidos:

- Cuestionario y entrevistas
- Observación directa o indirecta
- Análisis de contenido de documentos (episódicos o sistemáticos)
- Análisis de contenido de producciones visuales y audiovisuales
- Registros sistemáticos administrativos (cotizaciones, huelgas)
- *big data*

Sin importar quien los produce:

- producido nosotros para nuestros propios objetivos de investigación (**primarios**)

Producción de un orden

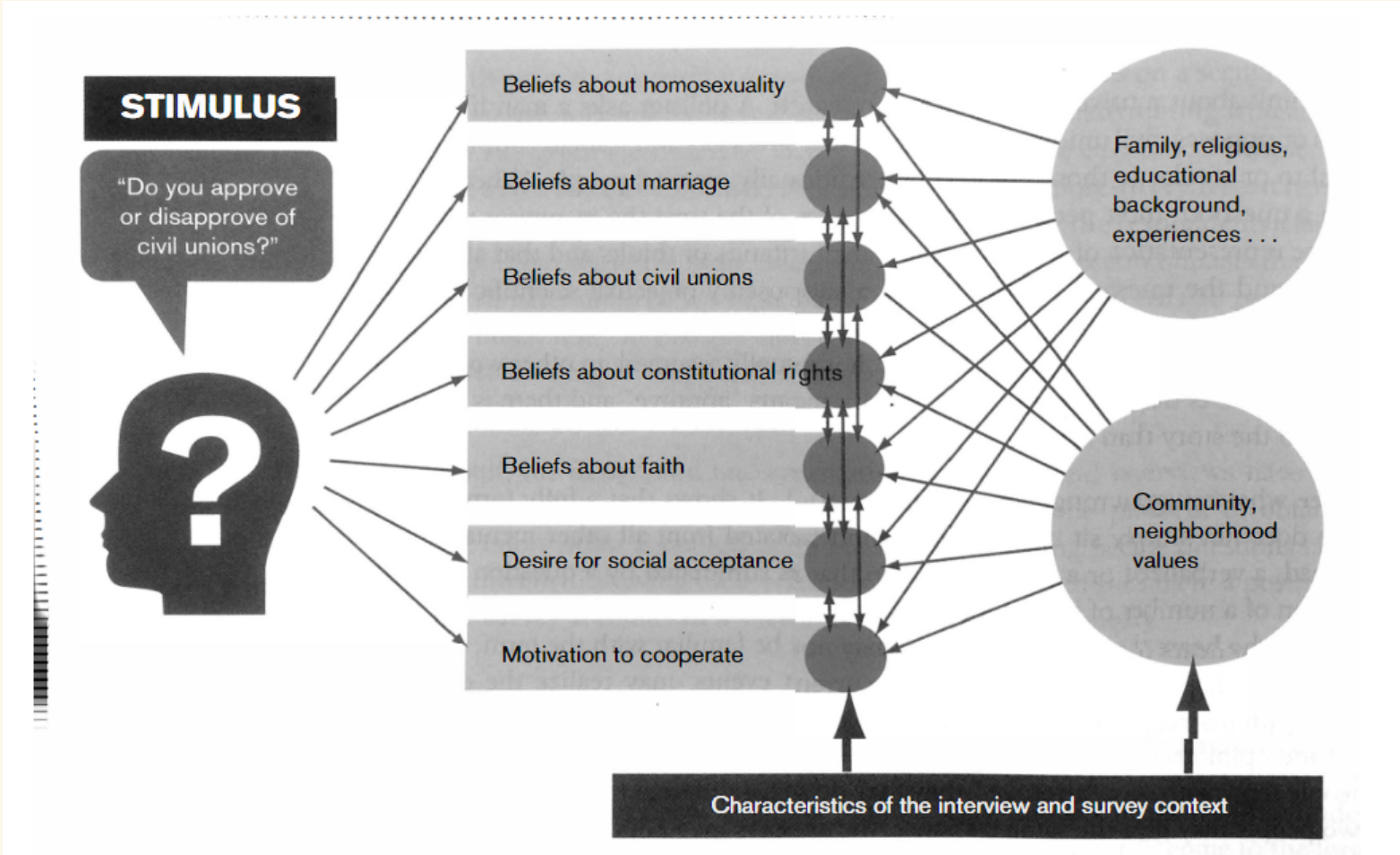
El dato ordenado y de calidad no es fortuito ni está en estado natural.

