Pontificia Universidad Católica de Valparaíso
Escuela de Ingeniería Industrial

Data Science Aplicada
Entregable 3b

Informe
Análisis exploratorio de datos (AED)

Integrantes
Andrés Arenas Rodríguez.
Indy Navarro Vidal.
Juan Saavedra Jeria.
Equipo
Mauricio Huerta Aguiar
Pablo Zúñiga Carvajal
Víctor Leiva Sánchez

Índice

Introducción.

En el presente informe se detalla el primer *análisis exploratorio de datos* (AED) realizado al ETL, el cual fue presentado anteriormente. Este análisis tiene por objetivo realizar una primera medición y descripción del comportamiento de los datos correspondiente a los votos obtenidos por los candidatos correspondientes a las elecciones presidenciales de los años 2013 y 2017 (tanto en primera como segunda vuelta). Junto con lo anterior, se realiza una pequeña comparación en función de la cantidad de votos obtenidos en las diferentes comunas país, por candidato, por periodo y vuelta electoral.

El análisis realizado a lo largo de este informe fue llevado a cabo con diversos software como R, Excel y Tableau que permitieron llevar adelante una comparación consistente respecto de los diversos datos que componen el *dataset* utilizado, el cual tiene la característica de contar solamente con una variable cuantitativa correspondiente a la cantidad de votos obtenidas por cada candidato, mientras que, todo el resto de variables definidas para el conjunto corresponden a variables de carácter cualitativo, como lo son el nombre del candidato, la región, etc.

Este informe entrega una primera aproximación del trabajo a realizar bajo el objetivo del proyecto, el cual busca responder la siguiente pregunta de investigación:

¿Es Santiago una muestra representativa de las tendencias políticas de todo el territorio nacional?

Finalmente, se entregan las principales conclusiones y decisiones futuras emanadas a partir del tratamiento de los datos realizados en este estudio.

Metodologías Utilizadas

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:

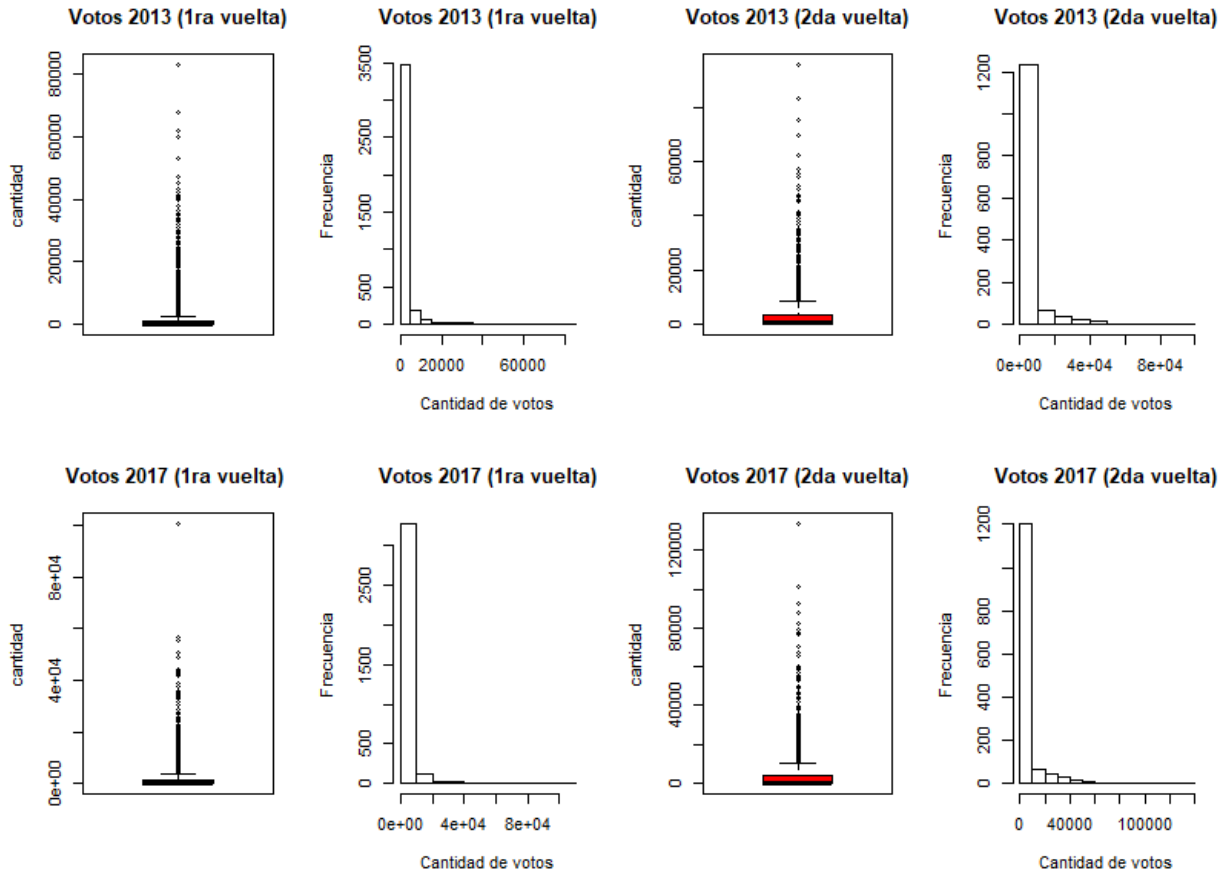

Figura 1.1- Boxplot e Histograma correspondiente a cada vuelta y su elección respectiva

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:

