Proyecto final HIGI: base de datos del INE

Daniela Jiménez Pano
NUA:427292

17 de junio de 2021

Resumen

En el presente documento se reportará el desarrollo y resultados del proyecto final para la asignatura de herramientas informáticas y gestión de la información, el cual solicitó el análisis estadístico de dos bases de datos, una que encontraramos los alumnos por cuenta propia y otra que nos indicara la profesora. Estos estudios fueron realizados en python, herramienta nueva que aprendimos el último parcial del semestre. Sin embargo, este documento mostrará solamente el análisis de la base de datos que indicó la maestra de la asignatura.

1. Introducción

Para el proyecto final de la materia mencionada, se pidió analizar dos bases de datos, los primeros datos obtenidos de la página del Instituto Nacional Electoral, estos llamados: Estadísitica de Padrón Electoral y Lista Nominal de Electores [1]. Un análisis estadísitico se basa en recopilar, explorar y presentar cantidades de datos en los que se descubren ciertos patrones y tendencias. [2] El objetivo del análisis de este conjunto de datos fue obtener la predicción del número de personas dentro de la lista nominal 2 meses después de los últimos datos obtenidos, además de encontrar cómo afectaba el incremento (tasa de crecimiento/variación) de este número de personas para cada una de las secciones de los municipio del estado de Guanajuato.

2. Desarrollo

2.1. Primer análisis: Predicción del número de casillas y lista nominal.

Para esta actividad se llevaron a cabo ciertos pasos, lo primero a realizar fue descargar el conjunto de datos desde el mes de septiembre de 2019 hasta diciembre de 2020. Este análisis ya había sido realizado con anterioridad con ayuda de la herramienta Excel, sin embargo, la manera de trabajar en python no es tan cómodo trabajar debido a que sus comandos no son tan accesibles como en excel. Como los documentos venían organizados por estado además de involucrar los municipios y secciones de estos, se tenían que filtrar cada uno de los documentos para que pudieran organizarse solamente por el estado de nuestro interés, es decir, Guanajuato, para obtener la información específica de este:

```
data3=pd.read_csv(files[2],usecols=[0,1,2,3,9])
data3_GTO=data3[data3['ENTIDAD']==11]

```

Donde en cada uno de los datos con la terminación "GTO.^{er}an los datos a utilizar para llevar a cabo el análisis anteriormente mencionado, en cada uno de estos datos se eligieron las columnas que nos brindaran la información específica a utilizar, en este caso, la última columna definía el total de la lista nominal de cada municipio. Además, como se muestra en el código, la palabra 'ENTIDAD' fue de suma importancia para poder filtrar los datos y el número 11 especificaba a Guanajuato.

Y cada uno de los documentos fueron agrupados por muncipio, de manera que nos ayudara a obtener un mejora acomodo de los datos, ya que también la asignación así lo pedía, con eso se utilizó lo siguiente en donde al final se imprimía cada dataframe para observar el resultado:

```
    LISTA_NAL_MPO=data1_GTO[1:].groupby(['MUNICIPIO']).sum()
    LISTA_NAL_MPO

```

Al terminar de filtrar cada documento, lo siguiente a realizar fue agrupar los documentos en orden debido a que los nombres de los documentos, como eran nombrados numericamente, no se acomodaban del menor al mayor, por lo tanto, fue necesario ejecutar un código que nos permitiera llevar a cabo eso y desplegarnos los documentos en el acomodados en el orden mencionado:

```
date=[]
date_=[]
files_=[]

for i,file in enumerate(files):
    date.append(re.findall(r'\d+',file)[0])


temp=sorted(range(len(date)), key=date.__getitem__)

for i in temp:
    date_.append(date[i])
    print(date[i],files[i])
    files_.append(files[i])

```

A continuación, se juntaron todos los dataframes creados a partir de cada uno de los documentos en una tabla para poder observar de una manera más clara cómo es que había un cambio en el aumento o decremento de cada uno de los meses con relación a la lista nominal y así ayudarnos a poder realizar la predicción deseada, es decir, cuántas personas estarían en la lista nominal para 2 meses después, en este caso para febrero de 2021.

