Proyecto final de Herramientas Informáticas y Gestión de la Información

Hernandez Serrano Oscar Kariel: ok.hernandezserrano@ugto.mx
NUA:427340

División de Ciencias e Ingenierías, Campus León, Universidad de Guanajuato

(17 de Junio de 2021)

I Introducción

Con la finalidad de fomentar y reforzar lo aprendido a el estudiante se le solicitó realizar el análisis de datos de dos distintas bases de datos está vez en el lenguaje de programación de Python: la primera fue una actividad que se había realizado con anterioridad; calcular el número de casillas que eran necesarias en el estado de Guanajuato a partir de datos que proporcionaba el INE y la segunda el alumno debia de seleccionar una base de datos de acuerdo a su criterio.

II Base de datos sobre los investigadores activos en el SNI durante el periodo 2012 a 2018

II.1 Motivación

Se ingresó a la base de datos del Gobierno de México[[1]], el principal objetivo era analizar la deserción escolar a nivel universitario, sin embargo los datos que proporcionaba la SEP tenían una diferencia de 5 años por cada periodo, por lo que realizar un seguimiento no era del todo fiable para determinar que alumnos pertenecían al periodo anterior. Posteriormente se encontró en la misma página, las bases de datos que proporciona el CONACyT, una base de datos que proporciona es acerca de los investigadores que se encontraban activos durante el periodo 2012 a 2018 en el SNI.

II.2 ¿Qué se realizó?

Se descargaron 7 archivos, cada uno representaba un año del periodo 2012 a 2018. Los códigos que se mostrarán de aquí en adelante solo pertenecen al año 2012, sin embargo en el archivo de Python (HIGI_HernandezSerrano_ProyectoFinal.ipynb) Se procedió a cargar el archivo que venía en formato csv, pero al compilar no se detectaba, por lo cuál se tuvo que agregar la dirección del archivo de la siguiente manera:

[language=Python] C:/Users/oscar/Downloads/HIGI$_{H}ernandezSerrano_{P}royectoFinal/Proyecto_Final_Investigadores_$
$pd.read\_csv(path + ' 2012.csv')$

Al realizar esto no se podía leer en su totalidad el archivo en Jupyter. Se abrió el archivo en Excel en donde marcaba un error de lectura de datos, los cuales fueron eliminados sin afectar la información esencial que sería utilizada por el estudiante. Una vez resuelto el problema se podia visualizar la tabla de datos:

	Nobilis	Apellido paterno	Apellido materno	Nombre	Nivel	Institución	Área
0	DR	MORENO	DE ALBA	JOSE GUADALUPE	3.0	CENTRO DE INVESTIGACION Y ESTUDIOS SUPERIORES ...	4
1	DR	PASCUAL	BUXO	JOSE	3.0	CENTRO DE INVESTIGACION Y ESTUDIOS SUPERIORES ...	4
2	DR	VERMA	JAISWAL	SURENDRA PAL	2.0	UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO	2
3	DR	ALONSO	SANCHEZ	JORGE	3.0	INSTITUTO DE ECOLOGIA, A C	2
4	DR	AZAOLA	GARRIDO	ELENA	3.0	UNIVERSIDAD AUTONOMA DE SAN LUIS POTOSI	1
...	...	...	...	...	...	...	...
18549	DR	DOMINGUEZ	OVIEDO	AGUSTIN	NaN		
18550	DR	HUITZIL	MELENDEZ	FIDEL DAVID	NaN		
18551	MTRO	KINDL	SIN INFORMACIÓN	OLIVIA SELENA	NaN		
18552	DR	PLASCENCIA	VILLA	GERMAN	NaN		
18553	DR	REYES	LOPEZ	ALFONSO	NaN		

Figure 1: Tabla de datos original de acuerdo a cómo la realizó el CONACyT.

