Informe avance Data Science "Grupo Politics".

Metodologías utilizadas para la construcción del ETL:

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.

Criterios de selección o eliminación de filas o columnas:

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 de app 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.

*****La nomenclatura de los resultados es la siguiente: Presidenciales*Numero vuelta*-Año.csv

Diseño del *dataset*:

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.

Finalmente, el archivo ETL tiene 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.