

## **Segunda Entrega Proyecto de Título - Diplomado Data Science (USACH)**

# **Propuesta para formulación de un instrumento de evaluación para el desarrollo de nuevas comunas**

Martín Bonnefoy Valdés

### **Resumen de entrega anterior**

El trabajo actual se inscribe dentro del contexto de la indagación en torno a mecanismos mediante los cuales son evaluadas las propuestas de generación de nuevas comunas por parte de la SUBDERE, organismo institucional encargado de realizar los estudios de factibilidad para dichas solicitudes. Al respecto, se destacó en la entrega anterior la existencia de un interés por modernizar dichos procesos debido a la antigüedad de las metodologías existentes para ello (Ramírez, 2019), así como también de la necesidad de generar mecanismos que permitan sistematizar la información debido a la heterogeneidad presente tanto en los territorios en cuestión como en el tipo y magnitud de variables que son consideradas para ello (PUC-Instituto de Geografía, 2003). Es para ello que en la entrega anterior se presentó como propuesta el “... diseño de un sistema que permita integrar nuevas variables de estudio para ser posibles aristas de incorporar a las metodologías de evaluación existentes, incluyendo en el análisis aspectos provenientes de redes sociales y las imágenes satelitales para poder realizar estudios más dinámicos y ágiles que el único uso del Censo como fuente de información...”<sup>1</sup>, y específicamente se escogió como caso de estudio la localidad de Laguna Verde. El presente informe dará cuenta de los avances realizados en esa dirección.

### **Esquema institucional**

Debido a que el trabajo busca indagar sobre la revisión de un proceso metodológico en curso, en una primera instancia se buscó poder acceder al estado actual de este, ante lo cual fue posible ponerse en contacto con el Centro de Inteligencia Territorial de la Universidad Adolfo Ibáñez (CIT-UAI), quienes facilitaron una presentación realizada ante la SUBDERE el año 2018 en la que se presentaba una propuesta metodológica para la evaluación de generación de nuevas regiones y comunas (CIT-UAI, 2018). Este documento detalla un diagrama de flujo compuesto por diversas etapas y sub-etapas, las que detallarían una hoja de ruta revisada para la modificación de la DPA (División Político Administrativa), la que regiría las divisiones territoriales existentes y las propuestas tanto para el caso de regiones como comunas.

---

<sup>1</sup> Señalado en la entrega anterior.

En lo que es el caso de la propuesta de nuevas comunas, la primera instancia consistiría en una etapa de pre-factibilidad (ver Diagrama 1), la que a su vez consistiría de cuatro sub-etapas: siendo la primera el ingreso de la solicitud de generación de la nueva entidad territorial, la segunda un estudio de pre-factibilidad técnica, la tercera la evaluación por parte de la autoridad de los antecedentes recopilados, y una cuarta y última etapa en la que se encargaría el inicio de un estudio de factibilidad.

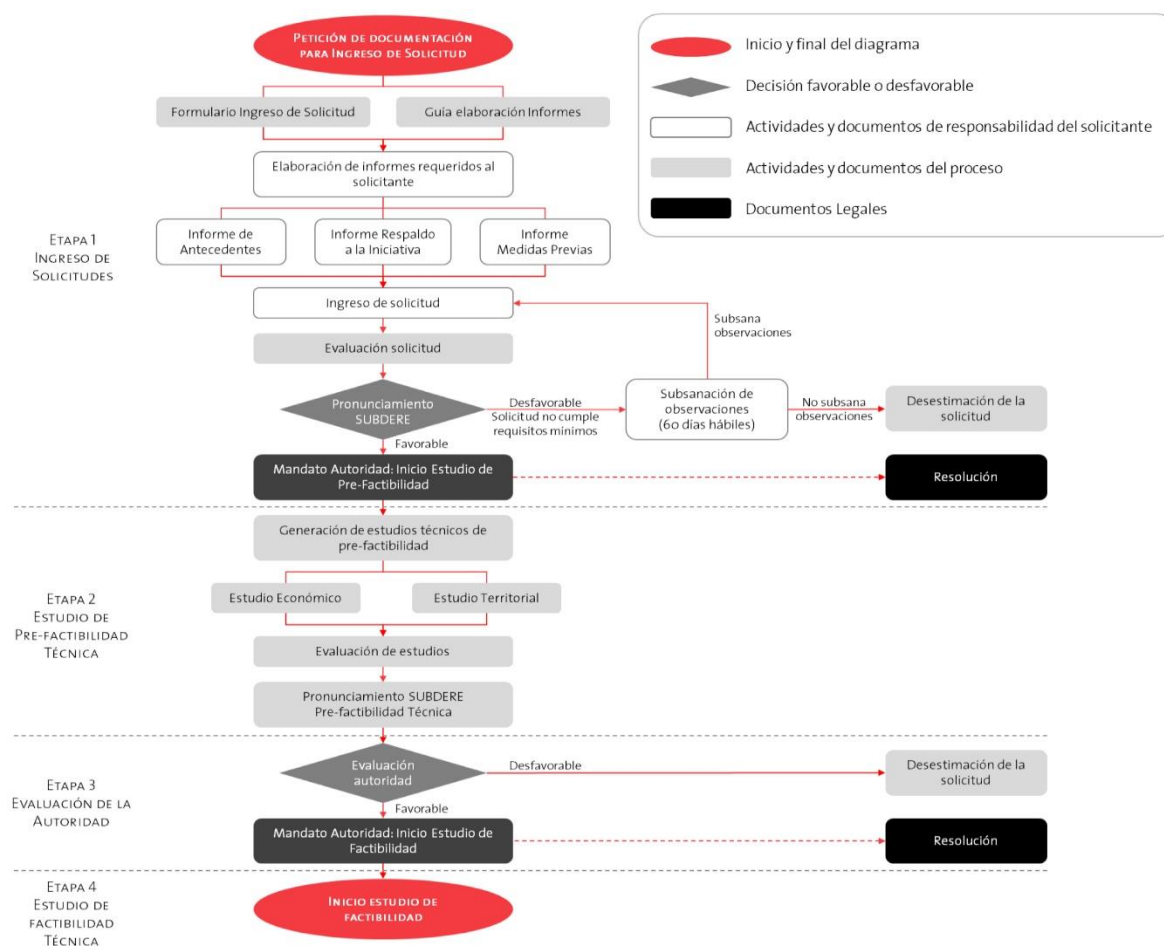


Diagrama 1. Esquema de estudio de prefactibilidad

Atingente a los objetivos planteados en el presente trabajo se identifica la existencia de un Estudio Territorial, sin embargo, al respecto de este el documento no hace referencia a en qué consiste ni las variables que considera para el caso de la generación de nuevas comunas, como sí lo hace para el caso de las nuevas regiones (CIT-UAI, 2018, pp. 17–20), por lo que esta no pudo ser considerada en el desarrollo de este estudio.

Posteriormente, una vez aprobada la pre-factibilidad, el documento señala que se iniciaría un estudio de factibilidad, el que a su vez consistiría de tres sub-etapas: en primer lugar un

estudio de factibilidad técnica, en segundo lugar una “Evaluación de la Autoridad y la ciudadanía”, y en tercer y último lugar correspondería la creación de la nueva comuna.