Todos los datos proporcionados no eran necesarios, por lo tanto se removieron los datos personales, usando el siguiente código:

[language=Python] df_2012.iloc[:, [0,4,5,6]]

	Nobilis	Nivel	Institución	Área
0	DR	3.0	CENTRO DE INVESTIGACION Y ESTUDIOS SUPERIORES ...	4
1	DR	3.0	CENTRO DE INVESTIGACION Y ESTUDIOS SUPERIORES ...	4
2	DR	2.0	UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO	2
3	DR	3.0	INSTITUTO DE ECOLOGIA, A C	2
4	DR	3.0	UNIVERSIDAD AUTONOMA DE SAN LUIS POTOSI	1
...	...	...	...	...
18549	DR	NaN	UNIVERSIDAD AUTONOMA DE NAYARIT	1
18550	DR	NaN	INSTITUTO POLITECNICO NACIONAL	3
18551	MTRO	NaN	UNIVERSIDAD JUAREZ AUTONOMA DE TABASCO	6
18552	DR	NaN	CORROSION Y PROTECCION INGENIERIA, S C	7
18553	DR	NaN	INSTITUTO NACIONAL DE ENFERMEDADES RESPIRATORIAS	3

18554 rows × 4 columns

Figure 2: Tabla con los datos que eran relevantes para el análisis de datos (Nobilis, nivel, institución y área).

Una vez aplicado el filtro dentro de la tabla de datos, se procedió a realizar una serie de tablas con relación a los datos seleccionados.

La primera tabla que se realizó fue una en donde se buscaba mostrar el número de investigadores de acuerdo a su preparación y en qué institución laboraban de acuerdo al año, para realizar esta tabla fue necesario emplear el siguiente código:
[language=Python] pd.crosstab(df_2012.Institución, df_2012.Nobilis)

Nobilis	DR	LIC	MED	MTRO	SIN INFORMACIÓN
Institución					
AALBORG UNIVERSITY	1	0	0	0	0
ABBOTT GMBH & CO KG	1	0	0	0	0
ADVANTICAL LTD	1	0	0	0	0
AGRICULTURE AND AGRI FOOD CANADA	1	0	0	0	0
ANGLIA RUSKIN UNIVERSITY	1	0	0	0	0
...	...	...	...	...	...
WASEDA UNIVERSITY	1	0	0	0	0
WASHINGTON STATE UNIVERSITY	1	0	0	0	0
WORLD HEALTH ORGANIZATION	1	0	0	0	0
YALE UNIVERSITY	1	0	0	0	0
YESHIVA UNIVERSITY	1	0	0	0	0

591 rows × 5 columns

Figure 3: Tabla con la relación entre las instituciones en donde labora un investigador y su respectivo titulo.

La segunda tabla proporciona la relación entre el nivel del SNI del Investigador y la institución en donde laboraba en el año correspondiente, para esto se utilizaba el mismo código que el anterior, pero cambiando las variables:

[language=Python] pd.crosstab(df_2012.Institución, df_2012.Nivel)

A partir de aquí, todos los años contenían las subsecuentes tablas, exceptuando el año 2014 debido a que no proporcionaba el área del investigador. Así que la tercera tabla relacionaba el área del investigador con su grado de estudios. Para realizar esta tabla y las siguientes, se uso de la misma manera la función «pd.crosstab» la tercera tabla quedó así:

Nivel	1.0	2.0	3.0	4.0
Institución				
AALBORG UNIVERSITY	1	0	0	0
ABBOTT GMBH & CO KG	1	0	0	0
ANTHONY NOLAN RESEARCH INSTITUTE	0	0	1	0
ASOCIACION PARA EVITAR LA CEGUERA EN MEXICO, I A P	2	0	0	0
AUSTRALIAN NUCLEAR SCIENCE AND TECHNOLOGY ORGANISATION	1	0	0	0
...	...	...	...	...
VITRO TEC FIDEICOMISO	1	0	0	0
WAKE FOREST UNIVERSITY	0	0	1	0
WASEDA UNIVERSITY	1	0	0	0
WORLD HEALTH ORGANIZATION	1	0	0	0
YESHIVA UNIVERSITY	1	0	0	0

440 rows × 4 columns

Figure 4: Tabla con la relación entre las instituciones en donde labora un investigador y su respectivo nivel en el SNI.

Nobilis Área	DR	LIC	MED	MTRO	SIN INFORMACIÓN
1	2875	31	19	78	0
2	3023	27	17	95	0
3	1813	22	9	69	1
4	2643	31	19	80	0
5	2641	23	19	64	0
6	2104	15	11	47	0
7	2647	28	16	87	0

Figure 5: Tabla con la relación entre el área del investigador y su respectivo grado académico.

La cuarta tabla relacionó los datos del Nivel de SNI del investigador y su área de estudio.

Finalmente la quinta tabla muestra la relación entre la institución y el área de investigación.

II.3 ¿Qué se obtuvo a partir del análisis de datos?

