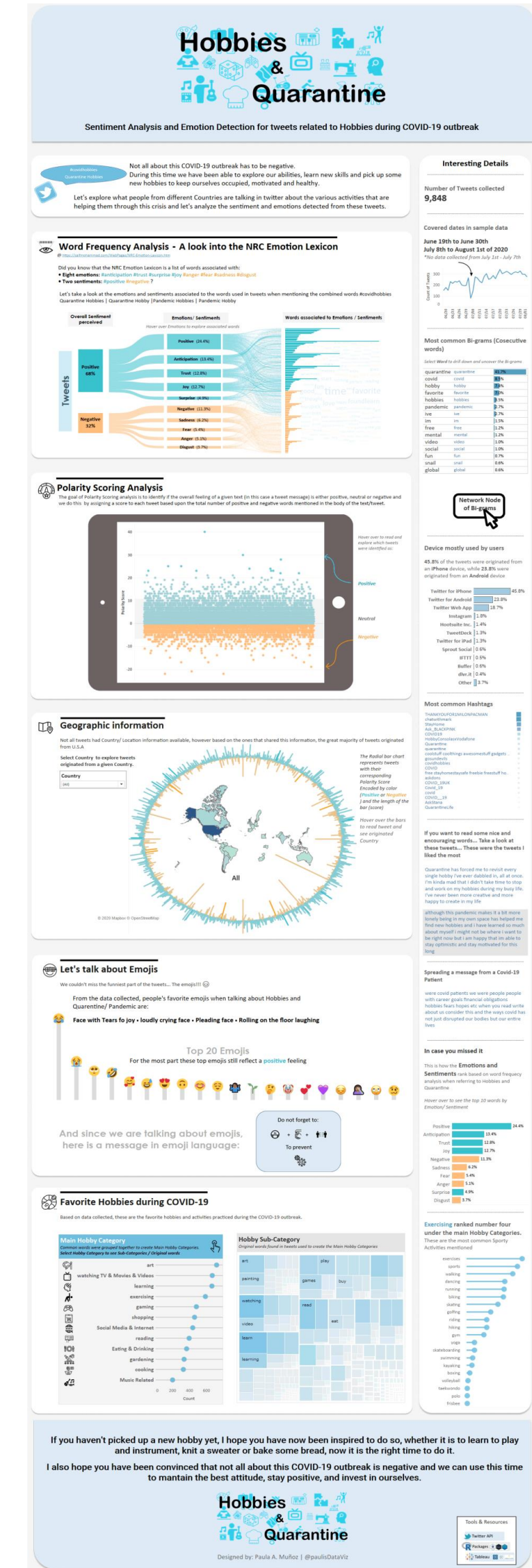
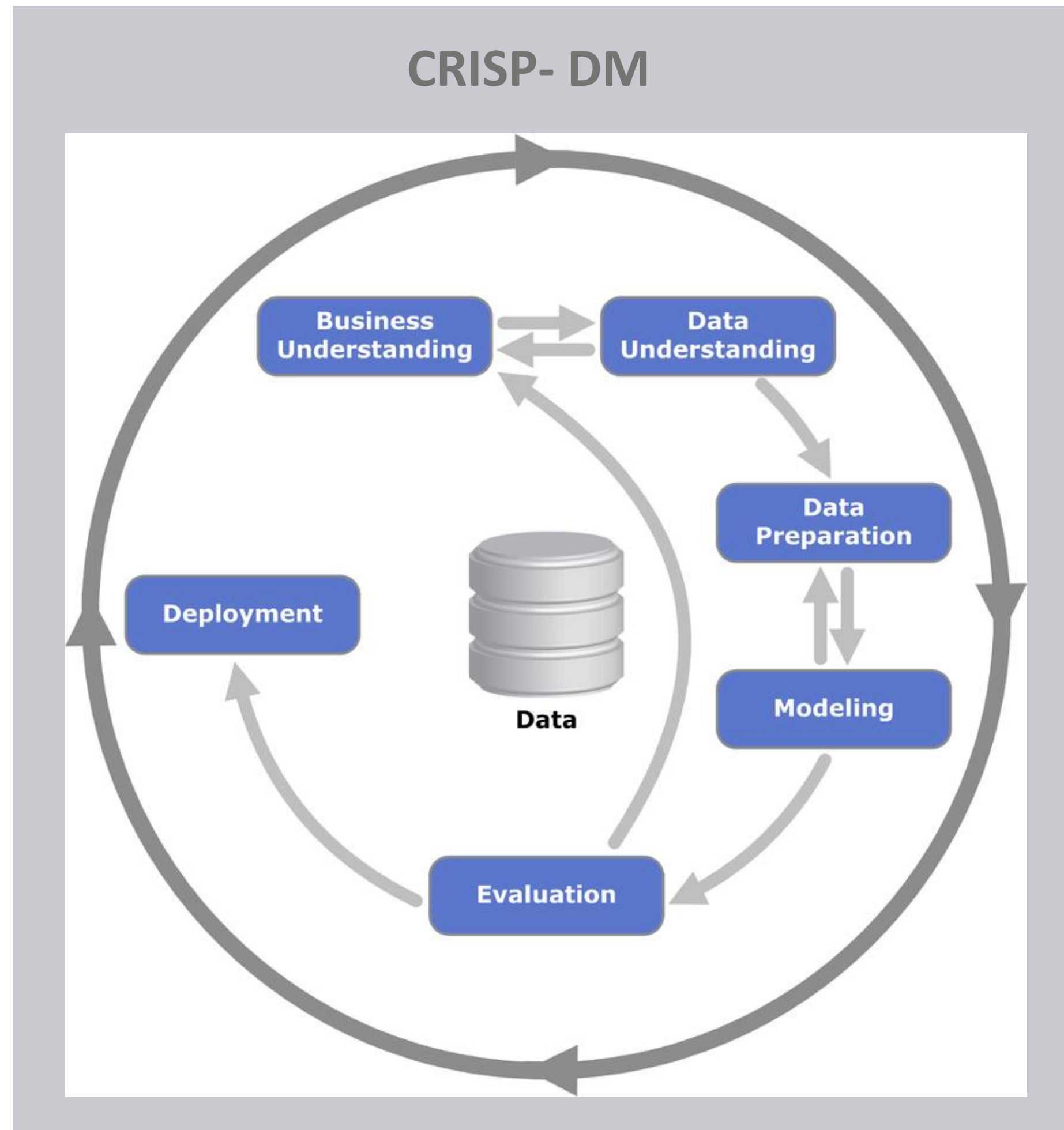


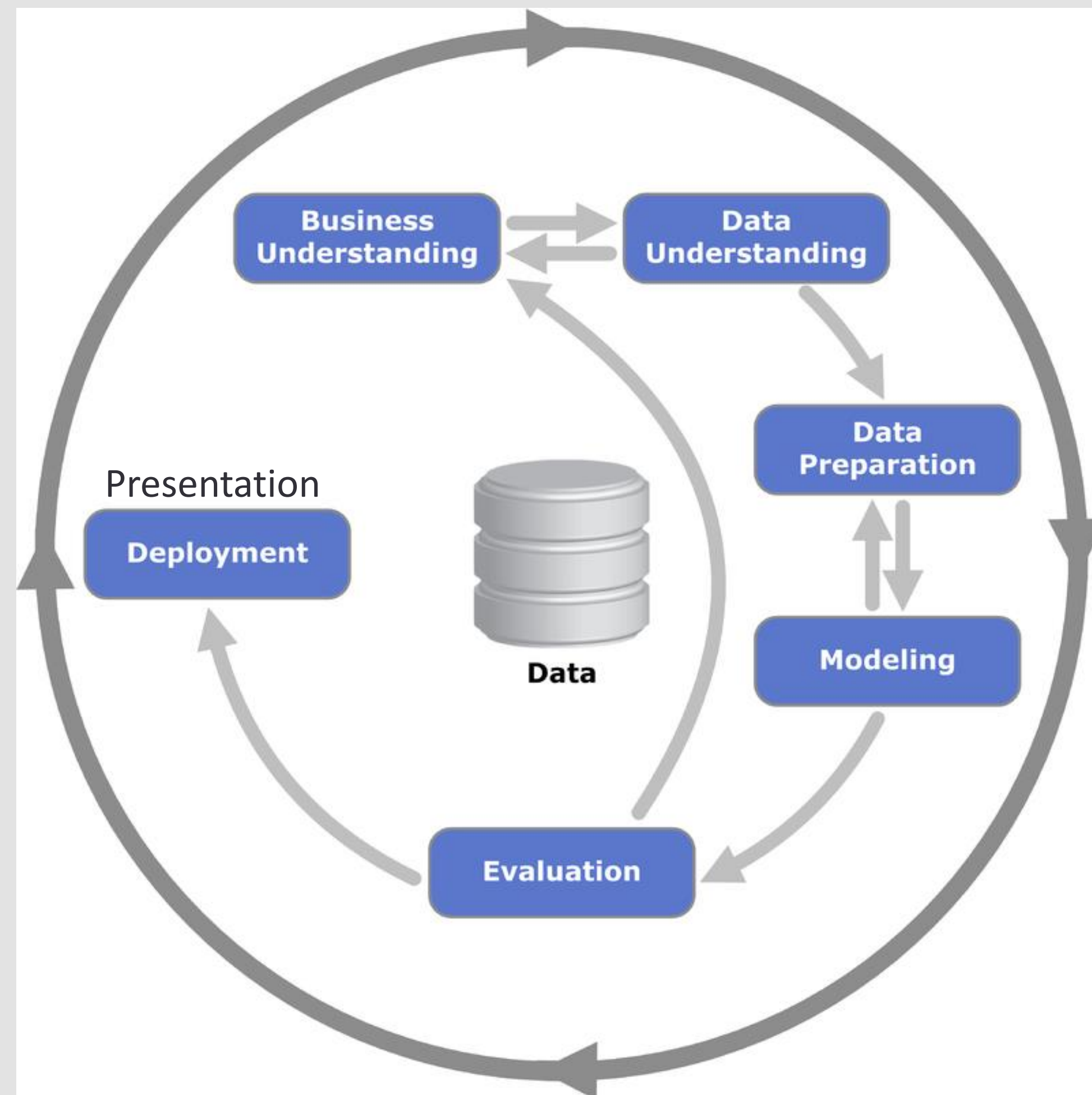
Entendiendo el ciclo de un Proyecto de Análisis de Datos

Usando diferentes herramientas como R, Tableau Prep Builder y Tableau



CICLO DE UN PROYECTO DE ANALISIS DE DATOS

METODOLOGIA CRISP - DM



Cross
Industry
Standad
Process
for
Data
Mining

"A data mining process model that describes commonly used approaches that data mining experts use to tackle problems" - Wikipedia

Diagram by Kenneth Jensen - Own work based on:
ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPD_M.pdf (Figure 1), CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=24930610>
Reference: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

CICLO DE UN PROYECTO DE ANALISIS DE DATOS

METODOLOGIA CRISP - DM

1. COMPRENSION DEL NEGOCIO/ CASO

- Plan del Proyecto
- Objetivos
- Información necesaria
- Tipo de Análisis
- Alcance
- Producto final

2. COMPRENSION DE LOS DATOS

- Recolectar datos iniciales
- Exploración y características de los datos
- Calidad de los datos

