Pontificia Universidad Católica de Valparaíso
Escuela de Ingeniería Industrial

Data Science Aplicada
Entregable 5

Informe Final:
Polytics.

Integrantes
Andrés Arenas Rodríguez.
Indy Navarro Vidal.
Juan Saavedra Jeria.

Equipo
Mauricio Huerta Aguiar
Pablo Zúñiga Carvajal
Víctor Leiva Sánchez

Índice

1. INTRODUCCIÓN 3

2. METODOLOGÍAS UTILIZADAS 4

2.1 Extract 5
2.1.1 Modificación al archivo Regiones-Comunas 6
2.1.2 Modificación archivo candidatos 7

2.2 Transform 9
2.2.1 ETLpolitics_solo_id 9

2.3 Load 10

3. CRITERIOS DE GESTIÓN DE FILAS Y COLUMNAS 11

4. DISEÑO DE *DATASETS* 13

5. ANÁLISIS EXPLORATORIO DE DATOS 14

6. METODOLOGÍA APLICADA 30

6.1 Metodología *K-means* 31
Definiciones 31

7. RESULTADOS 34

8. CONCLUSIONES 43

Interpretación de resultados 42
Propuesta de gestión 43

9. BIBLIOGRAFÍA 44

1. Introducción

En los últimos años, ha sido de vital relevancia el estudio y generación de encuestas para conocer con un cierto nivel de certidumbre las tendencias políticas arrojadas en las elecciones presidenciales, como es de común conocimiento, la conectividad, la proliferación de nuevas tecnologías, la construcción de un nuevo paradigma donde los datos y la información son activos importantes para cualquier organización. Dado esto, se ha generado una especial atención a las encuestas que logran predecir de manera más acertada los resultados de las elecciones oficiales, a tal punto, que la difusión de los resultados de las encuestas, tienen impactos económicos y sociales importantes.

Si bien no es solamente uno el instrumento aplicado cuando se avecinan épocas de elecciones presidenciales, entre los más reconocidos se encuentran las encuestas CEP correspondiente al centro de estudios públicos y la CADEM, siendo este último ampliamente criticado en el último proceso de elecciones presidenciales que enfrentó nuestro país el año pasado. Lo anterior dejó en jaque y además puso en duda la metodología de aplicación de uno de los instrumentos de medición más importantes y de mayor impacto las expectativas de estos procesos fundamentales para la definición política de un país.

El error principal observado en la encuesta y que generó una ola de críticas desde diversos medios de comunicación, se encontraba en la aplicación de la metodología de realización del instrumento, específicamente en el criterio de muestreo seleccionado, en lo cual CIPER declara lo siguiente: *"Nuestra estimación difiere de la que realizó CADEM en algunas comunas importantes. Por ejemplo, en Maipú y Puente Alto, cuya población estimada es de 410.650 y 447.615 respectivamente, CADEM tiene muestras de 50 y 45 personas, respectivamente. Si se respetara el peso relativo de cada comuna en la población nacional—o regional— el número de entrevistados debió ser 53 y 48, respectivamente"* (CIPER, ¿En qué se equivocó la encuesta CADEM?, 2017).

Dada la situación anterior, y la importancia que poseen las encuestas en la política actual el objetivo de nuestro trabajo buscará principalmente contribuir a reflejar de manera adecuada las preferencias de la gente a partir de una fuente totalmente anónima y *confiable* como son los resultados de las escrutaciones del proceso electoral mismo vivido en los últimos años dado la inserción del sistema de sufragio voluntario (año 2012), y de esta forma generar un nuevo enfoque para la visualización del comportamiento de la población con respecto a las últimas elecciones presidenciales realizadas, y así poder comprender la relación que tiene la Región Metropolitana con respecto al resto del país.

2. Metodologías Utilizadas

En el presente resumen ejecutivo se pretende señalar las metodologías utilizadas para la construcción del ETL, para esto es importante definir el alcance y la delimitación del problema a través de la pregunta de investigación. Luego fue determinante enfocarse en el objetivo del análisis, planteando la siguiente interrogante:

***¿Es Santiago representativo de las tendencias políticas de todo el territorio nacional*?**

Entonces, en función de la pregunta antes mencionada, el equipo decidió trabajar únicamente con aquellos datos correspondientes a las 2 últimas elecciones presidenciales nacionales, realizadas el año 2013 y 2017 respectivamente, considerando primera y segunda vuelta realizada en ambos años. Lo anterior, se justifica dado que son el único tipo de elecciones donde los candidatos participantes son los mismos a lo largo del territorio, característica que no se cumple en las elecciones municipales, diputado y senadores. Esto permite una simplificación en el análisis, al eliminar la gran cantidad de variables que puede estar sujeto cada candidato en cada distrito electoral.

Además de lo anterior, no se consideraron los partidos políticos por los que militaba cada candidato siendo este poco influyente principalmente por dos razones:

1. La cantidad de candidatos postulantes al cargo presidencial en ambos casos no supera las 10 personas (8 específicamente en las últimas elecciones), por ende, esta información es posible tratarla de forma manual.
2. Se considera que el espectro político, de manera subyacente posee una clasificación cualitativa, que si bien existen 2 extremos notorios como la derecha y la izquierda existe un sector *"centro"* el cual no posee limites definidos.

A continuación, se especifica de manera detallada el proceso de generación del ETL a través de sus tres etapas: Extract, Transform y Load.

2.1 Extract

Para la generación de nuestro ETL, se utilizaron 3 tipos de fuentes de datos explicadas a continuación:

Elecciones: Son 4 archivos de extensión *.csv* obtenidos de datachile, que contienen los datos de las elecciones correspondientes a los periodos 2013 y 2017, tanto primera como segunda vuelta. Estos datos están agrupados por los votos correspondiente por comuna y por candidatos. Estos archivos son:

-*Presidenciales1-2013.csv*: Datos de las elecciones presidenciales del 2013 correspondiente a la primera vuelta, como análisis exploratorio de datos se verifica que la columna *"electo"*, está en 0 para todos los registros, lo que indica que en esa instancia ningún candidato fue electo.

-*Presidenciales2-2013.csv*: Datos de las elecciones presidenciales del 2013 correspondiente a la segunda vuelta, como análisis exploratorio de datos se verifica que la columna *"electo"*, está en 1 para algunos de los registros, lo que indica que en esa instancia hubo un candidato que fue electo.

-*Presidenciales1-2017.csv*: Datos de las elecciones presidenciales del 2017 correspondiente a la primera vuelta, como análisis exploratorio de datos se verifica que la columna *"electo"*, está en 0 para todos los registros, lo que indica que en esa instancia ningún candidato fue electo.

-*Presidenciales2-2017.csv*: Datos de las elecciones presidenciales del 2017 correspondiente a la segunda vuelta, como análisis exploratorio de datos se verifica que la columna *"electo"*, está en 1 para algunos de los registros, lo que indica que en esa instancia hubo un candidato que fue electo.

Candidatos: Corresponde a un archivo de extensión *.csv,* obtenido de datachile, contiene un listado de 1.772 candidatos. Este archivo contiene los candidatos que participan en la elección del 2013, pero no tiene todos los candidatos de la elección del 2017, por lo que más adelante se explica el procedimiento que se realizó para poder completar este archivo. La estructura de este archivo solo contiene una columna con el nombre y otra con el id del candidato, esta última columna nos permite hacer la relación con los archivos de elecciones.
Como análisis exploratorio de datos, nos damos cuenta que este archivo no tiene caracteres especiales, es decir no hay ni letras con tilde, ni uso de letra *'ñ'*.

Regiones-Comunas: Corresponde a un archivo de extensión .csv, obtenido de datachile, que contiene la información por comuna, indicando el nombre de la misma y la región a la que pertenece. Este archivo es de vital importancia para el desarrollo del proyecto, ya que permite poder asociar las votaciones a la región que corresponde.

Todos estos datos, fueron cargados en DataFrame de R, con la función *'read.csv2'*, utilizando un delimitador de una coma ',' . Todo esto fue corroborado con la función View, para comprobar la correcta carga de los archivos en el entorno de trabajo.

2.1.1 Modificación al archivo Regiones-Comunas

Haciendo un análisis exploratorio de datos, nos damos cuenta que este archivo 'Regiones-Comunas' contiene caracteres especiales, tales como tildes y el uso de la ñ, considerando también como situaciones especiales el uso de mayúsculas.

Para evitar cualquier posible problema con el manejo de estos caracteres especiales, se decide cambiar estos caracteres por la vocal sin tilde, además de la 'ñ' pasarla a 'n'.

A continuación, se detalle la modificación que se hace a la tabla de 'Regiones-Comunas', con la finalidad de evitar cualquier problema producido por los caracteres especiales:

1	region_id,region_name,comuna_datachile_id,comuna_customs_id,comuna_tax_office_id,comuna_name
2	1,TarapacÃ¡,226,1401,1204,Pozo Almonte
3	1,TarapacÃ¡,217,1405,1203,Pica
4	1,TarapacÃ¡,113,1101,1201,Iquique
5	1,TarapacÃ¡,108,1404,1206,Huara
6	1,TarapacÃ¡,58,1403,1210,Colchane
7	1,TarapacÃ¡,26,1402,1208,CamiÃ±a
8	1,TarapacÃ¡,5,1107,1211,Alto Hospicio
9	2,Antofagasta,321,2301,2101,Tocopilla
10	2,Antofagasta,313,2104,2202,Taltal
11	2,Antofagasta,309,2103,2206,Sierra Gorda
12	2,Antofagasta,297,2203,2303,San Pedro de Atacama
13	2,Antofagasta,189,2202,2302,Ollague
14	2,Antofagasta,170,2102,2203,Mejillones
15	2,Antofagasta,165,2302,2103,MarÃ-a Elena
16	2,Antofagasta,19,2201,2301,Calama
17	2,Antofagasta,9,2101,2201,Antofagasta
18	3,Atacama,330,3301,3301,Vallenar
19	3,Atacama,317,3103,3203,Tierra Amarilla
20	3,Atacama,109,3304,3303,Huasco
21	3,Atacama,93,3303,3302,Freirina
22	3,Atacama,81,3202,3102,Diego de Almagro

Figura 2.1.1.1- archivo original de Regiones-Comunas

Se hace el cambio con la función de Excel "reemplazar", con el siguiente criterio:

Carácter Original	Corresponde	Cambio
Ã¡	á	a
Ã±	ñ	n
Ã-	í	i
Ã³	ó	o
Ã©	é	e
Ã¼	ü	u
Ãº	ú	U
Ã	Á	A
A‘	Ñ	N

Figura 2.1.1.2- corresponde al criterio para cambiar los caracteres especiales.

Al hacer el análisis exploratorio de datos, solo se encontraron esos caracteres especiales, los cuales fueron reemplazos para evitar problemas futuros en el tratamiento de los datos.

2.1.2 Modificación archivo candidatos

Al empezar a trabajar con los datos, se nota que, para la elección del 2017, existen los candidatos con id: 2550, 2551, 2552. Los cuales no existen en el archivo candidatos, ya que solo llega hasta el id 1779. Esto se puede explicar como que el archivo candidatos solo contiene los candidatos de la elección del 2013, pero no así la del 2017, y existen tres nuevos candidatos que no participaron en la elección del 2013.

Es por esto que existe el desafío de investigar cual es el nombre del candidato 2550, 2551 y 2552. Tras probar distintas técnicas en R para poder averiguar los nombres, se optó por usar SQL, ya que tiene la función agrupar por algún criterio. A continuación, se detalla el proceso.

Se carga el archivo .csv de '*Presidenciales1-2017'* como una base de datos, con el programa SQL-Front, utilizando el delimitador de '**,**'.

comuna_datachi...	candidato...	votos_candid...	electo	partido_id	year	election_id
1	**3**	247	0	5	2017	1
1	**8**	36	0	0	2017	1
1	**9**	24	0	0	2017	1
1	**147**	338	0	8	2017	1
1	**535**	12	0	30	2017	1
1	**561**	405	0	9	2017	1
1	**567**	1170	0	8	2017	1
1	**2550**	2974	0	7	2017	1
1	**2551**	1099	0	8	2017	1
1	**2552**	24	0	28	2017	1

Figura 2.1.2.1- datos cargados como base de datos

Con la siguiente query, se obtienen las cantidades totales de votación por cada candidato

```
select sum(votos_candidato) as votos,candidato_id from votaciones group by candidato_id order by votos
```

A continuación, se muestra el resultado obtenido

votos	candidato_id
23880	535
33468	2552
39315	9
64859	8
375762	3
386394	561
521983	147
1331237	2551
1490549	567
2409993	2550

Figura 2.1.2.2- resultado de la query

Esta información la contrastamos con los candidatos que ya se conocen el nombre según su *id*. Por lo tanto, es posible construir la siguiente tabla:

Votos	candidato_id	Porcentaje	Nombre Candidato
23.880	535	0,36	Alejandro Navarro Brain
33.468	2552	0,50	
39.315	9	0,59	Votos En Blanco
64.859	8	0,97	Votos Nulos
375.762	3	5,63	Marco Enriquez-Ominami Gumucio
386.394	561	5,79	Carolina Goic Boroevic
521.983	147	7,82	Jose Antonio Kast Rist
1.331.237	2551	19,94	
1.490.549	567	22,32	Alejandro Guillier Alvarez
2.409.993	2550	36,09	
Total Votos	6.677.440		

Figura 2.1.2.3- Porcentaje de votos por candidato

Para poder obtener el nombre de los candidatos 2550, 2551, 2552 se procede a comparar los porcentajes de las votaciones, con los datos oficiales de la elección. Con lo que es posible asociar los *id's* de candidatos a los nombres. De esta forma se concluye lo siguiente para llenar la tabla:

- El candidato 2550 es Sebastian Pinera Echenique
- El candidato 2551 es Beatriz Sanchez Munoz
- El candidato 2552 es Eduardo Artes Brichetti

Votos	candidato_id	Porcentaje	Nombre Candidato
23.880	535	0,36	Alejandro Navarro Brain
33.468	2552	0,50	Eduardo Artes Brichetti
39.315	9	0,59	Votos En Blanco
64.859	8	0,97	Votos Nulos
375.762	3	5,63	Marco Enriquez-Ominami Gumucio
386.394	561	5,79	Carolina Goic Boroevic
521.983	147	7,82	Jose Antonio Kast Rist
1.331.237	2551	19,94	Beatriz Sanchez Munoz
1.490.549	567	22,32	Alejandro Guillier Alvarez
2.409.993	2550	36,09	Sebastian Pinera Echenique
Total Votos	6.677.440		

Figura 2.1.2.4- Tabla completa de los candidatos por nombre e id.

La información de estos tres candidatos se agrega de forma manual al archivo *candidatos.csv*

1771	1777,Jose Nunez Gonzalez
1772	1778,Daniel Esteban Godoy Mendez
1773	1779,Maria Paz Espinoza Carvajal
1774	2550, Sebastian Pinera Echenique
1775	2551,Beatriz Sanchez Munoz
1776	2552,Eduardo Artes Brichetti

Figura 2.1.2.5- Final del archivo *"candidatos.csv"*, contiene los últimos 3 candidatos que requerimos para el proyecto.

2.2 Transform

Con los 6 archivos cargados correctamente como DataFrame en la memoria de R, se procede a transformar los datos para poder generar el archivo ETL.