Los datos responden a una necesidad de información y son producidos en el marco de un diseño.

Posteriormente los datos pasan por un proceso de **validación**, **ordenamiento** y **análisis**.

La subjetividad del encuestado debe pasar a una hoja estándar (cuestionario), el cuestionario marcado por el encuestador debe pasar a una tabla, el texto codificado, el registro debe ser depurado, los valores no deben ser contradictorios entre sí ni imposibles, no todo se libera, etcétera.

Producción de un orden



Producción de un orden

Cerca de 1.300 trabajadores de la División Andina de Codelco podrían llegar a estar en huelga si es que se confirma la decisión del Sindicato de Unión Plantas (Suplant) de rechazar las propuestas de la empresa.

Hasta el cierre de esta edición, se esperaba que la empresa presentara una nueva oferta -cerca de las 20:30 horas- para intentar desactivar el conflicto en el marco de la negociación colectiva de la minera, después de que la organización rechazara una propuesta previa, lo que hasta el mediodía de este lunes hacía pensar que la huelga se haría efectiva a partir de las 8 horas de este martes.

Un punto que mantenía trabadas las conversaciones, era el traspaso de beneficios para trabajadores nuevos de la estatal, según explicó el presidente del sindicato.

Esta negociación colectiva se da casi en paralelo a la que llevan el Sindicato Industrial de Integración Laboral (SIIL) y Sindicato Unificado de Trabajadores (SUT) de esa misma división de la estatal, los que ya habían rechazado las últimas ofertas de la compañía y habían comenzado sus paralizaciones el jueves de la semana pasada.

De confirmarse la decisión de Suplant, los trabajadores paralizados en esa faena llegarían hasta las 1.300 personas movilizadas.

Además, se está a la espera de la negociación del Sindicato de Supervisores, el que recibiría este martes la última oferta por parte de la estatal.

Presión en la industria

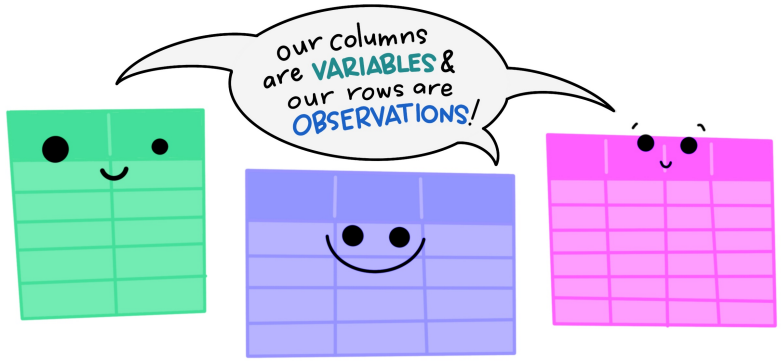
Consultado por los motivos que los alejan de la empresa, el presidente de Suplant, Clodomiro Vásquez, señaló que se pretende eliminar los beneficios en salud de los trabajadores nuevos, entre otros aspectos como la indemnización por año de servicio.