La cantidad de Investigadores activos en el SNI aumentó considerablemente al igual que las instituciones a las que pertenecían. En 2012 se contaba con un total de 18,554 investigadores y 591 instituciones participantes, para el año de 2018 dichas cantidades pasaron a ser 28,578 y 941 respectivamente. Por lo tanto hubo un crecimiento de 10,024 investigadores y 350 instituciones. Aunque el crecimiento en su mayoría fue lineal, había demasiadas inconsistencias en los datos, tal es el caso de investigadores que no contaban con nombre u otros que no proporcionaban el nivel y el área de estudio, este caso se empeoró en el 2014, año en el cual hubo más inconsistencias. A continuación se muestran las principales gráficas que se obtuvieron:

Área	1	2	3	4	5	6	7
Nivel							
1.0	1339	1772	1180	1529	1524	1241	1473
2.0	689	493	285	598	522	326	398
3.0	403	270	162	243	233	126	109
4.0	9	15	3	5	1	1	1

Figure 6: Tabla con la relación entre el nivel del SNI del investigador y su respectiva área de estudio.

Área	1	2	3	4	5	6	7
Institución							
AALBORG UNIVERSITY	0	0	0	0	0	0	1
ABBOTT GMBH & CO KG	0	0	1	0	0	0	0
ADVANTICAL LTD	0	0	0	0	0	0	1
AGRICULTURE AND AGRI FOOD CANADA	0	0	0	0	0	1	0
ANGLIA RUSKIN UNIVERSITY	0	0	0	1	0	0	0
...	...	...	...	...	...	...	...
WASEDA UNIVERSITY	0	0	0	0	0	0	1
WASHINGTON STATE UNIVERSITY	0	0	0	0	0	1	0
WORLD HEALTH ORGANIZATION	0	0	1	0	0	0	0
YALE UNIVERSITY	0	1	0	0	0	0	0
YESHIVA UNIVERSITY	0	1	0	0	0	0	0

591 rows × 7 columns

Figure 7: Tabla con la relación entre la institución del investigador y su respectiva área de estudio.

III Base de datos de la lista nominal del INE

III.1 Introducción

Dentro de la página del Instituto Nacional Electoral [2] se descargaron datos de "Estadística de Padrón Electoral y Lista Nominal de Electores", fueron utilizados los de la lista nominal desde septiembre de 2019 hasta diciembre de 2020, se utilizaron únicamente los datos del estado de Guanajuato (estado=11) y se desea predecir los de febrero de 2021.

III.2 Desarrollo

El análisi se realizó en Python, por lo que lo primero que se realizó fue importar las librerías que se usarían para el mismo.

Después de esto, se cargaron los archivos necesarios para realizar la predicción y se creó una variable que los almacenara a todos.

A continuación, se eligió uno de los archivos, del que se seleccionaron únicamente 5 columnas: "ENTIDAD", "DISTRITO", "MUNICIPIO", "SECCION Y "LISTA_NAL", este también se filtró por entidad, eligiendo únicamente la 11, correspondiente a Guanajauto. Esto fue realizado para analizar el orden y comportamiento de los datos.

Figure 8:

