Pontificia Universidad Católica de Valparaíso
Escuela de Ingeniería Industrial

Data Science Aplicada
Entregable 2

Informe 1:
Extract, *Transform*, *Load* (ETL)

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

Índice

1. Introducción

A lo largo de los años el espectro político no ha estado exento de diversos escándalos de diferentes tipos que abarcan desde temas como la corrupción, uso de información privilegiada, entre otros. Parte de estos diversos escándalos han llegado a poner en duda la correcta realización de diversos instrumentos utilizados para sondear las preferencias ciudadanas, como lo son las encuestas.

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 instrumento 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, y de esta forma generar un aporte en el camino de la transparencia lo cual es fundamental, ya que además de generar expectativas en la población estos resultados pueden determinar el futuro político de un periodo de gobierno, y junto con ello el futuro de un 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. Para ello fue determinante enfocarse en el objetivo del análisis, planteando la siguiente interrogante:

***¿Es Santiago una muestra representativa 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,TarapacÃi,226,1401,1204,Pozo Almonte
3	1,TarapacÃi,217,1405,1203,Pica
4	1,TarapacÃi,113,1101,1201,Iquique
5	1,TarapacÃi,108,1404,1206,Huara
6	1,TarapacÃi,58,1403,1210,Colchane
7	1,TarapacÃi,26,1402,1208,CamiÃ±a
8	1,TarapacÃi,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 de 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, notamos 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.

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.

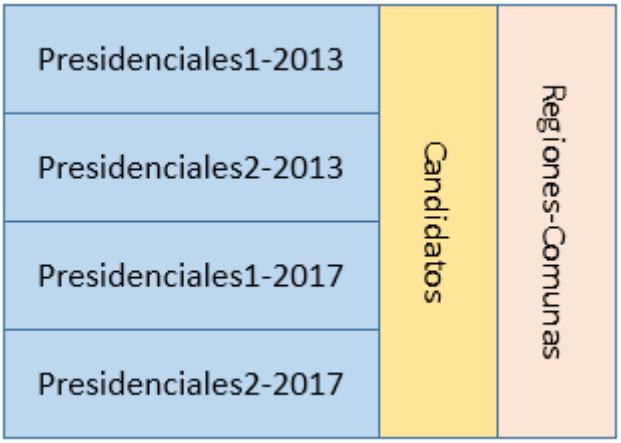

Figura 2.2.1 Diagrama de la estructura del ETL

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.

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, o algún otro problema.

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).

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.

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

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

5. Conclusiones

En primera instancia, se procedió a realizar una profundización del diseño del Data Warehouse, dado los comentarios realizados por el profesor se incluyeron una serie de subcarpetas y archivos para complementar el trabajo realizado, donde se encuentran: artículos relacionados sobre encuestas políticas (como el artículo de la CEPAL sobre la crítica a la CASEN), manual de gestión de proyectos, códigos de R, datasets, archivos README en formato txt para la explicación de estos, etc.

Por otra parte, se construyó el ETL, generando una matriz única que incluye todas las votaciones de las elecciones presidenciales tanto primera como segunda vuelta de los años 2013 y 2017, el procesamiento de datos se hizo de tal forma, que se eliminó ciertas variables que no generarán valor para el análisis (como es el caso de los ID de partidos políticos asociados a cada candidato), y agregando otras como el nombre de las comunas que tienen asociado su ID respectivo, en caso alternativo, se generó una segunda ETL que saca algunas variables como el nombre de los candidatos, comunas y regiones, en caso hipotético de que la gran cantidad de variables con su formato String puedan afectar los tiempos de procesamiento.

Con este trabajo desarrollado, se ha creado un único dataset, que se encuentra listo y dispuesto a ser procesado para realizar el primer análisis exploratorio tanto global como por cualquiera de las variables consideradas por el equipo: candidatos, comunas, años, etc. Por lo tanto, el equipo Politycs está en condiciones para comenzar la siguiente fase.

6. Bibliografía (si usaron alguna en este informe, eliminar si no usaron)

¿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