D	E	F	G	H	I	J	K	L
nnot	prensa	yr	mes	inicio	fin	duracion	ddpp	leg
0		2015	11	2015-11-27	2015-12-04	8,0	6,0	1
0		2015	11	2015-11-30	2015-12-11	12,0	10,0	1
0		2015	11	2015-11-30	2015-12-18	19,0	15,0	1
0		2015	11	2015-11-30	2015-12-04	5,0	5,0	1
6	1	2015	11	2015-11-30	2015-12-17	17,0	13,0	2
0		2015	12	2015-12-03	2015-12-14	12,0	10,0	1
0		2015	12	2015-12-03	2015-12-04	2,0	2,0	1
0		2015	12	2015-12-04	2015-12-10	7,0	5,0	1
0		2015	12	2015-12-04	2015-12-04	1,0	1,0	1
0		2015	12	2015-12-04	2015-12-18	15,0	11,0	1
0		2015	12	2015-12-04	2015-12-07	4,0	4,0	1
0		2015	12	2015-12-07	2015-12-16	10,0	8,0	1
0		2015	12	2015-12-07				1
3	4	2015	1	2015-12-09	2016-01-15	38,0	28,0	1
0		2015	12	2015-12-09	2016-01-20	42,0	30,0	1
0		2015	12	2015-12-10	2015-12-14	5,0	5,0	1
2	2	2015	12	2015-12-10	2015-12-11	0,1	0,1	2
2	2	2015	12	2015-12-11	2015-12-13	2,0	2,0	2
1	1	2015	12	2015-12-12	2015-12-12	1,0	1,0	2
1	2	2015	12	2015-12-12				2
0		2015	12	2015-12-14	2015-12-22	9,0	7,0	1
0		2015	12	2015-12-15	2015-12-17	3,0	3,0	1
1	4	2015	1	2015-12-16	2016-01-07	23,0	21,0	1
0		2015	12	2015-12-16	2016-01-06	20,0	16,0	1
1	2	2015	12	2015-12-16	2015-12-16	1,0	1,0	2
184	3	2015	12	2015-12-16	2015-12-20	4,0	4,0	2
0		2015	12	2015-12-17	2015-12-30	14,0	10,0	1
23	3	2015	12	2015-12-17	2016-01-17	27,0	17,0	2
1	1	2015	12	2015-12-17	2015-12-17	1,0	1,0	2
1	1	2015	12	2015-12-18	2015-12-22	4,0	4,0	2
0		2015	12	2015-12-21	2016-01-12	23,0	17,0	1
0		2015	12	2015-12-21	2016-01-07	17,0	13,0	1

Producción de un orden

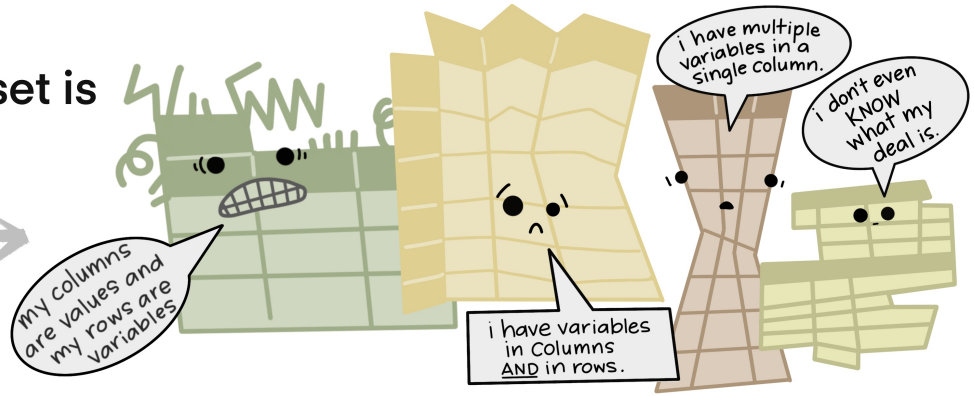
Producción de un orden

The standard structure of tidy data means that "tidy datasets are all alike..."



"...but every messy dataset is messy in its own way."

–HADLEY WICKHAM



El proceso estadístico

Nada garantiza que el dato que nos llega pasó por estos controles (preguntarse por qué institución lo produjo y con que fines es un buen filtro inicial).

La producción del dato cuantitativo está estandarizada por las ONEs.

El GSBPM (*Generic Statistical Business Process Model*) es un estándar internacional consolidado y adoptado por numerosas oficinas de estadística y organismos internacionales que propone una estructura de procesos y subprocesos del modelo de producción de estadísticas.

Pese a su esquematismo el marco es flexible y adaptado.