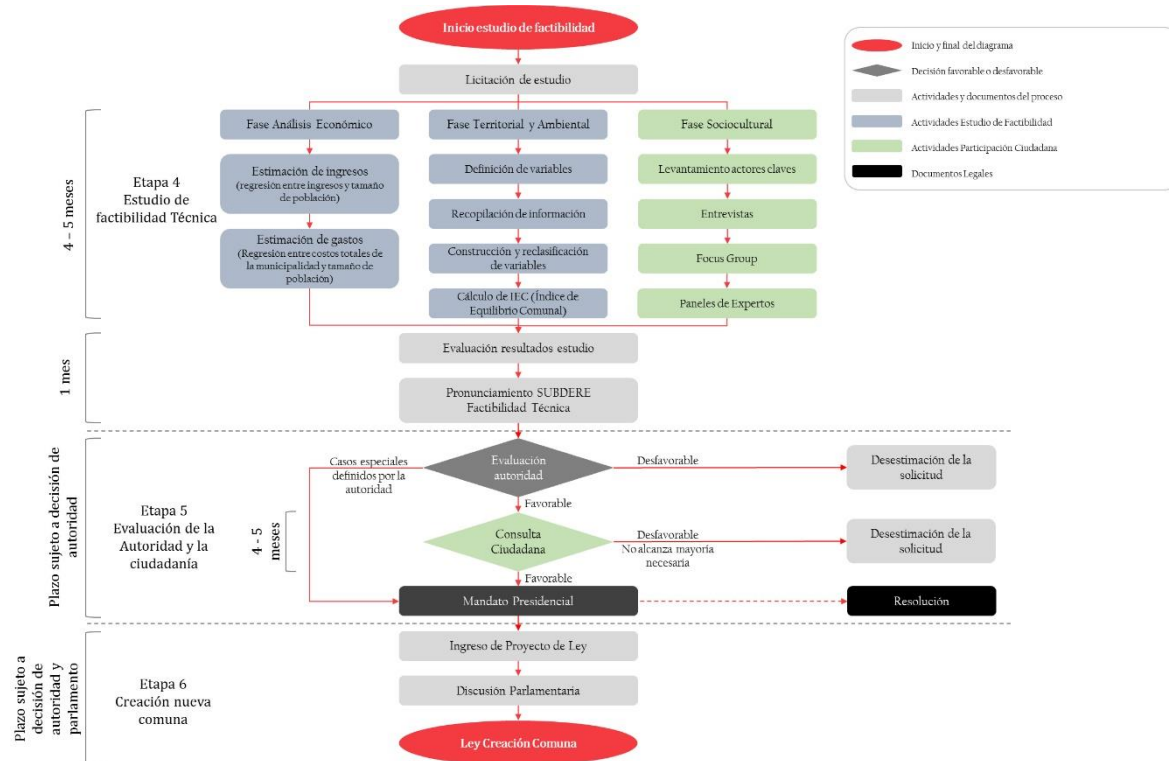


Diagrama 2. Esquema de estudio de factibilidad

## Propuesta

De acuerdo a lo señalado anteriormente, el presente trabajo buscó tanto identificar nuevas variables para sumar a las metodologías existentes como nuevas maneras de analizar y sistematizar las variables ya trabajadas. Para estos fines se pudieron identificar dos áreas de desarrollo posibles, estando ambas presentes tanto en los objetivos iniciales del presente trabajo como en aquellos declarados por la metodología desarrollada por el CIT-UAI en relación a poder realizar “... estudios de factibilidad de creación de nuevas regiones y comunas, considerando las dinámicas territoriales existentes...” (CIT-UAI, 2018, p. 3), las que se propondría podrían ser integradas como variables para el estudio de factibilidad.

### 1.- Fase Territorial y Ambiental - Proyección y caracterización territorial

En la etapa de factibilidad la propuesta analizada toma como variable o indicador una serie de indicadores en lo que llama la “Fase Territorial y Ambiental” (CIT-UAI, 2018, p. 43),

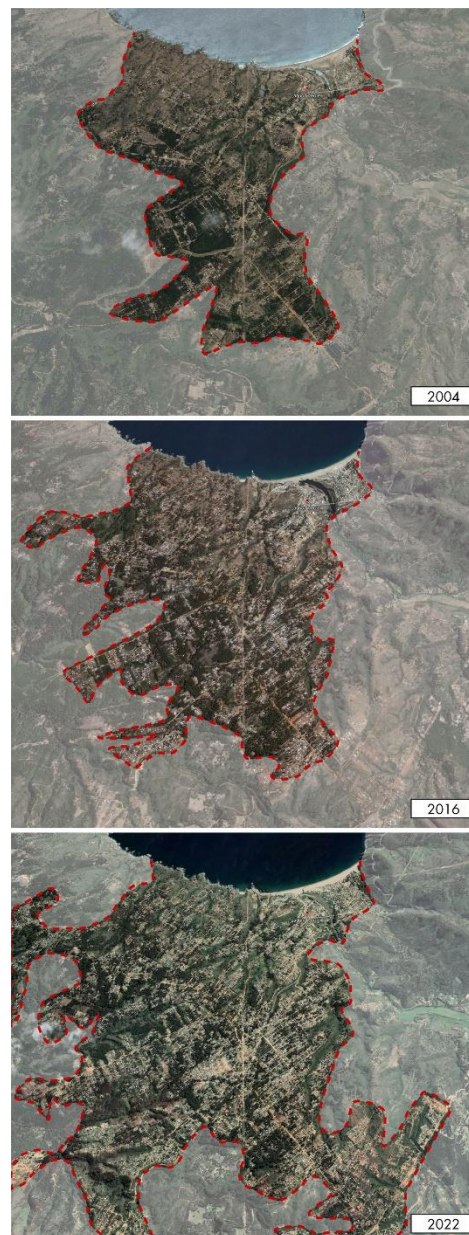
entre las que considera la evaluación del estado actual de factores demográficos como la “Población Comunal” y la “Extensión territorial de la comuna solicitada” (CIT-UAI, 2018, p. 43). Si bien la propuesta no profundiza sobre los instrumentos que serían ocupados para levantar dichos indicadores, da a entender que estas serían métricas realizadas sobre la realidad presente al momento de realizado el estudio de factibilidad.

Debido a la manera en que muchos de los territorios en cuestión están en un proceso de expansión territorial desregulado y acelerado, como lo ha sido en los últimos años en el caso de estudio elegido (ver Imagen 1), se considera esencial poder realizar una proyección del crecimiento de estas no solo atendiendo al aumento demográfico sino también previendo la futura demarcación de un posible límite urbano, incorporando el análisis de lo espacial en el estudio.

Adicionalmente, se identifican dos áreas a considerar: en primer lugar identificar instrumentos que permitan realizar mediciones ágiles y sensibles al tiempo, considerando el intervalo temporal entre cada Censo como insuficiente para estudiar la expansión territorial en determinados casos de estudio, y también poder profundizar en lo que la metodología define como “Características de ocupación del Territorio”, con un énfasis en poder identificar y separar aquello que corresponde a la población de habitantes del lugar respecto a lo que sería la

construcción de segundas viviendas, casos en los que se considera tienen distintas necesidades y compromisos hacia los territorios en expansión.

Para ello es que se propone integrar como variable un estudio de la expansión urbana a la vez que logre caracterizar los sectores en los que esta ocurre, logrando discriminar entre polos de desarrollo urbano residencial y aquellos que se correspondan a viviendas para personas que no residen ahí. Esto se lograría a través de un análisis de cobertura y uso de suelo (en inglés *Land Cover and Land Use*, abreviado como LCLU) el que separa las imágenes satelitales en



*Imagen 1. Crecimiento urbano de Laguna Verde. La línea roja describe la expansión de la “mancha urbana” a partir de identificación visual de los límites del territorio construido.*



cuadrantes y posteriormente los clasifica a partir de reconocimiento automático a través de un *dataset* de tipologías, lo que luego puede sobreponerse a un estudio de iluminación nocturna mediante fotografías satelitales que captan la luz visible e infraroja-cercana (en inglés *Visible and Near-infrared*, abreviado como VNIR). Al momento de la realización de este informe, para el caso del estudio LCLU, si bien fue posible encontrar aproximaciones a la manera en que es posible de realizarse<sup>2</sup> es considerada fuera de las capacidades técnicas y los tiempos del diplomado, por lo que es dejado como algo a realizarse a posterioridad. Y por otra parte, en cuanto al estudio de iluminación nocturna, si bien fue posible visualizarla<sup>3</sup> (ver Imagen 2 y 3) se constató que los registros existentes cuentan con una muy baja resolución para el caso de estudio, por lo que se descarta para esta instancia.

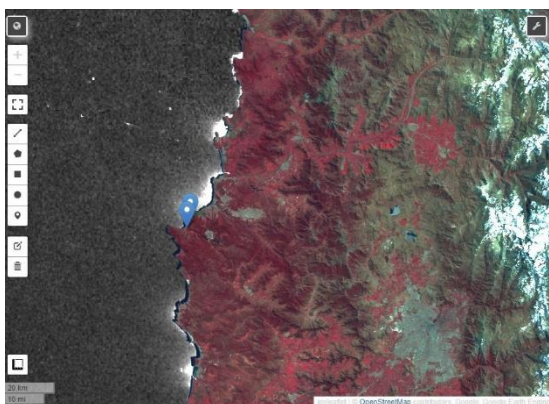


Imagen 3. Visualización de capas de imágenes satelitales mediante Google Earth Engine.  
Fuente: Elaboración propia

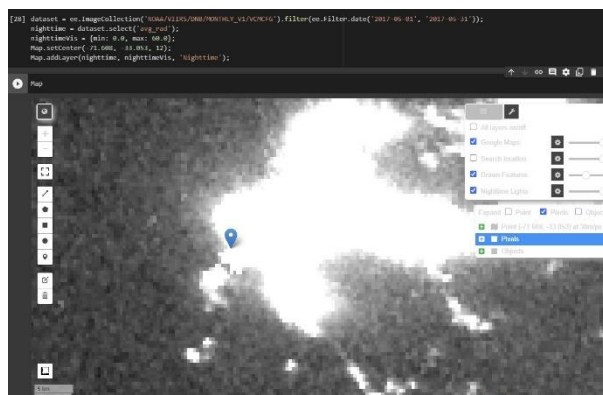


Imagen 2. Acercamiento a imagen satelital nocturna (VNIR) en Laguna Verde  
Fuente: Elaboración propia.

## 2.- Fase Sociocultural - Integración de variables desde RRSS

La metodología propuesta por el CIT-UAI integra dentro de sus componentes y variables a analizar un componente de participación ciudadana tanto en las instancias de prefactibilidad como de factibilidad (CIT-UAI, 2018, p. 39) (ver Tabla 1), estando dentro de sus objetivos el crear “... un mayor capital social, empoderando a los ciudadanos de comunidades como líderes de los procesos de desarrollo y cambio, lo cual fortalece la representación y la cohesión social de las comunidades.” (CIT-UAI, 2018, p. 37), aspectos clave dentro del reconocimiento de la necesidad de participación activa y constante en la conformación de nuevas delimitaciones territoriales y su gestión política y administrativa. Dicho proceso sería parte de lo que el estudio de factibilidad considera parte de la “Fase Sociocultural”, en la que la propuesta indica una serie de técnicas (CIT-UAI, 2018, p. 38). Dentro de estas, en una instancia inicial, es que la metodología indica “Técnicas para la Recolección de

<sup>2</sup> Land Use and Land Cover Classification using Pytorch.  
<https://colab.research.google.com/drive/1Wj0LsIuotZssoQUw0QUryWk9WbOFIzV8?usp=sharing>

<sup>3</sup> <https://colab.research.google.com/drive/1DLj-k42ox4GDsBRLXm0kZGYnr90iLPfn?usp=sharing>

Información”, la que plantea como objetivos “... recopilar datos o información nueva acerca del entorno, la comunidad, las demandas y necesidades existentes, y los conflictos en torno a un proyecto a desarrollar” (CIT-UAI, 2018, p. 38), y consistiría en una serie de instancias de levantamiento de información de tipo cualitativo tanto desde la interacción con la comunidad como a cargo de la persona o equipo a cargo del estudio (ver Tabla 2).