3. PREPARACION DE LOS DATOS

- Seleccionar
- Limpiar
- Estructurar
- Integrar
- Muestreo

6. IMPLEMENTACION/ PRESENTACION

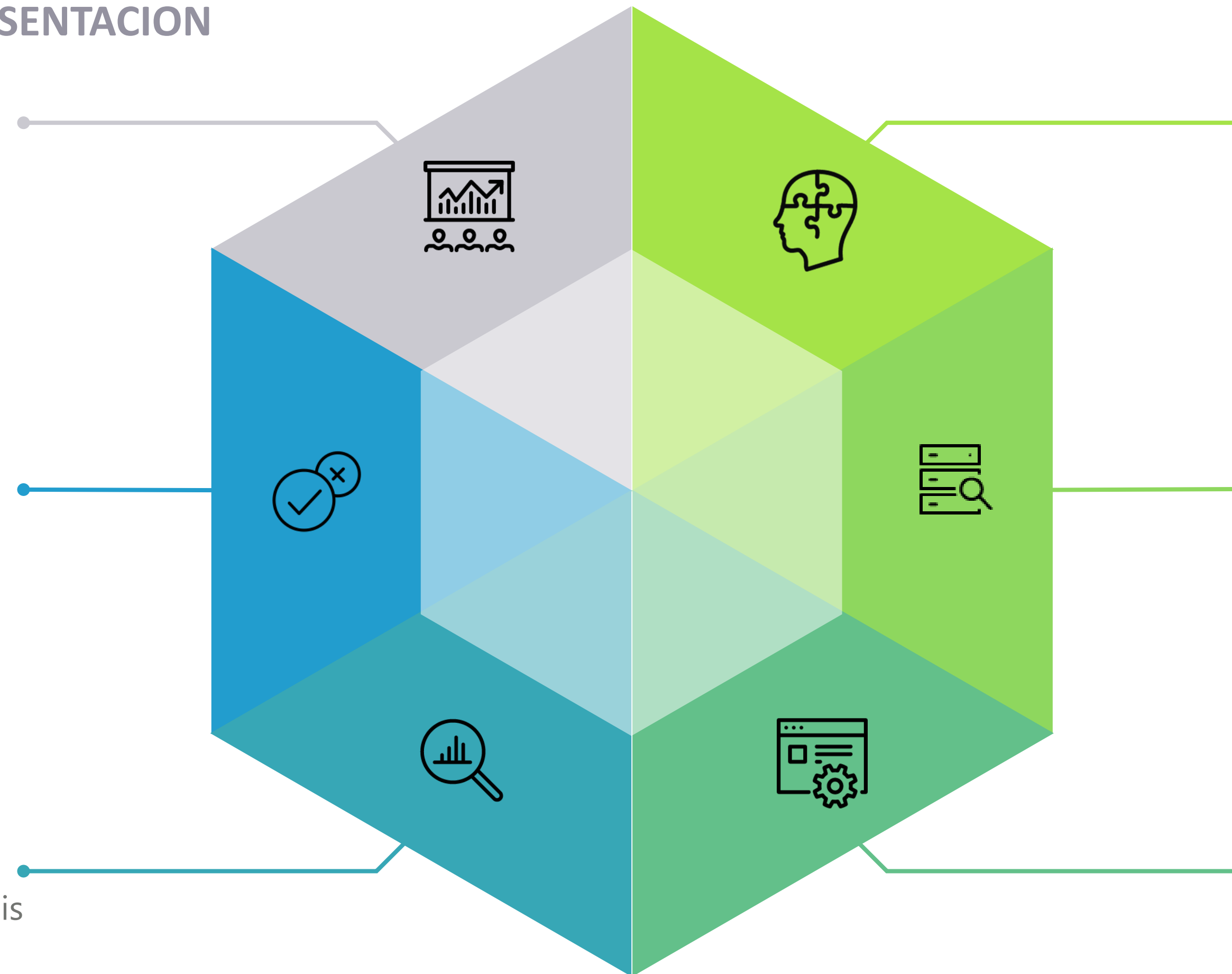
- Comunicar resultados
- Hacer recomendaciones
- Entrega producto final

5. EVALUACION

- Evaluar resultados
- Revisar proceso
- Determinar próximos pasos
 - OK ?
 - NOT OK ?

4. ANALYSIS Y/O MODELADO

- Seleccionar tipo análisis
- Desarrollar metodología/ análisis
- Construir modelo
- Buscar respuestas, contenido interesante
- Testear modelo



PASO 1. COMPRESION DEL NEGOCIO / CASO

- Explorar y entender que tipo de actividades y Hobbies las personas han desarrollado durante la pandemia.
- Analizar las emociones y polaridad de sentimientos de tweets recolectados

- Publicaciones de Twitter

- Descriptivo
- Clasificación basado en reglas específicas para calcular la polaridad de sentimientos

- Recolectar y analizar Tweets publicados aproximadamente durante un mes

- Dashboard/ Tablero Interactivo



<https://public.tableau.com/profile/paula.munoz#!/vizhome/IronViz2020-HobbiesandQuarantine-SentimentAnalysis/HobbiesQuarantine-SentimentA>

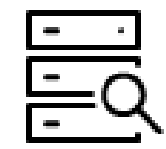
PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



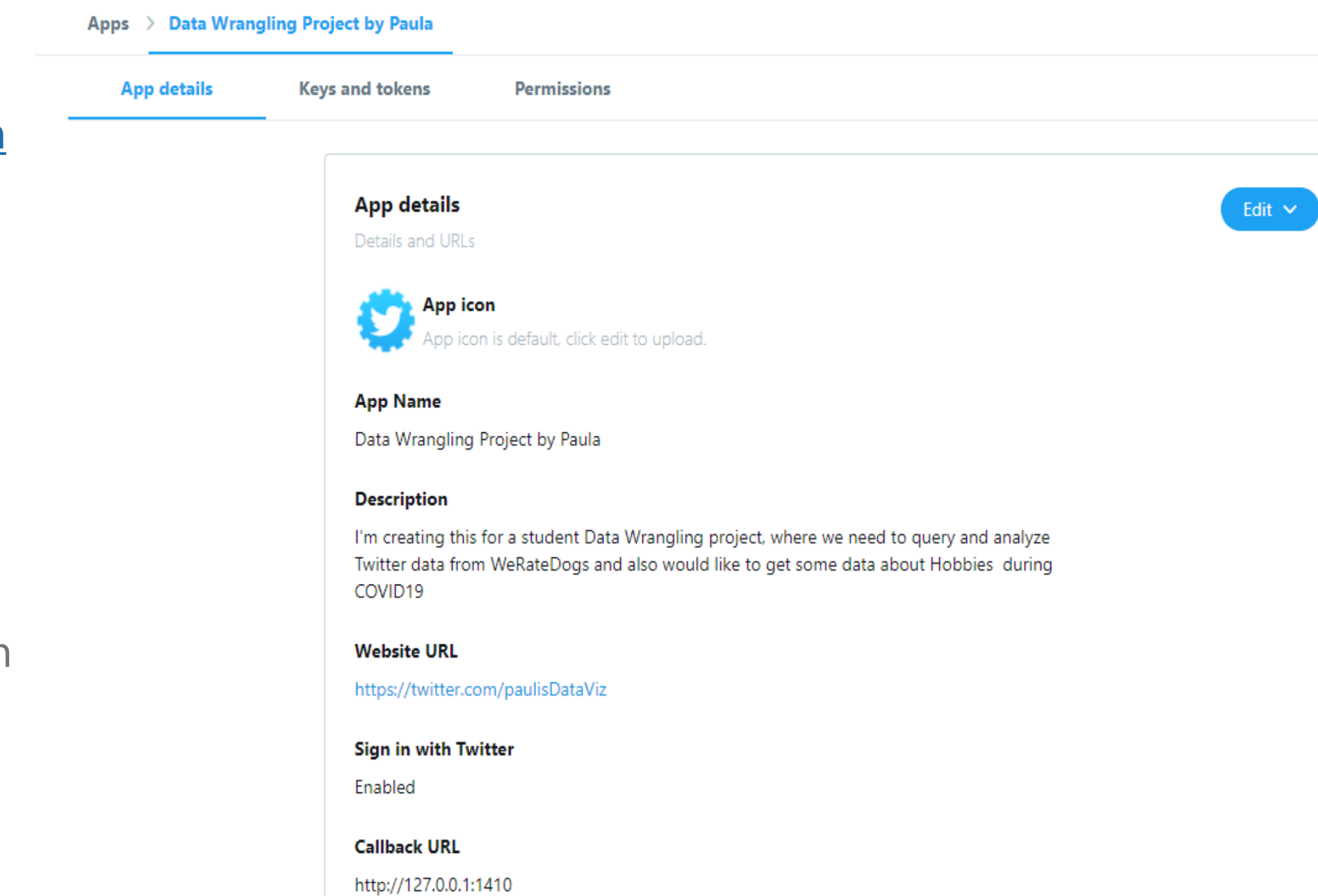
FUENTES DE DATOS DISPONIBLES

PASO 2. COMPRESION DE LOS DATOS



- **Fuentes de datos disponibles**
- Recolectar datos iniciales
- Exploración y características de los datos
- Calidad de los datos

- Publicaciones de Twitter por medio del Twitter API
- Twitter developer API: <https://developer.twitter.com/en>
- Crear cuenta de desarrollador de Twitter
- Pedir autorización diligenciando la forma online para crear una app
- Una vez autorizad@ vas a tener tu propi@ información de acceso (Key and tokens) para poder interactuar con el API
- Pasos detallados de como llenar la forma para pedir autorización: <https://cran.r-project.org/web/packages/rtweet/vignettes/auth.html>

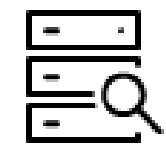


PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



PASO 2. COMPRESION DE LOS DATOS



- Fuentes de datos disponibles
- **Recolectar datos iniciales**
- Exploración y características de los datos
- Calidad de los datos

TECNOLOGIA PARA RECOLECTAR DE DATOS INICIALES

Para la recolección inicial los datos vamos a usar:



Lenguaje de programación R (R Studio IDE)

<https://cran.r-project.org/>

<https://rstudio.com/products/rstudio/download/>



Librería rtweet

Interactuar con la API de Twitter

<https://cran.r-project.org/web/packages/rtweet/readme/README.html>



Librería tidyverse

Minería de datos

<https://www.tidyverse.org/>

INSTALAR Y CARGAR LIBRERIAS EN R STUDIO:

```
## 2.1 INSTALAR Y CARGAR LIBRERIAS ----
```

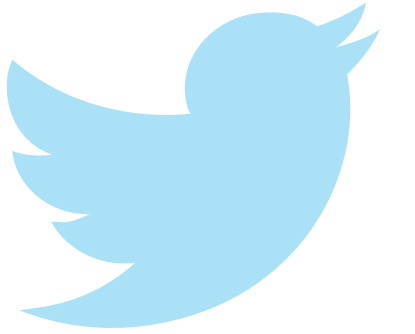
```
#install.packages("rtweet") # Quitar comentario "#" si necesita instalar libreria
```

```
library(rtweet)|  
library(tidyverse)
```

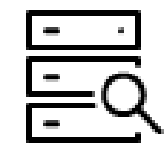
| Task | rtweet | twitterR | streamR | RTwitterAPI |
|-----------------------------|--------|----------|---------|-------------|
| Available on CRAN | ✓ | ✓ | ✓ | ✗ |
| Updated since 2016 | ✓ | ✗ | ✓ | ✗ |
| Non-'developer' access | ✓ | ✗ | ✗ | ✗ |
| Extended tweets (280 chars) | ✓ | ✗ | ✓ | ✗ |
| Parses JSON data | ✓ | ✓ | ✓ | ✗ |
| Converts to data frames | ✓ | ✓ | ✓ | ✗ |
| Automated pagination | ✓ | ✗ | ✗ | ✗ |
| Search tweets | ✓ | ✓ | ✗ | ? |
| Search users | ✓ | ✗ | ✗ | ? |
| Stream sample | ✓ | ✗ | ✓ | ✗ |
| Stream keywords | ✓ | ✗ | ✓ | ✗ |
| Stream users | ✓ | ✗ | ✓ | ✗ |
| Get friends | ✓ | ✓ | ✗ | ✓ |
| Get timelines | ✓ | ✓ | ✗ | ? |
| Get mentions | ✓ | ✓ | ✗ | ? |
| Get favorites | ✓ | ✓ | ✗ | ? |
| Get trends | ✓ | ✓ | ✗ | ? |
| Get list members | ✓ | ✗ | ✗ | ? |
| Get list memberships | ✓ | ✗ | ✗ | ? |
| Get list statuses | ✓ | ✗ | ✗ | ? |
| Get list subscribers | ✓ | ✗ | ✗ | ? |
| Get list subscriptions | ✓ | ✗ | ✗ | ? |
| Get list users | ✓ | ✗ | ✗ | ? |
| Lookup collections | ✓ | ✗ | ✗ | ? |
| Lookup friendships | ✓ | ✓ | ✗ | ? |
| Lookup statuses | ✓ | ✓ | ✗ | ? |
| Lookup users | ✓ | ✓ | ✗ | ? |
| Get retweeters | ✓ | ✓ | ✗ | ? |
| Get retweets | ✓ | ✓ | ✗ | ? |
| Post tweets | ✓ | ✓ | ✗ | ✗ |
| Post favorite | ✓ | ✗ | ✗ | ✗ |
| Post follow | ✓ | ✗ | ✗ | ✗ |

PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



PASO 2. COMPRESION DE LOS DATOS



- Fuentes de datos disponibles
- **Recolectar datos iniciales**
- Exploración y características de los datos
- Calidad de los datos



INTERACTUAR CON LA API DE TWITER

- Asegurarse de estar conectado en su cuenta de desarrollador de Twitter
- Referirse a las credenciales proporcionadas por Twitter (Consumer API Keys)

CREAR TWITTER TOKEN POR PRIMERA VEZ Y COMO CARGARLO POSTERIORMENTE

- El procedimiento para crear el token solo se hace una vez por computador, ya que este queda guardado internamente y en un futuro solo hay que llamarlo

1. GUARDAR API_KEY Y SECRET_KEY EN DOS VARIABLES:

```
# Guardar API keys (Borrar despues de Cargar)
api_key <- "reemplazar con key"
api_secret_key <- "reemplazar con secret key"
```

2. CREAR TOKEN :

```
# Crear Token
twitter_token <- create_token(
  app = "Data wrangling Project by Paula", ## << Nombre de su APP
  consumer_key = api_key,
  consumer_secret = api_secret_key)
```

3. CARGAR/ OBTENER TOKEN

Este es la función que se va a usar cuando quiera volver a interactuar con la API en el mismo Computador

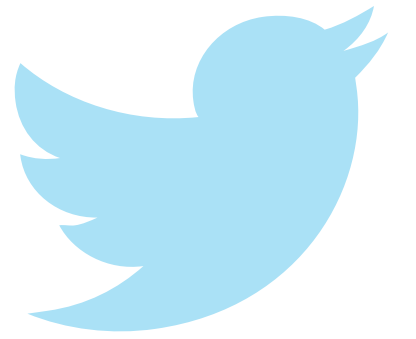
```
# * 2.2.2 Obtener Token ----
# Se abrirá una ventana en su web browser para comprobar la autenticación con Twitter

get_token()
```

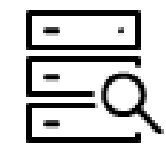


PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



PASO 2. COMPRESION DE LOS DATOS



- Fuentes de datos disponibles
- **Recolectar datos iniciales**
- Exploración y características de los datos
- Calidad de los datos



BUSCAR PUBLICACIONES

Funciones:

- **search_tweets** (Una sola búsqueda)
- **search_tweets2** (búsqueda de múltiples términos)

Hay que tener en cuenta que con la API estándar, solo va a retornar tweets de los últimos 6 a 9 días.

Los parámetros disponibles son:

- **q** - Query, termino, hashtag de interés
- **n** - Número de tweets a retornar. Hay una restricción de máximo 18,000 tweets por 15 min, para obtener más, hay que usar el

parámetro **retryonratelimit**

- **geocode** - Latitud, Longitud y radio en Millas o kilómetros
- **type** - Tipo de tweet a retornar: recent (Recientes), Mixed (Mixtos), Popular (Populares)
- **include_rts** - Incluir Retweets
- **lang** - Lenguaje de los Tweets
- **retryonratelimit** - Como se mencionó anteriormente este parámetro va a determinar si debe intentar obtener tweets cada 15 minutos para volúmenes por encima de 18,000

```
# * 2.3.1 Ejemplo busqueda individual ----  
# Vamos a hacer una busqueda de tweets  
quarantine_hobby_01 <- search_tweets(  
  q = "quarantine hobby",  
  n = 100000,  
  #lang = "en",  
  type = "mixed",  
  include_rts = FALSE,  
  retryonratelimit = TRUE  
)
```

```
# * 2.3.2 Busqueda de multiples combinaciones ----  
covid_hobbies_multiple_01 <- search_tweets2(  
  c("#CovidHobbies\\", "quarantine hobbies", "covid hobbies",  
    "covid hobby", "quarantine hobby", "#StayHome Hobby",  
    "#StayHome Hobbies", "#quedatenencasa hobbies", "cuarentena hobbies",  
    "cuarentena hobby", "#cuarentena hobby", "#cuarentena hobbies",  
    "cuarentena #hobbies", "cuarentena #hobby", "#yomequedoencasa hobby",  
    "cuarentena Hobby", "#pandemiahobbies", "pandemia hobbies",  
    "pandemia hobby", "#Pandemia hobby", "#Pandemia hobbies",  
    "pandemia #hobbies", "pandemia #hobby", "#pandemicHobbies",  
    "pandemic hobbies", "pandemic hobby"),  
  n = 18000,  
  #lang = "en",  
  type = "recent",  
  include_rts = FALSE,  
  retryonratelimit = TRUE  
)
```

Downloading [=====>-----] 17%retry on rate limit...
waiting about 12 minutes...

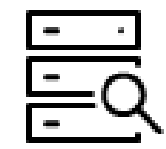
PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



EXPLORACION INICIAL DE LOS DATOS

PASO 2. COMPRESION DE LOS DATOS



- Fuentes de datos disponibles
- Recolectar datos iniciales
- **Exploración y características de los datos**
- **Calidad de los datos**



- El objetivo es familiarizarnos con los datos disponibles para entender su calidad, granularidad, escoger las variables de interés y tomar decisiones acerca de valores nulos, duplicados y demás.

Las funciones más comunes que personalmente yo uso para hacer el primer análisis exploratorio son:

- **dim()** Dimensiones del objeto/ dataframe (# de observaciones x # de variables)
- **head()** & **tail()** Preview de las cinco primeras o cinco últimas observaciones
- **View()** Ver los datos en estilo spreadsheet si se dificulta verlos en la consola
- **glimpse()** Preview los datos, pero la vista es transpuesta, es de especial ayuda cuando el dataset tiene muchas columnas

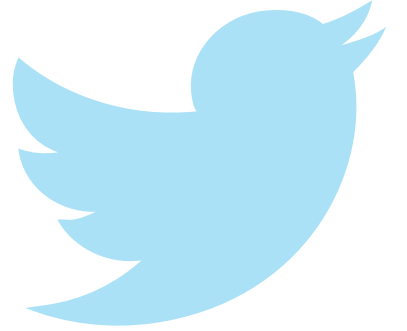
- La librería **dataexplorer** tiene muchas funciones interesantes que nos ayudan con la exploración y análisis descriptivo:

- **Plot_missing()** Ver la distribución de valores nulos (missing values) en todas las variables
- **Plot_histogram()** Ver histogramas de todas las variables continuas
- **Plot_bar()** Ver la frecuencia de variables discretas

Demo en R Studio: 2.4
Exploración inicial de los datos

PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



PASO 3. PREPARACION DE LOS DATOS

- **Seleccionar**
- Estructurar
- Limpiar
- Integrar
- Muestreo

SELECCIONAR VARIABLES DE INTERES

Hay 91 Variables disponibles



- Dimensiones de dataframe antes de seleccionar las variables de Interés

```
> # Dimensiones (Numero de observaciones x numero de columnas)
> covid_hobbies_multiple_01 %>%
+   dim()
[1] 1932    91
```

- **Seleccionando** Variables de interés (9 Variables) y **Filtrando** Observaciones de Interés, en este caso solo Tweets en **Ingles**

```
covid_hobbies_multiple_02 <- covid_hobbies_multiple_01 %>%
  select(user_id, created_at, text, source, hashtags, lang, country, location, query) %>%
  filter(lang == "en") %>%
  # Eliminar observaciones que se repiten
  unique()
```

```
> covid_hobbies_multiple_02 %>%
+   dim()
[1] 1690     9
```

de observaciones es menor después de eliminar varias columnas, Lo que significa que existían y todavía pueden existir observaciones que se repiten

- **En** este punto ya tenemos nuestro primer dataset con todas las variables de interés (sin limpiar)

```
> covid_hobbies_multiple_all_03 <- covid_hobbies_multiple_02
> covid_hobbies_multiple_all_03
# A tibble: 1,690 x 9
  user_id      created_at      text source      hashtags lang country location query
  <chr>      <dtm>      <chr> <chr>      <list> <chr> <chr> <chr>      <chr>
1 18993777 2020-10-17 17:31:37 "Any trip to @traderjoes us not complete without flowers! #covidhobbies #flowers #flowerstagram #pret~ Instagram <chr [4]> en NA "Philadelph~ "\"#CovidHo~
2 183449410 2020-10-17 01:48:06 "@lindarchilders I bought a house and moved early June so since then I have been working in the yard ~ Twitter for ~ <chr [1]> en NA "" "\"#CovidHo~
3 124862498380~ 2020-10-16 13:32:53 "During the pandemic, it can get boring #stayingatHome. Here are ideas of things you can do at home w~ Twitter web ~ <chr [8]> en NA "Northern o~ "\"#CovidHo~
4 124056116929~ 2020-10-16 08:53:34 "@ronysdreamz Now on to district population :) \n#covidhobbies" Twitter for ~ <chr [1]> en NA "India" "\"#CovidHo~
5 35072366 2020-10-14 06:45:59 "Needed a new hobby so a buddy and myself found a way to keep busy. #woodworking #covidhobbies #forsa~ Instagram <chr [4]> en NA "" "\"#CovidHo~
6 2989348043 2020-10-14 02:17:10 "One of the best so far! \n\u0001f342\u0001f3a8\n\n#watercolorpainting #fallvibes #covidhobbies https~ Instagram <chr [3]> en NA "Toronto" "\"#CovidHo~
7 161836448 2020-10-11 03:16:08 "OMG, who the heck am I? #covidhobbies #planting #homegardening https://t.co/SNrxbrvWH" Twitter for ~ <chr [3]> en NA "New York, ~ "\"#CovidHo~
8 265636267 2020-10-10 22:26:02 "I made another cemetery. It's still technically a cemetery when it's an ancient ritual site, yea~ Instagram <chr [5]> en NA "Austin, TX" "\"#CovidHo~
9 734888035310~ 2020-10-18 19:52:27 "wanna know when you are enjoying your quarantine? when you're so excited you get to be home doing th~ Twitter for ~ <chr [2]> en NA "San Antoni~ "quarantine~
10 3322038432 2020-10-18 19:36:49 "@PKMTrainerKayla Sounds interesting. My quarantine isn't too bad, just juggling my hobbies really" Twitter web ~ <chr [1]> en NA "" "quarantine~"
```



Demo en R Studio:
3. Preparacion de los datos

PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



PASO 3. PREPARACION DE LOS DATOS

- Seleccionar
- **Estructurar**
- Limpiar
- Integrar
- Muestreo

ESTRUCTURACION

- Crear un dataset enfocado al análisis exclusivo del texto de las publicaciones
- Duplicar variable "text" y llamarla "text_clean" con el fin de usar esta variable para limpiar el texto de las publicaciones
- En este punto tenemos un dataset con todas las variables de interés (sin limpiar)

```
covid_hobbies_multiple_text_04 <- covid_hobbies_multiple_all_03%>%  
  mutate(text_clean = text)
```

- Seleccionar las variables relevantes al texto de los tweets
- Este dataset solo va a contener cuatro variables: "user_id", "created_at", "text" y "text_clean"

```
covid_hobbies_multiple_text_04 <- covid_hobbies_multiple_text_04 %>%  
  select(user_id, created_at, text, text_clean) %>%  
  unique()
```

```
> covid_hobbies_multiple_text_04 %>%  
  + dim()  
[1] 1612 4
```