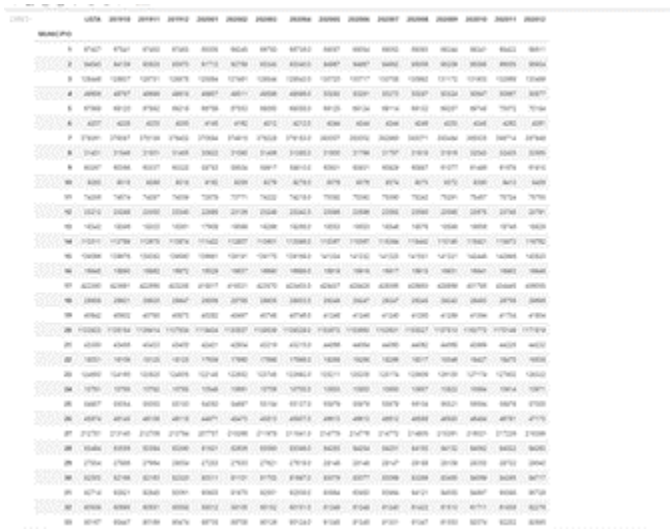

Figura 1: Tabla (dataframe) donde se acomodaron los datos de todos los documentos de la base de datos

Además, con ayuda de la librería matplotlib. se añadió una gráfica para poder ver el comportamiento de ciertos muncipios con respecto a la lista nominal, en este caso se pudo notar que los datos reflejaban un comportamiento lineal, por lo que sí era factible realizar un análisis correcto por mínimos cuadrados.

Ahora bien, para poder hacer la regresión lineal que se ha mencionado a lo largo del documento, se tuvo que convertir el dataframe en un array, esto para poder aplicar un ciclo for que permitiera

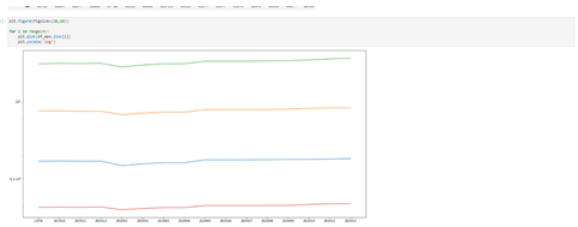

Figura 2: Graficación del comportamiento de la lista nominal por cada municipio.

al programa poder comprender qué datos estaba tomando en cuenta para poder hacer el análisis estadístico necesario:

```
municipios=np.asarray(df_mpo)
```

Cabe mencionar que la variable *dfmpo* muestra relación al dataframe creado a partir de los datos de los municipios. Teniendo esto, como el proyecto nos solicitaba observar el crecimiento (tasa de crecimiento/variación) de las personas en la lista nominal por cada una de las secciones de cada uno de los municipios de Guanajuato. Por lo que se decidió obtener la regresión utilizando un código que nos permitiera obtener la predicción en un año después por cada uno de los 46 municipios:

```
fits=[]
prediction_lnal=[]

for i in range(len(municipios)):
    xx=np.arange(len(municipios[i]))
    ma, ba = np. polyfit(xx, municipios[i],1,w=municipios[i])
    fits.append([ma,ba])
    pred=ma*(xx[-12]+12)+ba

    prediction_lnal.append(pred)
```

Donde los datos que muestan la variable municipios hacen una referencia (un poco obvia) a los datos de los municipios de cada documento. Después de haber ingresado el código anterior, se imprimió la predicción como una columna más dentro del dataframe que contenía la información de todos los documentos, permitiéndonos así visualizar la cantidad de personas en la lista nominal predicha.