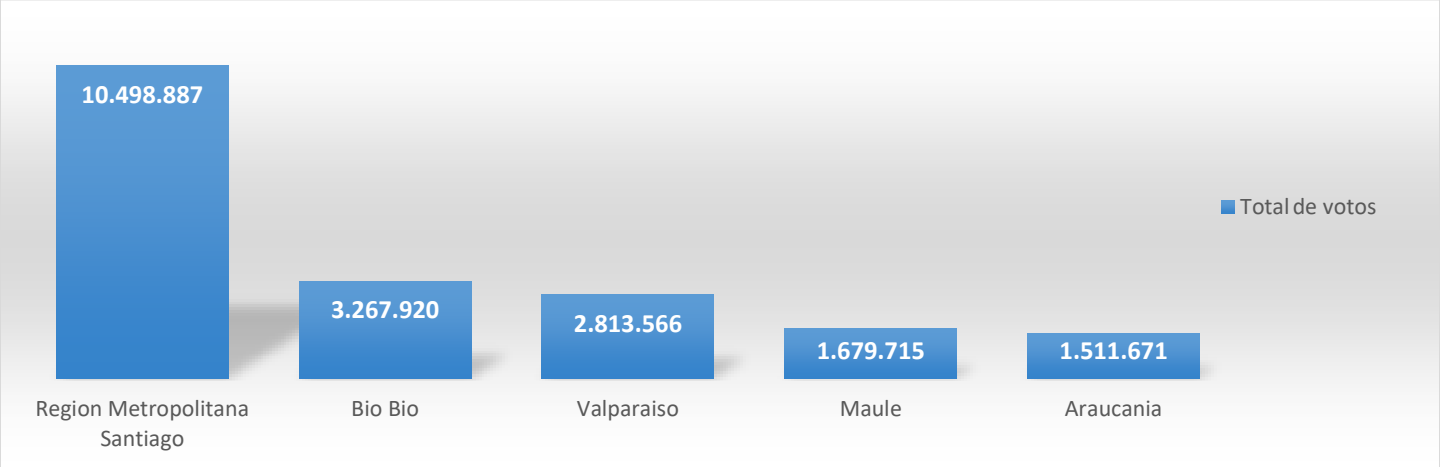

Figura 1.2- Top 5 de las regiones con mayor cantidad de votantes.

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:

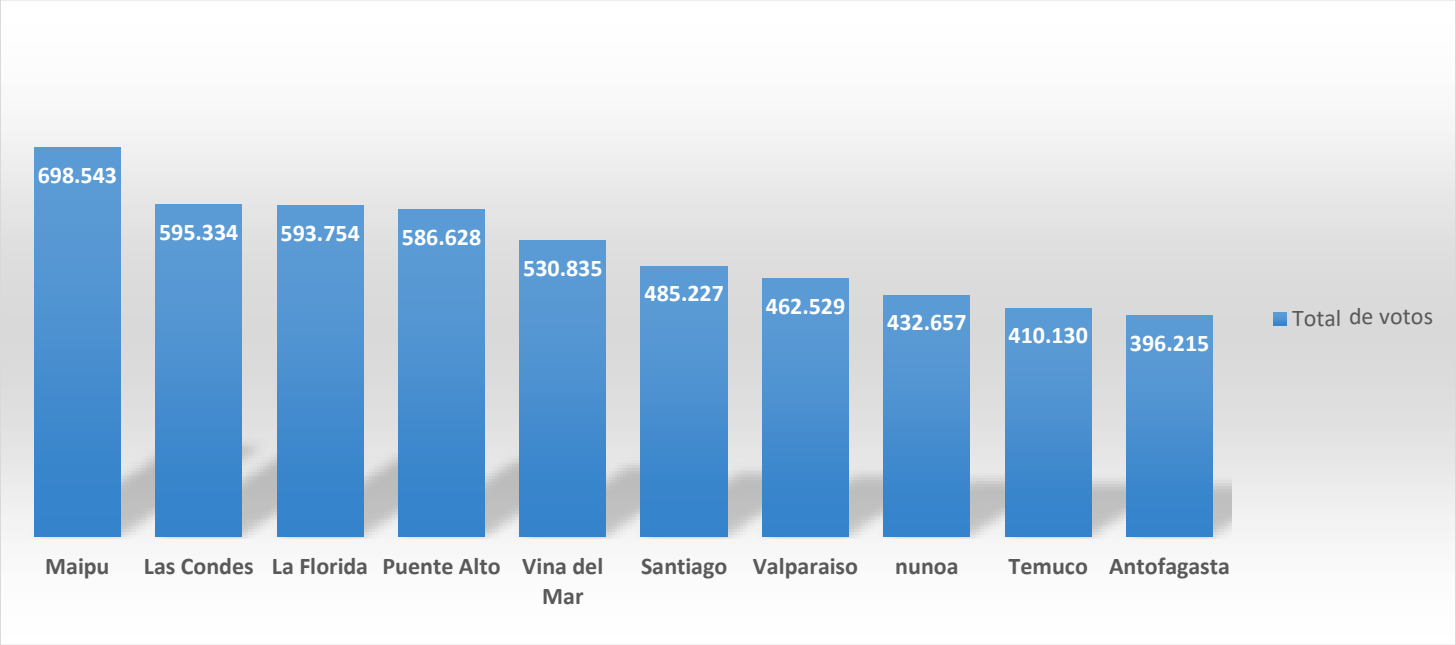

Figura 1.3- Top 10 de las comunas con mayor cantidad de votantes.