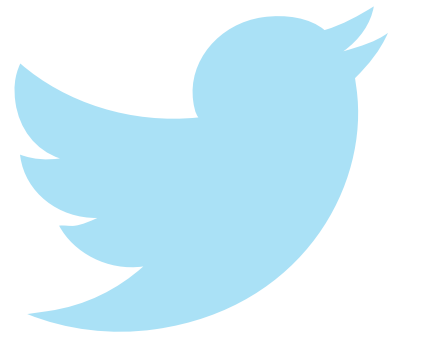
de observaciones es menor después de eliminar varias columnas, Lo que significa que existían y todavía pueden existir observaciones que se repiten

```
> covid_hobbies_multiple_text_04  
# A tibble: 1,612 x 4  
  user_id    created_at      text      text_clean  
  <chr>      <dtm>      <chr>      <chr>  
1 18993777 2020-10-17 17:31:37 "Any trip to @traderjoes us not complete without flowers! #covidhobbies #flowers ~ "Any trip to @traderjoes us not complete without flowers! #covidhobbies #flowers #f~  
2 183449410 2020-10-17 01:48:06 "@lindarchilders I bought a house and moved early June so since then I have been ~ "@lindarchilders I bought a house and moved early June so since then I have been wo~  
3 12486249838~ 2020-10-16 13:32:53 "During the pandemic, it can get boring #StayingAtHome. Here are ideas of things ~ "During the pandemic, it can get boring #StayingAtHome. Here are ideas of things yo~  
4 12405611692~ 2020-10-16 08:53:34 "@ronysdreamz Now on to district population :) \n#covidhobbies" "@ronysdreamz Now on to district population :) \n#covidhobbies"  
5 35072366 2020-10-14 06:45:59 "Needed a new hobby so a buddy and myself found a way to keep busy. #woodworking ~ "Needed a new hobby so a buddy and myself found a way to keep busy. #woodworking #c~  
6 2989348043 2020-10-14 02:17:10 "One of the best so far! \n\u0001f342\u0001f3a8\n\n#watercolorpainting #fallvibes~ "One of the best so far! \n\u0001f342\u0001f3a8\n\n#watercolorpainting #fallvibes ~  
7 161836448 2020-10-11 03:16:08 "OMG, who the heck am I? #covidhobbies #planting #homegardening https://t.co/SNrx~ "OMG, who the heck am I? #covidhobbies #planting #homegardening https://t.co/SNrx~  
8 265636267 2020-10-10 22:26:02 "I made another cemetery. It's still technically a cemetery when it's an anci~ "I made another cemetery. It's still technically a cemetery when it's an ancien~  
9 73488803531~ 2020-10-18 19:52:27 "Wanna know when you are enjoying your quarantine? when you're so excited you get t~ "Wanna know when you are enjoying your quarantine? when you're so excited you get t~  
10 3322038432 2020-10-18 19:36:49 "@PKMTrainerKayla Sounds interesting. My quarantine isn't too bad, just juggling ~ "@PKMTrainerKayla Sounds interesting. My quarantine isn't too bad, just juggling my~  
# ... with 1,602 more rows
```

Demo en R Studio:
3. Preparacion de los datos



TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



LIMPIAR LOS DATOS

PASO 3. PREPARACION DE LOS DATOS

- Seleccionar
- Estructurar
- **Limpiar**
- Integrar
- Muestreo

- Limpiar texto de la variable "text_clean" usando regular expressions

```
covid_hobbies_multiple_text_04$text_clean <- covid_hobbies_multiple_text_04$text_clean %>%  
  gsub("http://t.co/[a-z,A-Z,0-9]*{8}", "", .) %>% # Eliminar http links  
  gsub("https://t.co/[a-z,A-Z,0-9]*{8}", "", .) %>% # Eliminar https links  
  iconv(from = "latin1", to = "ASCII", sub="") %>% # Eliminar Emojis y caracteres especiales  
  gsub("(RT|via)((?:\\b\\W*@\\W+)+)", "", .) %>% # Eliminar re-tweet entities  
  gsub("@\\W+", "", .) %>% # Eliminar @persona  
  gsub("[[:punct:]]", " ", .) %>% # Eliminar puntuaciones  
  gsub("[[:digit:]]", " ", .) %>% # Eliminar digitos  
  gsub("\\n", " ", .) %>% # Eliminar \n  
  gsub("[ \\t]{2,}", " ", .) %>% # Eliminar espacios innecesarios  
  gsub("^\\s+|\\s+$", "", .) %>% # Eliminar espacios innecesarios  
  tolower()
```

Comparativo de text vs text_clean, las observaciones en variable text_clean ya no contienen links, caracteres especiales, etc



```
# A tibble: 1,612 x 2  
  text  
  <chr>  
1 "Any trip to @traderjoes us not complete without flowers! #covidhobbies #flowers #flowerstagram #p~ any trip to us not complete without flowers covidhobbies flowers flowerstagram pretty  
2 "@lindarchilders I bought a house and moved early June so since then I have been working in the ya~ i bought a house and moved early june so since then i have been working in the yard every free mome~  
3 "During the pandemic, it can get boring #stayingathome. Here are ideas of things you can do at hom~ during the pandemic it can get boring stayingathome here are ideas of things you can do at home whi~  
4 "@ronysdreamz Now on to district population :) \n#covidhobbies" now on to district population covidhobbies  
5 "Needed a new hobby so a buddy and myself found a way to keep busy. #woodworking #covidhobbies #fo~ needed a new hobby so a buddy and myself found a way to keep busy woodworking covidhobbies forsale ~  
6 "One of the best so far! \n\u00001f342\u00001f3a8\n\n#watercolorpainting #fallvibes #covidhobbies ht~ one of the best so far watercolorpainting fallvibes covidhobbies  
7 "OMG, who the heck am I? #covidhobbies #planting #homegardening https://t.co/SNrxBxrvWH" omg who the heck am i covidhobbies planting homegardening  
8 "I made another cemetery. It's still technically a cemetery when it's an ancient ritual site, ~ i made another cemetery its still technically a cemetery when its an ancient ritual site yeah c~  
9 "Wanna know when you are enjoying your quarantine? when you're so excited you get to be home doing~ wanna know when you are enjoying your quarantine when you re so excited you get to be home doing th~  
10 "@PKMTrainerKayla Sounds interesting. My quarantine isn't too bad, just juggling my hobbies really" sounds interesting my quarantine isn t too bad just juggling my hobbies really  
# ... with 1,602 more rows
```

Demo en R Studio:
3. Preparacion de los datos

PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



PASO 3. PREPARACION DE LOS DATOS

- Seleccionar
- Estructurar
- **Limpiar**
- Integrar
- Muestreo



LIMPIAR LOS DATOS

- Exportar primeras versiones de nuestros datasets para continuar limpieza en Tableau Prep

```
# * Dataset 1: Enfocado al analisis de texto
# **** IMPORTANTE: AJUSTAR EL NOMBRE DEL ARCHIVO EN CADA ITERATION
covid_hobbies_multiple_text_04 %>%
  save_as_csv("step02_and_step03_data_understanding_and_prep/Data_collection_clean_01/covid_hobbies_clean_text_01_20201018pm.csv", prepend_ids = TRUE, na = "",
             fileEncoding = "UTF-8")

# * Dataset 2: Dataset con todas las variables (Antes de limpiar)
covid_hobbies_multiple_all_03 %>%
  save_as_csv("step02_and_step03_data_understanding_and_prep/Data_collection_not_clean_01/covid_hobbies_not_clean_all_01_20201018pm.csv", prepend_ids = TRUE, na = "",
             fileEncoding = "UTF-8")
```

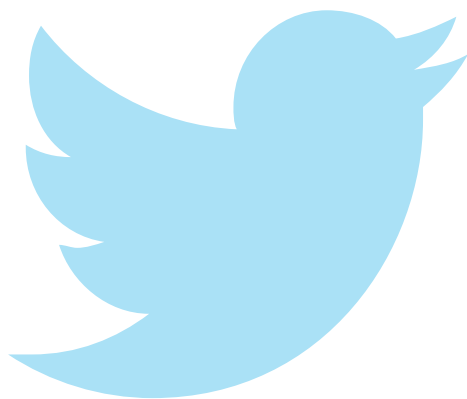
Importante mantener la organización de los archivos durante el desarrollo del proyecto

| | | | | |
|---|--|--------------------|-------------|-------|
| Data_collection_clean_01 | | 10/18/2020 7:28 PM | File folder | |
| Data_collection_not_clean_01 | | 10/18/2020 7:29 PM | File folder | |
| step02_and_step03_twitter_covid_hobbies.R | | 10/18/2020 7:30 PM | R File | 10 KB |

Demo en R Studio:
3. Preparacion de los datos

PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



EN QUE NOS
VAMOS A
ENFOCARNOS
DURANTE EL DEMO
EL DIA DE HOY?

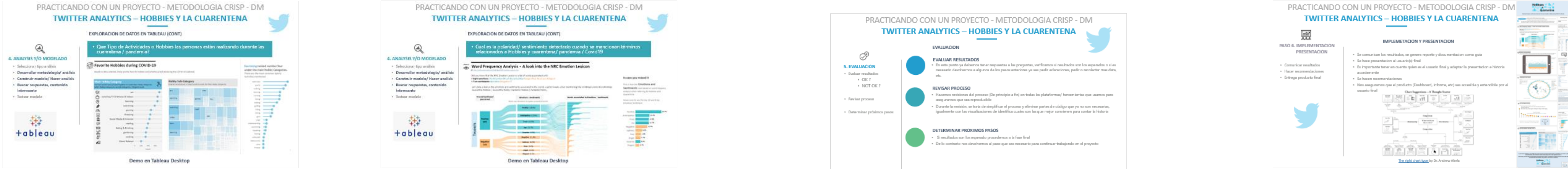
MOTIVACION PROYECTO | COMO INTERACTUAR CON LA API DE TWITTER | R- STUDIO



TABLEAU PREP BUILDER| COMO USAR R-SCRIPT EN TABLEAU PREP



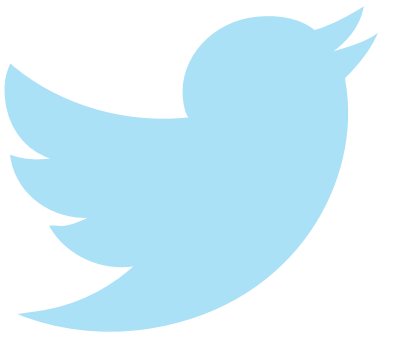
VISUALIZACIONES TABLEAU DESKTOP | PRODUCTO FINAL



POR FAVOR REVISAR PRESENTACION PARA VER PASO A PASO COMO SE APLICO LA METODOLOGIA CRISP-DM EN ESTE PROYECTO

PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



LIMPIAR LOS DATOS – FASE 2 EN TABLEAU PREP



PASO 3. PREPARACION DE LOS DATOS

- Seleccionar
- Estructurar
- **Limpiar**
- Integrar
- Muestreo



Funcionalidades/ pasos mas comunes en Tableau Prep que nos ayudan a preparar y limpiar nuestros datasets:

Pasos

- Paso **Agregate**, para eliminar líneas idénticas
- Paso **Pivot** para cambiar estructura del dataset (Columns to rows) (Rows to columns)
- Paso **Join** para integrar varios datasets
- Paso **Union** para unir varios datasets

A nivel de variable:

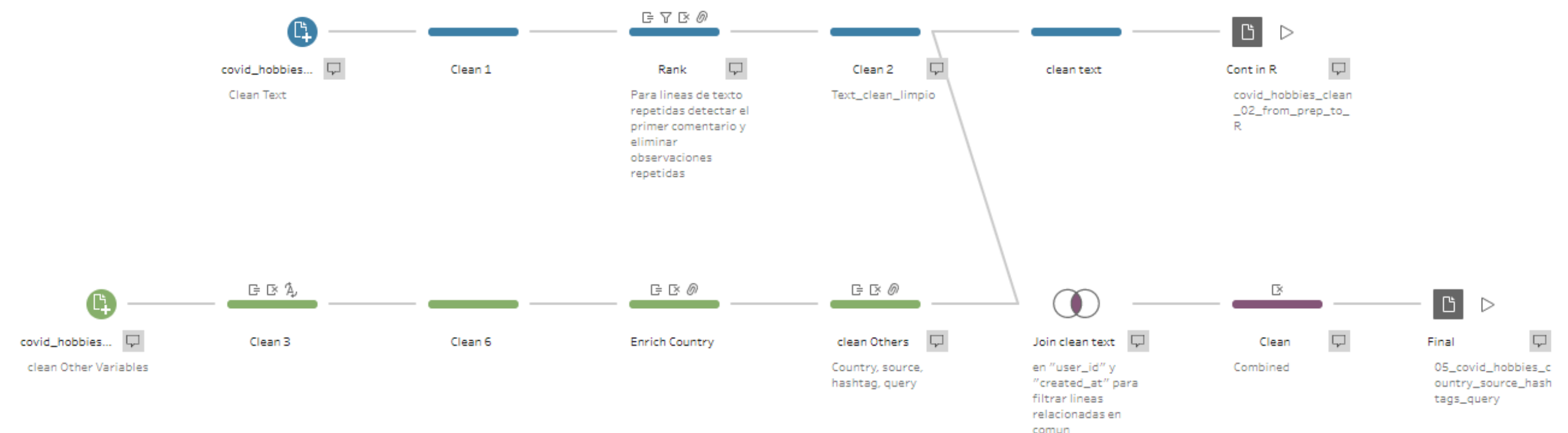
- **Filter** - Filtrar
- **Clean** - Eliminar Puntuaciones, números y mas
- **Group Values** - Agrupar
- **Split values** - Separar valores