Lo primero es unir las 4 elecciones presidenciales a través de operaciones fila. Se utiliza la función **rbind**, que permite agregar al DataFrame *'Presidenciales'* al final del mismo.

```
22  Presidenciales = rbind(datos2013vuelta1, datos2013vuelta2)
23  Presidenciales = rbind(Presidenciales, datos2017vuelta1)
24  Presidenciales = rbind(Presidenciales, datos2017vuelta2)
25
```

Figura 2.2.2 -Operaciones fila de las elecciones

Se elimina la columna *partido_id*, ya que no es parte de nuestro análisis, se utiliza la función **subset** para este propósito.

A continuación, se agrega los nombres de los candidatos como columnas utilizando la función **merge**, la cual realiza la unión por columna de los DataFrame, utilizando la columna que funciona como índice que tiene el mismo nombre para ambos DataFrame: *candidato_id*.

```
Presidenciales = merge(Presidenciales,candidatos)
```

Figura 2.2.3 - Función merge que permite unir por columna elecciones con candidatos.

Para agregar las columnas correspondientes a la región y comuna, se procede con la función **merge**, que utiliza como índice el nombre de columna.

Finalmente se eliminan las columnas que no aportan valor para nuestro análisis, que corresponden a comuna*_customs_id* y *comuna_tax_office_id*, con ayuda de la función **subset**.

2.2.1 ETLpolitics_solo_id

Dado que el ETL que se menciona anteriormente tiene como campo *String* las columnas: candidato, región_name y comuna_name, lo que nos facilita el análisis a la hora de identificar cada candidato y región. Pero esto también nos puede traer inconvenientes a la hora de procesar el archivo, ya que manejar los nombres que se

van repitiendo en los distintos registros, aumenta el tamaño del archivo, y con ello un posible problema de tiempo de proceso, o algún problema con los nombres que hay, que luego se comprobará que no existe mayor incidencia.

Es por esto que, ante un eventual problema, se prepara otro ETL que aporta la misma información, pero solo incorporando *id's* relevantes, pero no contiene los campos que corresponde a los *String* de nombres, con lo que se espera mejorar el tiempo de proceso.

Este archivo será usado solo en caso de los problemas antes descritos, ya que nuestro ETL original y con el que trabajaremos contiene los nombres en String.

2.3 Load

Con el DataFrame '*Presidenciales'* ya completado, se genera el archivo físico en el directorio de trabajo. Se utiliza la función **'write.cvs2'** que utiliza en forma nativa el separador ";".

```
write.csv2(x=Presidenciales, file="ETLpolitics.csv")
```

Figura 2.3.1 generación del ETL 'ETLpolitics.cvs'

Por último, se sube el archivo del ETL al DataWarehouse.

.

3. Criterios de Gestión de Filas y Columnas

Para establecer los criterios para gestionar las filas y columnas del ETL generado, se debe considerar en todo momento el objetivo del proyecto y que variables serán relevantes para obtenerlo. Dado lo anterior es que para facilitar el trabajo se generó solamente un archivo consolidado que corresponde a todos los resultados de las votaciones presidenciales, ayudando de esta forma a realizar el análisis de manera global con la finalidad de lograr una mayor representatividad de los datos y en el trabajo de los mismos.

Además de lo anterior, se considera que cada fila representa el total de los votos obtenidos de cada uno de los candidatos para una comuna/ciudad en específico, clasificado por año y por tipo de elección (si es de primera o segunda vuelta), también se considera dentro de las filas las posibilidades de que se encuentren votos nulos y blancos (de manera excluyente) para cada comuna, e incluyendo las 4 elecciones presidenciales realizadas, que son las consideradas debido a que en el año 2012 comenzó a regir el voto voluntario en nuestro país.

En el criterio de columnas se consideró los siguientes campos finalmente: id_comuna, comuna, candidato_id, nombre del candidato, tipo de elección, año, cantidad de votos, electo. Posteriormente, se removería el campo partido_id, que relacionaba el partido político del candidato en cuestión, esto debido que frente a cualquier tipo de análisis podía ser relacionado directamente al candidato, además, considerando que en las elecciones presidenciales son muchos menos candidatos, la variable pierde importancia en el análisis, por lo que se remueve por simplicidad.

A continuación, se presenta un esquema de la organización que va a tener los datos. A través de operaciones filas se agregarán los datos de las votaciones presidenciales. Y con operación columna se agrega los datos de los candidatos y los datos de las regiones-comunas.

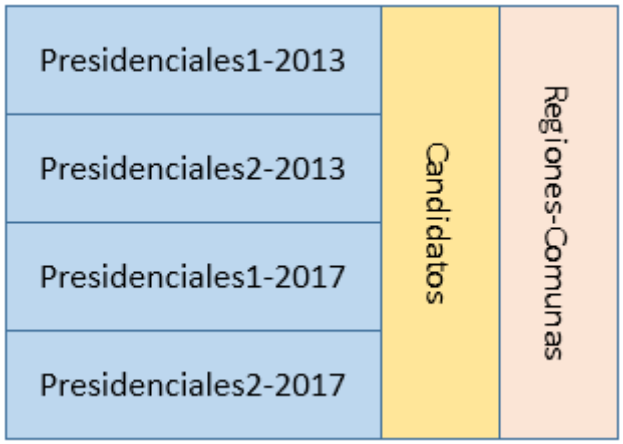

Figura 3.1- Diagrama de la estructura del ETL

En función de lo descrito en el punto anterior, a continuación, se especificará los pasos realizados para generar el archivo correspondiente a nuestro ETL:

1. El primer paso para generar el ETL consistió en unir todos los resultados* de las elecciones presidenciales correspondientes al año 2013 y 2017 (tanto primera como segunda vuelta), cabe mencionar que cada columna de nuestra base de datos representa la cantidad de votos obtenidos por un candidato en una comuna en específico. De esta forma se unieron los 4 archivos de elecciones disponibles con operaciones de agregar fila.
2. Se procedió a eliminar la columna llamada *partido_id* correspondiente al código identificador de cada partido.

3. Luego se procedió a incorporar la lista de *candidatos* con sus nombres correspondientes, en este punto fue requerido completar la base de datos original correspondiente a los nombres de los candidatos dado que esta se encontraba incompleta, generando una pérdida aproximada de 2000 resultados diferentes, específicamente de 3 candidatos correspondientes a las elecciones del año 2017 y que en la tabla de resultados adoptaban la siguiente nomenclatura:
    - 2550, Sebastian Pinera Echenique.
    - 2551, Beatriz Sanchez Munoz.
    - 2552, Eduardo Artes Brichetti.
4. Una vez corregida la inconsistencia anterior, se procedió a unir el archivo generado con los resultados presidenciales con los nombres de aquellas comunas y regiones correspondientes. Esta operación se lleva a cabo con la función Merge que permite añadir por columna de acuerdo al id correspondiente.
5. Posteriormente, antes de generar el archivo final se eliminaron las columnas *comuna_custom_*id y *comuna_tax_office_id* correspondientes a información de las comunas y regiones, estas se eliminaron dado que solamente era necesario identificar las comunas y su ubicación.
6. Finalmente se genera el archivo final correspondiente al ETL.

Como criterio de calidad, se decidió generar 2 ETL, en el cual uno además de manejar *id's* considera los *nombres* de candidatos como de comunas para su mayor comprensión, dado que es el que se espera trabajar, y también una segunda versión indexada solamente con *id's,* que en caso de que los nombres complejicen el tratamiento de los datos el ETL anterior será reemplazado por esta versión más sencilla. Finalmente, esta propuesta de dataset, permitirá que generar con algunos procesamientos obtener votos totales por regiones, comunas en cada elección, o totales para cada candidato, variaciones de cada zona en las diversas elecciones, etc.

4. Diseño de *Datasets*

Como resultado final se obtuvo un ETL el cual posee las siguientes columnas:

identificador	Índice del archivo [0 al 10.020]
comuna_datachile_id	Id de la comuna que corresponde los votos
candidato_id	Id del candidato que corresponde los votos
votos_candidato	Votos del candidato
electo	[0=no electo; 1=electo]
year	Año de la elección
election_id	[1=primera vueta;2=segunda vuelta]
candidato	Nombre del candidato
region_id	Id de la región de la comuna
region_name	Nombre de la región
comuna_name	Nombre de la comuna

Tabla 4.1 – Atributos ETL generado.

En esta versión del ETL los campos correspondientes a *comuna, región y candidato* se manejan tanto por su *nombre* como por su *id* numérico, esto con la finalidad de poder identificar y representar los datos fácilmente. Lo anterior se realizó como una medida de seguridad ante la posibilidad de enfrentar inconvenientes con la velocidad de procesamiento del archivo, eventualmente será eliminada aquella fila que no será utilizada.

Como resultado, el archivo correspondiente al ETL posee 11 columnas y 10.021 filas (incluyendo el encabezado), en la cual cada una de las filas corresponde a los votos totales obtenidos por un candidato (en esta variable también se identifican además de los candidatos votos nulos y blancos) en una comuna/ciudad para un tipo de elección (primera o segunda vuelta) en particular.

Para el archivo ETLpolitics_solo_id, que se menciona en el apartado 2.2.1, se tiene la siguiente estructura:

identificador	Índice del archivo [0 al 10.020]
comuna_datachile_id	Id de la comuna que corresponde los votos
candidato_id	Id del candidato que corresponde los votos
votos_candidato	Votos del candidato
electo	[0=no electo; 1=electo]
year	Año de la elección
election_id	[1=primera vueta;2=segunda vuelta]
region_id	Id de la región de la comuna

Tabla 4.2 – Atributos ETL generado solo id.

5. Análisis Exploratorio de Datos

En la presente sección se indican las metodologías utilizadas para la realización del *análisis exploratorio de datos* (AED). En un principio, una característica del ETL elaborado, la cual, además es determinante para el desarrollo de este análisis, corresponde a que la única variable *cuantitativa* dentro del set corresponde a la columna *votos_candidato* la cual contiene los votos obtenidos por cada candidato.

Dada las características mencionadas para el ETL, fue que en una primera instancia y antes de comenzar el análisis exploratorio se confirmó la consistencia de los datos, en este caso se realizó por medio del uso de la librería *DataExplorer* mediante el software R y el comando *introduce*, como resultado se obtuvo que ningún valor se encontraba vacío.

Además, cabe destacar que el desarrollo de este AED se centró principalmente en una variable correspondiente al ETL realizado, esta fue la correspondiente a la columna de los *votos_candidato* dado que es la única columna presente correspondiente a una variable numérica y cuantitativa.

En primera instancia, se procedió a graficar el comportamiento de los datos emitidos por los votantes para ambos procesos de elecciones y sus respectivas vueltas. Para este análisis se realizaron dos tipos de gráficos: Boxplot e histograma. Para el primer caso el Boxplot presentado para cada una de las situaciones (definida por elección y vuelta) se observó un comportamiento similar en los tipos de vuelta similares, ya que ya que en ambos casos de primera vuelta las medias correspondientes a las cantidades de votos obtenidas por los candidatos fueron menores que en el caso de la segunda vuelta, situación que se podría esperar dado que la cantidad de candidatos en una segunda vuelta es mucho menor que para el caso de la primera y por ende es de esperar que aquellos que pasen a segunda vuelta concentren una mayor cantidad de votos.

Además, a razón de lo anterior, ocurre que en el caso de la segunda vuelta la mediana se encuentra más alejada del tercer cuartil dado que al ser una cantidad menor de candidatos estos concentran de manera individual una mayor cantidad de datos, como se puede apreciar en la zona roja del grafico presentado a continuación:

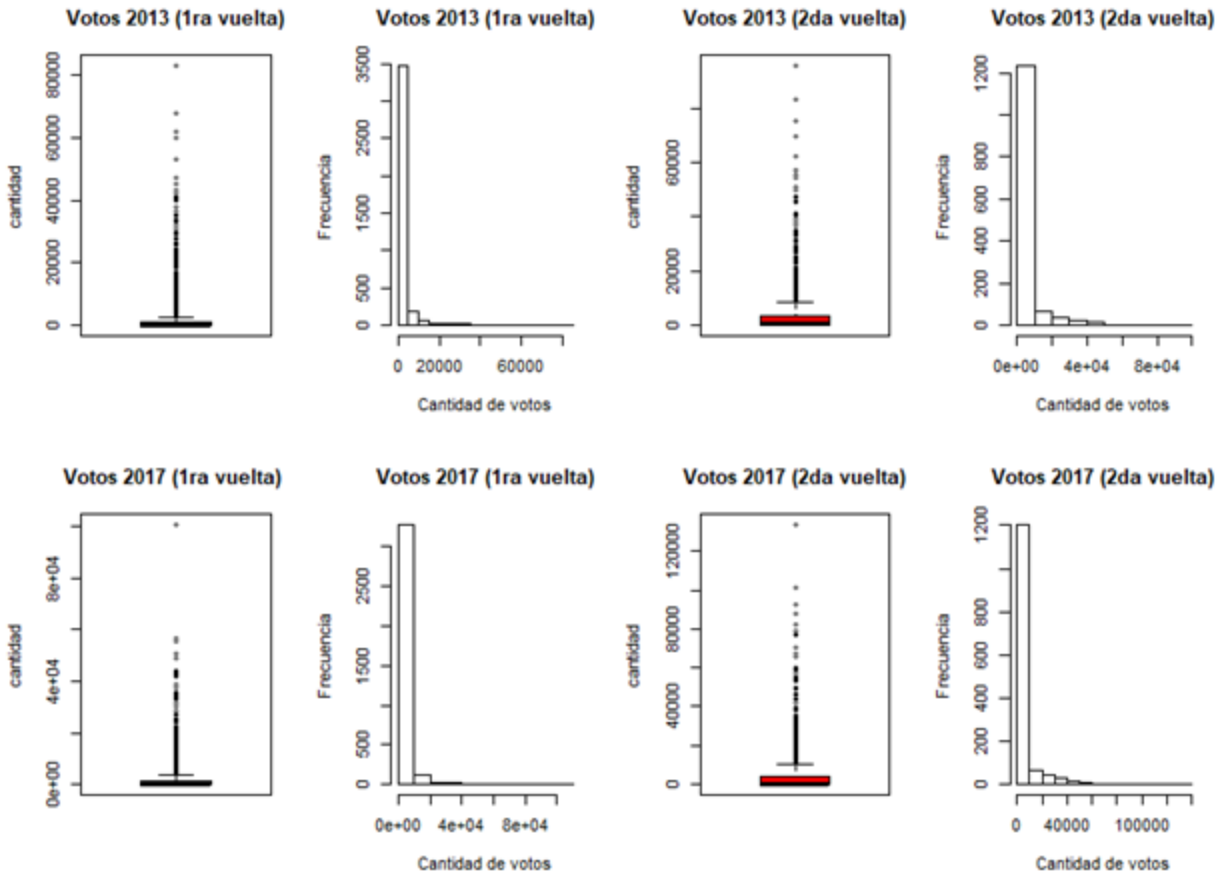

Figura 5.1- Boxplot e Histograma correspondiente a cada vuelta y su elección respectiva

En el caso de los histogramas presentados anteriormente, el sesgo positivo que presentan hacia la derecha se explican principalmente por la estructura del proceso de elecciones dado que este proceso es altamente descentralizado y coordinado, las personas para realizar su voto se acercan al lugar de votación asignado para ellos, esto se traduce que, existen muchas localidades y comunas con cifras de hasta 10.000 votantes.

En una segunda instancia, dada las características de nuestro ETL, se realizó un AED a través de la herramienta de Excel con la finalidad de contrastar nuestra principal variable cuantitativa con el resto de columnas y variables cualitativas. Dentro de este análisis se clasificaron las regiones Top 5 regiones que contaron con una mayor participación y donde es la Región Metropolitana la cual concentra la mayor cantidad de votantes, como se muestra a continuación:

Figura 5.2- Top 5 de las regiones con mayor cantidad de votantes.