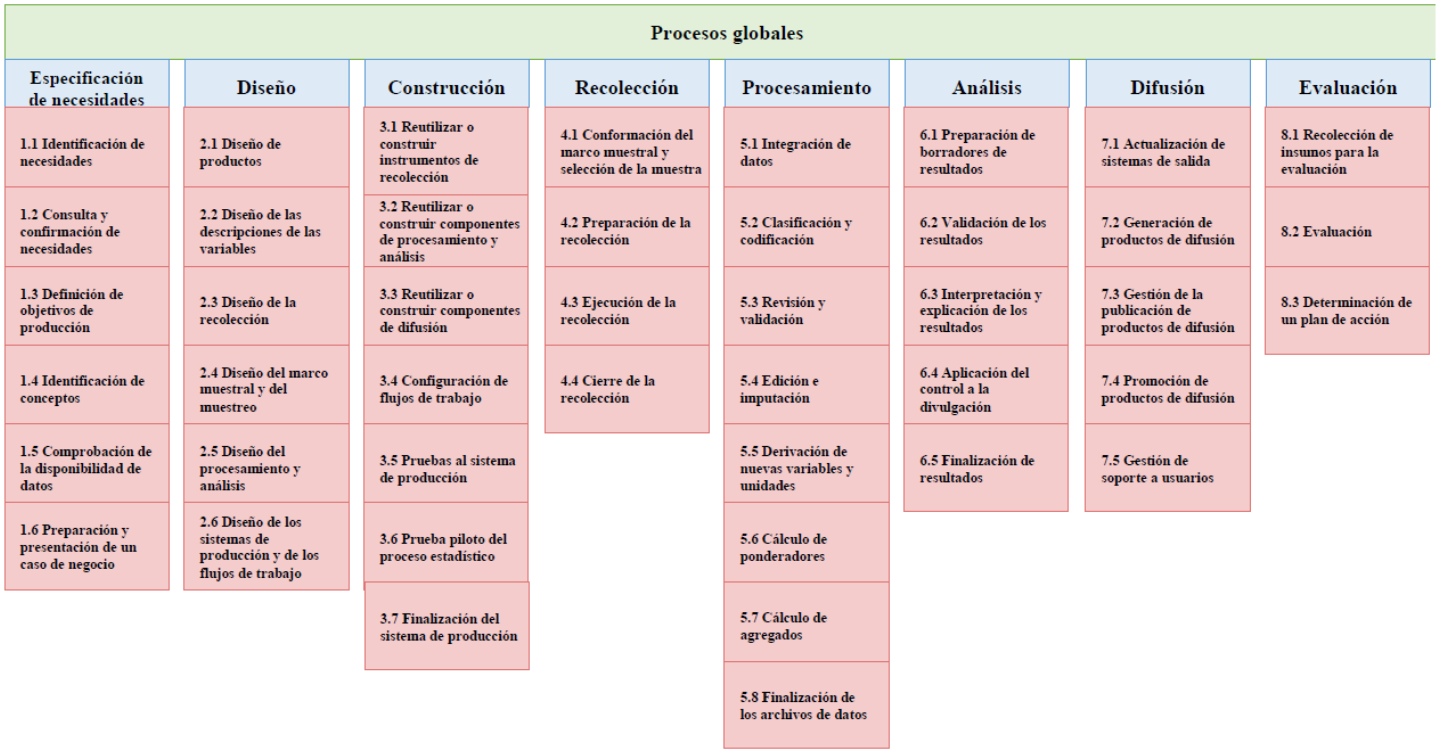
No son las fases de un diseño de investigación ni el modelo de la ciencia de datos.

Es el marco en el cuál se piensan y ordenan los procesos de producción estadística nacionales.

Como usuarios de los datos producidos por las ONEs es relevante comprender el proceso de producción detrás del dato.

El proceso estadístico

Figura 3. Las fases (nivel 1) y subprocesos (nivel 2) del GSBPM



El proceso estadístico

El proceso estadístico llega a tal punto que no solo produce las *data frame* que como investigadores/as ocuparemos.