Crear variables a base de calculo:

- Custom calculation
- Fixed LOD
- Rank

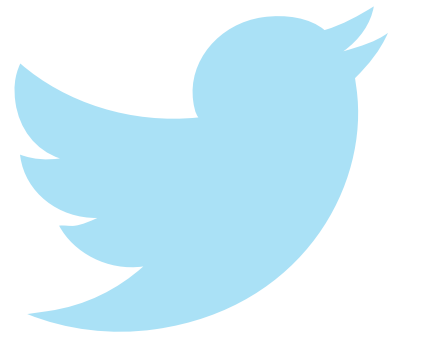
Recomendations

- Geograficas
- Agrupaciones



PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



PASO 3. PREPARACION DE LOS DATOS

- Seleccionar
- Estructurar
- **Limpiar**
- Integrar
- Muestreo



LIMPIAR LOS DATOS – FASE 2 EN TABLEAU PREP

Configurar Tableau Prep para aceptar varios datasets de la misma carpeta por medio de “wildcard unión”

| Type | Field Name | Original Field Name | Changes | Sample Values |
|------|------------|---------------------|---------|---|
| Abc | user_id | user_id | | x18993777, x183449410, x1248624983804239872 |
| Abc | created_at | created_at | | 10/17/2020, 05:31:37 PM, 10/17/2020, 01:48:06 AM, 10/16/2020, 05:31:37 PM |
| Abc | text | text | | Any trip to @traderjoes us not complete without flowers! #covidhobbies |
| Abc | text_clean | text_clean | | any trip to us not complete without flowers covidhobbies flo |
| Abc | File Paths | File Paths | | covid_hobbies_clean_text_01_20201018pm.csv |

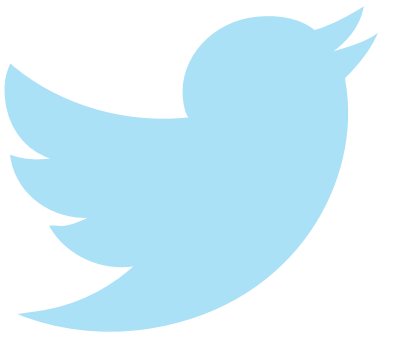
Va a leer todos los datasets en la carpeta especificada que empiezan con: covid_hobbies_clean_text*

| | | | |
|---|--------------------|-------------|-------|
| Data_collection_clean_01 | 10/18/2020 7:28 PM | File folder | |
| Data_collection_not_clean_01 | 10/18/2020 7:29 PM | File folder | |
| step02_and_step03_twitter_covid_hobbies.R | 10/18/2020 7:30 PM | R File | 10 KB |

Demo en Tableau Prep:
3. Preparacion de los datos

PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



PASO 3. PREPARACION DE LOS DATOS

- Seleccionar
- Estructurar
- **Limpiar**
- Integrar
- Muestreo



LIMPIAR LOS DATOS – FASE 2 EN TABLEAU PREP

Demostrar como se pueden interactuar con R scripts en Tableau Prep

1. En R Studio:

- Instalar librería **Rserve**
- Cargar librería **Rserve**
- Inicializar conexión con **Rserve**

```
install.packages("Rserve")  
library(Rserve)  
Rserve()
```

2. Crear script en R para realizar una tarea especifica y guardar archivo

Con propósito de demostración voy a repetir uno de los pasos de limpieza realizado previamente en R en la variable "text_clean" el cual elimina @persona, esta vez voy a duplicar la variable original "text" y aplicar el paso de limpieza

```
library(tidyverse)  
remove_at <- function(df){  
  df <- df %>%  
    mutate(text_clean2 = gsub("@\\w+", "", df$text));  
  return(df)  
}  
  
getOutputSchema <- function() {  
  return(data.frame(  
    user_id = prep_string(),  
    text = prep_string(),  
    text_clean2 = prep_string()  
  ));  
}
```

Nombre de la función:
remove_at

Crea variable
text_clean2 a la cual le
aplica el paso de
limpieza para eliminar
@persona

Variables de salida

Guardar Script

| 01_Presentacion_Oct_23 > Scripts | | | | |
|----------------------------------|---------------------|--------|--------------------|--------|
| <input type="checkbox"/> | Name | Status | Date modified | Type |
| | remove_at.R | | 10/18/2020 9:44 PM | R File |
| | rserve_connection.R | | 10/18/2020 9:38 PM | R File |

PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



PASO 3. PREPARACION DE LOS DATOS

- Seleccionar
- Estructurar
- **Limpiar**
- Integrar
- Muestreo



LIMPIAR LOS DATOS – FASE 2 EN TABLEAU PREP

Demostrar como se pueden interactuar con R scripts en Tableau Prep (Cont)

1. En Tableau Prep:

- Ir a Help > Settings and Performance > Analytics Extension Connection
- Seleccionar Rserve
- server: localhost, port 6311
- Añadir al Tableau Prep workflow un paso de script
- Seleccionar connection type : Rserve
- Buscar/ Cargar el archivo con el script
- Escribir el nombre de la función
- Si la conexión es exitosa, el script debe hacer la tarea especificada



Antes de script

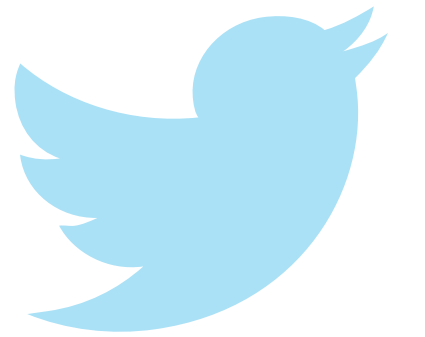
| user_id | created_at | text |
|----------------------|-------------------------|---|
| x18993777 | 10/17/2020, 05:31:37 PM | Any trip to @traderjoes us not complete without flowers! #covidhobbies #flowers #flowerstagram #prett |
| x183449410 | 10/17/2020, 01:48:06 AM | @lindarchilders I bought a house and moved early June so since then I have been working in the yard every |
| x1248624983804239872 | 10/16/2020, 01:32:53 PM | During the pandemic, it can get boring #StayingAtHome. Here are ideas of things you can do at home while |
| x1240561169296814080 | 10/16/2020, 08:53:34 AM | @ronysdreamz Now on to district population :) \n#covidhobbies |
| x25072265 | 10/14/2020, 06:45:58 AM | Needed a new hobby so a buddy and myself found a way to keep busy #woodworking #covidhobbies #f |



Después de Script
no @persona

| user_id | text | text_clean2 |
|----------------------|---|---|
| x18993777 | Any trip to @traderjoes us not complete without flowers! #cov | Any trip to us not complete without flowers! #covidhobbies #flowers #flo |
| x183449410 | @lindarchilders I bought a house and moved early June so sinc | I bought a house and moved early June so since then I have been working i |
| x1248624983804239872 | During the pandemic, it can get boring #StayingAtHome. Here | During the pandemic, it can get boring #StayingAtHome. Here are ideas of |
| x1240561169296814080 | @ronysdreamz Now on to district population :) \n#covidhobbie | Now on to district population :) \n#covidhobbies |

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



PASO 3. PREPARACION DE LOS DATOS

- Seleccionar
- Estructurar
- **Limpiar**
- **Integrar**
- Muestreo



LIMPIAR LOS DATOS – FASE 2 EN TABLEAU PREP

Usar el segundo dataset para aplicar pasos de limpieza y obtener:

- Información Geográfica (**País**)
- **Source** (Plataforma usada)
- **Hashtags**
- **Query** (Términos de Búsqueda, basados en el parámetro q cuando se hizo la búsqueda inicial en R con la función search_tweets2)
- Uno de los pasos de limpieza interesante consistió en :
 - Usar la variable **location** para **enriquecer** variable **country** ya que la variable location tiene mas información Geográfica que se puede usar para extraer el País (La variable country paso de tener ~ 97% valores nulos a 56%)

Demo en Tableau Prep

Abc Country/Region

country 6

| |
|----------------|
| null |
| Canada |
| India |
| Mexico |
| United Kingdom |
| United States |

Antes

Tooltip: null, 1,643 (97%) rows

Abc Country/Region

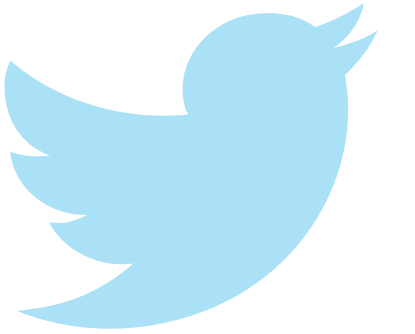
country 32

| |
|----------------|
| null |
| United States |
| United Kingdom |
| Canada |
| India |
| Philippines |
| Netherlands |
| Australia |
| Ireland |
| Germany |
| UAE |
| Switzerland |

Después

Tooltip: null, 847 (54%) rows

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



PASO 3. PREPARACION DE LOS DATOS

- Seleccionar
- Estructurar
- Limpiar
- Integrar
- Muestreo



Después de realizar los pasos de limpieza en Tableau Prep, tenemos dos datasets:

- Con el primer dataset relacionado a los mensajes de tweet vamos a guardarlo en formato /csv e importarlo en R Studio para continuar con el paso 4 y trabajar en el análisis de sentimiento y polaridad.
- El segundo dataset lo vamos a usar directamente en Tableau para empezar a extraer datos interesantes y hacer visualizaciones.

PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



EXPLORACION DE DATOS PARA ANALISIS DE SENTIMIENTO EN R (CONT)

CALCULAR PUNTAJE DE POLARIDAD



4. ANALYSIS Y/O MODELADO

- Seleccionar tipo análisis
- Desarrollar metodología/ análisis
- Construir modelo/ Hacer analisis
- Buscar respuestas, contenido interesante
- Testear modelo



- El objetivo del puntaje de polaridad es identificar que tan positivo o negativo es el texto en general
- **Puntaje de Polaridad** = Numero total de palabras positivas - Palabras negativas mencionadas en el tweet

```
# 4.5 PUNTAJE DE POLARIDAD (Polarity Score) ----  
  
# 4.5.1 Contar # de palabras positivas vs palabras negativas por row_id ----  
covid_hobbies_polarity <- tweets_nrc_overall_sentiment %>%  
  count(row_id, overall_sentiment) %>%  
  
# Hacer un Pivot horizontal  
pivot_wider(names_from = overall_sentiment, values_from = n, values_fill = 0) %>%  
  
# Calcular el polarity score por row_id = positive - negative  
mutate(polarity_score = positive - negative ) %>%  
  
# Unir con covid_hobbies_data para traer el dataframe con todo el data  
left_join(covid_hobbies_data)
```

Técnica que aprendí de
Matt Dancho
(Business Science)

En este punto tenemos los datasets necesarios para continuar
análisis en Tableau

Demo en R Studio: 4 Analisis de datos

PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



4. ANALYSIS Y/O MODELADO

- Seleccionar tipo análisis
- **Desarrollar metodología/ análisis**
- **Construir modelo/ Hacer analisis**
- **Buscar respuestas, contenido interesante**
- Testear modelo



EXPLORACION DE DATOS EN TABLEAU (CONT)

- Teniendo en cuenta los objetivos iniciales, construir visualizaciones que den respuesta a dichas preguntas.
- En mi caso yo tenia dos preguntas:
 - Que Tipo de Actividades o Hobbies las personas están realizando durante las cuarentena / pandemia?
 - Cual es la polaridad/ sentimiento detectado cuando se mencionan términos relacionados a Hobbies y cuarentena/ pandemia / Covid19 ?
- Después de dar respuesta a las preguntas iniciales, se sigue analizando las otras variables para encontrar mas información que pueda ser de interés.