Además, se realizó una clasificación similar para el caso de las comunas, en esta instancia se seleccionaron el top 10 de aquellas que concentraron una mayor cantidad de votos, en este caso cabe destacar que una de las comunas que más votos aglomera corresponde a Antofagasta, que, si bien no pertenece a una de las regiones con mayor participación, en su calidad de comuna y capital regional, concentra una participación importante:

Figura 5.3- Top 10 de las comunas con mayor cantidad de votantes.

Para el mismo caso se realizó un análisis detallado para cada una de las elecciones presidenciales por separado, para contrastar específicamente la importancia de las comunas que no pertenecen a Santiago, con ello se encontró que otras comunas y/o ciudades como Concepción también poseen un peso relativo importante respecto la cantidad de votantes:

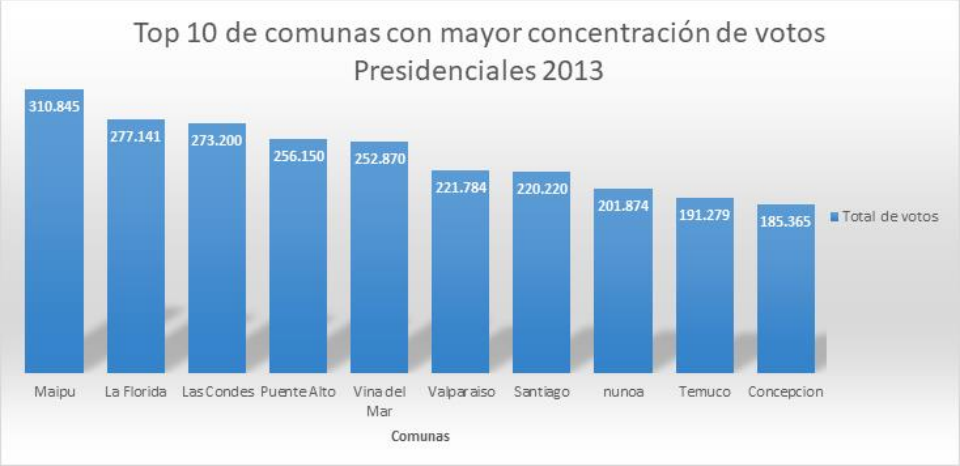

Figura 5.4- Top 10 de comunas que concentran votaciones (Presidenciales 2013)

En una tercera instancia se analiza la cantidad de votos obtenida por cada uno de los candidatos, en este análisis se consideró la cantidad total obtenida por cada una de las elecciones y no se discriminó por cada una de las vueltas de las elecciones correspondientes, cabe destacar que se menciona los 4 candidatos con mayor cantidad de votos, el resto de los candidatos (incluyendo votos nulos y blancos) que consideran cantidades marginales de votos fueron aglomerados dentro de la opción "*Otros*" para ambos gráficos presentados a continuación:

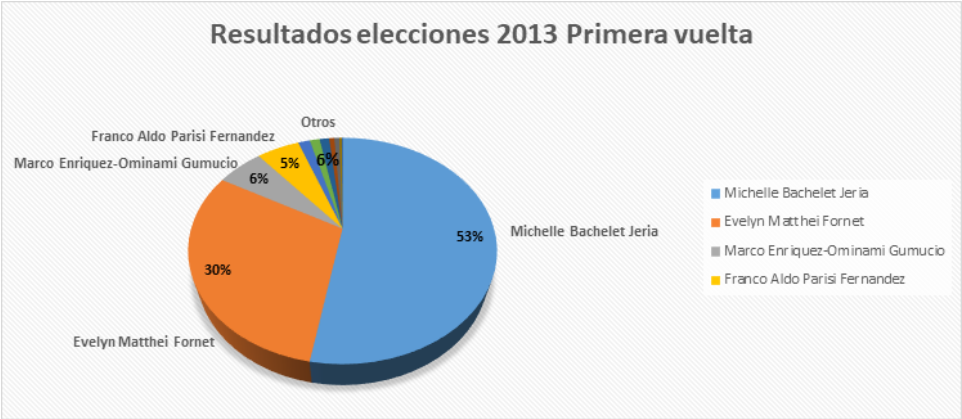

Figura 5.5- Porcentaje de votos obtenidos por candidato (Presidenciales 2013)

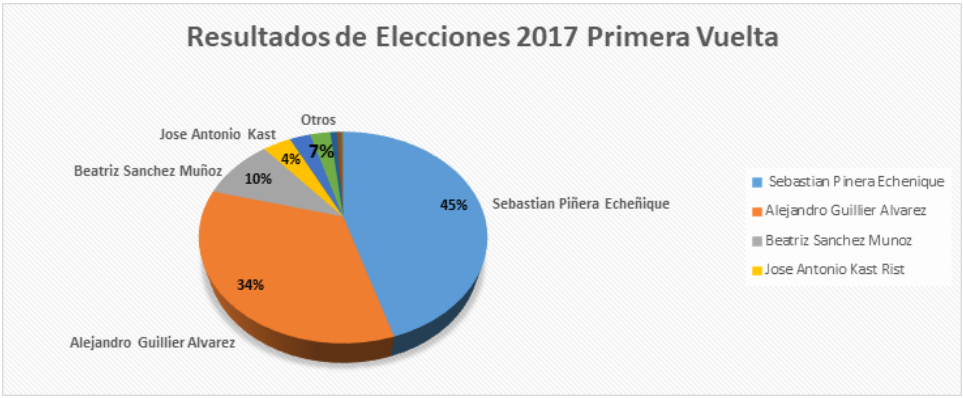

Figura 5.6- Porcentaje de votos obtenidos por candidato (Presidenciales 2017)

Para profundizar en el análisis de las variables es deseable el uso de las correlaciones, con énfasis especial en el coeficiente de correlación de Pearson, sin embargo, como se mencionó anteriormente es que la gran mayoría de las variables son del tipo cualitativas, por lo tanto, se queda la variable *votos_ candidato* como la principal variable numérica. Para lograr una correlación se creó un nuevo dataset que describe cada región junto con su población y número de votantes por cada elección dividiéndose en la primera y segunda vuelta, este nuevo dataset proviene de dos fuentes principales: La primera es el actual dataset en el que se ha desarrollado todo el análisis anterior y por otra parte se obtuvo los datos de la población de acuerdo a la información que provee el INE respecto al CENSO del 2017, quedando en la siguiente tabla:

.

Región	**Población**	**total votos 2013-1° vuelta**	**total votos 2013-2° vuelta**	**total votos 2017-1° vuelta**	**total votos 2017-2° vuelta**
Arica y Parinacota	226.068	71188	55582	74818	71215
Tarapacá	330.558	86117	65299	96330	93327
Antofagasta	607.534	176671	131326	180300	182263
Atacama	286.168	99162	80579	100868	102749
Coquimbo	757.586	259764	223985	257222	266737
Valparaíso	1.815.902	723231	614798	725844	749693
Región Metropolitana Santiago	7.112.808	2658373	2252394	2737395	2850725
O'Higgins	914.555	366505	321995	353873	381798

Maule	1.044.950	438073	386240	410182	445220
Bio Bio	2.037.414	853913	733425	808190	872392
Araucanía	957.224	390034	340178	373599	407860
Los Ríos	384.837	161729	138160	155538	164302
Los Lagos	828.708	313901	271558	304215	324595
Aysén	103.158	37731	32095	38047	38039
Magallanes	166.533	62619	50137	61019	60288

Tabla 5.1- Población total y votos realizados por región

Cada variable representa un número de personas, y al aplicar el coeficiente de correlación se logra los siguientes resultados:

	población	**total votos 2013-1° vuelta**	**total votos 2013-2° vuelta**	**total votos 2017-1° vuelta**	**total votos 2017-2° vuelta**
población	1				
total votos 2013-1° vuelta	0,998535178	1			
total votos 2013-2° vuelta	0,997791244	0,999887641	1		
total votos 2017-1° vuelta	0,999527454	0,999549068	0,999112746	1	
total votos 2017-2° vuelta	0,999153262	0,999863419	0,999613996	0,999870467	1

Tabla 5.2- Correlación entre población y cantidad de votos.

Como se representa en la tabla anterior, descartando la diagonal (entendiéndose que cada variable tiene correlación 1 consigo misma), todas las variables tienen una fuerte correlación positiva con las demás, lo que significa una clara influencia entre estas variables, es decir, si aumenta alguna de estas variables, las demás deberían aumentar en una proporción constante. Esta correlación entre la población y las votaciones están sujetos a un supuesto, dado que se obtuvo datos del INE sobre la población del 2017, se espera que en general no exista una variación significativa de la población en el año 2013.

Finalmente, esta correlación puede ser descrita en un gráfico de la siguiente forma:

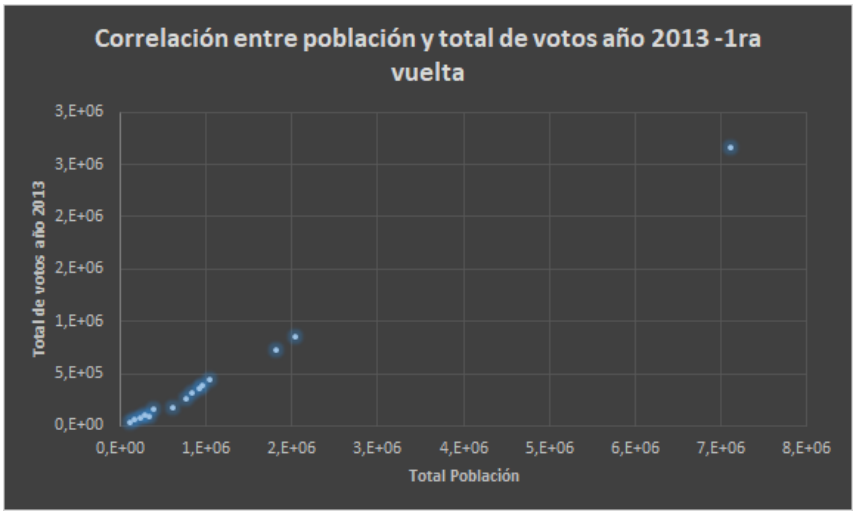

Figura 5.7- Relación entre las poblaciones de cada región y total de votaciones.

Como se observa en el gráfico, la percepción de la correlación entre la población frente a cualquiera de las elecciones es muy fuerte, para cualquier otra elección que desee ser puesta a prueba junto a la población u otras elecciones presidenciales, la generación de gráficas tendrá un comportamiento muy similar, y las tendencias estarán muy marcadas.

Ya dentro de la última recta de nuestro análisis exploratorio de datos, se decide utilizar el software Tableau Public, que permite trabajar gran cantidad de datos de manera cómoda e intuitiva. Mediante este software es posible aplicar distintos filtros y criterios, obteniendo información en forma ordenada y explicativa. Gracias a las distintas herramientas de visualización, es posible combinar distintas dimensiones en un mismo gráfico, permitiendo generar un entendimiento de la situación.

Al igual que en algunos análisis ya vistos en el presente informe, se continúa trabajando el archivo ETL de acuerdo al año de elección y a que vuelta corresponde, para estos efectos se utilizan los filtros que proporciona Tableau.

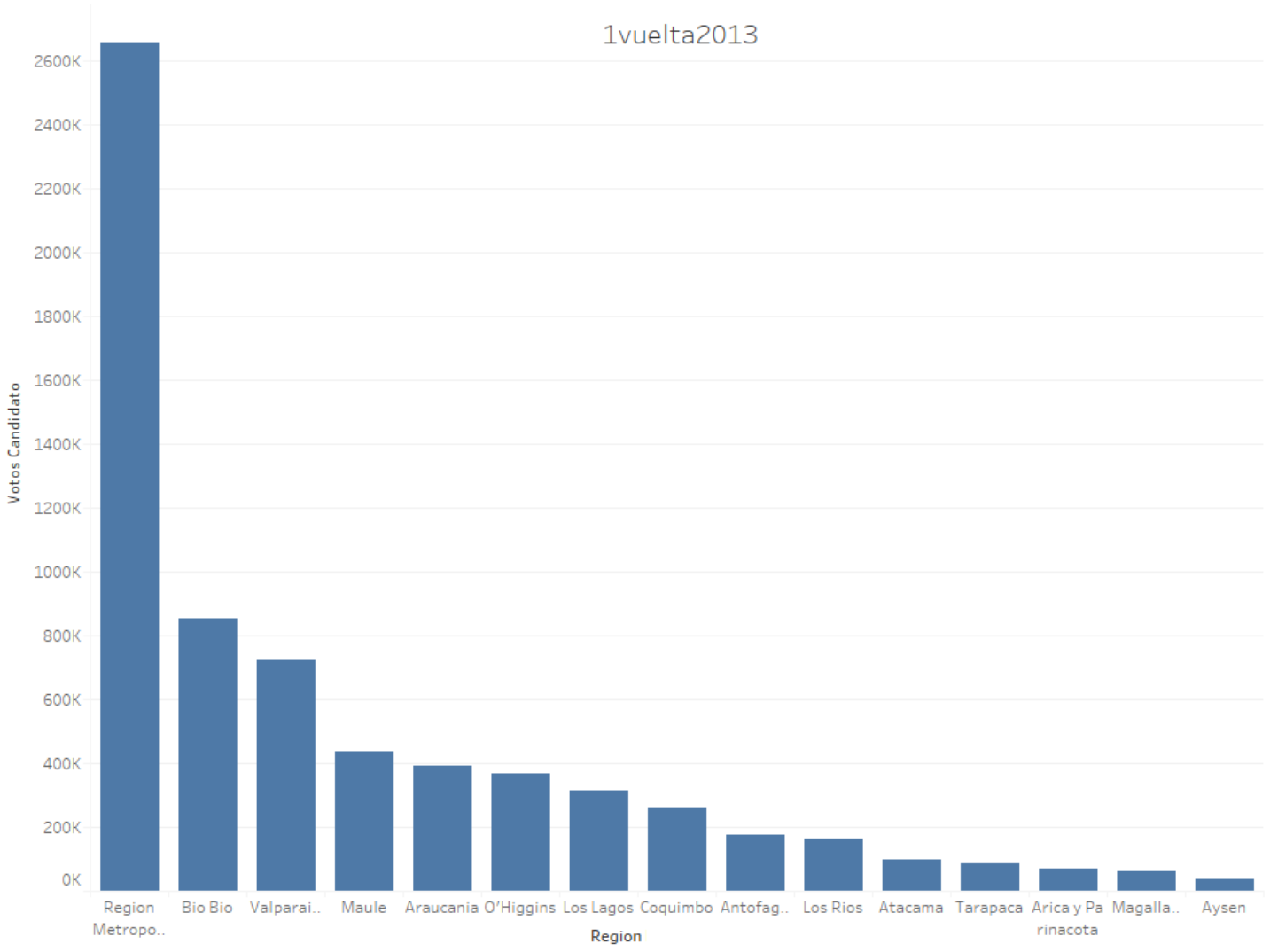

Figura 5.8 - Votos primera vuelta, agrupados por región (Presidenciales 2013).

Se incluye la variable del candidato, para ver cómo se comportan los votos en cada región.