Llega a analizar e incluso interpretar y explicar los resultados.

Al hacer este proceso las instituciones **transforman** y **visualizan** los datos según sus objetivos, produciendo **tabulados**, **cuadros estadísticos** y **gráficos** que resumen la información.

No hay que confundir los productos estadísticos:

- Microdato o data
- Cuadro o tabulado

Si bien ambos se pueden leer como "data frames" en R, las filas del microdato corresponden a individuos u otras unidades de observación y/o análisis (hogares, empresas, eventos de protesta).

Algunas definiciones

Microdata: *"Conjunto de registros que contienen información sobre encuestados individuales o entidades económicas. Dichos registros pueden contener respuestas a un cuestionario de encuesta o formularios administrativos."* (SDC Practice guide).

Unidades de análisis: *El qué o quién está siendo estudiado. En la investigación en ciencias sociales, las unidades de análisis más típicas son las personas individuales (...) Es importante distinguir entre la unidad de análisis y los agregados sobre los que generalizamos.* Babbie (2014).

Unidades de observación: Las unidades de análisis en un estudio suelen ser también las unidades de observación. Las unidades de observación corresponden a quien(es) nos entregan la información de las unidades de análisis.

(Micro) datos ordenados

Combinación de observaciones y variables.

Cada fila es una observación.

Cada variable es una propiedad de las observaciones.

Cada celda es un solo valor.

El término variable hace referencia a que el valor que asume **varía** entre las distintas observaciones.

Los valores pueden ser milímetros de precipitaciones, o respuestas "sí" o "no" a la pregunta de sí llovió.

Las variables se pueden clasificar por su **nivel de medición**. Cada nivel de medición incorpora distintas propiedades de los números.

Variables categóricas

La variable será **categórica** si cada observación pertenece a un set de diferentes categorías.

Aunque ocupemos números para representar categorías como un "sí" (1) o un "no" (2), la variable sigue siendo categórica.

Para las categóricas será clave saber el número de observaciones en cada categoría.

Existiendo protestas con demandas económicas, políticas y culturales, ¿Cuál fue el porcentaje de protestas económicas en 2019?

Las variables categóricas se pueden usar tanto para **ordenar** (ordinales) como para **clasificar** (nominales).

Variables categóricas

En el **nivel nominal** la función de los números es sólo distinguir entre sujetos que posean la propiedad de manera igual o diferente.

La asignación del número de cada categoría es arbitraria (bien podría ser 1, 2 y 3, o 567, 89 y 77)

La variable puede indicar la presencia o ausencia de una propiedad (**nominal dicotómica**).

En el nivel **ordinal* la asignación de números respeta el orden en que los sujetos poseen la propiedad medida.

- Nivel educacional (1 = Básica incompleta, 2 = Básica completa, 3 = Media Incompleta, etcétera)
- Nivel de acuerdo con afirmaciones (e.g. *la democracia es siempre preferible a un régimen de gobierno autoritario*).

Variables cuantitativas

La variable será **cuantitativa** si la observación toma valores numéricos que representan diferentes magnitudes de una variable.

Este nivel incorpora la noción de distancia entre las magnitudes en que los objetos poseen las diversas propiedades.

Características claves de las **variables cuantitativas** serán su **centro** y **variabilidad**,

¿cuántas personas iban al día en las protestas de 2019?, ¿Cómo varió la participación desde octubre hasta fin de año?

Pueden ser **discretas** o **continuas**.

Variables cuantitativas

La variable es **discreta** si sus valores forman un set de número separados (1, 2, 3, 0, 10).

Si tiene un número finito de valores es discreta (hijos, protestas, terremotos, etc.).

Un decimal no tiene sentido (no pueden haber 3,5 terremotos en un año)

La variable es **continua** si sus valores posibles forman un intervalo (1.2, 3.5, 100.1).

Los valores posibles que puede tomar forman un continuo infinito (peso, ingresos)

Tanto las discretas como las continuas pueden ser “de razón” o “de cociente”. Este nivel incorpora la existencia de un valor “0” de carácter absoluto.