Técnicas de participación	Etapas de solicitud de DPA	Etapas de pre-factibilidad de DPA	Etapas de factibilidad de DPA
Formulario o cuestionario auto-administrado	X	X	X
Actores y redes para la participación	✓	✓	X
Redes	✓	✓	✓
Grupo focal	X	✓	✓
Grupo de discusión	X	✓	✓
Taller de trabajo	✓	✓	✓
Conferencia de búsqueda	✓	✓	✓
Casa abierta	X	✓	✓
Talleres de modelación/diseño	X	✓	✓
Entrevista	X	✓	✓
Recorrido por la dpa solicitada	✓	✓	✓
Análisis de documentos	X	✓	X
Consulta pública	X	X	✓
Protocolo de acuerdos	X	✓	✓
Comités permanentes	✓	✓	✓

Tabla 1. Mecanismos de participación propuestos. Fuente: (CIT-UAI, 2018, p. 39)

TÉCNICAS	OBJETIVOS	EJEMPLOS DE ENFOQUE METODOLÓGICO O TÉCNICA
Técnicas para la Recolección de Información	Éstas, se encuentran destinadas a recopilar datos o información nueva acerca del entorno, la comunidad, las demandas y necesidades existentes, y los conflictos en torno a un proyecto a desarrollar.	Grupos focales Grupos de discusión Taller de trabajo Conferencia de Búsqueda Casa Abierta Talleres de modelación Entrevistas Encuesta Cuestionario Auto-administrado Observación directa Observación participante Recorrido por el barrio Análisis de Documentos Análisis de Discurso
Herramientas para Sistematización de la Información	Éstos, permiten ordenar la información recolectada, facilitando su comprensión por parte de los distintos actores, simplificando la información disponible para la toma de decisiones, y permitiendo la incorporación de nuevos datos.	Análisis de Actores Análisis de Redes Rankings

Tabla 2. Técnicas y objetivos de participación. Fuente: (CIT-UAI, 2018, p. 38)

Es en esta instancia que se observa como espacio a desarrollar de manera complementaria a las técnicas indicadas, y de acuerdo al objetivo declarado de determinar la “... pertinencia de consultar la opinión a los actores claves en la creación de comunas y regiones, mecanismos para realizar las consultas, forma de financiamiento y cualquier otro elemento que permita levantar información de la participación de autoridades y comunidades en este aspecto” (CIT-UAI, 2018, p. 37) que se observa la posibilidad de integrar un elemento proveniente desde redes sociales que permita nutrir el análisis mediante la integración de variables e indicadores extraídas desde las mismas interacciones generadas en el cotidiano por los y las habitantes, y desde lo cual podrían extraerse insumos o variables a ser integrados en las dimensiones “Territorial” y “Sociocultural y Organizativo” (CIT-UAI, 2018, p. 43).

Para ello se elige trabajar con la red social Instagram, siguiendo metodologías de trabajo ejecutadas con anterioridad (Bonney, 2020; Bonney y Chaná, 2017), las que permitirían levantar y analizar información tanto desde los metadatos de las publicaciones compartidas (nombres de usuario/a, fechas y horas de subida, manejo de georreferencias, texto en comentarios, etc.) como aspectos relacionados con los contenidos de las imágenes en sí a través de un análisis exploratorio mediante la visualización de estas según sus características visuales de acuerdo a metodologías y herramientas previamente aplicadas en otros contextos (Manovich et al., 2015; Manovich y Hochman, 2013). El presente trabajo incluiría, a diferencia de las ejecuciones anteriores mencionadas, el procesamiento de texto a través de la aplicación de los métodos y técnicas vistas en el diplomado.

### **Adquisición de datos iniciales**

Para la obtención de datos se ocupó el *software* propietario 4k Instagram<sup>4</sup> debido a que a diferencia de alternativas como *instaloader*<sup>5</sup> o similares permite simplificar el procesamiento de la información al incrustar los metadatos de cada post en cada archivo de imagen. Para la obtención de las publicaciones se ocupó la georreferencia de Instagram *Laguna Verde, Valparaíso*<sup>6</sup>, la hasta la fecha contiene 34.770 publicaciones de imágenes fijas subidas desde el año 2012. Posteriormente desde estas se generaron dos dataset: uno con los metadatos y otro con las mediciones de características visuales, realizadas a través del *software* ImageJ<sup>7</sup>.

---

<sup>4</sup> <https://www.4kdownload.com/-ad0p9/es/stogram>

<sup>5</sup> <https://instaloader.github.io/>

<sup>6</sup> <https://www.instagram.com/explore/locations/228794947623537/laguna-verde-valparaiso/>

<sup>7</sup> <https://imagej.nih.gov/ij/index.html>

## Descripción de los datos

El procesamiento de datos fue realizado en Google Colab<sup>8</sup>, y en una primera instancia consistió en juntar en un mismo conjunto de datos (*dataframe*) los valores obtenidos desde las dos fuentes señaladas. Se obtuvieron así dos tipos de datos (ver Tabla 3):

- (1) **Variables categóricas** correspondientes a información en formato de texto extraída desde los metadatos de los archivos de imagen (nombre de archivo, nombre de usuario/a, texto escrito en el post, etc.), y también los valores de desviación estándar calculados por ImageJ en cada imagen, los que no fueron detectados como datos numéricos debido al formato en que son exportados.
- (2) **Variables numéricas** correspondientes a la información del peso de cada archivo de imagen, y los valores extraídos por ImageJ: un ID identificador de cada imagen y los cálculos de la media de brillo, saturación y matiz.

Variables Categóricas				
Nombre	Filename	URL	User	Dimensions
Etiqueta (qué)	Nombre de archivo asociado	Dirección de post	Nombre de usuario/a del post	Tamaño de la imagen
Ejemplo	2012-10-01 23.31.01 2929 72465788 38430922234700.jpg	https://www.instagram.com/p/QG2OfISXQ1/	ivanretamal	612 x 612
Tipo de Dato	str	str	str	str
Valor Mínimo	x	x	x	x
Valor Máximo	x	x	x	x
Unidad	x	x	x	x
Categorías	x	x	x	x
Cantidad de datos perdidos		7	7	0
Porcentaje de datos perdidos	0%	0.0201 31718961 2038	0.0201 31718961 2038	0%

Variables Categóricas					
Nombre	Description	Hashtags	brightness_stddev	saturation_stddev	hue_stddev
Etiqueta (qué)	Texto escrito en la publicación	Hashtags ocupados en la publicación	Desviación estándar del brillo	Desviación estándar de la saturación	Desviación estándar del matiz
Ejemplo	Camino a #lasdocas #lasdocaschile	lasdocas; lasdocaschile	476.681	24.932	810.837
Tipo de Dato	str	str	str	str	str
Valor Mínimo	x	x	x	x	x
Valor Máximo	x	x	x	x	x
Unidad	x	x	x	x	x
Categorías	x	x	x	x	x
Cantidad de datos perdidos	0	0	0	0	0
Porcentaje de datos perdidos	0%	0%	0%	0%	0%

Variables Numéricas					
Nombre	ImageID	Size	brightness_median	saturation_median	hue_median
Etiqueta (qué)	Identificador de imagen	Peso del archivo	Media del brillo	Media de la saturación	Media del matiz
Ejemplo	1	140841 .0	120	44	124
Tipo de Dato	int64	float64	int64	int64	int64
Valor Mínimo	1	9875 .0	0	0	0
Valor Máximo	3477 1	1949995 .0	255	255	253
Unidad	x	byte	x	x	x
Categorías	x	x	x	x	x
Cantidad de datos perdidos	0	0	0	0	0
Porcentaje de datos perdidos	0%	0%	0%	0%	0%

Tabla 3. Diccionario de variables.