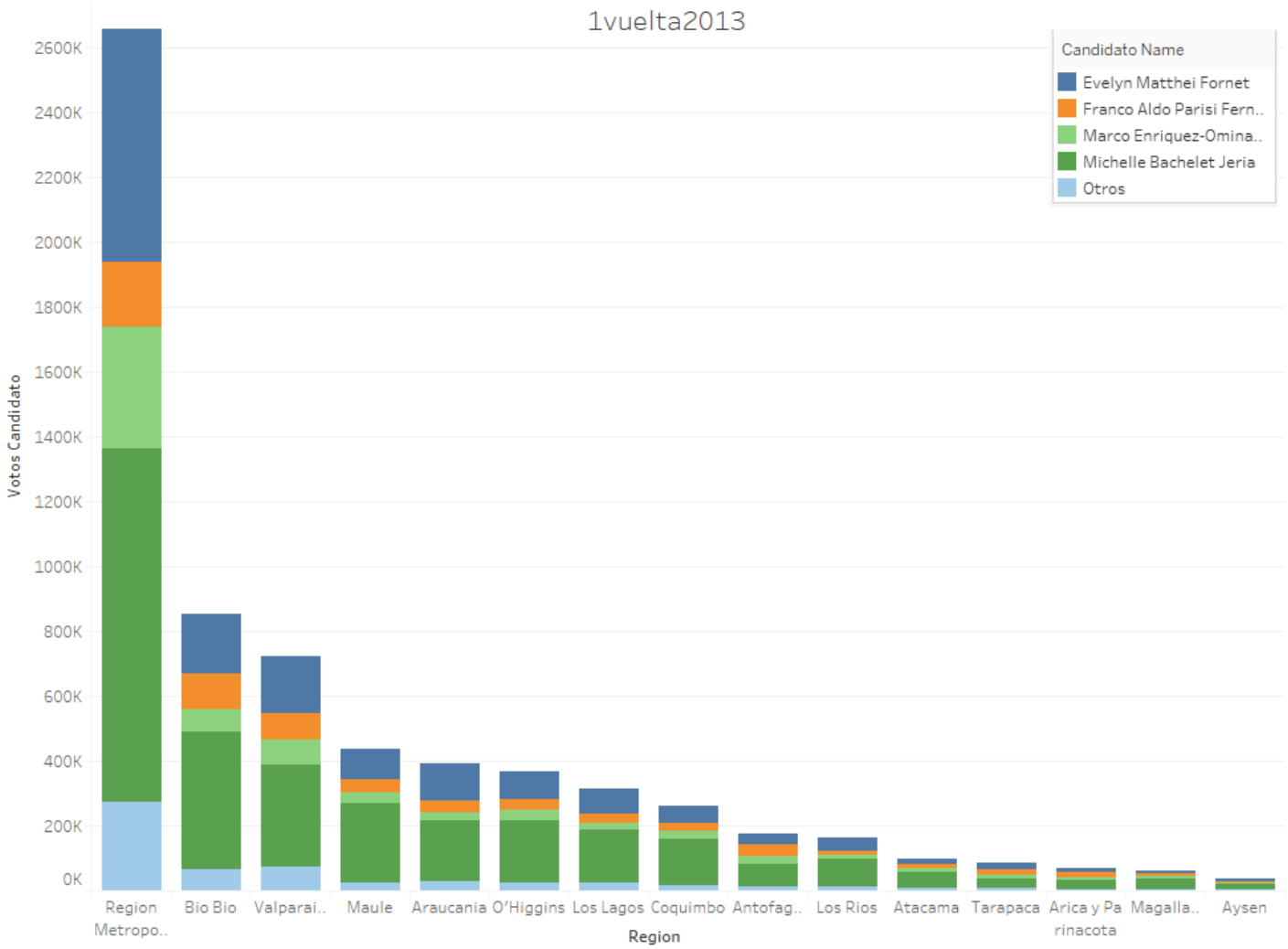

Figura 5.9- Votos primera vuelta, agrupado por región y candidatos (Presidenciales 2013)

Al incluir estas tres dimensiones en el gráfico, es decir Región, Votos y Candidatos, es posible visualizar a la mayoría de los candidatos presentes en todas las regiones, aunque no es posible realizar un análisis más acabado hasta esta parte. Dado que Santiago tiene la mayoría de los votos, tiene la columna más grande en el gráfico, con lo que es difícil comparar si las otras regiones tienen un comportamiento parecido. No obstante, se logra apreciar que Santiago concentra una inmensa cantidad de votos, que para esta elección contiene alrededor del 40% de las votaciones totales, e incursionando sobre el Principio de Pareto, considerando el 20% de las regiones que concentran la mayor cantidad de votos, en este caso Region Metropolitana, Región del Bio-bio y Región de Valparaíso, la concentración de votos alcanza un 63,22% lo que dista del 80% teórico, pero sigue siendo una gran concentración en solo 3 regiones.

Dado que el objetivo del presente proyecto es analizar si Santiago es representativo del territorio nacional, es necesario comparar el comportamiento de Santiago con el resto de las regiones, para lo cual se procederá a analizar los votos por región como porcentaje, para poder comparar comportamiento entre regiones.

Para los siguientes gráficos, se busca comparar el comportamiento porcentual de un candidato por región, para poder tener una base común para comparar, este análisis se hace de manera visual, ya que los gráficos están representados al 100%, de esta forma es posible identificar la proporción de votos que representa los candidatos por regiones. Se hace esta aclaración ya que el grafico tiene los ejes en cantidad de votos, los cuales no son los mismos números en las distintas regiones.

Figura 5.10- Colores de los candidatos presentes en los gráficos de 1° vuelta 2013

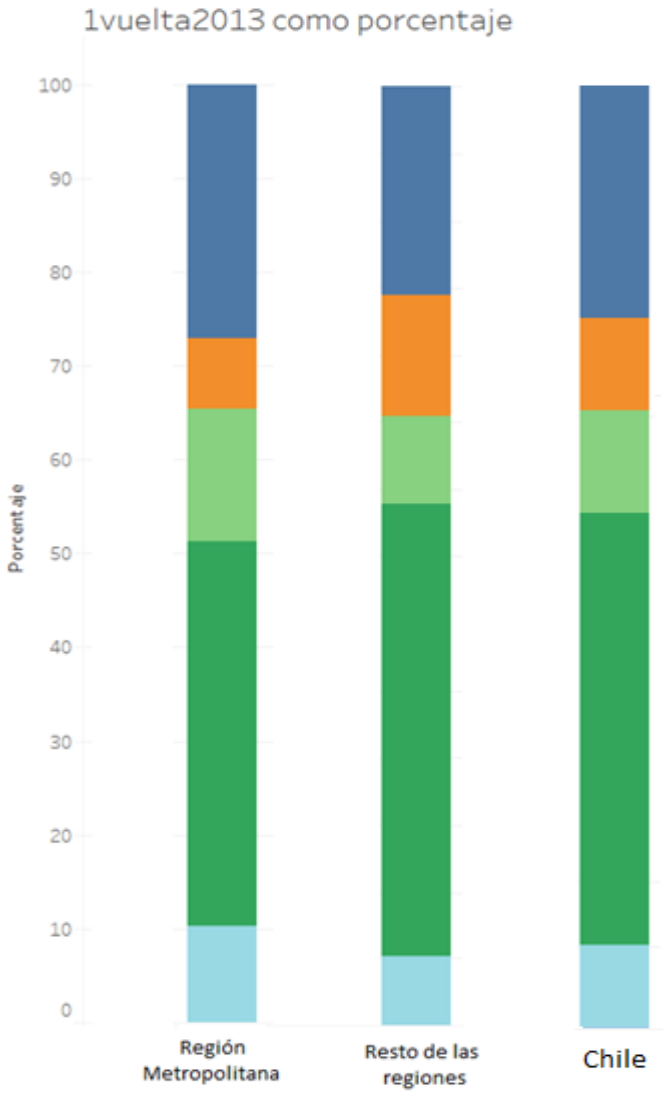

Figura 5.11- Comparación Santiago con resto de las regiones y Chile (1° vuelta 2013)

Al comparar la región Metropolitana con el resto de las regiones de manera proporcional, nos damos cuenta que en Santiago la candidata Evelyn Matthei (color azul oscuro) porcentualmente tiene más adeptos respecto al resto de las regiones, y Michelle Bachelet (color verde oscuro) tiene menos adeptos porcentualmente respecto al resto, si profundizamos en este gráfico y evaluamos mediante los porcentajes tenemos lo siguiente:

	Región Metropolitana	*Otras Regiones*	*Chile*
Evelyn Matthei	27,06%	22,09%	24,61%
Franco Parisi	7,55%	11,52%	9,94$
Marcos Ominami	14,12%	8,62%	10,8%
Michelle Bachelet	40,94%	49,19%	45,91%
Otros	10,33%	8,58%	8,74%

Tabla 5.3 Porcentajes de elecciones entre Región Metropolitana y el resto de Chile

Nos damos cuenta con la tabla y el gráfico anterior, poniendo énfasis en los gap que se generan porcentualmente, la Región Metropolitana con respecto al resto de las regiones, tiene una diferencia importante, es posible apreciar, como este puede llegar a ser casi un 9% (en el caso de Michelle Bachelet), para dimensionar esta diferencia, si las elecciones se hubieran realizado sin considerar la Región Metropolitana, la candidata Michelle Bachelet alcanza casi la mayoría de votos en la primera vuelta. Sin embargo, dado el peso relativo que posee la Región Metropolitana correspondiente a la enorme cantidad de votos que concentra, el gap entre la misma región y todo Chile, se reduce importantemente, llegando hasta un 5% aproximadamente.

A continuación, se compara la región metropolitana con algunas regiones en forma individual (para la primera vuelta 2013). Se utilizan los mismos colores respecto a la figura 5.9.

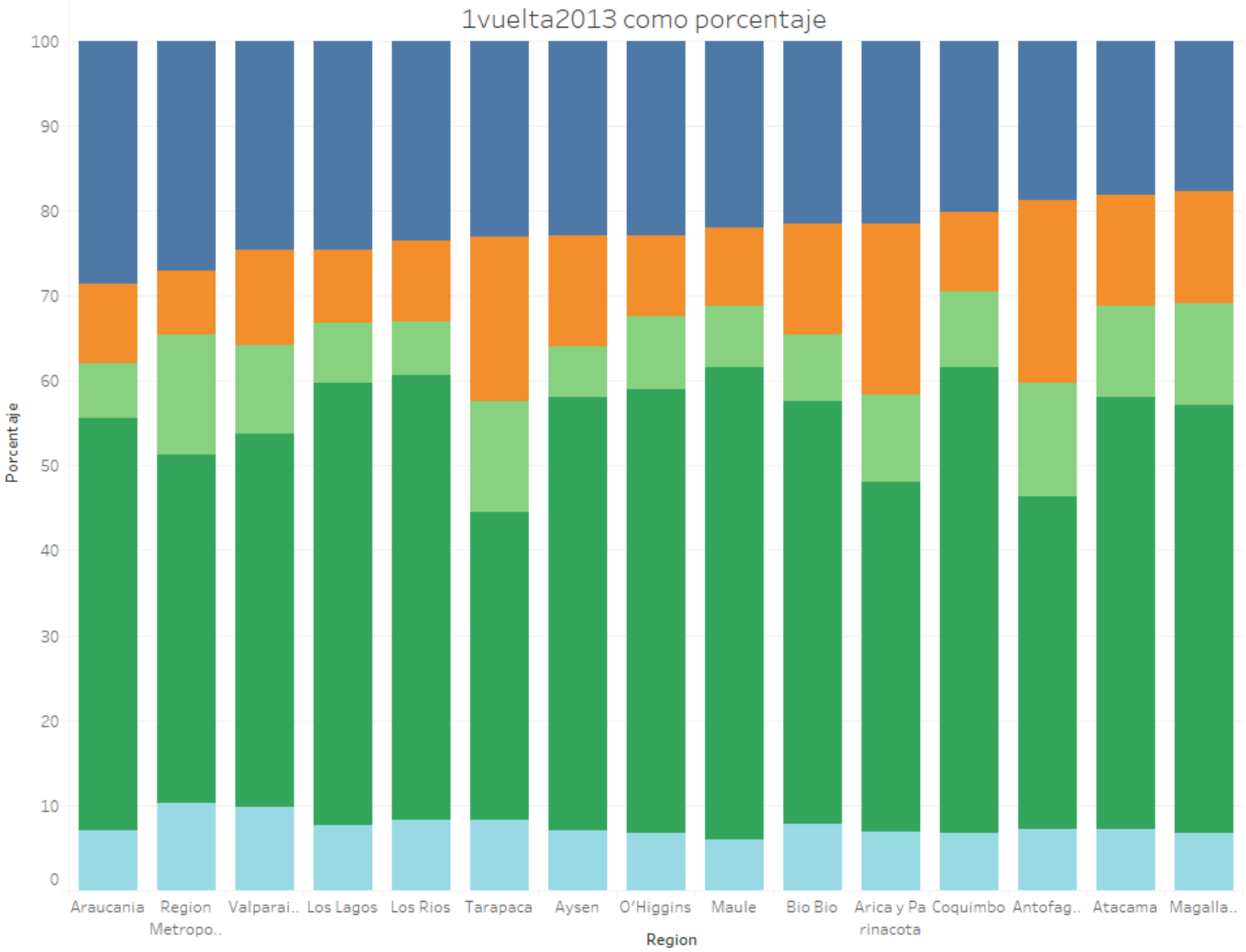

Figura 5.12- Comparación Santiago con distintas Regiones (1° vuelta 2013)

Nos damos cuenta que algunas regiones tienen un comportamiento muy parecido al de Santiago, como es el caso de la región de Valparaíso, pero hay otras regiones como Antofagasta, Araucanía tienen un comportamiento distinto en varios de sus candidatos, enfatizando por supuesto en los que lideran las elecciones (Evelyn Matthei y Michelle Bachelet).

Este último análisis se hace de manera visual respecto a los gráficos generados, pero es importante en una etapa posterior de desarrollo de metodología llegar a cuantificar el grado de semejanza entre las regiones.

Se realiza el mismo procedimiento para el caso de la segunda vuelta 2013, ya que para esta situación son solamente dos posibles candidatos, es posible apreciar más fácilmente los parecidos y diferencias entre regiones.