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:

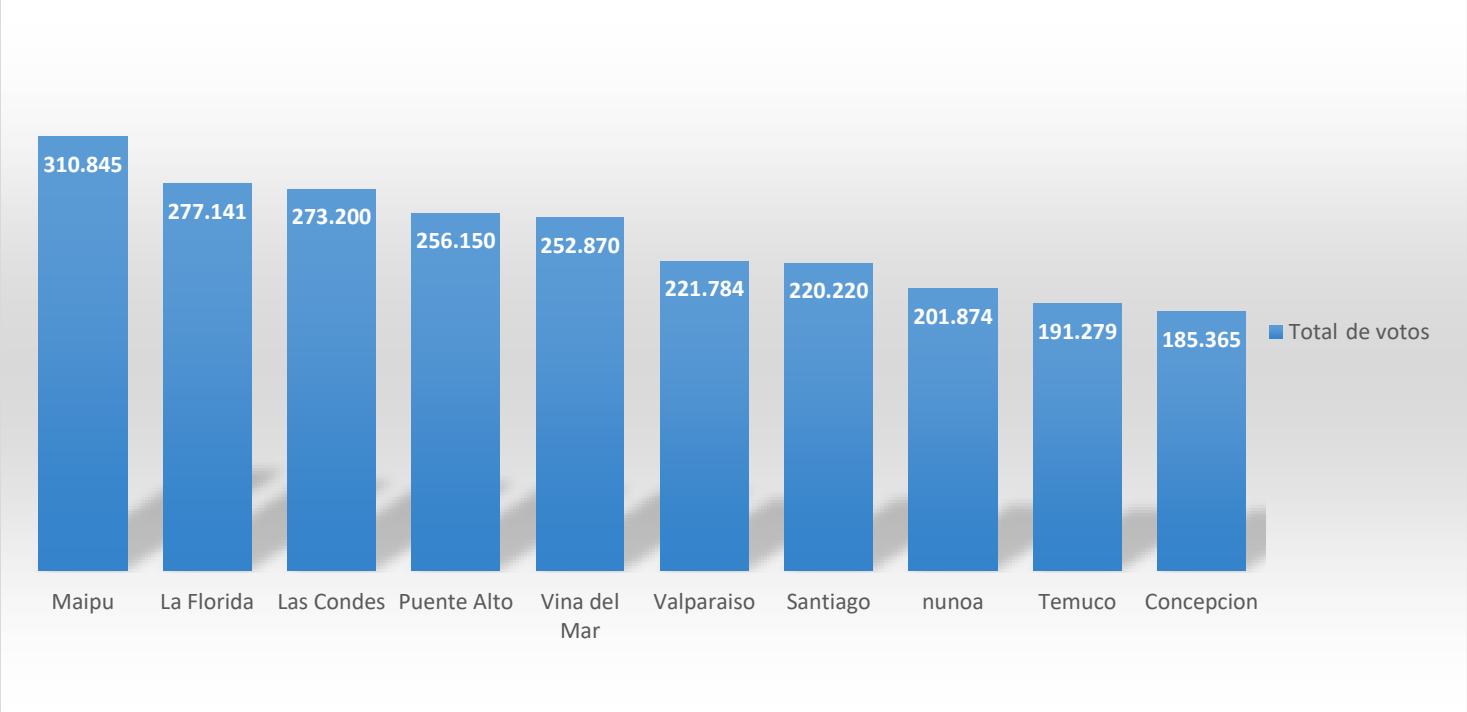

Figura 1.4- Top 10 de comunas que concentran votaciones (Presidenciales 2013)

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:

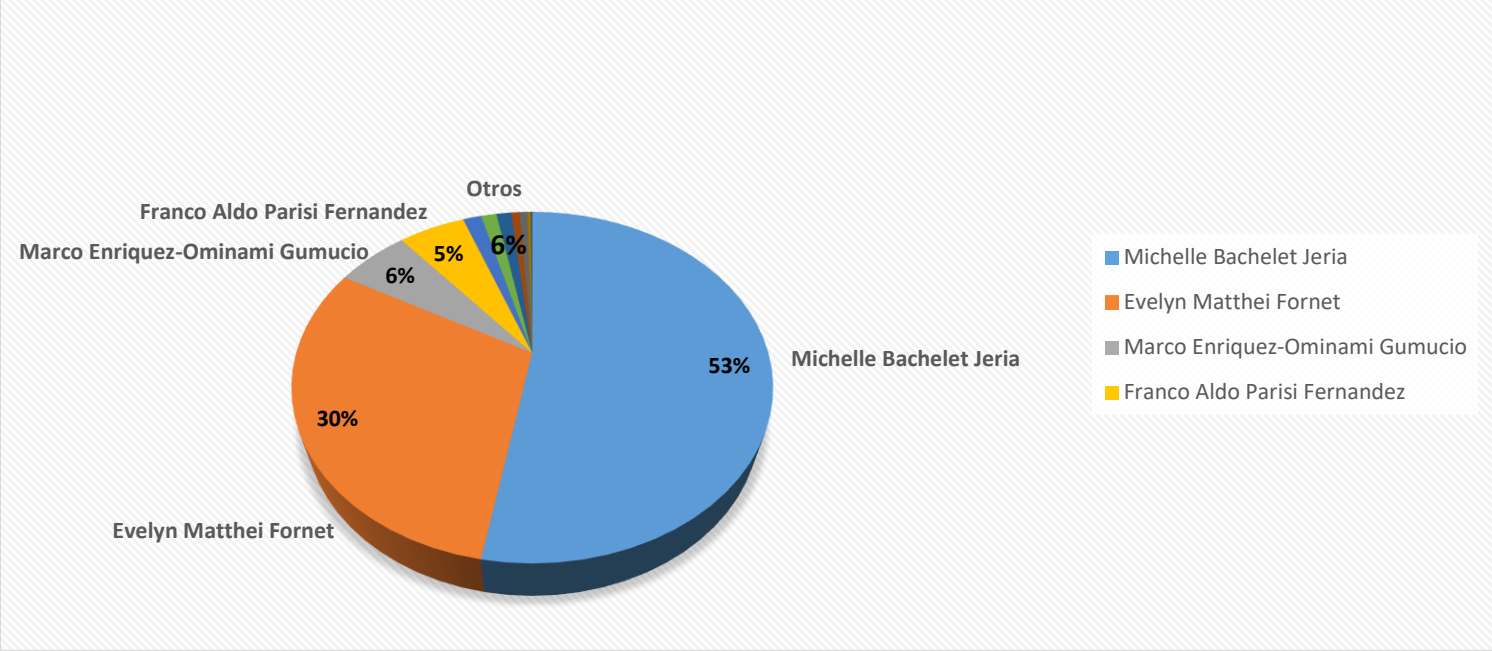

Figura 1.5- Porcentaje de votos obtenidos por candidato (Presidenciales 2013)