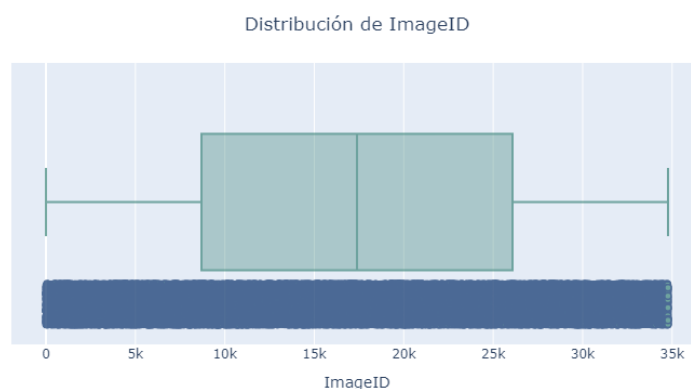
<sup>8</sup> <https://colab.research.google.com/drive/1bWIVjFmTU4NkqZntvD-FhcTuK6bIJeky?usp=sharing>



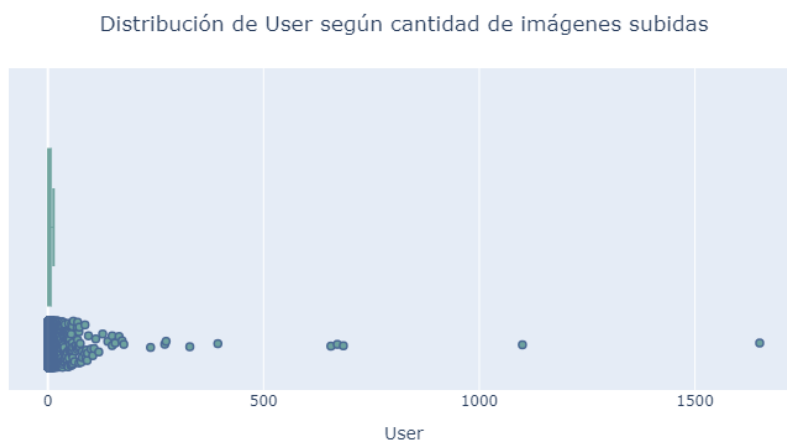
## Exploración de los datos iniciales

Durante la exploración de los datos iniciales sin ningún procesamiento se evidencia una serie de problemas respecto a la manera en que el conjunto de datos son generados, extraídos y luego importados.

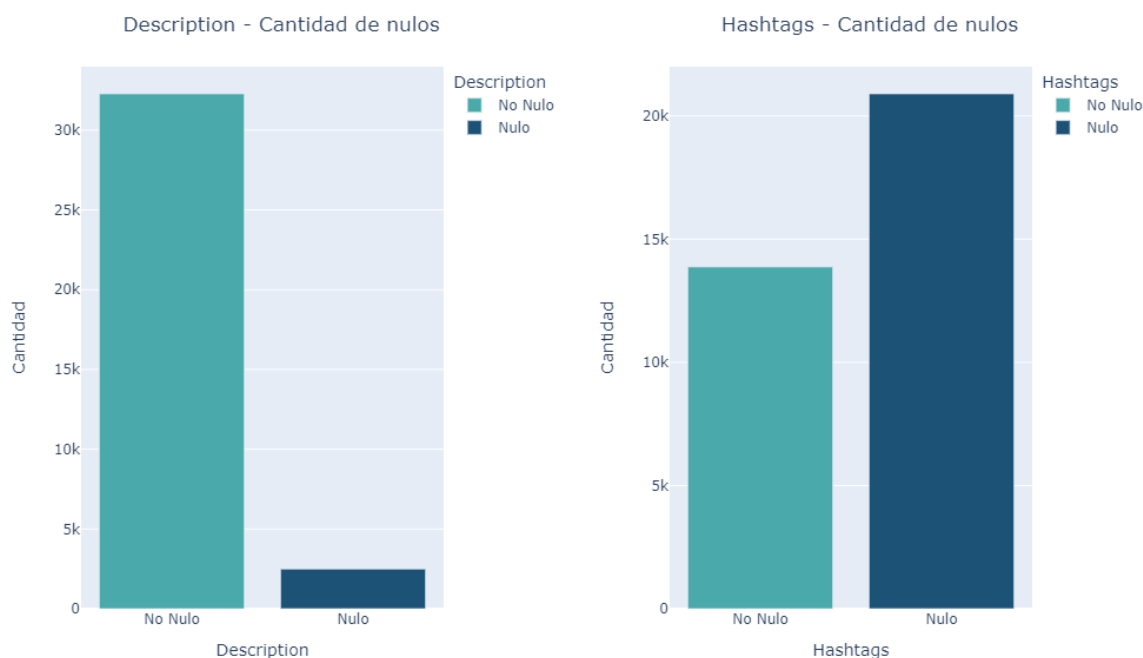
En primer lugar, para el ordenamiento de los datos, este puede realizarse secuencialmente a través del índice o de ImageID; sin embargo este ordenamiento solo responde a la indexación de cada archivo, no reflejándose ninguna información de día ni hora de subida de las imágenes ni características que puedan evidenciar algún tipo de distribución de los datos.



A partir del conteo de cantidad de veces en que cada User ha compartido imágenes (conteo de ImageID) sí es posible determinar una distribución en la que se evidencian algunos valores atípicos que es de utilidad identificar. Sin embargo, al igual que lo que fue señalado anteriormente, esta distribución sería absoluta y no permite representar la temporalidad de las publicaciones, ni tampoco discrimina entre cantidad de imágenes subidas versus la cantidad de posts, habiendo posts que contienen varias imágenes, cada una de las cuales suma al conteo.

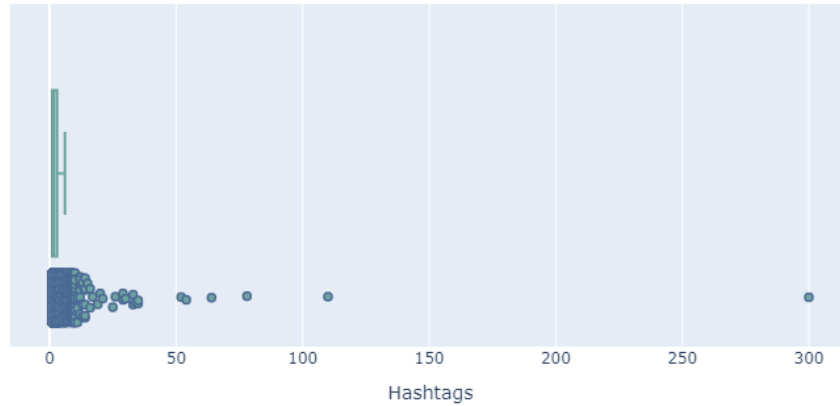


Tanto para el caso de Description como Hashtags estos están compuestos por cadenas de texto, y particularmente en el caso de Hashtags, estos son extraídos desde Description, por lo que se presentan todos los casos posibles: los posts sin descripción tienen valor nulo en Description y en Hashtags, y los posts con descripción pero sin uso de hashtags tienen nulos en Hashtags. A esta altura de la exploración se observa que ante la falta de una columna que detalle la presencia (o ausencia) de estos dos puede que resulte difícil una posterior agrupación.



En cuanto a la distribución de hashtags, al igual que en el caso anterior, se evidencia la presencia de una limitada cantidad de valores atípicos. Para esta instancia se evidencia un principal problema a trabajar la etapa de Integración: los hashtags están presentes como cadenas de texto, separadas (ver Tabla 4) pero sin un formato de lista, tupla o conjunto que permita realizar el conteo individual sobre ellos. Esto quiere decir que los hashtags, en la forma en que están presentes en el conjunto de datos, solo son una referencia a instancias donde fueron ocupados todos en la misma descripción de post, por lo cual contendrían la misma distribución que User.

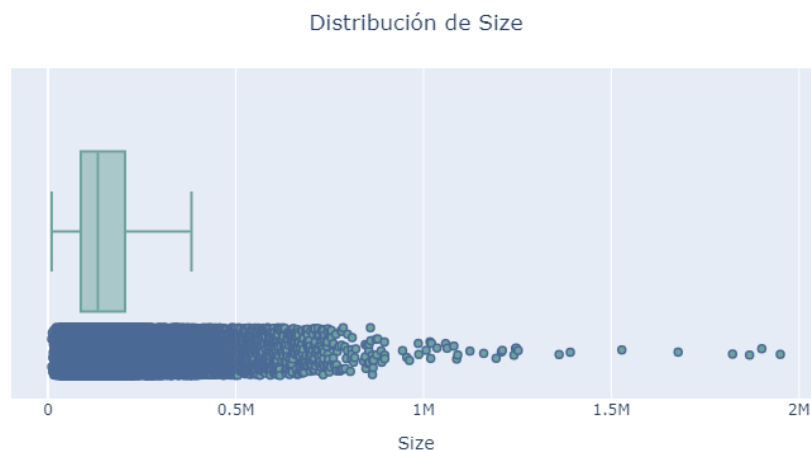
Distribución de Hashtags



index	ImageID	Hashtags
0	1	lasdocas; lasdocaschile
1	2	travel; loveislove; locura; whiskey; iloveyou; callmebaby; relax; lgbt; happy; loveme; disfrutando; libertad; naturaleza; felicidad; nearuda; bless; Moments; Perfect; amoramor; photography
2	3	travel; loveislove; locura; whiskey; iloveyou; callmebaby; relax; lgbt; happy; loveme; disfrutando; libertad; naturaleza; felicidad; nearuda; bless; Moments; Perfect; amoramor; photography
3	4	travel; loveislove; locura; whiskey; iloveyou; callmebaby; relax; lgbt; happy; loveme; disfrutando; libertad; naturaleza; felicidad; nearuda; bless; Moments; Perfect; amoramor; photography
4	5	travel; loveislove; locura; whiskey; iloveyou; callmebaby; relax; lgbt; happy; loveme; disfrutando; libertad; naturaleza; felicidad; nearuda; bless; Moments; Perfect; amoramor; photography
5	6	travel; loveislove; locura; whiskey; iloveyou; callmebaby; relax; lgbt; happy; loveme; disfrutando; libertad; naturaleza; felicidad; nearuda; bless; Moments; Perfect; amoramor; photography
6	7	marutichile; suzuki; pequeñosperopoderosos; clubmarutisuzukichile
7	8	marutichile; clubmarutisuzukichile; suzuki; maruti
8	9	NaN
9	10	laflajera
10	11	NaN
11	12	IronFest; Gilkl; Gileskls; SillaRIP
12	13	ig_chile; vicivisuals; arquitecturachile; theweekoninstagram; hallazgosemanal; way2ill; justgoshoot; valparaisohills; artofvisuals; heatercentral; createcommune; moodygrams; agameoffones; igshotz; hypnoticllm; theprecinct; dynamix_shots; leagueoffenses; rtephotography; thelensbible; fatalframes; ig_color; shotzdelight; theimaged; visualambassadors; acreativevisual; killersworldwide; valparaisotourism; valparaiso
13	14	NaN
14	15	huaweimate9; boramk4; vwchile; vwBora; volkswagenbora; mk4; jettamk4; vag; boraskyline; boramk4_5; vwjetta; vwlove; boraclubbr; volkswagen; Bulldog; englishbulldog; doglover; bulldoglovers