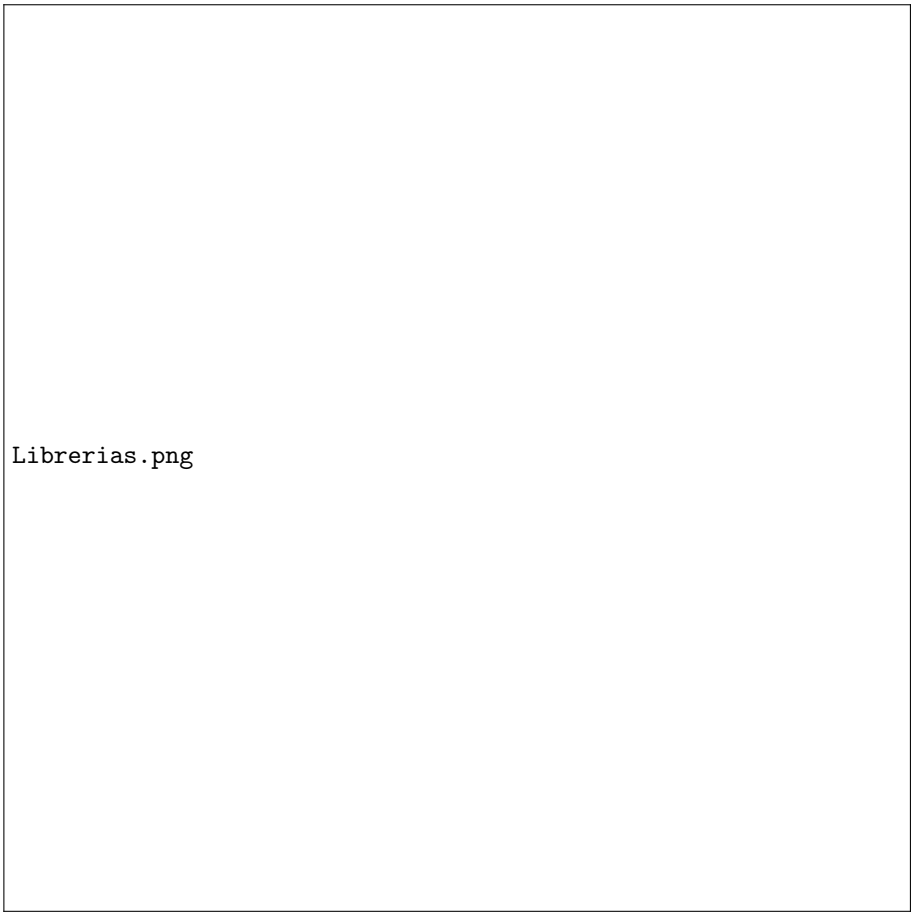

Figure 9: Importando las librerías.

Se elaboró un ciclo que le asignará a cada año, su archivo correspondiente. Primero, se agruparon los valores de cada fecha, en este caso de cada año, y se acomodaron cronológicamente, posteriormente, se le destino a cada una el archivo que coincidera con la fecha.

El siguiente paso fue graficar los datos para obeservar su comportamiento y comprobar que tuviera una tendencia lineal para poder aplicar una regresión lineal y ser capaces de realizar la predicción necesaria.

Se convirtió el DataFrame en un arreglo, de está manera podremos utilizarlo fácilmente en la función de la regresión. Volvemos a graficar para asegurarnos de que no hemos modificado los valores.

Se elaboró un arreglo que generó la regresión lineal por cada municipio y se guardaron los valores que arrojo la pendiente y ordenada al origen, se aplicó, utilizando la fórmula y=mx+b y guardamos los datos en la variable "predicción_lineal".

Se añadió la columna de predicción a la tabla general anteriormente.
Se realizó una suma con todos los valores dados por la predicción
.
Finalmente, se dividió entre 750.

IV CONCLUSIÓN

Se obtuvo un total de 6042.567323217931 actas, al igual que en la primera actividad del alumno por lo tanto se conisdera una cantidad que es correcta, pero no llega a la cantidad solicitada por la maestra y ya que se debe de ser un

Figure 10: Carga de los archivos ".txt".

número entero, resultaría en 6043 actas necesarias para febrero 2021.

References

[1] C. N. de Ciencia y Tecnología. Lista de investigadores activos en el sni del año 2012 a 2018. [Online]. https://datos.gob.mx/busca/dataset/sistema-nacional-de-investigadores.

[2] I. N. Electoral. Padrón electoral.