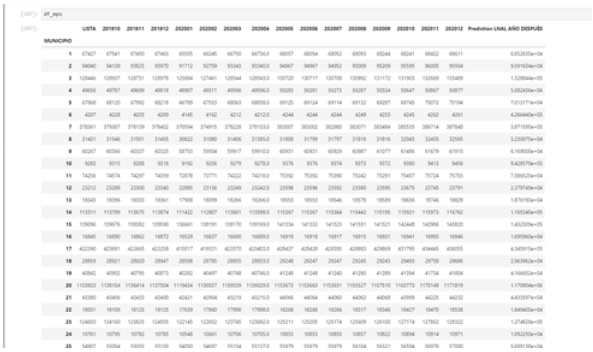

Figura 3: Predicción de la lista nominal para febrero de 2021

Finalmente para obtener el número de casillas a necesitar para el mes de febrero del año 2021 se sumaron cada una de las predicciones divididas en 750 (que es el número mínimo para abrir una casilla) y con eso finalizar el análisis estadísitico de esta base de datos.

3. Resultados y discusiones

Con los datos utilizados, se pudo llegar a que para el mes de febrero de 2021, es decir, unos cuántos meses después del último documento analizado, la cantidad de casillas a necesitar fue de 6038 casillas. Para obtener una vista más detallada de estos números, se adjunta la siguiente tabla:

Municipio	Número de casillas	Municipio	Número de casillas
1	91	14	155
2	128	15	191
3	177	16	23
4	68	17	579
5	94	18	40
6	6	19	60
7	516	20	1561
8	43	21	59
9	82	22	25
10	13	23	170
11	101	24	16
12	32	25	76
13	25	26	63

Cuadro 1: Predicción a Febrero de 2021: Primeros 23 municipios de Guanajuato.

Municipio	Número de casillas	Municipio	Número de casillas
27	287	40	18
28	113	41	67
29	38	42	156
30	112	43	21
31	127	44	64
32	83	45	11
33	123	46	85
34	6		
35	86		
36	9		
37	188		
38	14		
39	42		

Cuadro 2: Predicción a Febrero de 2021: Siguientes 23 municipios de Guanajuato.

Ahora, haciendo una comparación a la misma asignación que se realizó con la herramienta de excel, las casillas que nos arrojó por municipio fueron de 6829, con esto, se puede observar que la diferencia entre los resultados no es tan diferente y podremos aceptar la cantidad que python nos brindó.

Desafortunadamente para poder predecir y realizar una regresión lineal al apartado de secciones, debido a que se intentó aplicar un código que se utilizó anteriormente en el análisis, sin embargo, marcaba un error en el que la librería matplotlib. no permitía hacer el ajuste de mínimos cuadrados con los datos que se indicaron, sin embargo, se asume a que esto sucedió debido a que los datos de las secciones eran demasiados (el código fue marcado como comentario dentro de mi análisis).

4. Conclusión

Gracias a esta actividad, en lo personal pude observar qué tan importante es realizar este tipo de análisis ya que se pueden hacer proyecciones futuras acerca de datos tan importantes como lo son las listas nominales del Instituto Nacional Electoral. Además de que las herramientas que nos brinda python para este tipo de proyectos serán de mucha ayuda para proyectos futuros además de que el trabajo en este programa se hace de una manera más rápida.

Con respecto a los resultados obtenidos, en un principio me causó conflicto el hecho de que los valores de las casillas para los municipios fueran diferentes e al análisis realizado en excel, sin embargo, recordando el trabajo que hice, me percaté acerca de que se usó una función en la que se redondearan los valores de cualquier manera, por lo que provocó que el aumento de casillas fuera mayor a comparación

al presente proyecto. A pesar de esto, quedo satisfecha con los resultados ya que estos resultan bastante razonables.

Referencias

[1] Griselda Rojas Peña. Estadística de padrón electoral y lista nominal de electores. urlhttps://www.ine.mx/transparencia/datos-abiertos//archivo/estadistica-padron-electoral-lista-nominal-electores, 2020.

[2] TechTarget. Análisis estadístico. urlhttps://searchdatacenter.techtarget.com/es/definicion/Analisis-estadistico, s.f.