Tabla 4. Columna Hashtags en conjunto de datos inicial

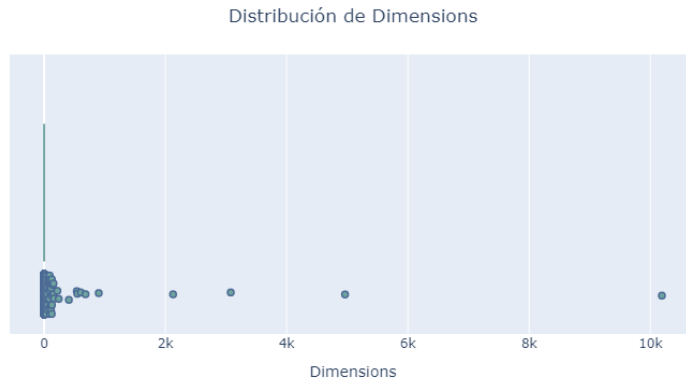
La variable Size, la que detalla el peso de cada archivo de imagen, en principio muestra una distribución cargada hacia pesos de archivo menores. Por el momento no se considera una variable que vaya a tener utilidad en el futuro, sin embargo, se decide mantener ante la posibilidad de que demuestre alguna correlación de interés en el futuro.



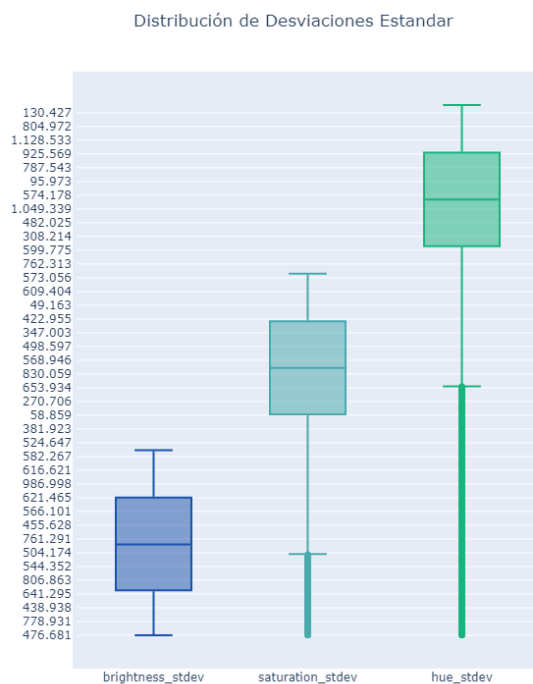
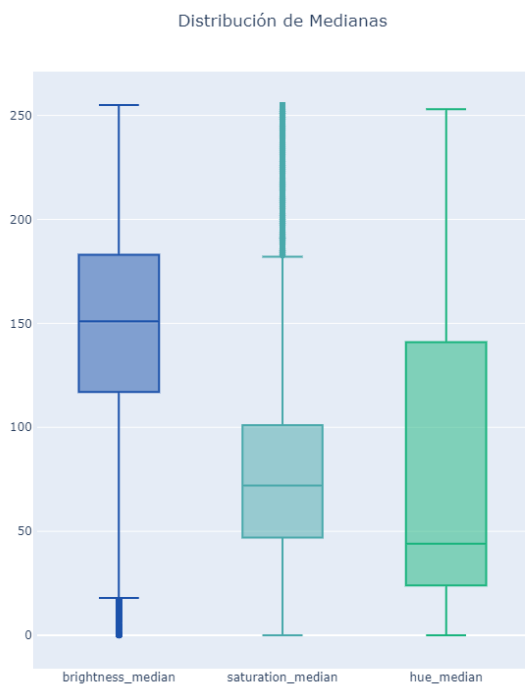
El caso de Dimensions sí se estima que pudiese ser un aspecto a considerarse dentro de un estudio sobre los tipos de registro de imagen compartidos, donde una proporción vertical suele corresponderse más a fotografías de registro de acciones, actividades y situaciones, mientras que las fotografías de proporciones horizontales se corresponden más a registro paisajístico, el que suele ser realizado más por personas ajenas al contexto, como lo son turistas. Se destaca sin embargo que el formato en que actualmente estaría presente la columna, como una cadena de caracteres con dos valores separados, lo que dificulta agrupaciones a partir de proporciones de imagen (ver Tabla 5), permite identificar que habría una distribución cargada hacia ciertos formatos; cuadrado, por ejemplo. Eventualmente sería lo óptimo el poder discernir si es el formateo automático realizado por Instagram o si es una decisión.

index	Dimensions	Count
246	1080 x 1080	10189
1169	1440 x 1440	4964
1369	1440 x 1800	3079
425	1080 x 1350	2128
2011	720 x 720	902
549	1080 x 810	684
1599	500 x 500	607
2634	960 x 960	549
1040	1440 x 1080	540
2056	720 x 900	409
1848	660 x 825	242
1368	1440 x 1799	220
1168	1440 x 1439	164
2186	780 x 780	157
427	1080 x 1352	156

Tabla 5. Columna Dimensions en conjunto de datos inicial



En cuanto a los datos visuales se observa que debido al formato en que estos fueron generados por ImageJ, el que realiza un cálculo sobre la mediana y la desviación estándar de los valores de todos los píxeles de cada imagen, los valores de medianas fueron correctamente importados como número entero pero el separador de decimal de punto de las desviaciones estándar no fue correctamente detectado. Esto implica que en el estado inicial del conjunto de datos no sería posible realizar evaluaciones ni cálculos tratando las desviaciones estándar como variables numéricas, al ser consideradas como categóricas.





## Calidad

Se puede señalar que en el estado inicial de los datos es poca la información que se puede extraer para fines del análisis a realizar: si bien se pueden identificar ciertos aspectos relativos a ciertos tipos de usuarios/as a partir de la cantidad de imágenes que han subido, esta presenta un sesgo al contabilizar cantidad de imágenes y no cantidad de posts, lo cual sería una métrica desproporcionada. Paralelamente, en el estado inicial de los datos, estos no contienen una marca de tiempo desde la cual se pudiera analizar la consistencia (o lo contrario) de usuarios/as a lo largo del tiempo y los análisis bivariados que se podrían realizar a través de ello, lo que se considera algo esencial en los objetivos planteados al lograr separar la actividad en redes sociales de personas que viven en el lugar respecto a la de aquellas que solo están de visita. También vinculado a los objetivos señalados, respecto a la posibilidad de incorporar el análisis de texto, se señala que por la forma en el que este estaría actualmente formateado tanto en descripciones de posts como en el listado extraído de hashtags, esta información no estaría en condiciones de ser ocupada y procesada.

Por lo mismo se plantean las siguientes estrategias:

- (1) En primer extraer la información relativa a la temporalidad de cada post, la que es posible de ser extraída desde *Filename*.
- (2) Extraer métricas respecto a la cantidad de publicaciones realizadas por cada persona en lugar de la cantidad de imágenes subidas, tal de poder identificar la frecuencia de su actividad en la red social.
- (3) Transformación todas las columnas en las que el formato de la información no sería el adecuado, de acuerdo a lo señalado en la exploración.
- (4) Cuando corresponda, desplegar una estrategia que permita normalizar los valores atípicos para transformar las columnas.

## Preparación de los datos

### 1. Selección

Debido al carácter exploratorio de la investigación no se decide generar ninguna selección ni filtrado *a priori* de la información, ocupándose todas las columnas y todos los registros presentes en el análisis. Se hace énfasis en que uno de los objetivos sería no perder la correlación entre el conjunto de datos y los archivos de imágenes descargados, con el fin de poder ocupar la información procesada como criterio para el ordenamiento de visualizaciones con ImageJ.

## 2. Limpieza y Construcción

### 2.1. Tratamiento de datos nulos

Se encontraron 7 registros donde no se registró correctamente la información de metadatos contenidos en el archivo: URL, User, y Description (ver Tabla 6). Se consideró que dado lo reducida que es esta muestra, la mejor alternativa sería eliminar los 7 registros junto con sus correlativos 7 archivos de imagen, por sobre alternativas como buscar reemplazar sus valores.