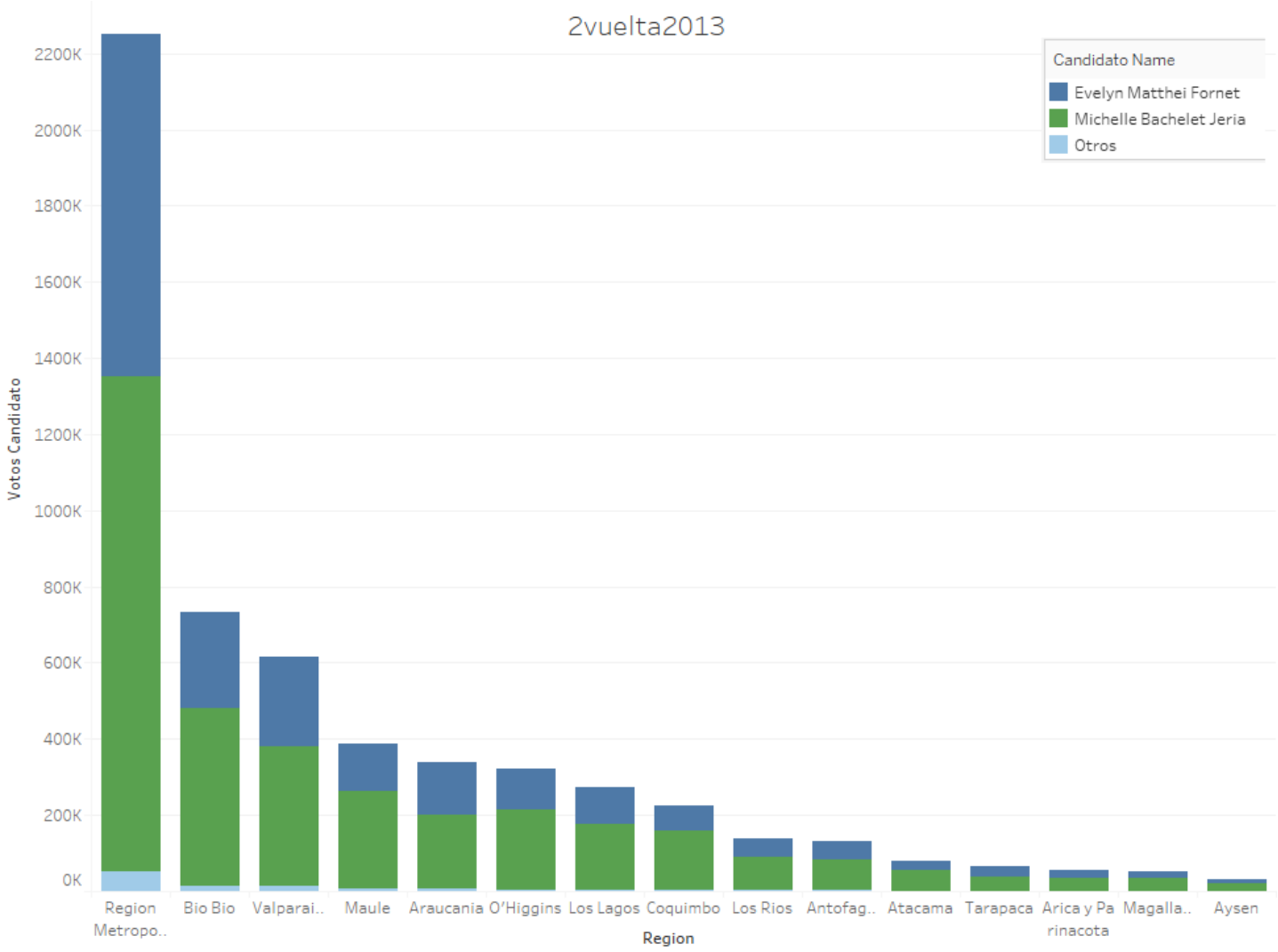

Figura 5.13- Votos segunda vuelta, agrupado por región y candidatos (Presidenciales 2013)

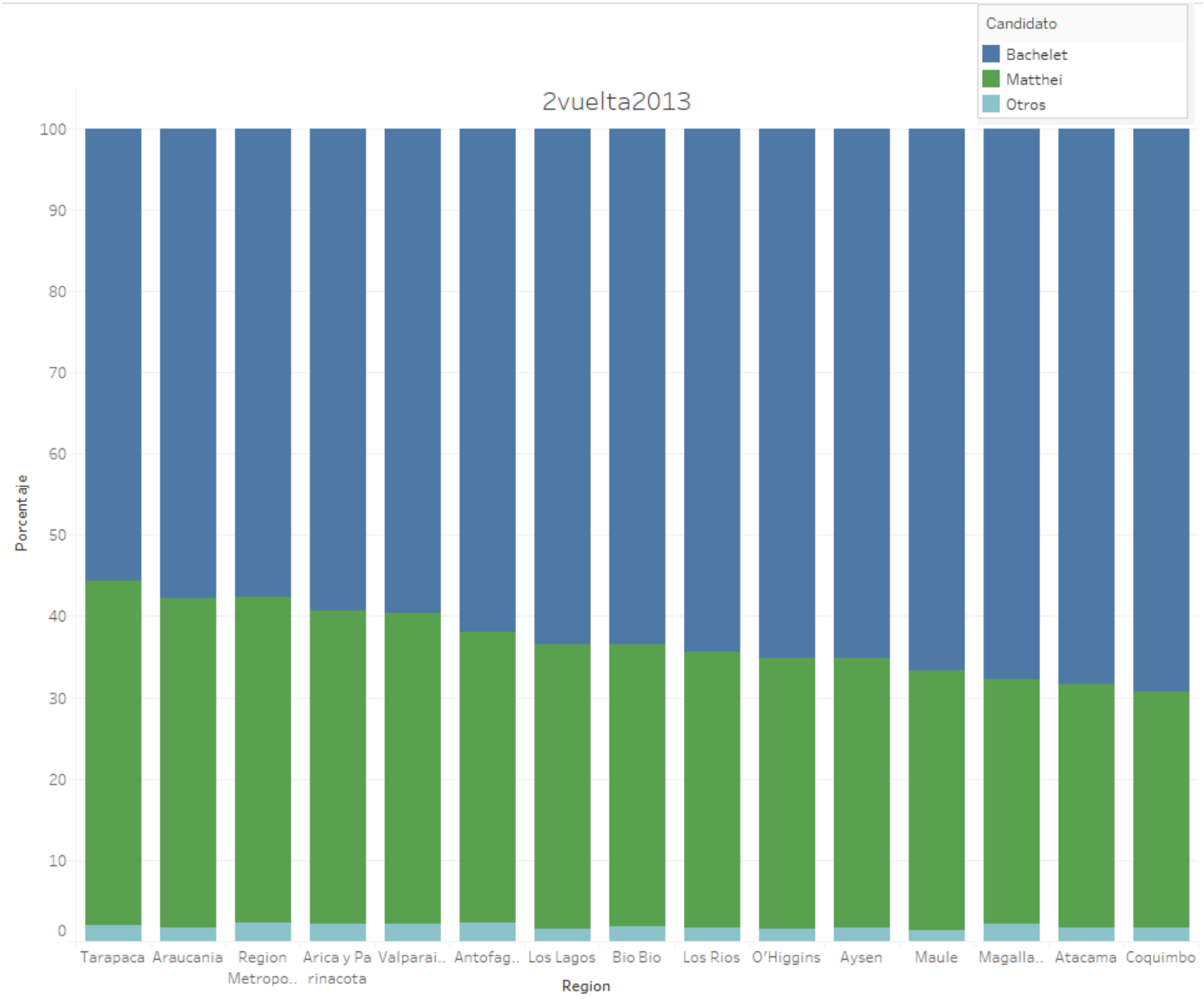

Figura 5.14- Comparación Santiago con distintas Regiones y total. 2 vuelta 2013

Al analizar la segunda vuelta del 2013, solo se tiene 2 opciones (descartando los votos nulos y los blancos), el análisis porcentual nos damos cuenta que Santiago se parece a la región de Arica y de la Araucanía.

A continuación, se realiza el mismo análisis en Tableau, para los datos de le elección del 2017

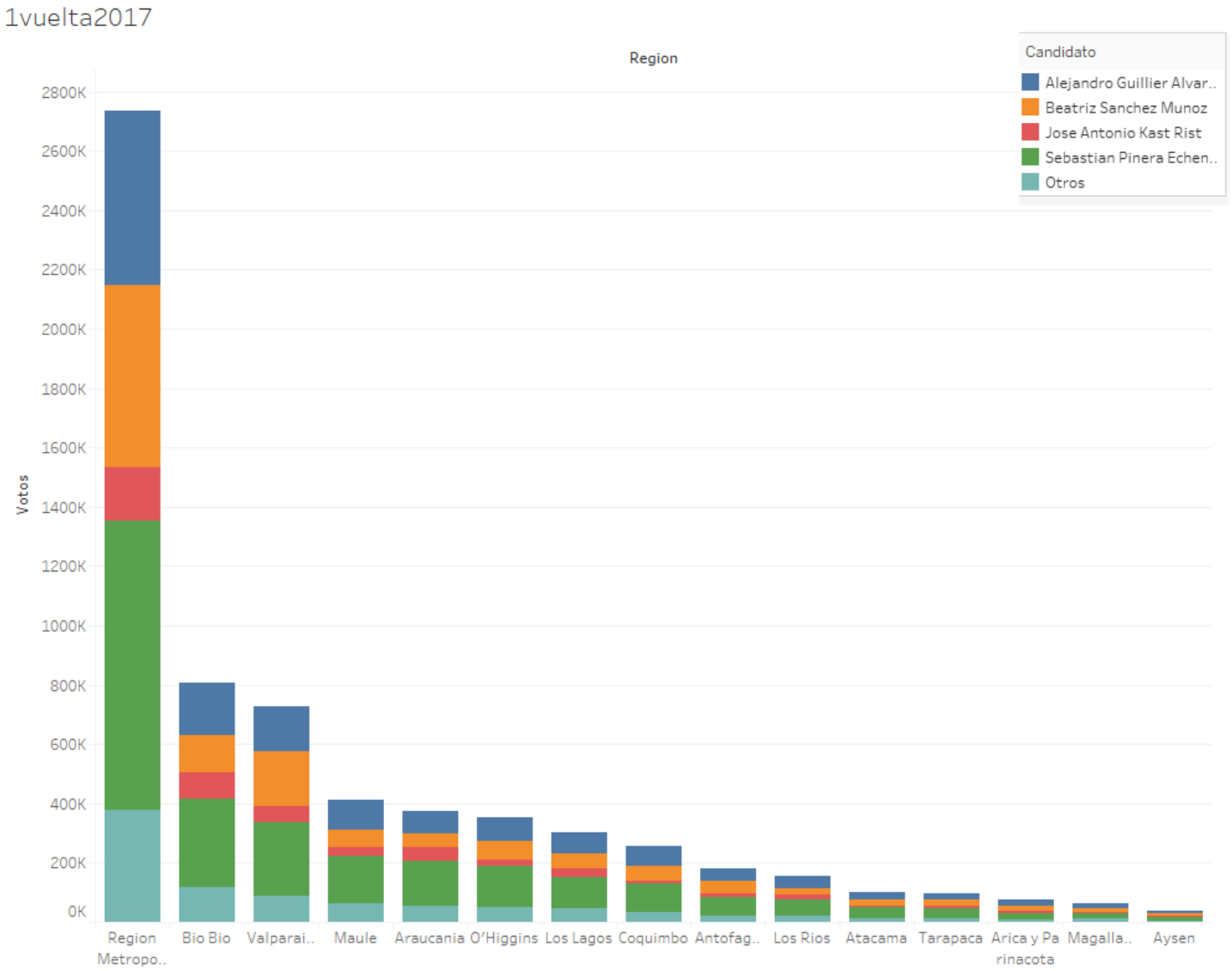

Figura 5.15: Votos agrupados por región, votos y candidatos (1ra vuelta 2017)

En esta nueva elección, se puede apreciar como existe nula variación en el orden de concentración de votos entre las regiones comparadas con la primera vuelta en el año 2013, de manera semejante las votaciones en Santiago nuevamente concentran un poco más de un 40% del total, y sumando Bío bío y Valparaíso se obtiene un 63,96%, valores similares a la elección anterior. Lo que supondría que para una siguiente elección no existiría una variación tan fuerte en las concentraciones de votos de las regiones.

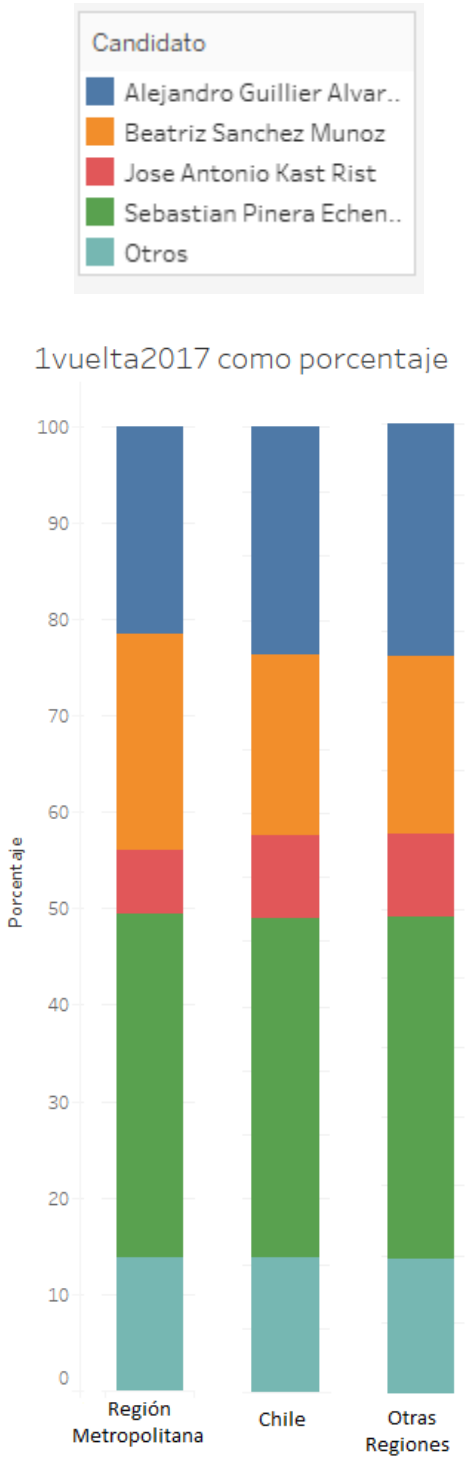

Figura 5.16: Comparación Santiago con el resto de las regiones (Presidenciales 2017 1ra vuelta)

En este gráfico y respaldándonos de la siguiente tabla, podemos apreciar que las diferencias porcentuales para cada candidato, a diferencia de la elección anterior no son tan grandes tanto en la Región Metropolitana como en Chile y el resto de las otras regiones.

	Región Metropolitana	*Otras regiones*	*Chile*
Sebastián Piñera	35,63%	36,41%	36,09%
Alejandro Guillier	21,55%	22,86%	22,3%
Beatriz Sánchez	22,44%	18,2%	19,14%
José Kast	6,63%	8,64%	7,82%
Otros	13,75%	13,89%	13,83%

Tabla 5.4 – Porcentajes obtenidos de cada candidato en Región Metropolitana y el resto de Chile (1ra vuelta 2017)

En esta tabla, podemos apreciar que las variaciones en las preferencias entre los candidatos para RM y el resto de Chile son menores, es decir los gap son menores comparados con la elección 2013 alcanzando en mayor medida a José Kast y Beatriz Sánchez con solamente un 2%, en esta sección pareciera que la RM es más representativa del resto de las regiones, y por supuesto de Chile. Pero esto emana de un caso particular dado que en las elecciones pasadas las diferencias fueron más notorias.

Finalmente, en la segunda vuelta 2017, se puede coincidir con el siguiente gráfico que las regiones siguen el mismo orden de concentración de votos, y no existen variaciones encontradas, desde la perspectiva de los gap porcentuales, las diferencias entre los candidatos Sebastián Piñera y Alejandro Guillier, desde la RM y el resto de las regiones alcanzó un máximo de un 3%, generando una diferencia final entre la RM y Chile de un máximo de 2% aproximadamente correspondiente a las preferencias de Sebastián Piñera.