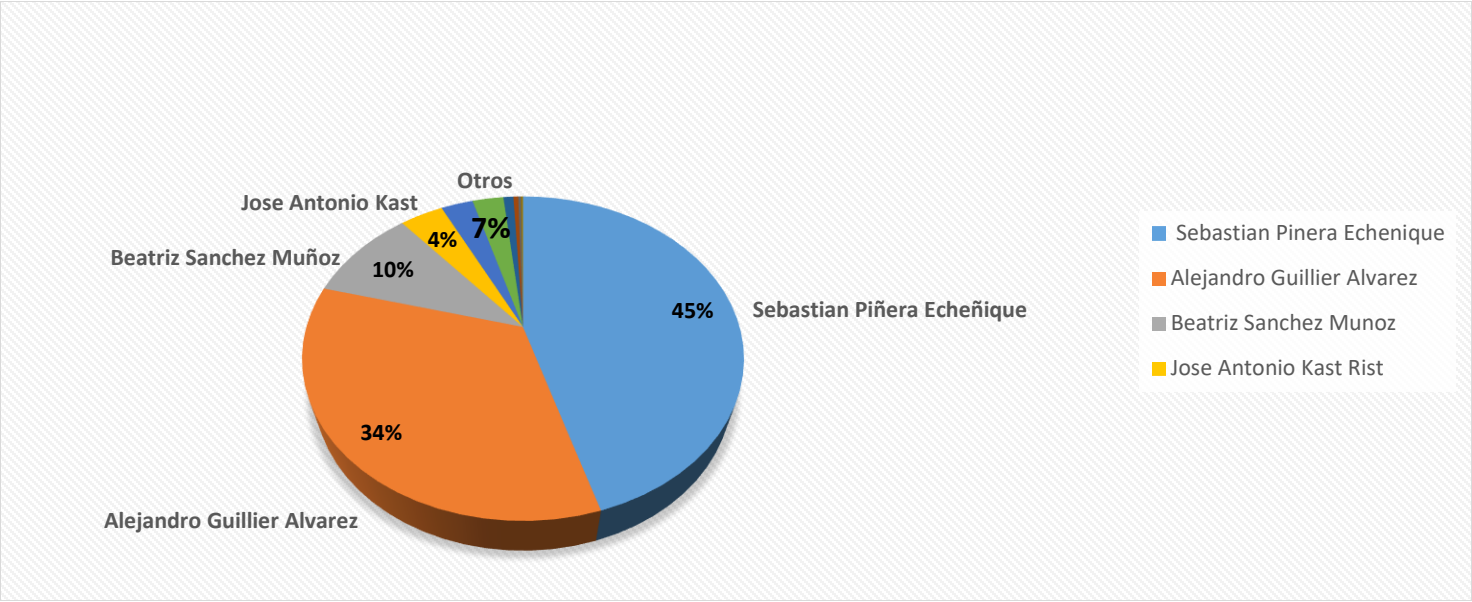

Figura 1.6- Porcentaje de votos obtenidos por candidato (Presidenciales 2017)

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, y 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:

Tabla 1- Población total y votos realizados por región.

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

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

Tabla 2- Correlación entre población y cantidad de votos.

	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

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.

Análisis en Tableau Public

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.

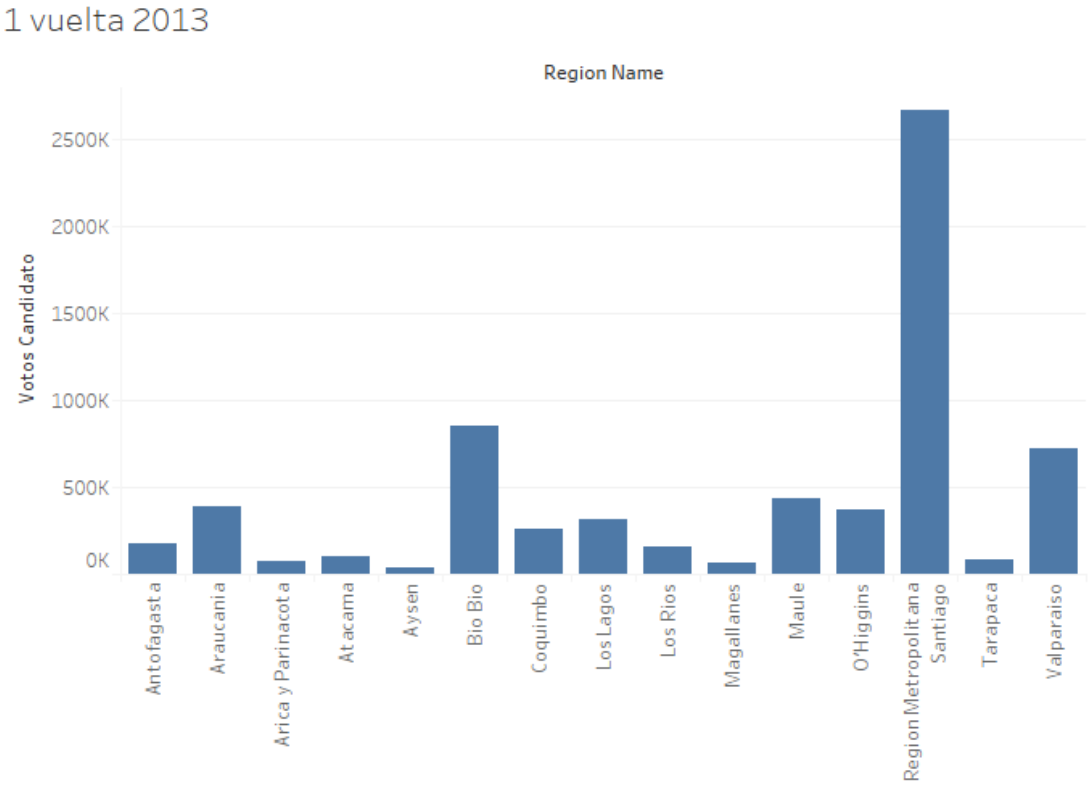

Figura 1.6 - Votos primera vuelta, agrupados por región (Presidenciales 2013).