Demo en Tableau Desktop

PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



EXPLORACION DE DATOS EN TABLEAU (CONT)



4. ANALYSIS Y/O MODELADO

- Seleccionar tipo análisis
- **Desarrollar metodología/ análisis**
- **Construir modelo/ Hacer analisis**
- **Buscar respuestas, contenido interesante**
- Testear modelo

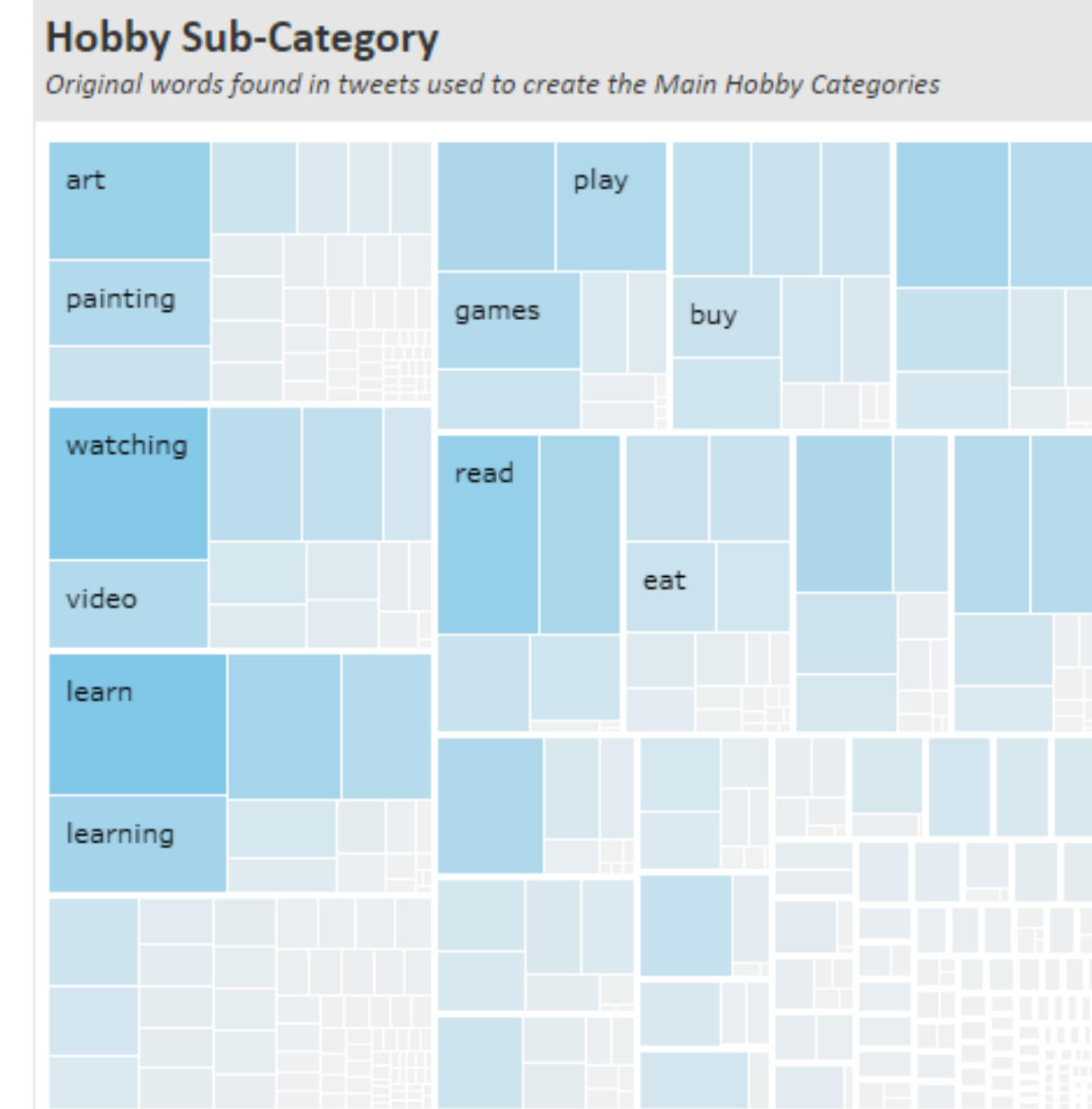
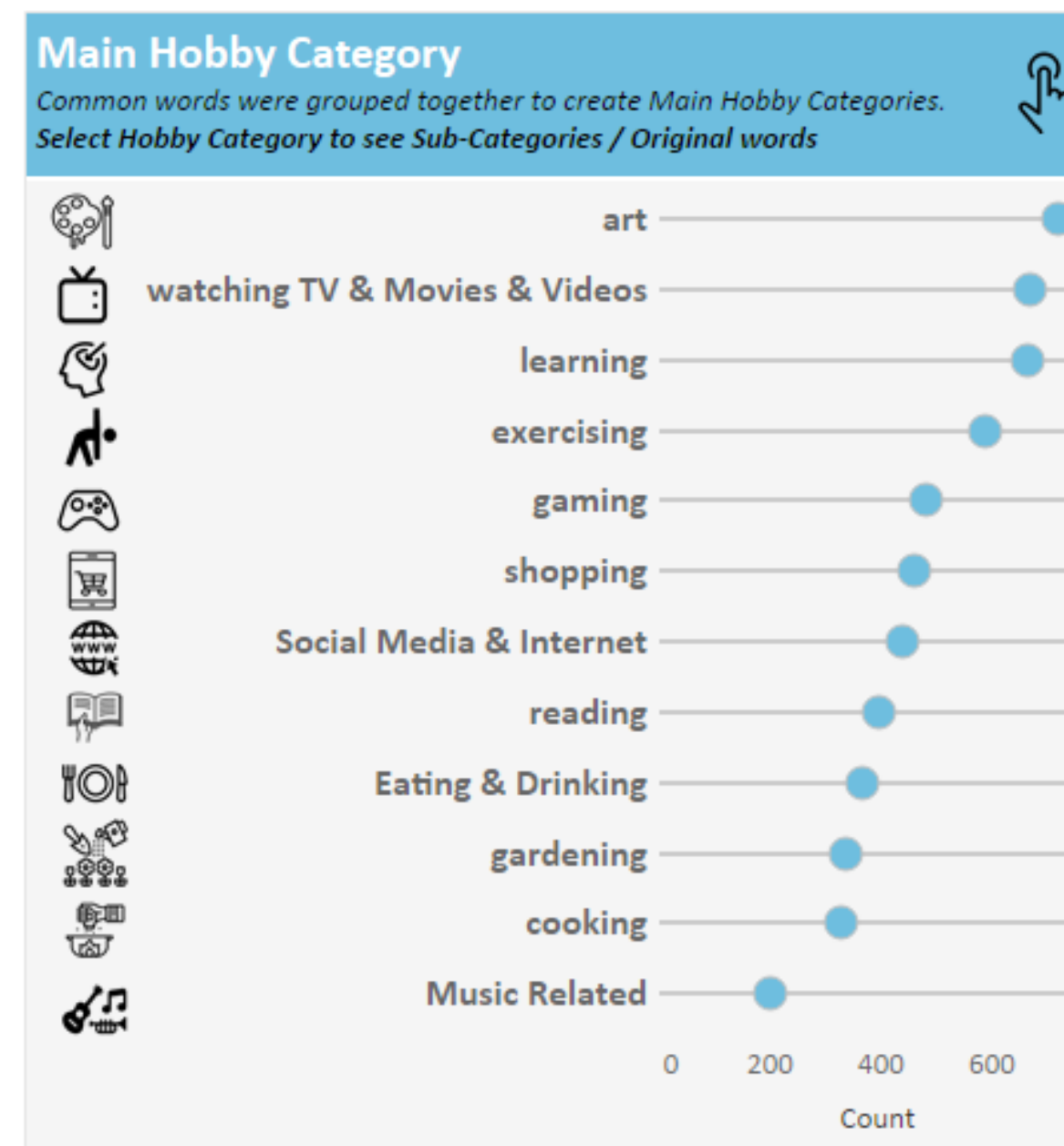


- Que Tipo de Actividades o Hobbies las personas están realizando durante las cuarentena / pandemia?



Favorite Hobbies during COVID-19

Based on data collected, these are the favorite hobbies and activities practiced during the COVID-19 outbreak.



Exercising ranked number four under the main Hobby Categories. These are the most common Sporty Activities mentioned



Demo en Tableau Desktop

PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



EXPLORACION DE DATOS EN TABLEAU (CONT)



4. ANALYSIS Y/O MODELADO

- Seleccionar tipo análisis
- **Desarrollar metodología/ análisis**
- **Construir modelo/ Hacer analisis**
- **Buscar respuestas, contenido interesante**
- Testear modelo



- Cual es la polaridad/ sentimiento detectado cuando se mencionan términos relacionados a Hobbies y cuarentena/ pandemia / Covid19



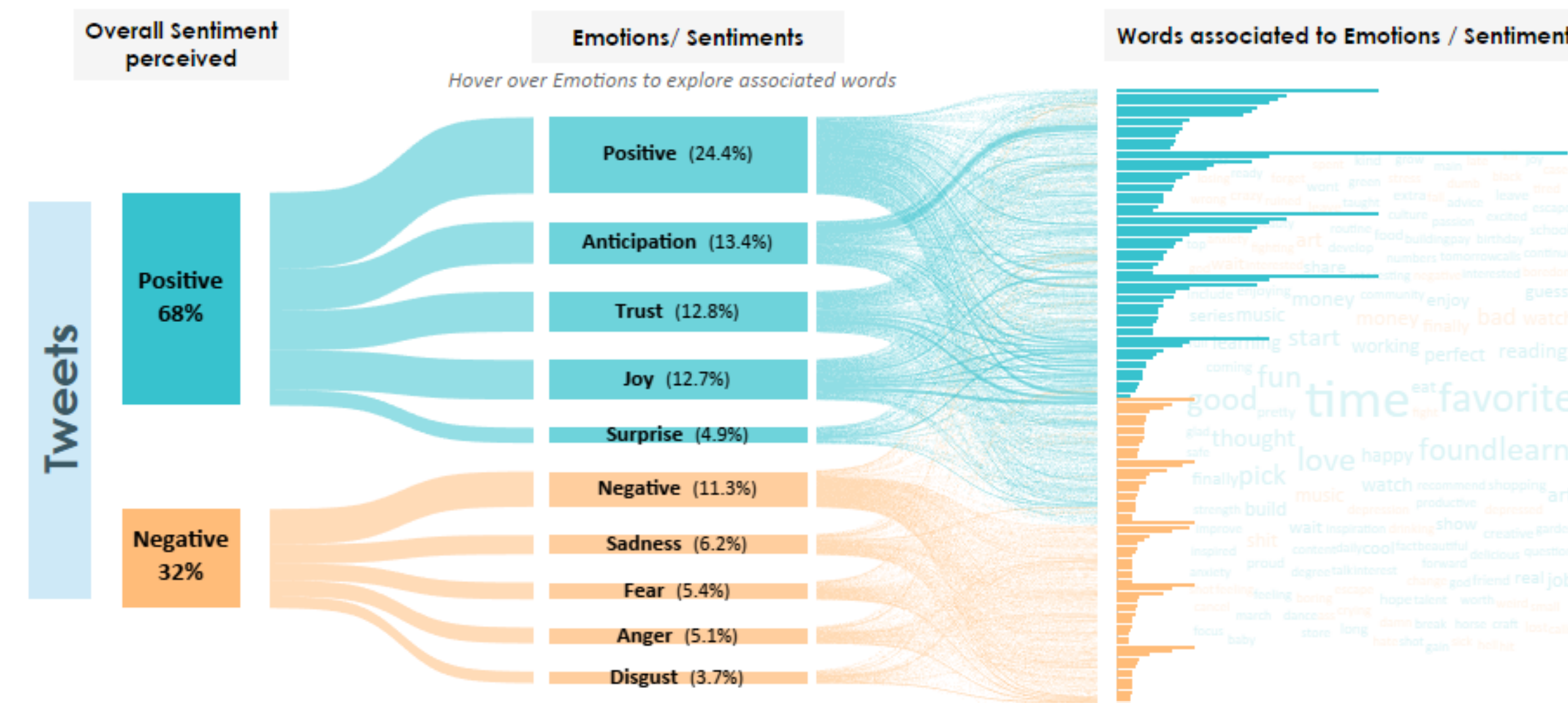
Word Frequency Analysis - A look into the NRC Emotion Lexicon

@ <https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>

Did you know that the NRC Emotion Lexicon is a list of words associated with:

- **Eight emotions:** #anticipation #trust #surprise #joy #anger #fear #sadness #disgust
- **Two sentiments:** #positive #negative ?