ImageID	Filename	Size	URL	User	Description	Hashtags	Dimensions	brightness_median	brightness_stddev	saturation_median	saturation_stddev	hue_median	hue_stddev
9341	2021-06-25 17.24.23 2604127940047036089_8121111627.jpg	68772.0	NaN	NaN	NaN	NaN	958 x 958	198	416969	16	769564	45	567887
11493	2021-07-28 21.18.07 2628163174874265596_10009967815.jpg	52729.0	NaN	NaN	NaN	NaN	1080 x 1080	92	80138	87	433802	15	624306
12608	2021-08-16 08.51.59 264158837618865192_25569231346.jpg	162508.0	NaN	NaN	NaN	NaN	1080 x 810	222	670652	167	878386	127	924924
15667	2021-10-09 12.46.26 2680784069107581493_4038118807.jpg	162254.0	NaN	NaN	NaN	NaN	1080 x 1080	158	654258	79	514731	44	549984
24070	2022-02-22 18.45.31 2779534306470495416_1360264757.jpg	180306.0	NaN	NaN	NaN	NaN	945 x 945	177	640233	103	480928	132	726277
24854	2022-03-10 11.20.36 2790906782459480977_7161250259.jpg	193697.0	NaN	NaN	NaN	NaN	1317 x 1317	227	671263	49	376144	20	57416
25584	2022-03-27 13.51.07 2803303720941067315_11857052694.jpg	101808.0	NaN	NaN	NaN	NaN	1080 x 1080	90	635154	56	374788	23	862079

Tabla 6. Registros incompletos

### 2.2. Extracción de fechas

Para la obtención de fechas se extrajo y transformó la información contenida en el nombre de archivo con el fin de obtener una columna en formato datetime. Posteriormente se realizó sobre esta operaciones para poder desagregar la información en conceptos de Año-Mes, Hora, Día de la semana, etc. (ver Tabla 7) con el fin de poder realizar evaluaciones y visualizaciones agrupadas por distintos formatos de tiempo.

ImageID	Filename	Date	Date_Year	Date_YearMonth	Date_Hour	Date_DOW
1	2012-10-01 23.31.01 292972465788384309_222234700.jpg	01-10-2012 23:31	2012	2012-10	23	1
2	2017-07-24 19.55.50 1566324167988813352_4193691355.jpg	24-07-2017 19:55	2017	2017-07	19	1
3	2017-07-24 19.55.50 1566324194572259642_4193691355.jpg	24-07-2017 19:55	2017	2017-07	19	1
4	2017-07-24 19.55.50 1566324222900693468_4193691355.jpg	24-07-2017 19:55	2017	2017-07	19	1
5	2017-07-24 19.55.50 1566324249467417713_4193691355.jpg	24-07-2017 19:55	2017	2017-07	19	1

Tabla 7. Columnas de tiempo

### 2.3. Cadenas de strings - hashtags

Para el manejo de cadenas de texto contenidas en Hashtags se generaron tuplas que contienen el conjunto de hashtags ocupados en cada descripción, al ser así conjuntos de datos sobre el cual luego se pueda iterar por sus elementos de manera individual, lo que no es posible con Pandas sobre las listas. Se generó también una columna en la que se contabilizó la cantidad de hashtags ocupados en cada publicación (ver Tabla 8).

index	ImageID	User	Description	Hashtags	N_Hashtags
10	11	about_you	And change the way in which I saw the life??	NaN	0
11	12	iamlucholi	Su Ironfest piola #IronFest #Giki #Gieskis #Sill aRIP	(IronFest,Giki,Gieskis,Sill aRIP)	4
12	13	lucas_g_acevedo	Laguna Verde es una bahía ubicada al sur de la ciudad de Valparaíso, si no lo conoces aún te aseguro que es lo más cercano a un lugar paradisíaco. Por cierto, lo primero que hice fue tirarme en la arena y al ver a mis amigos alejarse tuve un flash de fotógrafo en la mente y se me ocurrió esta idea... ¿Les gusta cómo quedó? #ig_chile #vicivisuals #arquitecturachile #heweeikoninstagram #hallazgosemanal #way2il #justgoshoot #valparaísohills #artofvisuals #theadcentral #createcommune #moodgrams #agameofones #ighotz #hypnoticdm #heprecind #dynamix_shots #leagueofenses #ephography #helenbible #fataframes #ig_color #shotdelight #heimaged #visualambassadors #creativevisual #killersworldwide #valparaísotourism #valparaíso	(ig_chile,vicivisuals,arquitecturachile,heweeikoninstagram,hallazgosemanal,way2il,justgoshoot,valparaísohills,artofvisuals,theadcentral,createcommune,moodgrams,agameofones,ighotz,hypnoticdm,heprecind,dynamix_shots,leagueofenses,ephography,helenbible,fataframes,ig_color,shotdelight,heimaged,visualambassadors,creativevisual,killersworldwide,valparaísotourism,valparaíso)	29
13	14	ctes.dr	A veces simplemente no nos fijamos en la lindas cosas que nos rodean?????	NaN	0
14	15	davidflowchannels	Posando para la Foto ??????#huaweimate9 #boramk4 #vwchile #vwBora #volkswagenbora #mk4 #jetamk4 #vag #boraskyline #boramk4_5 #vwjetta #vwlove #boracubbr #volkswagen #bulldog #englishbulldog #doglover #bulldoglovers	(huaweimate9,boramk4,vwchile,vwBora,volkswagenbora,mk4,jetamk4,vag,boraskyline,boramk4_5,vwjetta,vwlove,boracubbr,volkswagen,bulldog,englishbulldog,doglover,bulldoglovers)	18
15	16	criss.zamora	Mierda! #instalike #like4like #followtofollow #likeme #gay #gayusagays #gaylove #gaylife #gaylikes4likes #gaylikes #gayfollow #gay4t #gayfitness #gayboy #gaymalpu #maipugay #santiagogay #gaysantiago #chilegay #gaychile #gayman #gayusagays #gaygay #gay15yearolds #adolecentegay #gaypride #gaymen #gaystagram #gaylatino	(instalike,like4like,followtofollow,likeme,gay,gayusagays,gaylove,gaylife,gaylikes4likes,gaylikes,gayfollow,gay4t,gayfitness,gayboy,gaymalpu,maipugay,santiago,gay,gaysantiago,chilegay,gaychile,gayman,gaygay,gay15yearolds,adolecentegay,gaypride,gaymen,gaystagram,gaylatino)	29

Tabla 8. Columnas de hashtags

## 2.4. Columnas de Cantidad de Imágenes y Cantidad de Posts por User

De acuerdo a lo señalado anteriormente se buscó generar columnas que identifiquen, de manera separada, la cantidad de imágenes que ha subido cada usuario/a y la cantidad de posts que ha subido (ver Tabla 9), entendiendo que pueden haber usuarios/as que subieron solo unos pocos posts con muchas imágenes, y que habría que diferenciarlo de cuantas veces la persona subió un post, independiente de la cantidad de imágenes que este tenga.

index	ImageID	User	Date	N_Images	N_Posts
5	6	nea_ruda	24-07-2017 19:55	5	1
6	7	clubsuzukimarutichile	09-08-2017 18:54	2	2
7	8	clubsuzukimarutichile	09-08-2017 18:55	2	2
8	9	isialvear	27-08-2017 17:45	11	7
9	10	electroandblue	14-09-2017 0:31	5	5

Tabla 9. Columnas de Número de Imágenes y Número de posts

## 2.5. Reconstrucción columna ImageID

Se encontraron 566 casos donde el valor ImageID, el que es generado por ImageJ, se encontraba duplicado. Para ello se reconstituyó desde cero esta columna, asignando valores generados automáticamente desde 1 hacia arriba.

## 2.6. Transformación de columna Dimensions

Tal de poder más adelante generar evaluaciones y filtros a partir de las proporciones de cada imagen es que se generó una separación de los valores contenidos en la columna Dimensions (ver Tabla 10), los que consistían en una cadena de texto con ambos valores (ancho y alto) separados.

index	ImageID	Dimensions	Dimension(W)	Dimension(H)
9340	9341	958 x 958	958	958
11492	11493	1080 x 1080	1080	1080
12607	12608	1080 x 810	1080	810
15666	15667	1080 x 1080	1080	1080
24069	24070	945 x 945	945	945
24853	24854	1317 x 1317	1317	1317
25583	25584	1080 x 1080	1080	1080

Tabla 10. Desagregación de columna Dimensions

### 3. Formato

De acuerdo a los objetivos señalados anteriormente se hicieron transformaciones de los tipos de datos contenidos en las columnas y se generaron las columnas derivadas correspondientes (ver Tabla 11) con el fin de poder posteriormente generar visualizaciones y agrupaciones en base a criterios sensibles ante el tipo de información que contienen, especialmente en lo que es la situación de poder incorporar al análisis series de tiempo que permitan constatar la evolución de la actividad de la red social en torno a la georreferencia a lo largo de año, y poder discernir entre cantidad de imágenes y cantidad de posts.