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

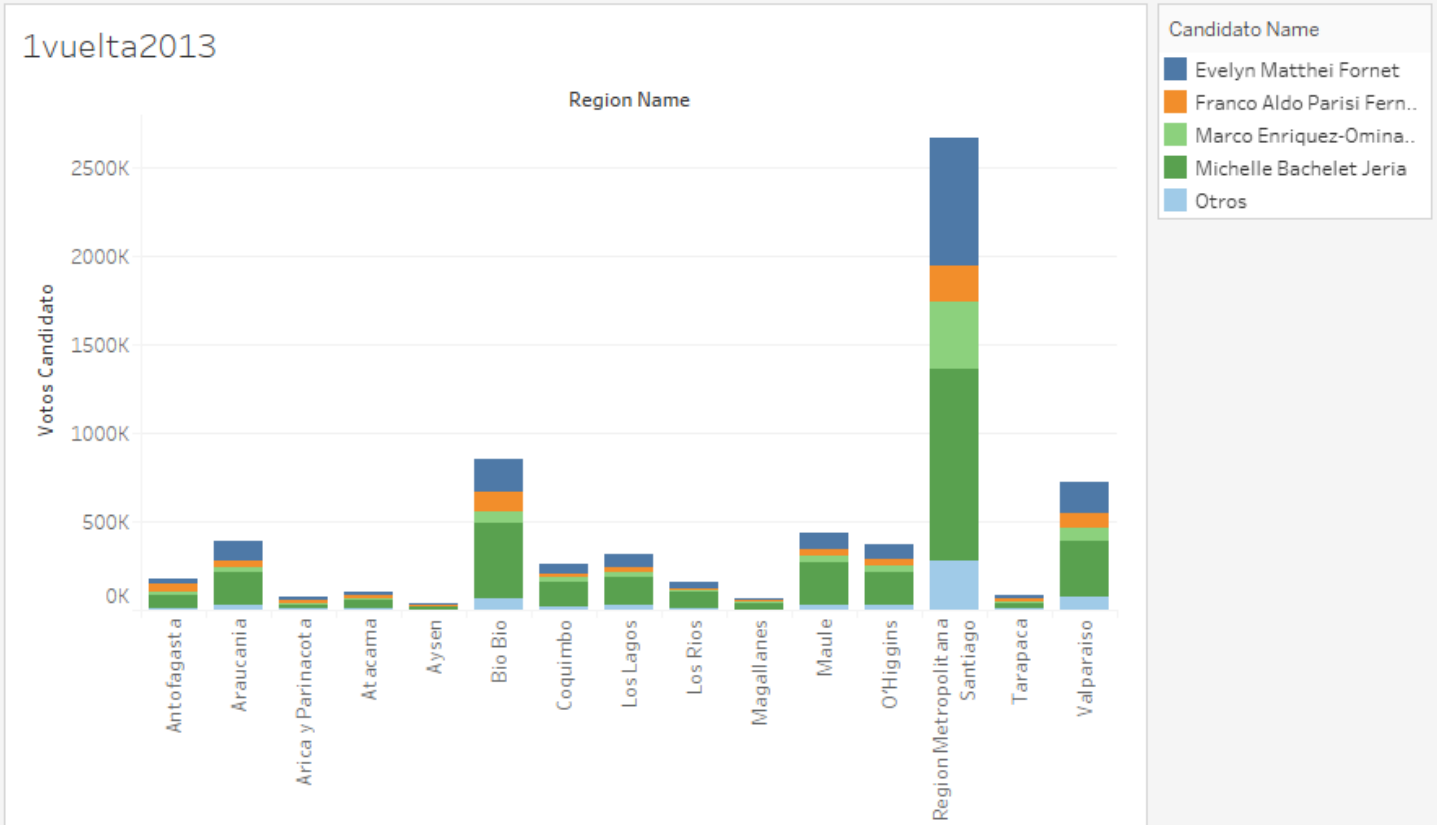

Figura 1.6- Votos primera vuelta, agrupado por región y candidatos (Presidenciales 2013)

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.

Dado que el objetivo del presente proyecto es analizar si Santiago es una muestra representativa 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 tienen el mismo alto, y de esta forma es posible identificar la porción 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 1.7- Colores de los candidatos presentes en los gráficos de 1° vuelta 2013