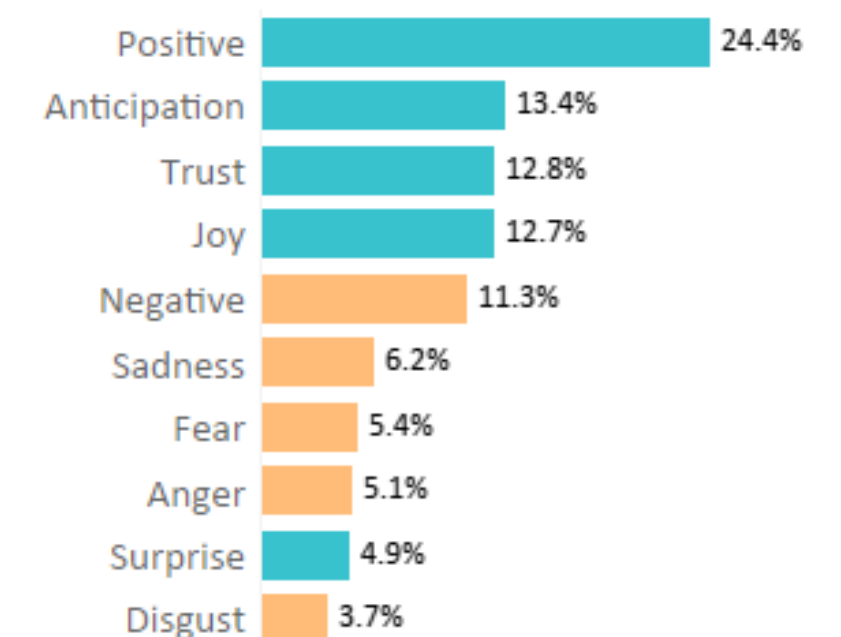
Let's take a look at the emotions and sentiments associated to the words used in tweets when mentioning the combined words #covidhobbies Quarantine Hobbies | Quarantine Hobby | Pandemic Hobbies | Pandemic Hobby



In case you missed it

This is how the **Emotions and Sentiments** rank based on word frequency analysis when referring to Hobbies and Quarantine

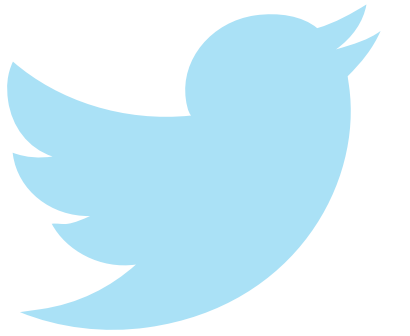
Hover over to see the top 10 words by Emotion/ Sentiment



Demo en Tableau Desktop

PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



EXPLORACION DE DATOS EN TABLEAU (CONT)



4. ANALYSIS Y/O MODELADO

- Seleccionar tipo análisis
- **Desarrollar metodología/ análisis**
- **Construir modelo/ Hacer analisis**
- **Buscar respuestas, contenido interesante**
- Testear modelo

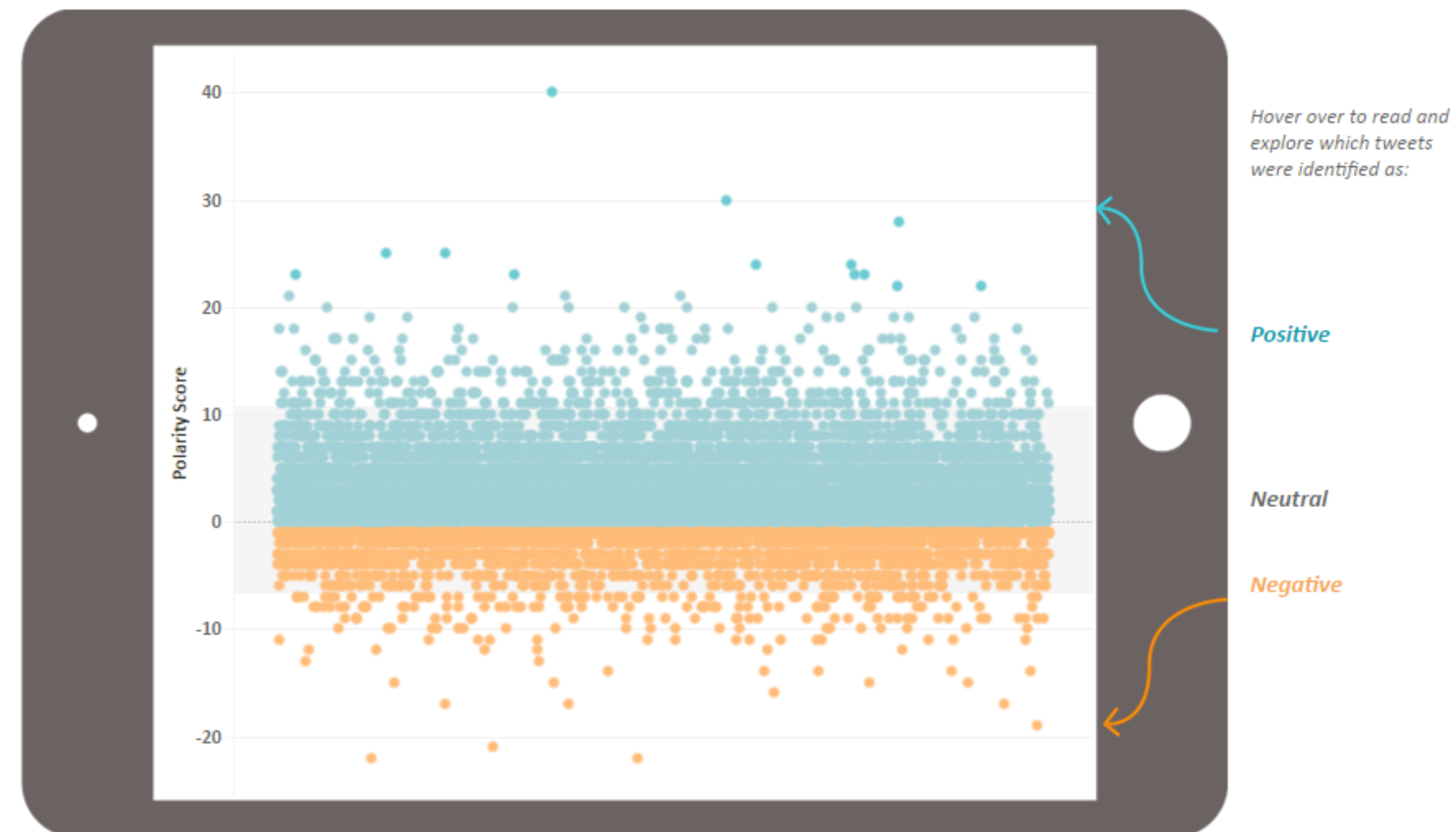


- Cual es la polaridad/ sentimiento detectado cuando se mencionan términos relacionados a Hobbies y cuarentena/ pandemia / Covid19



Polarity Scoring Analysis

The goal of Polarity Scoring analysis is to identify if the overall feeling of a given text (in this case a tweet message) is either positive, neutral or negative and we do this by assigning a score to each tweet based upon the total number of positive and negative words mentioned in the body of the text/tweet.



Demo en Tableau Desktop

PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



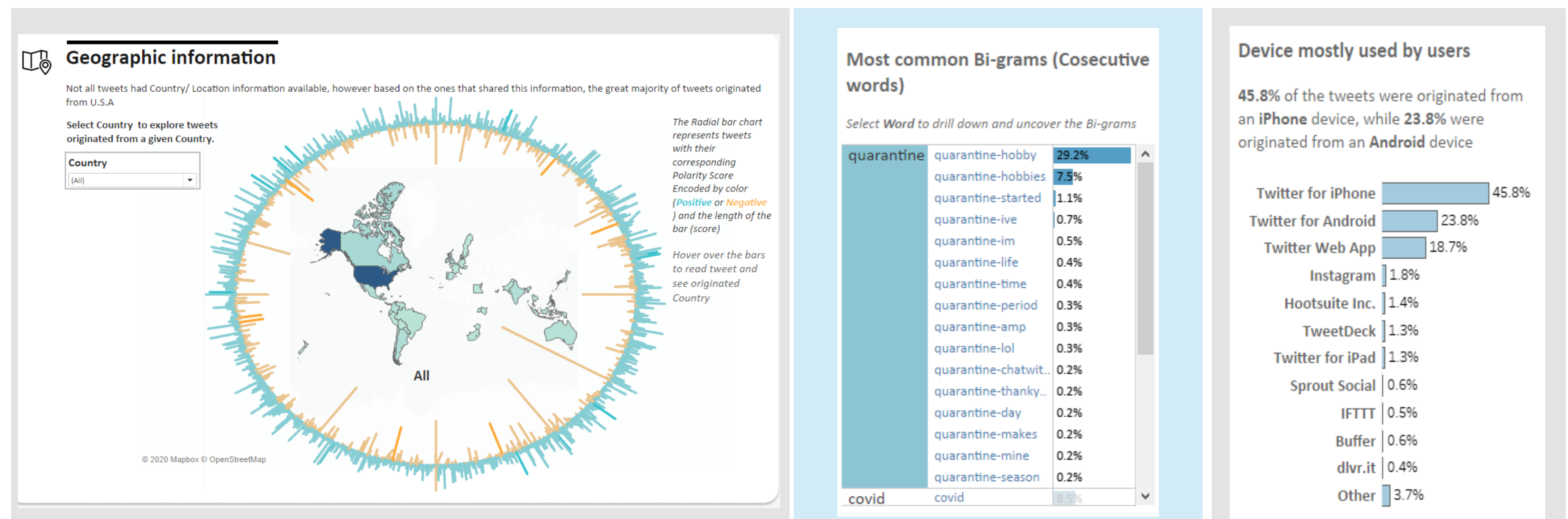
EXPLORACION DE DATOS EN TABLEAU (CONT)



4. ANALYSIS Y/O MODELADO

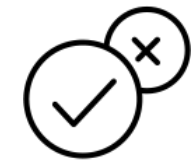
- Seleccionar tipo análisis
- **Desarrollar metodología/ análisis**
- **Construir modelo/ Hacer analisis**
- **Buscar respuestas, contenido interesante**
- Testear modelo

- Otros Datos interesantes:
- Información Geográfica/ País
- Bi-Grams
- Plataforma usada para enviar tweets
- Hashtags mas comunes
- Combinación de términos (q) que arrojaron mas resultados



Demo en Tableau Desktop

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA



5. EVALUACION

- Evaluar resultados
 - OK ?
 - NOT OK ?
- Revisar proceso
- Determinar próximos pasos

EVALUACION

EVALUAR RESULTADOS

- En esta punto ya debemos tener respuestas a las preguntas, verificamos si resultados son los esperados o si es necesario devolvemos a algunos de los pasos anteriores ya sea pedir aclaraciones, pedir o recolectar mas data, etc.