No se puede mover libremente la escala. El cero representa la ausencia total de la propiedad medida.

La arbitrariedad del nivel

Que una propiedad corresponda a uno u otro nivel no sólo depende de las características intrínsecas de dicha propiedad, sino sobre todo de las definiciones y operaciones teóricas y técnicas que hemos realizado para medirlas.

El cómo se operacionaliza y se llega a medir la propiedad es clave.

El caso de la edad es claro:

- ¿cuántos años tiene?
- ¿En cuál grupo de edad te ubicas? (0 a 20, 21 a 50, 51 o más)
- ¿Eres jóven o no (rango 15-29)

Distribución de una variable

Un paso fundamental para entender un conjunto de datos, es conocer la distribución de sus variables.

La distribución describe como las observaciones "caen" a lo largo de un rango posible de valores.

En las variables categóricas los valores posibles son las diferentes categorías. Cada observación cae en una categoría.

Se puede reportar el número de observaciones:

```
table(guaguas::guaguas$sexo)
```

```
##
##      F      M
## 523623 321777
```

Esto es una **tabla de frecuencias** (absolutas). Es un listado de los posibles valores de una variable junto al número de observaciones de cada una.

Distribución de una variable

También podemos observas una tabla de **frecuencias relativas**.

La proporción de observaciones en cada categoría corresponde al número de observaciones en dicha categoría dividido por el total del número de observaciones.

Porcentaje de mujeres en guaguas:

```
523623/(523623+321777)
```

```
## [1] 0.619379
```

Con función:

```
prop.table(table(guaguas::guaguas$sexo))
```

```
##
##        F        M
## 0.619379 0.380621
```

El porcentaje es cuando la proporción se multiplica por 100.

Distribución de una variable

La categoría que concentra la mayor frecuencia se llama la **categoría modal**.

En las cuantitativas también es importante visualizar la distribución, pero cuando el número de valores posibles es muy alto y cada valor lo asume un reducido número de observaciones **no es pertinente**.

Por ejemplo:

```
table(guaguas::guaguas$n)[1:100]
```

```
##
##      1      2      3      4      5      6      7      8      9     10     1
## 524247  91398  42081  25451  17365  12677   9888   7930   6694   5756   487
##     12     13     14     15     16     17     18     19     20     21     2
##   4299   3811   3469   2965   2827   2510   2284   2156   1929   1888   178
##     23     24     25     26     27     28     29     30     31     32     3
##   1606   1450   1445   1338   1181   1147   1079   1022   1020    957    87
##     34     35     36     37     38     39     40     41     42     43     4
##    888    841    825    784    762    729    656    652    703    605    61
##     45     46     47     48     49     50     51     52     53     54     5
##    579    585    555    533    528    521    510    466    428    492    45
```

Distribución de una variable

Para una variable cuantitativa resulta más pertinente observar la **forma** de la distribución, el **centro** y su **variabilidad**.

```
summary(guaguas::guaguas$n)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##     1.00     1.00     1.00    25.96     3.00 21448.00
```

En sesión de estadística descriptiva revisaremos en detalle estos conceptos y sus respectivos estadísticos.

Otras distinciones relevantes

Datos inválidos y perdidos

Datos pueden estar perfectamente ordenados, pero esconder errores y un alto número de valores perdidos.

Bibliografía utilizada

Agresti, A. and C. Franklin (2018). *Statistics the Art and Science of Learning from Data*. Pearson Education Limited.

Asún, R. (2006b). "Medir La Realidad Social: El Sentido de La Metodolog\'ia Cuantitativa". In: *Metodolog\'ias de Investigación Social: Introducción a los oficios*, pp. 31-60.

Babbie, E. (2014). *The Practice of Social Science Research*. 14th edition.

Wickham, H. (2021). *R Para Ciencia de Datos*.

Generic Statistical Business Process Model (GSBPM)

SDC Practice guide

Tidyr

Revisión tarea, un poco más de R base e introducción a dplyr

10:30