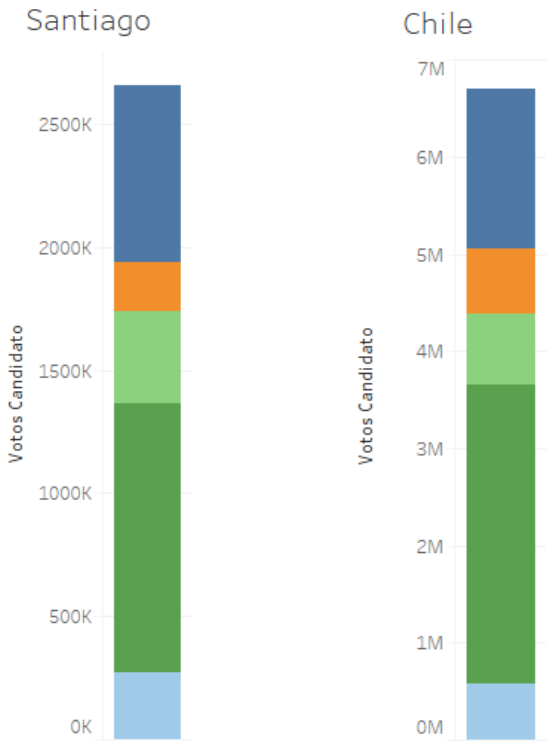

Figura 1.8- Comparación Santiago con votaciones totales (1° vuelta 2013)

Al comparar la región Metropolitana con las votaciones totales, nos damos cuenta que en Santiago la candidata Evelyn Matthei (color azul oscuro) porcentualmente tiene más adeptos respecto al total de las votaciones, y Michelle Bachelet (color verde oscuro) tiene menos adeptos porcentualmente respecto al total.

Nos damos cuenta que hay cambios entre los distintos candidatos, lo que a modo preliminar la región metropolitana no es representativa de lo que pasa en el resto del país.

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 1.7

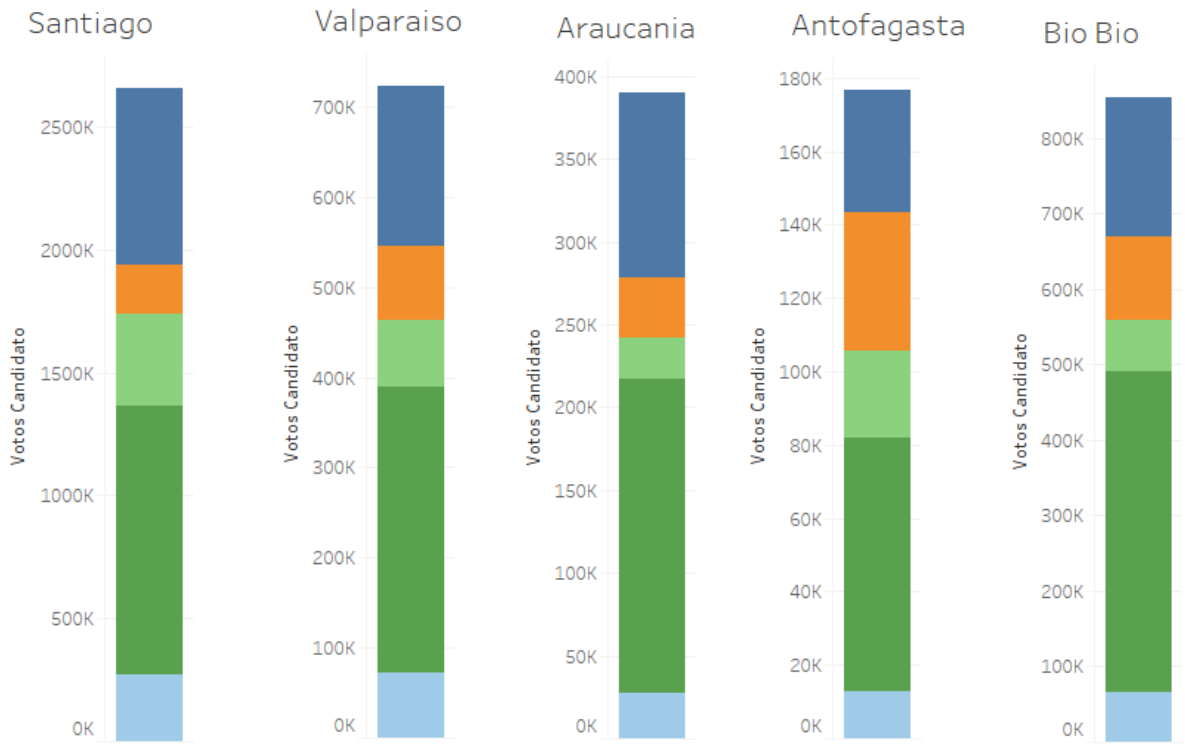

Figura 1.9- Comparación Santiago con distintas Regiones (1° vuelta 2013)

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.

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.

Para el caso de la segunda vuelta 2013, se obtienen los siguientes gráficos.