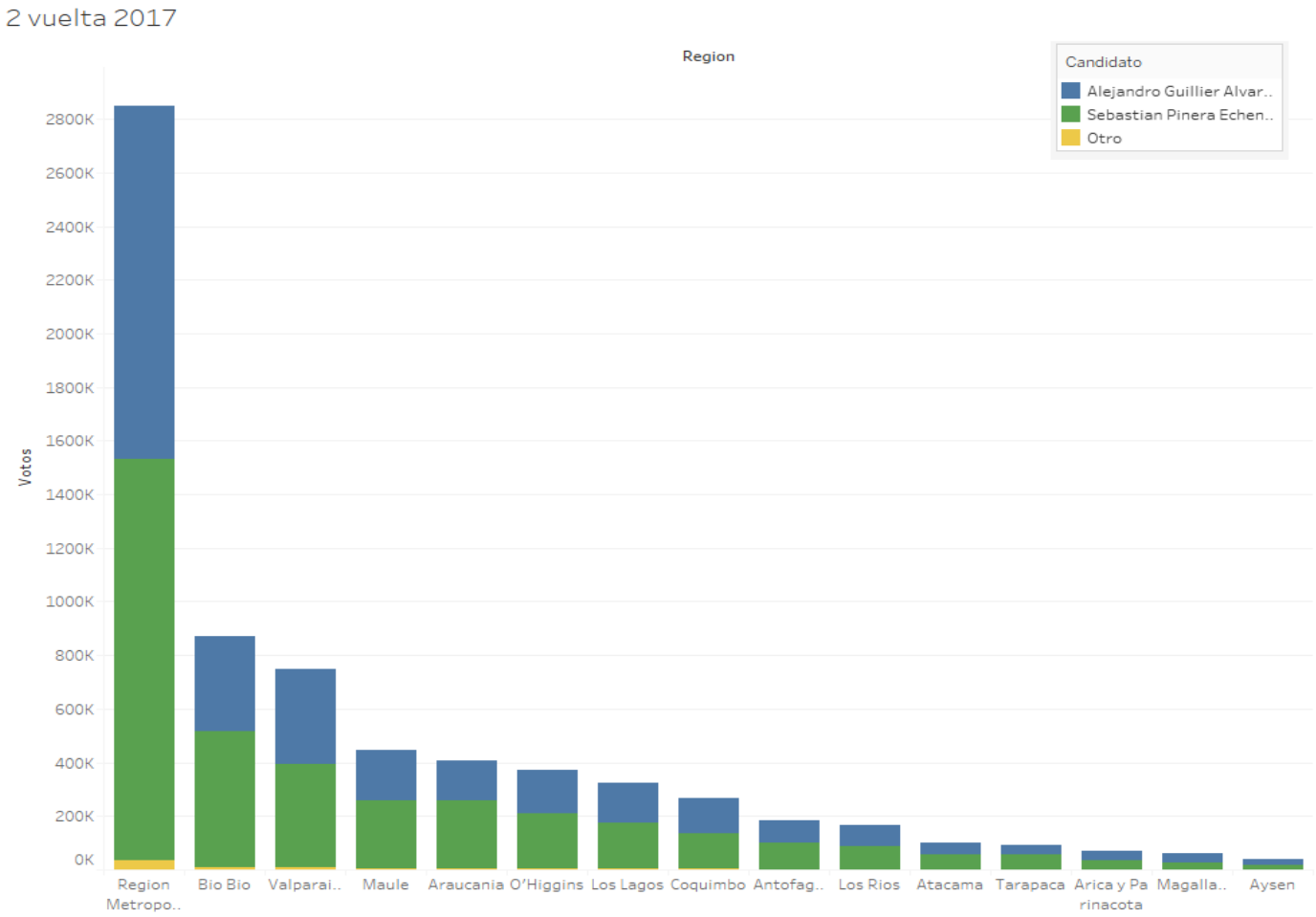

Figura 5.17 – Votos agrupados por región y candidatos (2da vuelta 2017)

Otras conjeturas emanadas a lo largo de la realización del AED que se pudieron percibir de diferentes situaciones, las cuales en primera instancia formaban parte de las suposiciones adoptadas por el equipo en la realización del proyecto las cuales afectan de manera directa la definición y obtención de resultados del mismo.

En primera instancia, el equipo tenía como idea casi irrefutable que la primera vuelta poseía una importancia relativa menor que la segunda vuelta, específicamente en lo que a cantidad de votantes se refiere. Lo anterior resultó ser totalmente falso, en el caso de las últimas dos elecciones utilizadas en el estudio.

Figura 5.18- Total de votos por año y vuelta.

Según lo mostrado en el grafico anterior, se puede observar que en el caso de las elecciones presidenciales del año 2013 existen una disminución de la cantidad de votantes en la segunda vuelta con respecto a la primera vuelta, si bien, existe un aumento de la cantidad de votantes entre las elecciones del año 2017 con respecto a las del 2017, no se puede asegurar que, en la instancia de una segunda vuelta esta tiene una mayor importancia relativa respecto desde la participación ciudadana.

6. Metodología aplicada

La primera aproximación para responder la pregunta de investigación es realizada con las pruebas de homogeneidad con Chi cuadrado.

Se aplica cuando se tiene muestras independientes de n individuos que se clasifican respecto a una variable cualitativa y se desconocer si provienen de la misma población. La prueba tiene como finalidad conocer si la distribución de la variable estudiada difiere en las "r" poblaciones subyacentes de las cuales se obtuvieron las muestras.

Para este test, se usan las pruebas de hipótesis, las cuales son descritas de la siguiente forma:

$H_0 : P_1 = P_2$ (Hay igualdad entre la proporción de elementos de cada grupo que caen la misma categoría de la variable).
$H_1 : P_1 \neq P_2$ (La proporción de los elementos de cada grupo que corresponden a la misma categoría de la variable difieren).

Para determinar el rechazo o no de la hipótesis nula, se usa el estadístico de chi cuadrado, que se describe de la siguiente forma:

$$\chi^2 = \sum \frac{(FO_i - FE_i)^2}{FE_i}$$

Donde:
FO_i: Corresponde al i-ésimo valor observado.
FE_i: Corresponde al i-ésimo valor esperado.

Luego, H_0 tendrá más posibilidades de no ser rechazados en la medida que las diferencias entre FO y FE sean pequeñas. Sin embargo, este mismo punto afecta el análisis para el trabajo realizado, debido a que las cifras que aglomeran el total de votos (para cualquiera de las elecciones), son valores relativamente altos, y por lo tanto las diferencias que se generan entre estos dos parámetros generan un valor de chi cuadrado de grandes proporciones, por lo que el valor p, independiente del grado de significancia que se considere, toma valores demasiado pequeños (casi 0), por lo que cualquier prueba realizada termina refutando la hipótesis nula.

```
                          Region Metropolitana Valparaiso
Sebastian Pinera Echenique              1497308     387282
Alejandro Guillier                      1319587     354326
> chi<-chisq.test(data_20172V, correct = FALSE)
> chi$p.value
[1] 1.724238e-46
> chisq.test(data_20172V, correct = FALSE)

        Pearson's Chi-squared test

data:  data_20172V
X-squared = 204.96, df = 1, p-value < 2.2e-16
```

Figura 6.1- Resultado Chi cuadrado en elecciones 2017 2da vuelta Piñera y Guillier.

Como se aprecia en la figura anterior, este es uno de los casos más simples donde se compara a la Región Metropolitana con Valparaíso, quien, en la percepción gráfica del análisis exploratorio, y como se muestra más adelante con la aplicación de K-means, tienen importantes semejanzas. Pero para la aplicación de este test, dicha semejanza no logra ser percibida por el chi-cuadrado, y su p-value es prácticamente 0, por lo que se refutaría la hipótesis de homogeneidad, por supuesto, el resto de las pruebas que se realizaron, englobando tanto a todas las regiones, o la totalidad de candidatos en otras elecciones, convergieron a resultados semejantes. Es por esto, que se cuestiona que el uso de este test, sea un buen indicador para medir homogeneidad entre regiones.

6.1 Metodología K-means.

La principal metodología utilizada por el equipo para realizar el análisis del *dataset* del proyecto, correspondió al método *K-means*. Esta metodología corresponde a un método de agrupamiento (*clustering*), además de ser un algoritmo no supervisado, es decir, no posee una variable dependiente.

El objetivo de esta metodología es particionar o segmentar un conjunto de datos en diferentes grupos que pueden ser disjuntos o no. Estos grupos se forman a partir de la similitud de los datos a partir de ciertas variables, y como estos grupos no son definidos a priori, el resultado se obtiene a partir de la interpretación de los grupos formados, es decir, por medio de un *análisis de conglomerados,* por ello se espera lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos.

Lo anterior se formaliza por medio del *criterio de inercia*, este criterio lo que busca es minimizar la distancia Intra-clúster (distancia que poseen los elementos que conforman el clúster) y maximizar la distancia Inter-clúster (distancia que poseen los clústeres entre ellos).

Criterio de inercia:

Por medio de este criterio se busca obtener clases lo más homogéneas posibles y que a su vez, se encuentren lo suficientemente separadas. Este objetivo se puede formalizar por medio de la siguiente propiedad:

Suponiendo que se posee una partición definida como $P = (C_1,\ C_2, C_3, \dots, C_k)$ de Ω y además se define $g_1, g_2, \dots\dots g_k$ como los centros de gravedad de cada una de las clases:

$$g_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$$

Además, el centro de gravedad total queda definido de la siguiente forma:

$$\boldsymbol{g} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Definiciones:

Las siguientes definiciones permiten evaluar el funcionamiento del modelo:

- **Inercia total:** corresponde a la inercia total formada por el conjunto de datos.

$$I = \frac{1}{n}\sum_{i=1}^{n} ||\boldsymbol{x_i} - \boldsymbol{g}\,||^2$$

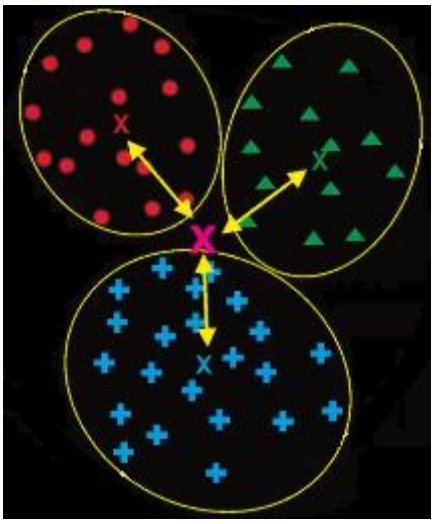

Figura 6.1.1- Inercia total.

- **Inercia inter-clúster:** corresponde a la inercia formada entre los *centros de masa* (centroides) de cada uno de los clústeres formados.

$$B(P) = \sum_{k=1}^{K} \frac{|C_k|}{n} ||\boldsymbol{g_k} - \boldsymbol{g}||^2$$

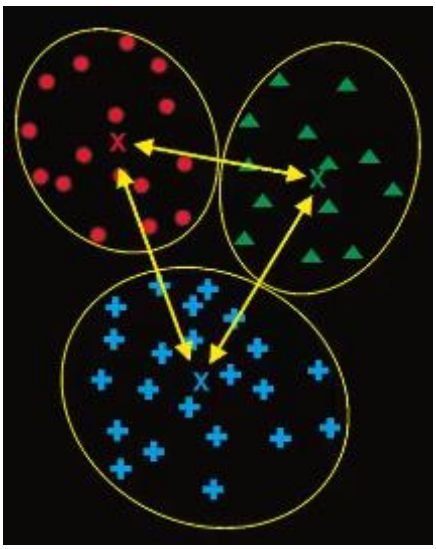

Figura 6.1.2- Inercia inter-clúster.

- **Inercia intra-clúster:** corresponde a la inercia generada al interior de cada clúster formado.

$$\mathrm{W(P)} = \sum_{k=1}^{K} I(C_k) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i \in C_k} ||x_i - g_k||^2$$

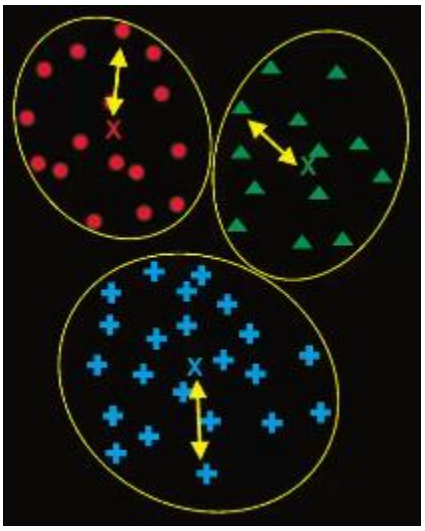

Figura 6.1.3 - Inercia intra-clúster.

Considerando las definiciones anteriores, y por medio del teorema de la igualdad de *Fisher*, se busca que la Inercia total se iguale a la suma de la inercia inter-clúster e intra-clúster como se muestra a continuación.

$$I = B(P) + W(P)$$

La ecuación anterior representa el objetivo del método de *K-means*, se busca que *B(P)* (inercia inter-clúster) sea máxima y que *W(P)* (inercia intra-clúster) sea mínima. Como la inercia total I es fija para el total de la nube de puntos, al maximizar la inercia inter-clúster automáticamente se minimiza la inercia intra-clúster.

Gráficamente la forma de operar del algoritmo se puede observar en la siguiente imagen.

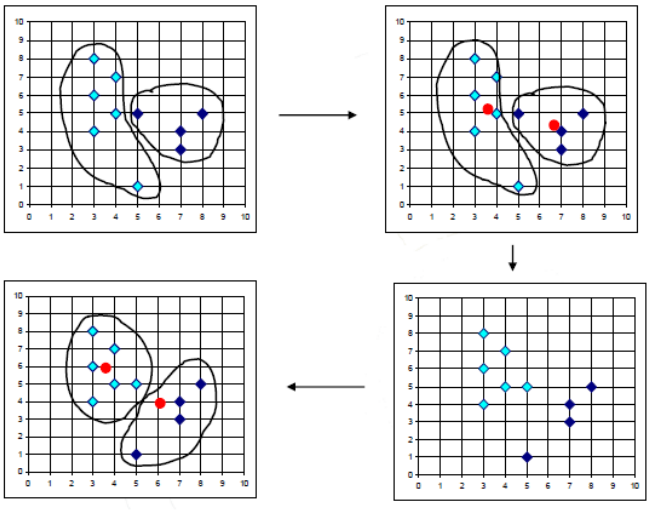

Figura 6.1.4- Algoritmo K-means.

7. Resultados

Para comenzar, el análisis presentado a continuación se resume utilizando solamente los resultados obtenidos de los procesos de elección correspondiente a las segundas vueltas presidenciales correspondientes a los años 2013 y 2017. Este análisis permite realizar una comparación con un mayor nivel de simplicidad dado que sugiere trabajar solamente con 2 candidatos a diferencia de los 10 app. existentes la primera vuelta de ambos procesos eleccionarios.

En el siguiente gráfico se presentan 4 dimensiones, estas corresponden: a los porcentajes de votos obtenidos por Sebastián Piñera y Alejandro Guillier, además consideran el nivel de participación de cada una de las regiones del país y su población correspondiente, para la segunda vuelta correspondiente al año 2017.

El gráfico presenta una correlación entre la participación de ambos candidatos, Guillier y Piñera. Por otra parte, presenta la participación correspondiente a cada región, en la cual, aquellas regiones de un tono azulado presentan una alta participación mientras que aquellas clasificadas con un color rojo presentan una menor participación, además, el porcentaje de participación oscila entre un 40% aproximadamente hasta un 60%.

Finalmente, el tamaño del circulo correspondiente a cada región hace referencia a la población total de la misma.