Columna (original)	Inicial	Transformada	Columna (derivada)	Inicial	Transformada
ImageID	int	string	Date	x	datetime
Filename	string	string	Date_Year	x	int
Size	float	int	Date_YearMonth	x	string
URL	string	string	Date_Hour	x	int
User	string	string	Date_DOW	x	int
Description	string	string	N_Hashtags	x	int
Hashtags	string	tuple	N_Images	x	int
Dimensions	string	x	N_Posts	x	int
brightness_median	int	int	Dimension(W)	x	int
brightness_stdev	string	int	Dimension(H)	x	int
saturation_median	int	int			
saturation_stdev	string	int			
hue_median	int	int			
hue_stdev	string	int			

Tabla 11. Transformaciones y columnas derivadas

## 4. Resultados preliminares

### 4.1. Periodo de estudio

La incorporación de información relativa al paso del tiempo permitió visualizar, a una escala general, la manera en que la red social ha aumentado su actividad transversalmente con el paso de los años, lo que coincidiría con el crecimiento urbano identificado y señalado en Laguna Verde, y también permite observar patrones de uso a lo largo de los meses. Esto

permitió la posibilidad de delimitar y acotar un periodo de estudio a través de la consideración por no tomar en consideración ciertos valores atípicos, particularmente posts subidos años atrás. Para ello se consideró el concepto de “*lower outer fence*”, el que se calcula con la fórmula  $Q1 - (3 * IQR)$ , el que devolvió como límite inferior la fecha 2018-12-31.

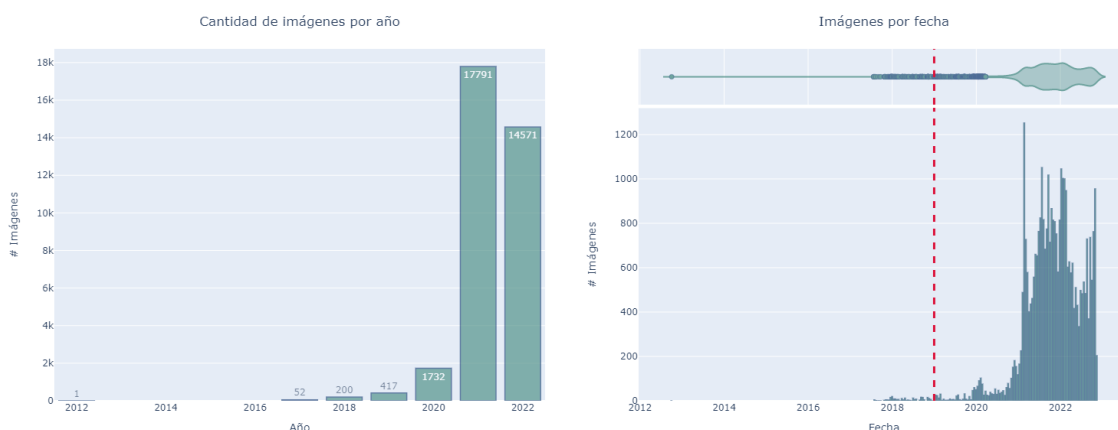


Imagen 4. Línea vertical roja: *lower outer fence* (2018-12-31)

Al cambiar a una medición a partir de cantidad de posts se evidenció un cambio radical en la escala; mientras que la máxima cantidad de imágenes del año 2021 fueron 17.791, para el mismo año la cantidad de posts fue de 7.498. Sin embargo, al buscar nuevamente un límite inferior para determinar un periodo de estudio, el nuevo límite correspondió a la fecha 2018-11-16; apenas un mes antes que la fecha calculada anteriormente, por lo que para conceptos de este análisis no se considera relevante.

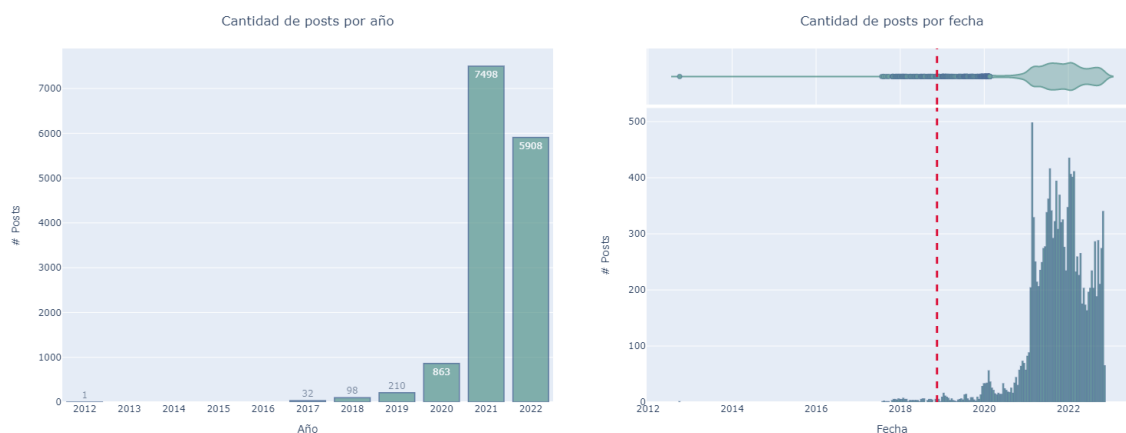


Imagen 5. Línea vertical roja: *lower outer fence* (2018-11-16)

Se decide así fijar como límite inferior para el periodo de estudio la segunda fecha obtenida; 2018-11-16.



## 4.2. Columnas de información visual

Una vez realizada la transformación sobre el formato de los datos de las desviaciones estándar y convertida esta en variable numérica los valores se equilibraron, resultando en una distribución más similar a la presente en la mediana. Se logra así poder disponer de la información de esta columna para cálculos y evaluaciones futuras.

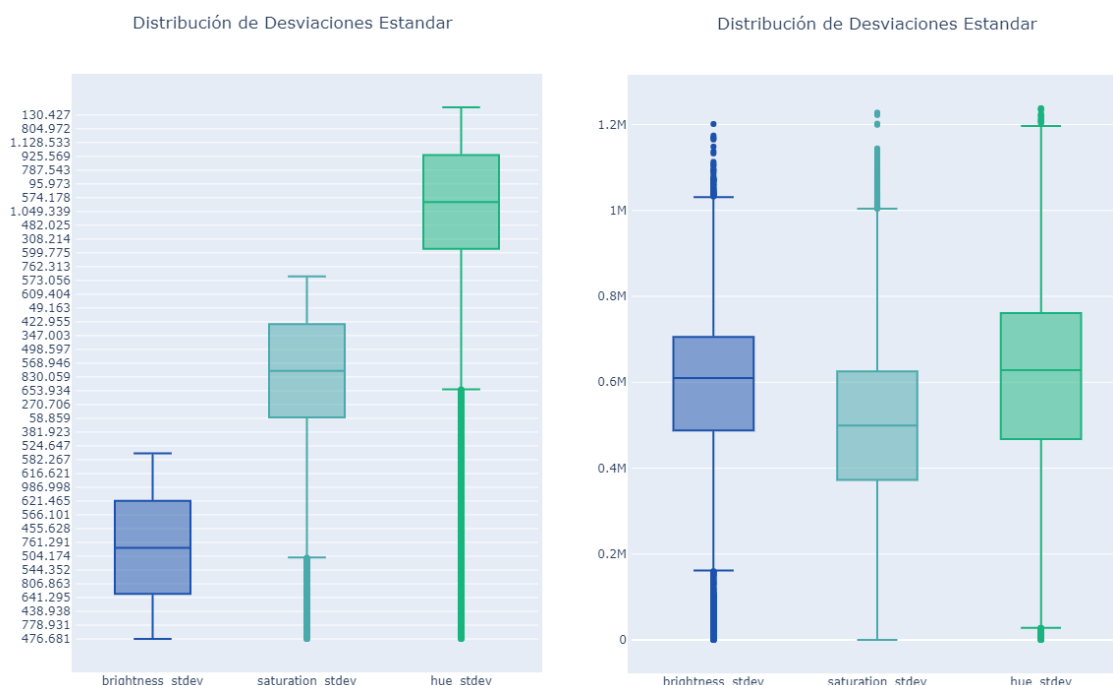


Imagen 6. Distribución de desviación estándar. Izq: Anterior / Der: Posterior

## 4.2. Distribución de actividad por User

En cuanto a la distribución de usuarios/as según cantidad de imágenes compartidas versus cantidad de posts no se observó una diferencia estadísticamente significativa, más allá del cambio de escala. En cuanto a la distribución en tiempo de los posts se observa que no existe una diferencia significativa entre cantidad de posts y cantidad de imágenes, por lo que hasta el momento se mantiene falsa la hipótesis de que esta separación sería significativa, pero sí siendo de interés investigar en más profundidad el cambio ocurrido posterior al año 2021.