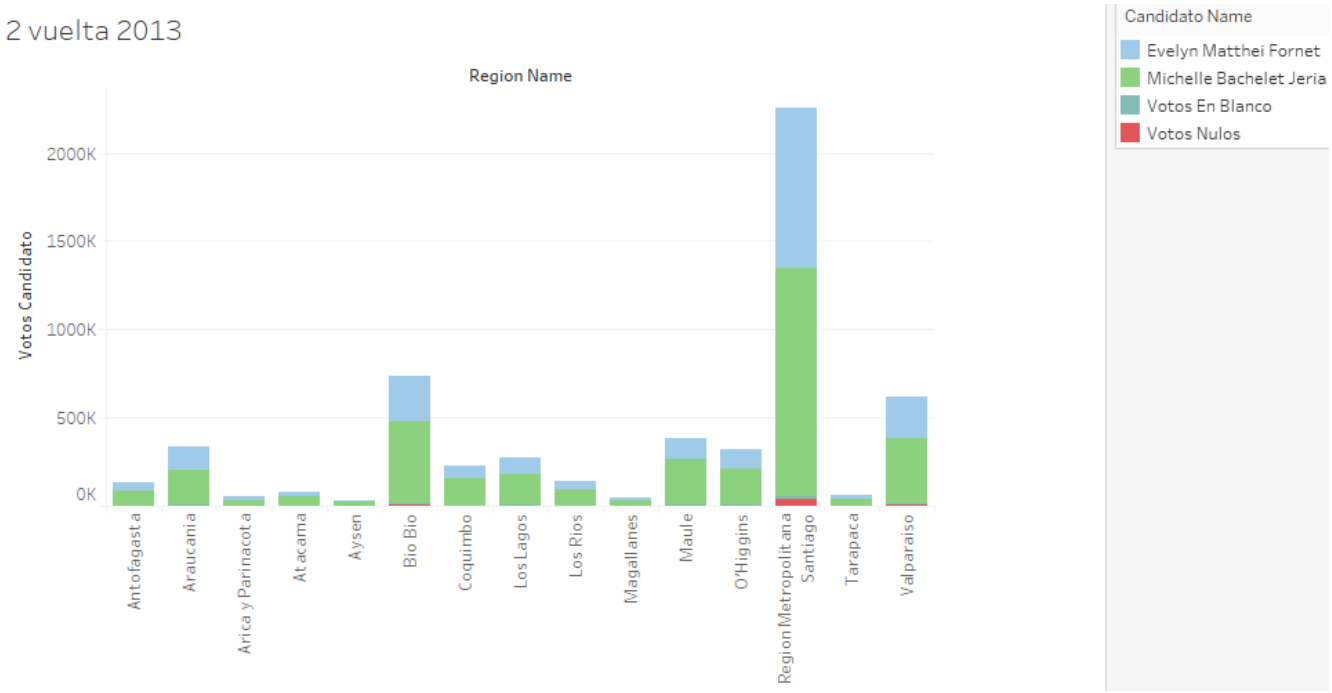

Figura 1.10- Votos segunda vuelta, agrupado por región y candidatos (Presidenciales 2013)

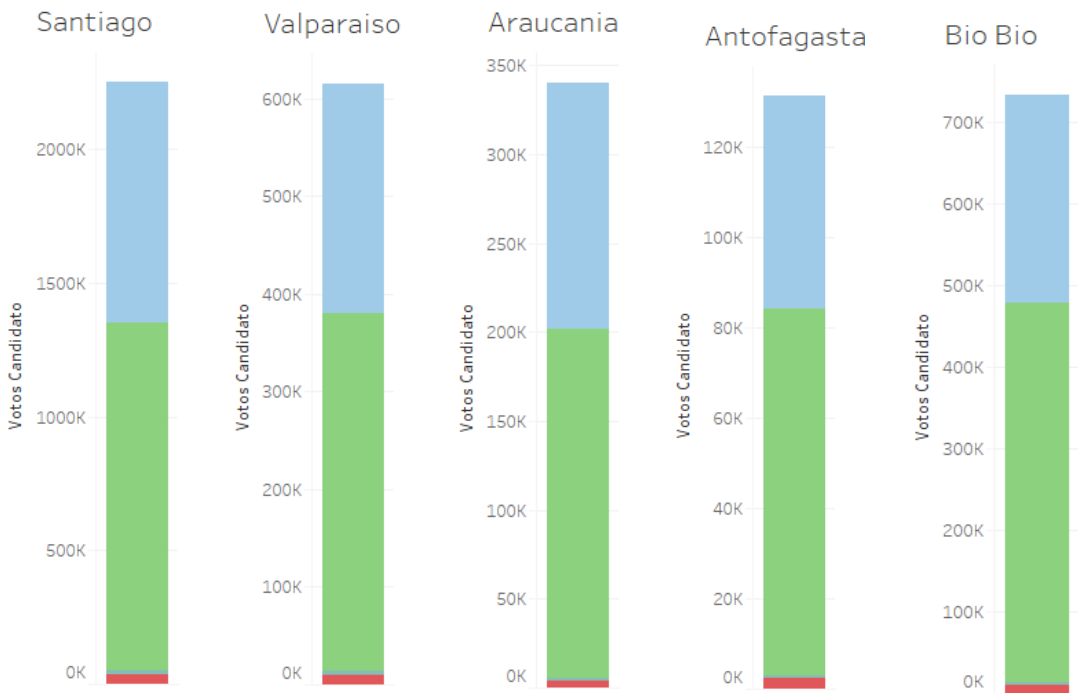

Figura 1.11- Comparación Santiago con distintas Regiones y total. 2 vuelta 2013

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 Valparaíso y de la Araucanía.

Conjeturas/decisiones y supuestos emanados del AED

A lo largo de la realización del AED se pudieron notar 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.

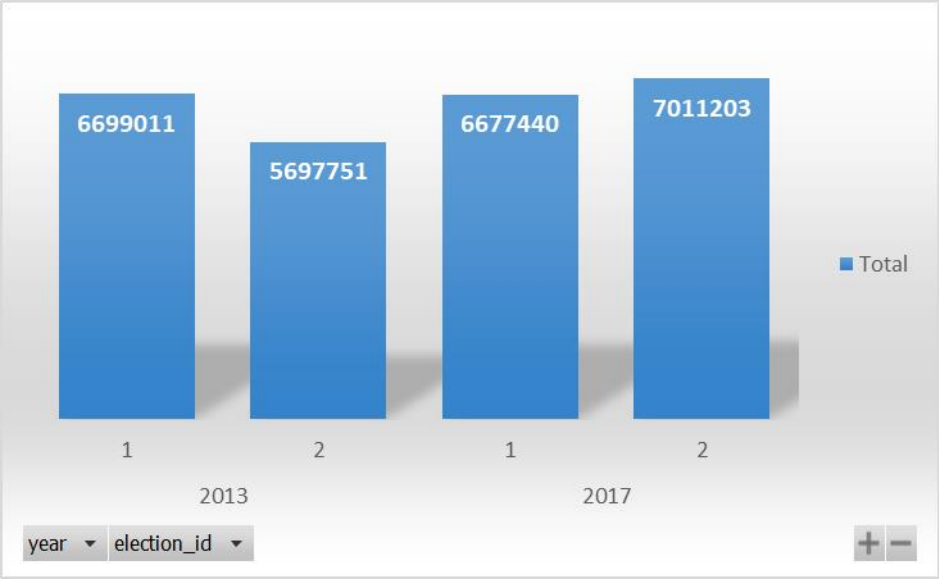

Figura 1.13- Total de votos por año y vuelta.