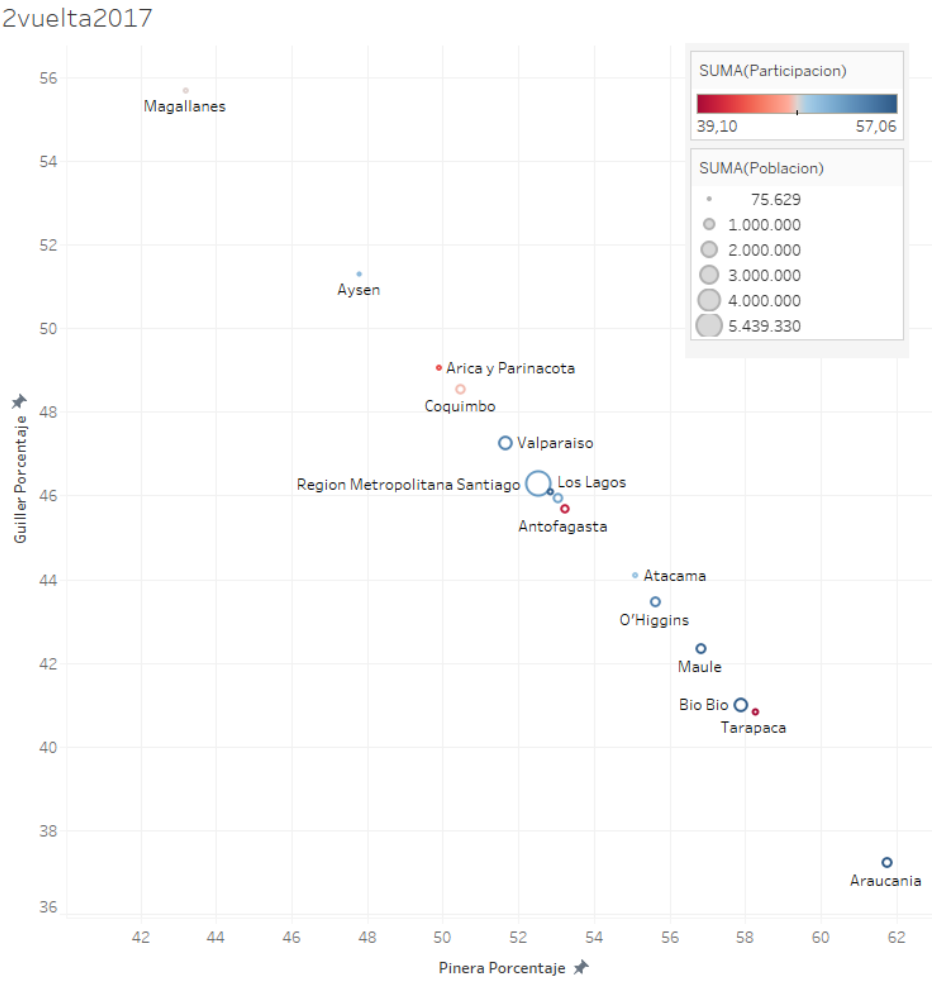

Figura 7.1- Análisis multivariable segunda vuelta elección año 2017

En segunda instancia, el análisis realizado también correspondió a la segunda vuelta del año 2017, en este caso se aplicó el método de clusterización *K-means*, para el caso se generaron 5 clúster distintos.

Dentro de los clústeres generados cabe destacar que el clúster 5 permite visualizar una cierta similitud entre algunas regiones y Santiago, un ejemplo de esto es la similitud existente con Valparaíso.

La clusterización presentada a continuación agrupa una clasificación a partir de 3 atributos relevantes; como la gráfica anterior utiliza el porcentaje de votos obtenido por Sebastián Piñera y Alejandro Guillier, considera también la participación de cada región.

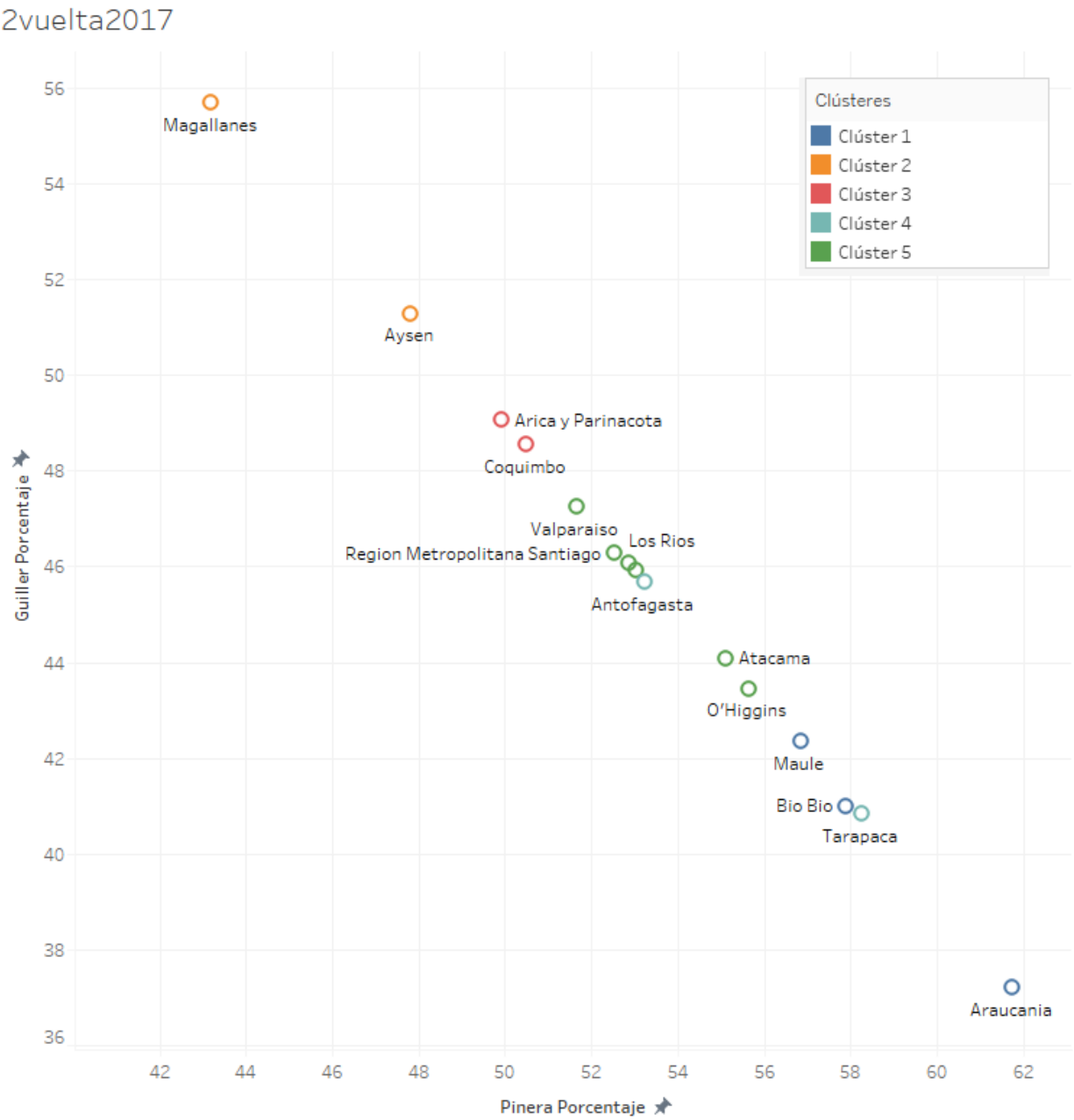

Figura 7.2-Clusterización segunda vuelta elección año 2017.

En la gráfica siguiente el análisis se concentra en el Clúster 5 (contiene Santiago), para el caso, se observa que la clusterización incluye además las regiones de: Valparaiso, Los Rios, Los Lagos, Atacama y O'Higgins. Esta agrupación demuestra que Santiago no tiene un comportamiento totalmente aislado del resto de las regiones, pero tampoco cumple con las características para representar a cada una de las regiones pertenecientes al territorio nacional, además si se considera este clúster, y se busca comparar el nivel de incidencia en las elecciones, se podrá observar que, considerando el total de votos de esta elección, el clúster número 5 aborda alrededor del 75,33%.

Figura 7.3 - Representación en detalle del Clúster 5.

Esto corresponde al clúster 5, que comparten los siguientes atributos en común:

Entradas para la agrupación en clústeres

Variables:	Suma de Guiller Porcentaje Suma de Participacion Suma de Pinera Porcentaje
Nivel de detalle:	Region
Escala:	Normalizada

Resumen de diagnósticos

Número de clústeres:	5
Número de puntos:	15
Suma de cuadrados entre grupos:	2.8257
Suma de cuadrados dentro de grupos:	0.40621
Suma de cuadrados total:	3.2319

		Centros		
Clústeres	**Número de elementos**	**Suma de Guiller Porcentaje**	**Suma de Participacion**	**Suma de Pinera Porcentaje**
Clúster 1	3	40.183	56.693	58.823
Clúster 2	2	53.49	49.18	45.5
Clúster 3	2	48.81	45.22	50.195
Clúster 4	2	43.25	39.635	55.745
Clúster 5	6	45.515	53.12	53.467
Sin clústeres	0			

Figura 7.4 – Resultados arrojados de la clusterización en Tableau Public (2017 segunda vuelta)

En segunda instancia, el análisis realizado anteriormente por medio de la metodología *K-means.* Se repite para la situación del proceso eleccionario 2013. Cabe destacar que este análisis no cuenta con un con información fidedigna de la población ya que corresponde a una proyección, para efectos prácticos del análisis y la fuente obtenida (INE) se consideran correctos.

En el siguiente gráfico se presentan 4 dimensiones, estas corresponden: a los porcentajes de votos obtenidos por Evelyn Matthei y Michelle Bachelet, además consideran el nivel de participación de cada una de las regiones del país y su proyección de población, correspondiente para la segunda vuelta correspondiente al año 2013.

El gráfico presenta una correlación entre la participación de ambas candidatas, Bachelet y Matthei. Por otra parte, presenta la participación correspondiente a cada región, en la cual, aquellas regiones de un tono azulado presentan una alta participación mientras que aquellas clasificadas con un color rojo presentan una menor participación, además, el porcentaje de participación oscila entre un 30% app. hasta un 50%, que comparado con el último proceso eleccionario se diferencia app en 10% por debajo de ambos límites, por lo que es clara la incidencia de una menor participación en el año 2013. Finalmente, el tamaño del circulo correspondiente a cada región hace referencia a la población total de la misma.

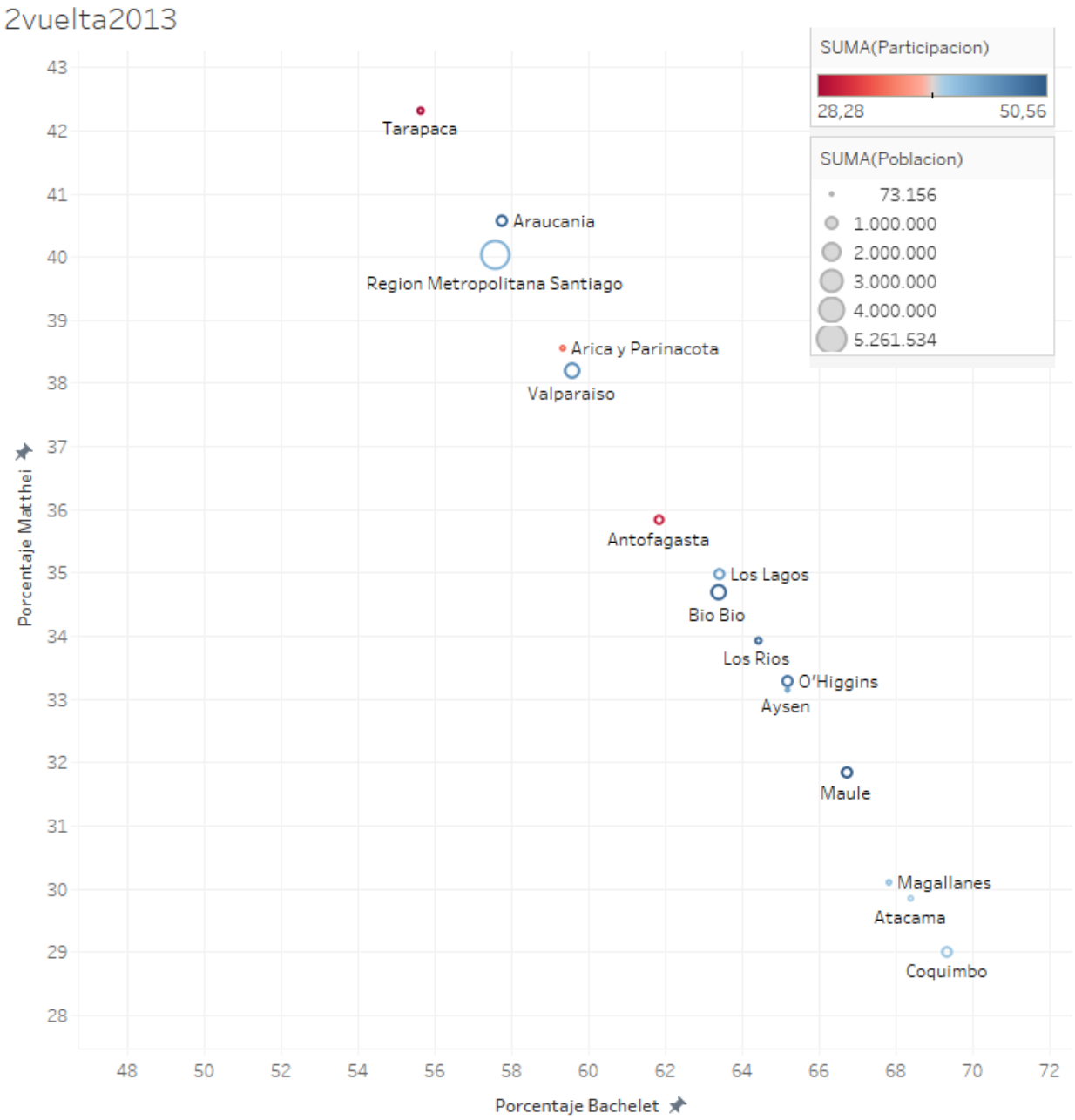

Figura 7.5- Análisis multivariable segunda vuelta elección año 2013.

A continuación, el análisis realizado también correspondió a la segunda vuelta del año 2013, también aplicando el método de clusterización *K-means*, que, a diferencia del análisis anterior, para el caso se generaron 3 clúster distintos.

Dentro de los clústeres generados cabe destacar que el clúster 3 permite visualizar una cierta similitud entre algunas regiones y Santiago, en la cual nuevamente existe una similitud con la región de Valparaíso.

La clusterización presentada a continuación agrupa una clasificación a partir de 3 atributos relevantes; como el en caso anterior utiliza el porcentaje de votos obtenido por Sebastián Piñera y Alejandro Guillier, considera también la participación de cada región.