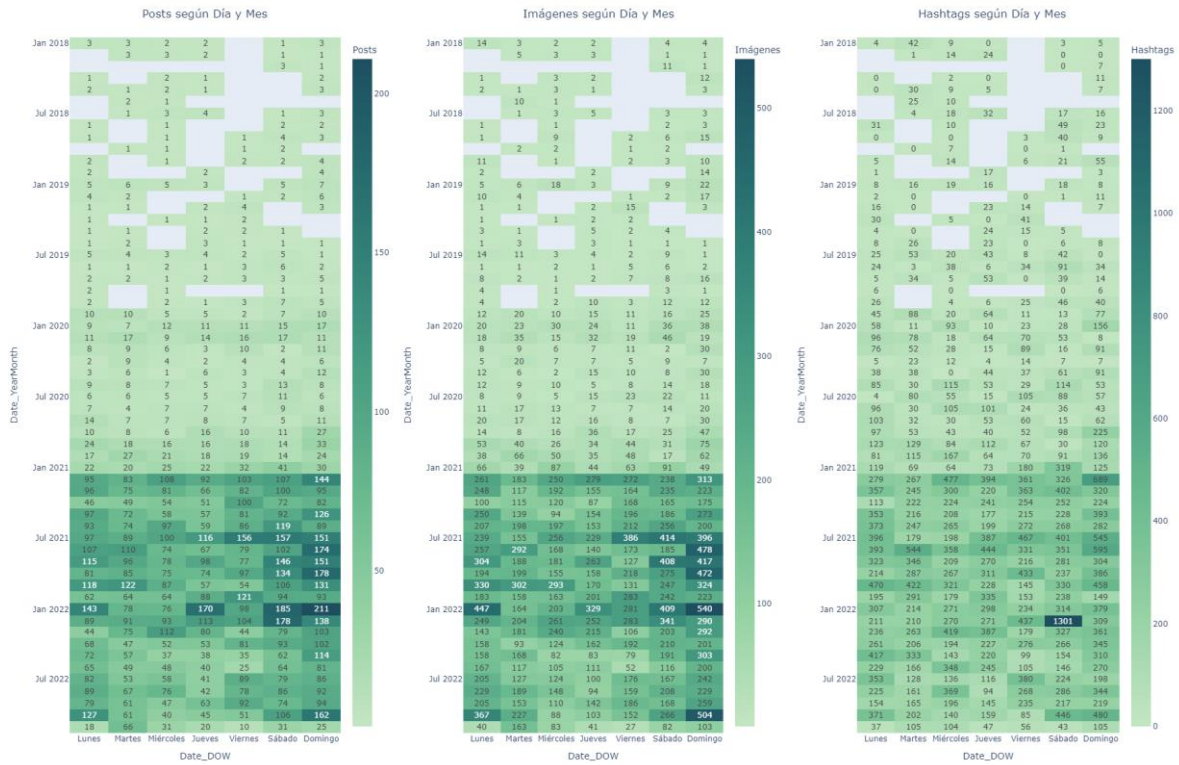
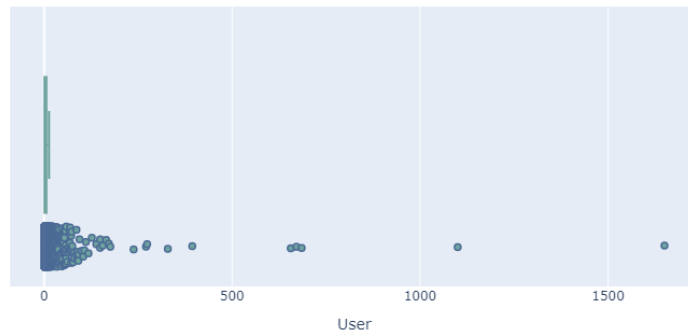


Imagen 7. Conteos según Día y Año-Mes

Distribución de User según cantidad de imágenes subidas



Distribución de User según cantidad de posts

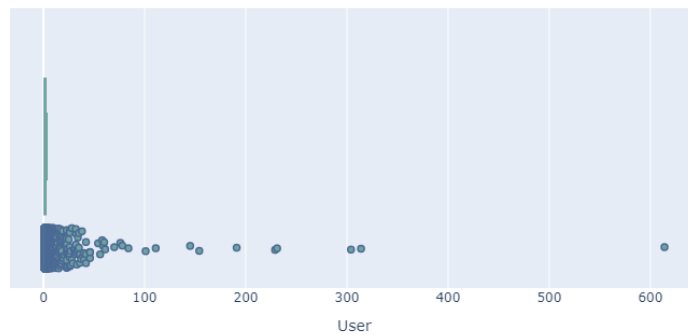


Imagen 8. Distribución de User

### 4.3. Análisis de descripción y hashtag

Correspondería a una futura instancia de desarrollo del trabajo el aplicar los contenidos vistos en el módulo de *Text Mining*, mediante los cuales se buscaría poder identificar patrones y temáticas recurrentes en el contenido escrito en las descripciones y los hashtags ocupados. Es de especial interés para esta investigación el poder jerarquizar dicha información en función del perfil de usuario/a que la compartió, con el fin de que pueda convertirse en un insumo a ser considerado como una etapa previa para la investigación socioterritorial, junto con identificar actores clave a los cuales acercarse en el contexto de las metodologías señaladas.

## 5. Visualizaciones mediante ImageJ

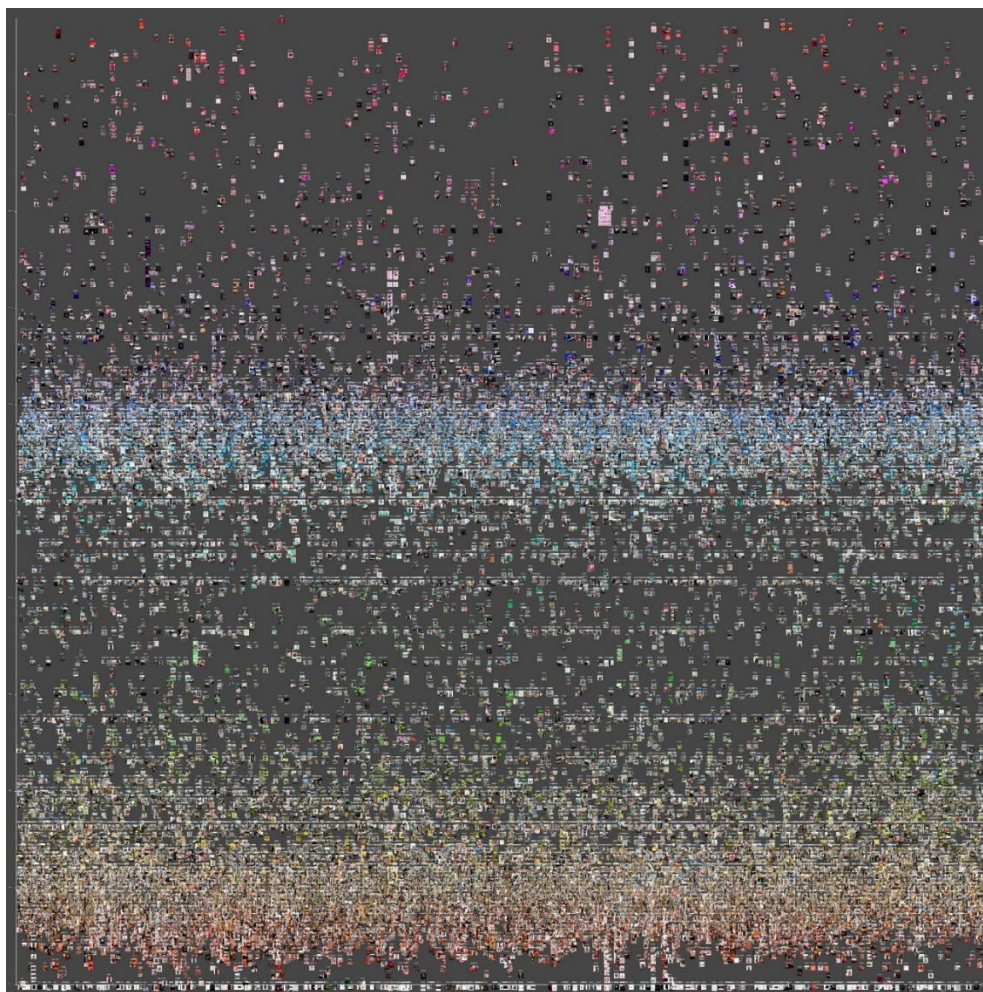


Imagen 9. Visualización de la totalidad de las imágenes obtenidas, ordenadas según orden de subida (horizontalmente) y matiz (verticalmente)

Fuente: Elaboración propia.

## Bibliografía

- Bonnefooy, M. (2020). *Instagram como plataforma de disputa simbólica y territorial : el caso de Plaza de la Dignidad en Santiago de Chile 2019-2020* [Universidad de Chile]. <https://repositorio.uchile.cl/handle/2250/177405>
- Bonnefooy, M., & Chaná, P. (2017, May 23). *Evidencias de un Mirador: Bajo la Alfombra del Smog*. LOFscapes. <http://lofscapes.com/evidencia-de-un-mirador/>
- CIT-UAI. (2018). *Evaluación Técnica de Modificaciones a la División Política Administrativa Nacional: Pre-factibilidad y Factibilidad*.
- Manovich, L., & Hochman, N. (2013). Zooming into an Instagram City: Reading the local through social media. *First Monday*, 18(7). <https://doi.org/https://doi.org/10.5210/fm.v18i7.4711>
- Manovich, L., Tifentale, A., Yazdani, M., & Chow, J. (2015). The exceptional and the everyday: 144 Hours in Kiev. *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, 72–79. <https://doi.org/10.1109/BigData.2014.7004456>
- PUC-Instituto de Geografía. (2003). *Diagnóstico y propuesta metodológica para modificar la división político administrativa del país*.
- Ramírez, N. (2019, May 10). El sueño de ser comuna: 64 barrios o localidades han ingresado una solicitud para convertirse en una de ellas. *Emol*.