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.

En segundo lugar, al realizar el análisis comparativo de las preferencias en diferentes regiones con Santiago (como se puede observar en la Figura 1.8), se puede notar a priori que Santiago no es una muestra representativa para el escenario nacional en su conjunto, ya que, para el caso, Evelyn Matthei cuenta con una preferencia considerablemente mayor que la presenta como la competidora directa de Michelle Bachelet en la Región Metropolitana, cuando en el resto del territorio nacional se comporta de una manera totalmente opuesta.

A pesar de lo anterior, si se realiza un análisis mucho más acabado y se compara específicamente la Región Metropolitana con una región en particular como en el caso de la Figura 1.12, en la cual se puede observar una semejanza con la región de Valparaiso respecto de las preferencias de votación, a pesar de ello, al realizar un análisis de manera global se puede notar que la Región Metropolitana no es totalmente representativa del territorio nacional.

Posteriormente, dada las características que presenta el *dataset* empleado en el desarrollo del proyecto, se realizó un análisis de correlación entre la población y la cantidad de votos realizados por cada elección y vuelta correspondiente, que se encuentra presentado en la Tabla 2, inicialmente, esta indica que a un aumento de la población sea para cualquiera de las elecciones, también aumentará la cantidad de votantes, dado que todas las correlaciones son cercanas a 1.

Finalmente, para concluir los diferentes análisis mencionados anteriormente no solo se encuentran afectados por diversos supuestos y factores externos a los mencionados en ellos. Uno de estos supuestos, fue considerado para realizar el análisis de las correlaciones, dado que fue necesario recurrir a fuentes externas de información, que para este caso fue el Censo (INE) que permitiera obtener un dato confiable respecto de las proyecciones poblacionales actuales, junto con ello

cabe destacar que además este análisis considera que el aumento de la población a lo largo de los años no ha variado de manera significativa, ya que la población utilizada para los años 2013 y 2017 corresponde a los datos obtenidos mediante la realización del último Censo realizado.

Junto con lo anterior también existen otros factores externos que pueden afectar las conclusiones emanadas de los indicadores obtenidos con anterioridad, un ejemplo de ello es el cambio de la Ley de inscripción automática y voto voluntario, ya que no exige que, dado un aumento de la población, exista con él, un aumento de la cantidad de votantes dado toma la característica de ser "*opcional"* para cada persona.

Conclusiones

Uno de los desafíos del análisis exploratorio de datos, es que la mayoría de los campos de nuestro datasets son de carácter cualitativo: nombre candidato, nombre de región, nombre de comuna, candidatos. Y por otro lado los datos cuantitativos son solamente los votos realizados. La situación anterior dificulta encontrar un método que permita desarrollar la pregunta de investigación, ya que se complica la comparación de los datos. Además, si se considera que se tiene los registros de dos años, es difícil sino imposible encontrar tendencias de comportamiento.

Al plantear la pregunta de investigación: *¿Es Santiago una muestra representativa de las tendencias políticas de todo el territorio nacional?*, existe la necesidad de establecer un criterio de comparación, para lo cual resulto pertinente comparar a través de porcentajes el comportamiento de las regiones y contrastarlo con Santiago. Este tipo de análisis se realizó de una manera visual por medio de gráficos, con los que fue posible identificar comportamientos porcentuales parecidos entre candidatos para distintas regiones, siendo la región de Valparaiso la que presento mayor semejanza con Santiago. Con este primer acercamiento es posible tener luces de la dirección que apunta la investigación, pero es necesario emplear una metodología para poder cuantificar el grado de semejanza entre las regiones, ya que pueden existir otros aspectos no considerados en el estudio como, por ejemplo, la geografía, ya que Valparaiso a pesar de ser región se encuentra considerablemente mucho más cerca que otras capitales regionales.

Una de las interrogantes durante la realización del análisis exploratorio de datos, es*: ¿hasta qué nivel de profundidad se debe desarrollar el análisis?* Esto se explica dado que se tienen cerca de 10.000 registros, y hay muchas formas de agrupar los datos, existiendo distintos criterios de selección, distintos tipos de gráficos, distintos tipos de comparaciones, etc. Esto nos lleva a iterar muchas veces en el AED, obteniendo cada vez más información respecto a la problemática, y que muchas veces puede ser redundante, pero se debe decidir qué información aporta realmente al desarrollo de la investigación. Uno de los criterios usados en la iteración del AED, fue la generación de valor de la información y hasta qué punto se genera valor en el análisis, y de ser pertinente se incluye como resultado.

Por último, parte del desarrollo del AED y dada la característica de los datos, se utilizó información obtenida por medio de fuentes externas con la finalidad de robustecer de mejor manera el análisis exploratorio. Específicamente, se utilizaron datos referentes a la magnitud de la población del país, obtenida por medio del último Censo y bajo el supuesto de un *crecimiento constante* de la misma, considerando que para el año 2013 no existe una estadística oficial sobre el aumento de la población desde el último censo (2002). Lo anterior si bien solo se sustenta en un supuesto, permite robustecer las conclusiones y suposiciones que podrían emanar a partir del desarrollo AED, ya que, dada las características de este, la inclusión de diversas variables cuantitativas permitiría el uso de nuevos modelos que permitan responder de manera más certera nuestra pregunta de investigación.

Bibliografía

1) Datos de Población Censo 2017, obtenidos en: https://resultados.censo2017.cl/Home/Download