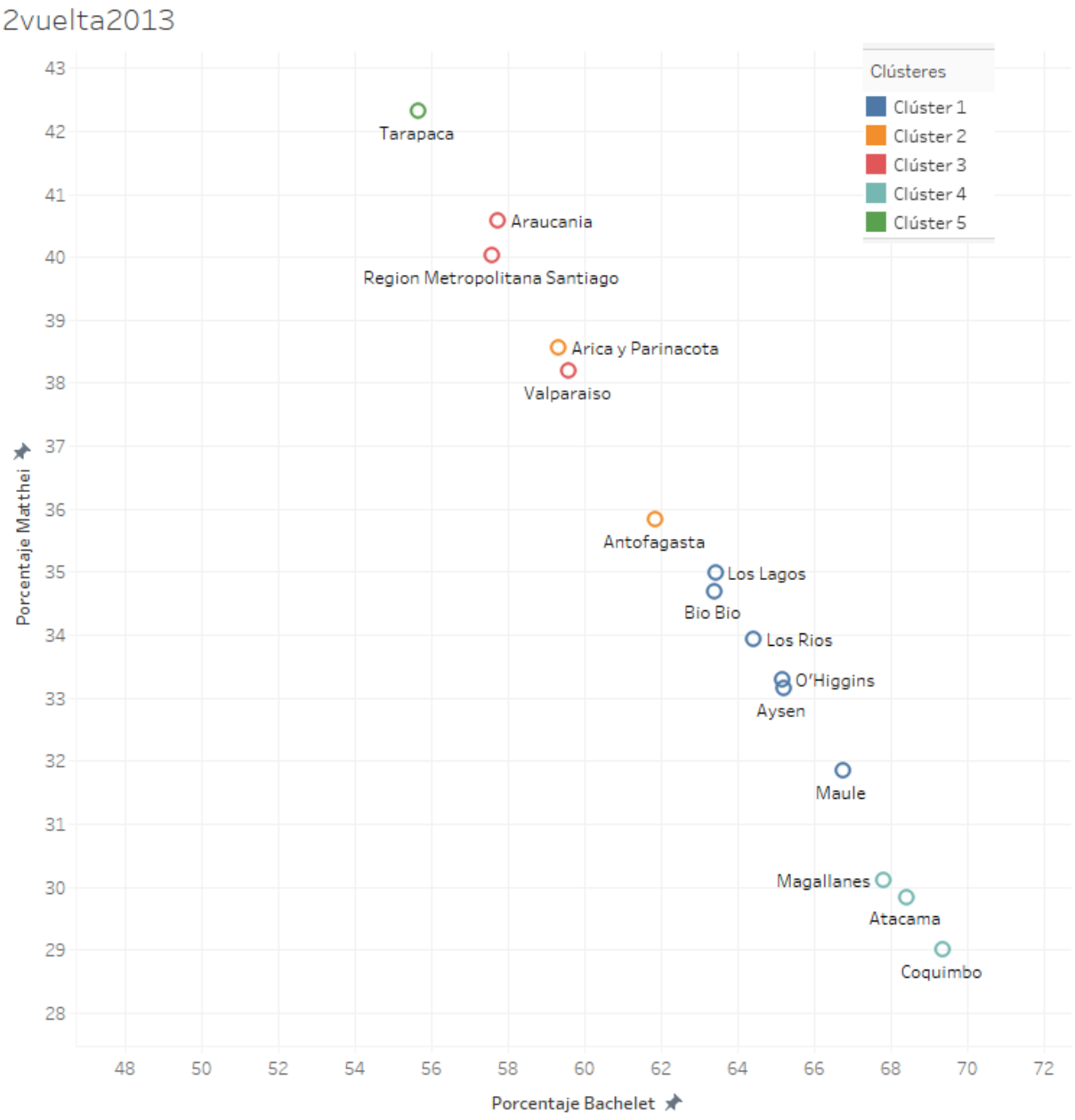

Figura 7.6-Clusterización segunda vuelta elección año 2013.

En la gráfica siguiente, el análisis se concentra en el clúster 3 (donde se ubica la RM), para el caso, se observa que la agrupación incluye además las regiones de: Valparaíso y la Araucanía. Esta agrupación dista bastante del análisis realizado de manera previa para el año 2017 el cual incluía 3 regiones más, además comparado con el 75% de la concentración de votos, el actual clúster reduce su concentración de votos a un 56,29%. Esta agrupación demuestra que Santiago no tiene un comportamiento totalmente aislado del resto de las regiones, pero que nuevamente tampoco cumple con las características para representar a cada una de las regiones pertenecientes al territorio nacional y menos que este comportamiento se vea mantenido en el tiempo.

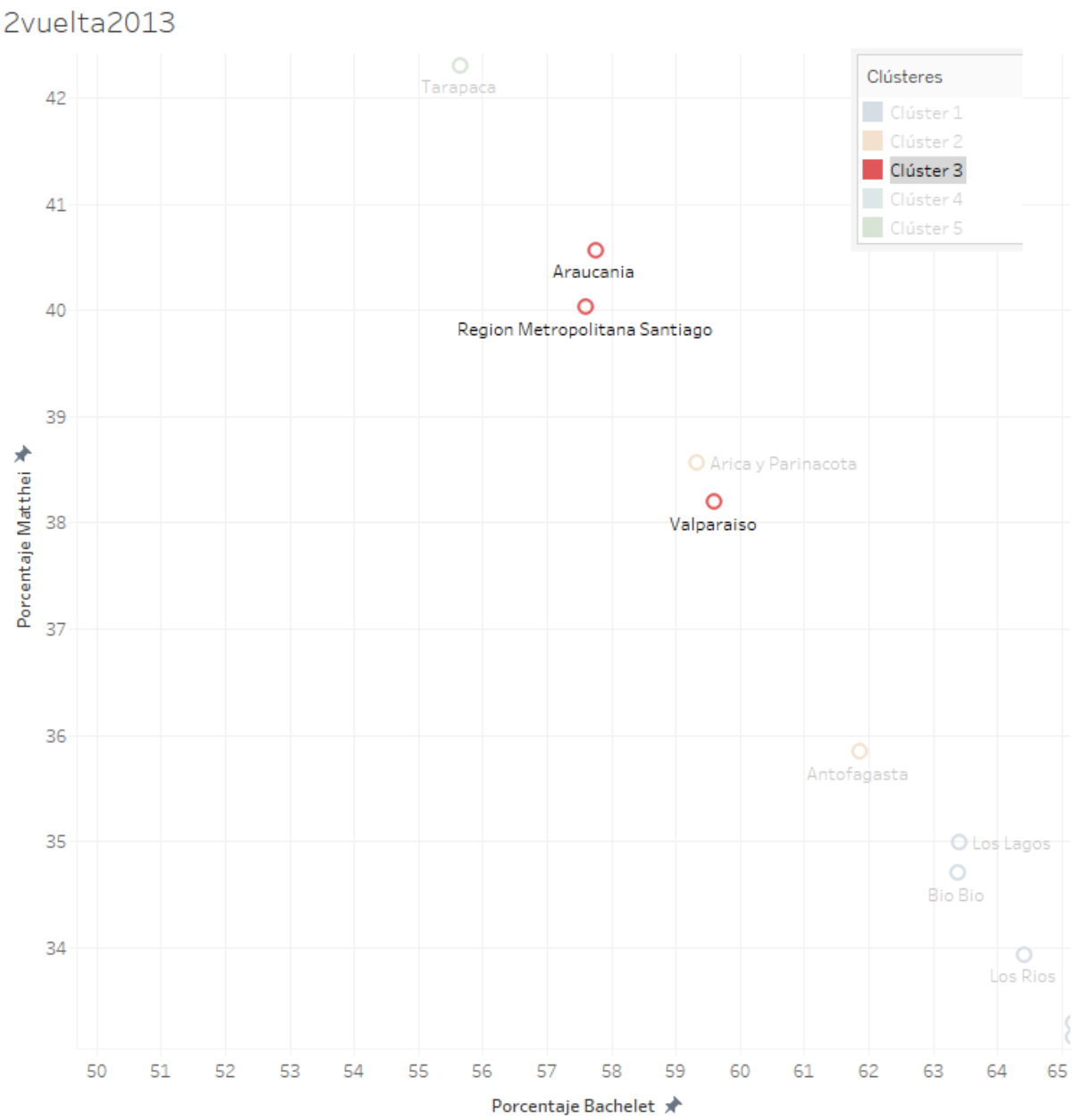

Figura 7.7- Representación en detalle del Clúster 3

.

Entradas para la agrupación en clústeres

Variables:	Suma de Participacion Suma de Porcentaje Bachelet Suma de Porcentaje Matthei
Nivel de detalle:	Region
Escala:	Normalizada

Resumen de diagnósticos

Número de clústeres:	5
Número de puntos:	15
Suma de cuadrados entre grupos:	3.7849
Suma de cuadrados dentro de grupos:	0.29611
Suma de cuadrados total:	4.081

		Centros		
Clústeres	**Número de elementos**	**Suma de Participacion**	**Suma de Porcentaje Bachelet**	**Suma de Porcentaje Matthei**
Clúster 1	6	47.697	64.725	33.648
Clúster 2	2	32.25	60.6	37.2
Clúster 3	3	45.947	58.327	39.6
Clúster 4	3	40.993	68.523	29.647
Clúster 5	1	28.28	55.66	42.31
Sin clústeres	0			

Figura 7.7 – Resultados arrojados de la clusterización de Tableau Public (2013 segunda vuelta)

8. Conclusiones:

Interpretación de resultados:

Respecto de los análisis realizados en la sección del análisis exploratorio y de la metodología utilizada, se desprenden varias observaciones importantes. En primer lugar, dado los gráficos 5.11 y 5.16, con las respectivas tablas 5.3 y 5.4, se puede apreciar que Santiago tiene un comportamiento diferente en cada elección con respecto al resto de las elecciones. Para el año 2013, se puede observar como las votaciones consideran diferencias importantes entre la RM y el resto de las regiones, pero como ya se ha abordado, la gran cantidad de votos que concentra la primera, compensa de manera importante la balanza en los resultados finales de las elecciones. Mientras, que para la elección 2017 se puede apreciar que Santiago está mucho más alineado con las tendencias de las otras regiones, y por lo tanto el resultado final del proceso se aprecia una mayor uniformidad.

Para el caso del proceso eleccionario presidencial correspondiente al año 2013, en el cual Santiago fue agrupado con las regiones de Valparaíso y la Araucanía, por otra parte, para el mismo proceso realizado el año 2017 en el cual la RM se agrupó con las regiones de: Valparaíso, Los Ríos, Los Lagos, Atacama y O'Higgins. Lo anterior nos indica que, si bien Santiago en cierta medida representa a otras regiones del territorio nacional en función de los votos obtenidos por candidato y la participación de la población, no se puede garantizar que esta representación sea constante en el tiempo, considerando que la RM está más alineada con las tendencias del resto en el año 2017, es un factor importante para considerar que la clusterización donde participa Santiago contenga más regiones que la del 2013. Por lo tanto, en el análisis se puede observar que Santiago es muy influyente en los resultados a nivel nacional dado la concentración de votos por efectos de la cantidad de población que agrupa, y de cierta manera hay una similitud con respecto del comportamiento observado en la región de Valparaíso, tanto en el nivel de participación (porcentual) y porcentaje de votos obtenidos por candidato en el 2013 y 2017, existen una amplia cantidad de variables que no fueron consideradas en este estudio, y que pueden afectar tanto la participación de la población como la preferencia por algún candidato, factores socioeconómicos de la población, situación política actual, economía, etc.

Finalmente, y dada la pregunta de investigación planteada de si la RM es representativa de todo el territorio nacional chileno, se concluye que tiene una influencia muy *importante*, pero *no absoluta*, es decir, dadas las apreciaciones del AED, las pruebas de chi cuadrado y K-means, nos dan resultados que evalúan a Santiago no es del todo representativo, si bien, junto con Valparaíso comparten una semejanza en el comportamiento, la cual además se ha repetido en los últimos dos procesos eleccionarios, estas distan de ser una mayoría y por supuesto, los comportamientos que existen en la capital y su región no son homogéneos y están totalmente influenciados al contexto y características de cada elección en particular.

Propuesta de Gestión:

Gracias al análisis de los gráficos, fue posible visualizar el comportamiento de las 15 regiones, correlacionando la participación porcentual de cada candidato, considerando la participación de la comuna. Gracias a esto, se analiza que algunas regiones tienen un comportamiento parecido, que fue posible agrupar con la metodología k-means, además de poder cuantificar los valores comunes que comparten cada clúster. Toda la generación de valor a partir de los datos, la canalizamos como una herramienta muy potente para actuales y futuros candidatos, además para los distintos partidos políticos.

Una de las decisiones que debe tomar un candidato o un partido político es seleccionar las regiones (y con un análisis más detallado considerar comunas) en que se invertirá dinero para publicitar una candidatura, si lo hiciera bajo una muestra representativa, puede que necesite estar presente en todas las comunas, de acuerdo al peso de la comuna, incurriendo en decisiones que impactan directamente en el presupuesto necesario para llevar a cabo la propaganda electoral. Pero gracias a nuestro análisis, es posible generar criterios de selección, y elegir una zona que potencialmente puede tener mejores resultados. Por ejemplo, un partido de derecha, gracias al análisis en k-means, puede seleccionar una de estas posibles regiones donde hay una preferencia política hacia la derecha, y además baja participación. En particular, respecto a los datos de la elección 2017 segunda vuelta, se podría elegir la región de Tarapaca para una futura campaña, ya que presenta un apoyo del 58% a la derecha, y presenta una participación de 40%, uno de niveles más bajo para esa elección. Con esto aprovechar la tendencia política favorable de la región, junto a una baja participación, e invertir en propaganda focalizada en esta región objetivo. Este criterio de selección queda aún más potenciado al utilizar los clústeres, ya que se identifica un potencial grupo de regiones. Con todo esto se contribuiría a la estrategia de una campaña, de acuerdo a los objetivos que se buscan.

Además, puede ser de especial interés lograr "democratizar" esta información, para que pueda llegar a candidatos y partidos políticos que manejan bajos presupuestos para sus propagandas electorales, generalmente los grandes conglomerados políticos pueden usar una gran cantidad de presupuesto para el marketing, pero para los pequeños, deben hacer un uso inteligente del acotado presupuesto que poseen, esto puede permitir reducir la brecha al otorgar un activo importante para cualquier persona o agrupación que quiera participar de los rituales democráticos.

Finalmente, se ha investigado sobre el proceso de muestreo que realizan la CADEM para ejecutar las encuestas, si bien, su proceso de muestreo al parecer es robusto, y va más allá de lo que nuestra investigación ha abordado, recomendamos ver el tema de las influencias que tiene preguntar directamente a las personas sobre sus preferencias políticas, actualmente esto se realiza con personas que preguntan directamente o a través de llamadas telefónicas, la decisión que tomarían en la urna, este método puede sesgar los resultados, dado que las personas pueden ser resistentes a responder auténticamente su preferencia a cierto candidato, por lo que un estilo de encuesta más impersonal, podría mejorar la asertividad de los resultados procesados por la encuesta.

En un trabajo futuro, sería interesante evaluar cómo se comportan las regiones extreman buscando algún factor basado en las tendencias políticas o participación. Por otro lado, se considera que un nuevo factor a considerar además de los resultados, y la participación para cada región y comuna, es considerar los factores socioeconómicos, ya que se ha podido apreciar desde un punto de vista empírico que las comunas con mayor nivel de participación y una tendencia marcada hacia la derecha, corresponden a comunas de un muy buen nivel socioeconómico, por ejemplo: Las Condes, Lo Barnechea, Vitacura, etc. Por lo que se considera que hay un interesante nicho de investigación ahí.

9. Bibliografía

¿En que se equivocó la encuesta CADEM? (Ciper, 2017) https://ciperchile.cl/2017/12/12/en-que-se-equivoco-la-encuesta-cadem/

Boletín público y solemne de resultado preliminares de la elección para presidente de la república 2017 https://www.servel.cl/wp-content/uploads/2018/08/4_1_Boletin_Detalle_Presidente.pdf

K-means Algorithm, Oldemar Rodriguez. (http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Presentaci%C3%B3n_-_k-means.13775252.pdf)

Test of independency and Homogeneity - Anthony Tanbakuchi, Et. Al (http://www.u.arizona.edu/~kuchi/Courses/MAT167/Files/LH_LEC.0640.HypTest.IndepHomog.pdf.)