REVISAR PROCESO

- Hacemos revisiones del proceso (De principio a fin) en todas las plataformas/ herramientas que usamos para asegurarnos que sea reproducible
- Durante la revisión, se trata de simplificar el proceso y eliminar partes de código que ya no son necesarias, igualmente con las visualizaciones de identifica cuales son las que mejor convienen para contar la historia

DETERMINAR PROXIMOS PASOS

- Si resultados son los esperado procedemos a la fase final
- De lo contrario nos devolvemos al paso que sea necesario para continuar trabajando en el proyecto

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA

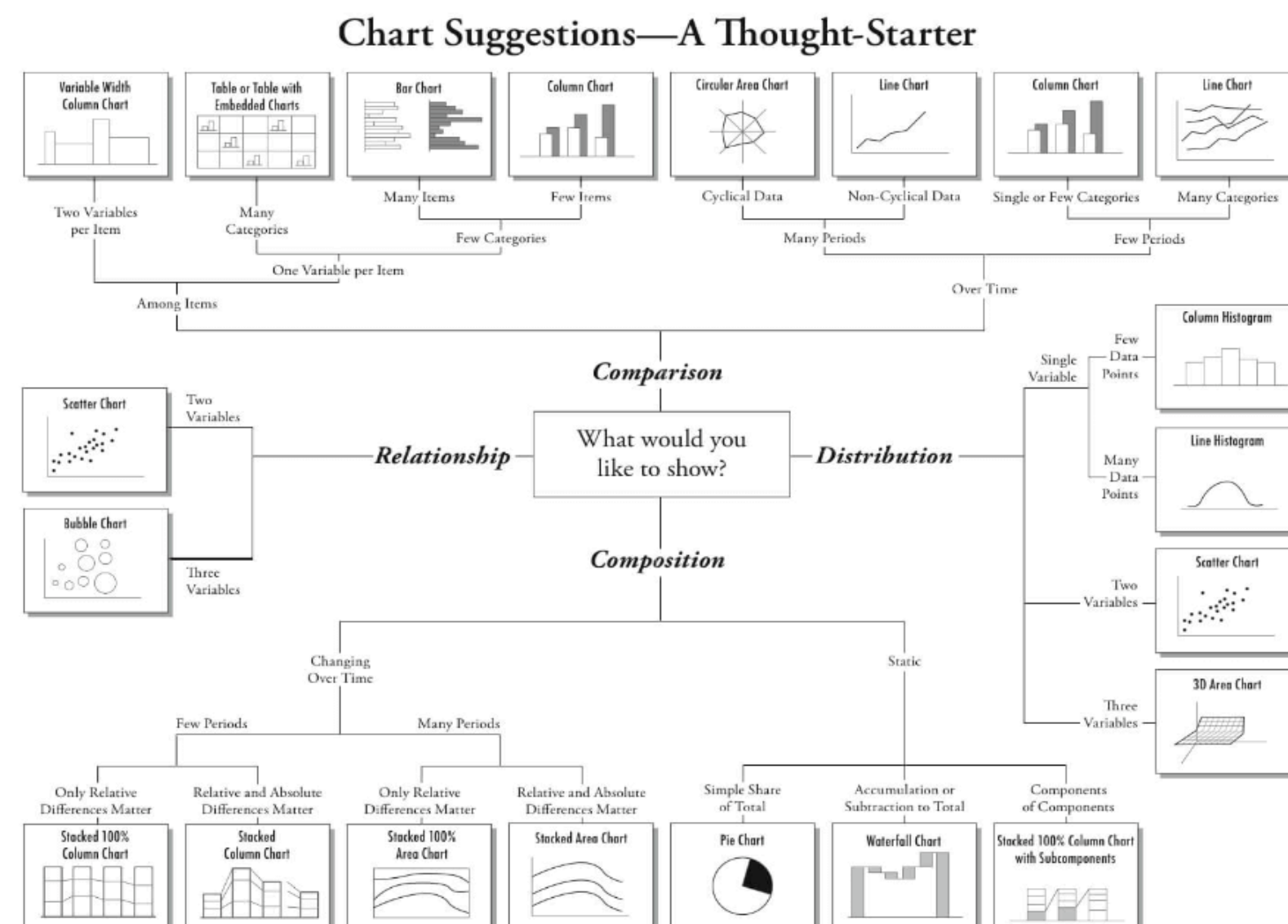


PASO 6. IMPLEMENTACION PRESENTACION

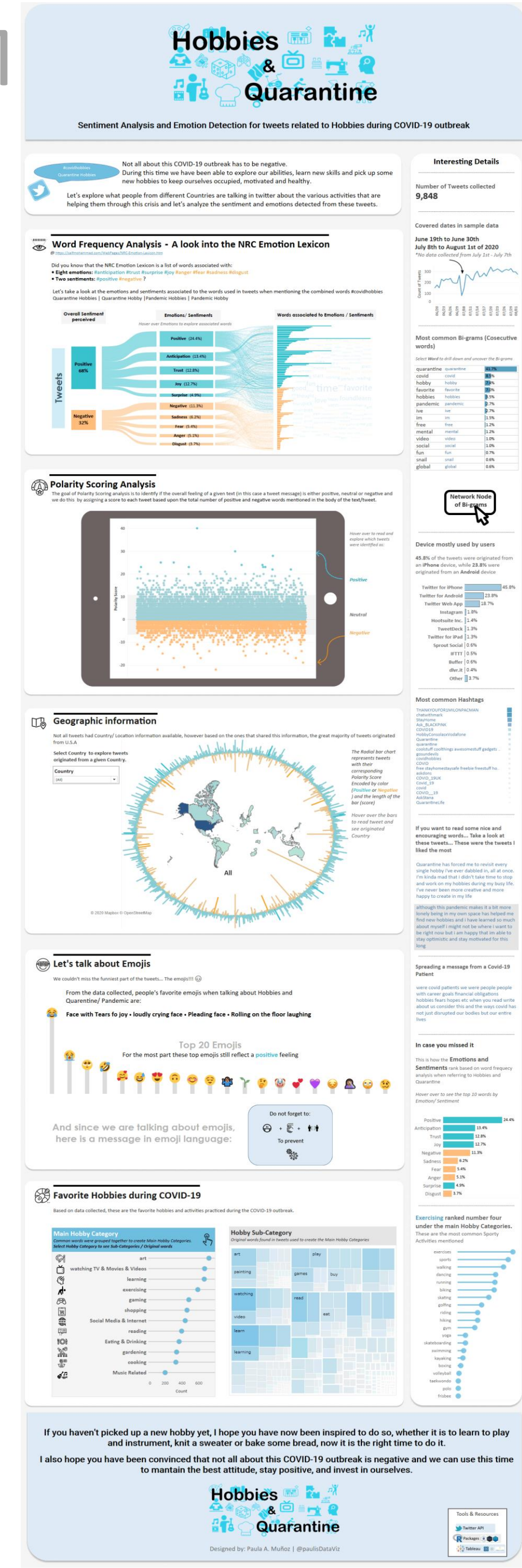
- Comunicar resultados
- Hacer recomendaciones
- Entrega producto final

IMPLEMETACION Y PRESENTACION

- Se comunican los resultados, se genera reporte y documentacion como guia
- Se hace presentacion al usuario(s) final
- Es importante tener en cuenta quien es el usuario final y adaptar la presentacion e historia acordemente
- Se hacen recomendaciones
- Nos aseguramos que el producto (Dashboard, informe, etc) sea accesible y entendible por el usuario final

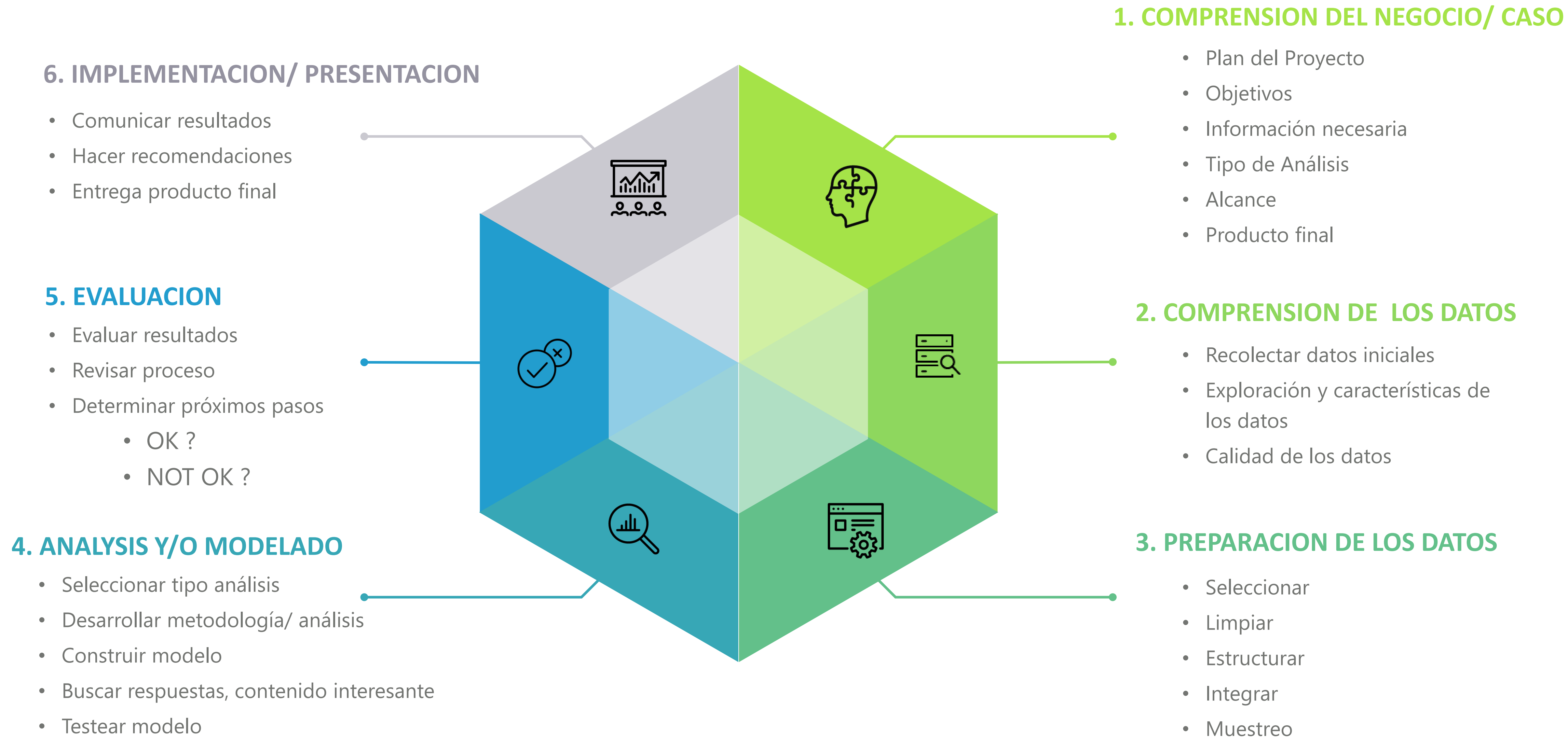


The right chart type by Dr. Andrew Abela



CICLO DE UN PROYECTO DE ANALISIS DE DATOS

METODOLOGIA CRISP - DM



PRACTICANDO CON UN PROYECTO - METODOLOGIA CRISP - DM

TWITTER ANALYTICS – HOBBIES Y LA CUARENTENA

REFERENCIAS

- **3-Stage Sankey Template:** <https://www.youtube.com/watch?v=ndvrj4drCB8>
- **Tableau Radial Bar Chart Tutorial:** <https://www.youtube.com/watch?v=d6-aptKLvgg>
- **NRC Word-Emotion Lexicon:** <http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm#:~:text=The%20NRC%20Emotion%20Lexicon%20is,were%20manually%20done%20by%20crowdsourcing.>
- **The Emoji Spotify Artists:** https://public.tableau.com/views/TheEmojiofSpotifyArtists/DescribeArtists?:embed=y&:display_count=yes&:showVizHome=no
- **Otros:**
 - <https://www.business-science.io/>
 - <https://www.campusbigdata.com/>
 - https://github.com/today-is-a-good-day/emojis/blob/master/emoji_analysis.R
 - <https://www.kdnuggets.com/2019/01/vazquez-2018-top-7-r-packages.html/2>
 - <http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know>
 - <https://towardsdatascience.com/create-a-word-cloud-with-r-bde3e7422e8a>
 - <https://stackoverflow.com/questions/31348453/how-do-i-clean-twitter-data-in-r>
 - <https://monkeylearn.com/sentiment-analysis/>
 - <https://www.tidytextmining.com/tidytext.html>
 - <https://www.youtube.com/watch?v=0KSI0WMGNRg>