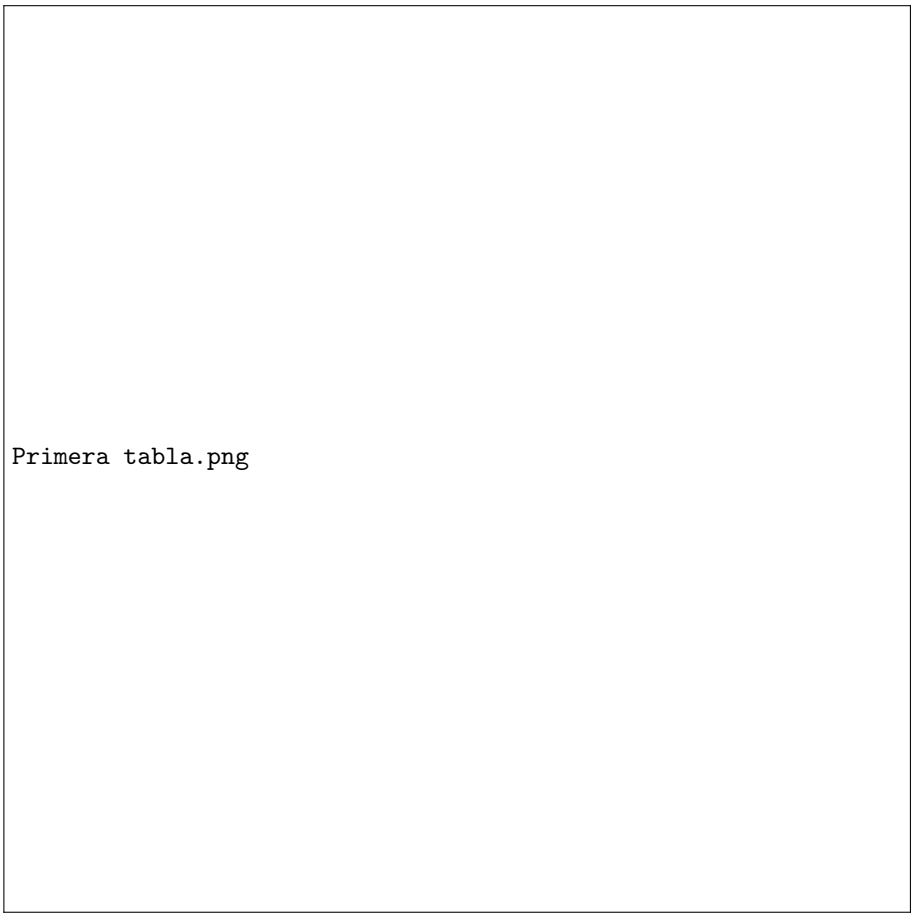

Figure 11: Creación de la primer tabla.

Figure 12: Ciclo de las fechas existentes en los archivos.

Figure 13: Grafica de 5 municipios.

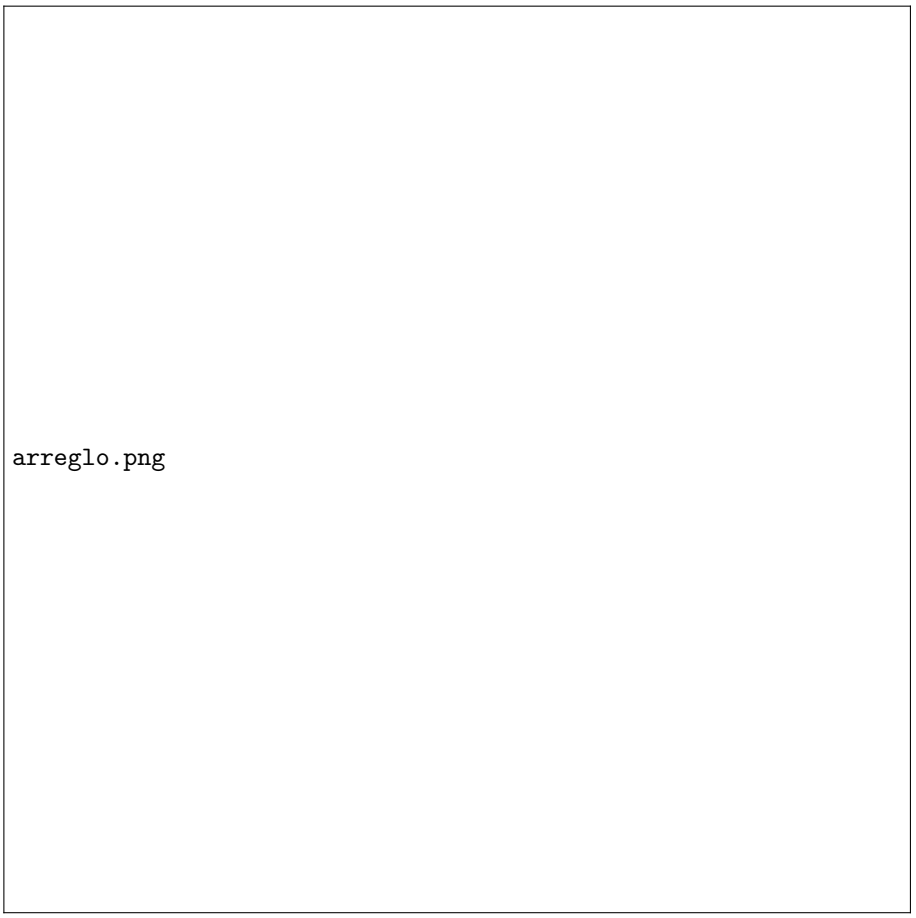

Figure 14: Conviertiendo nuestra tabla en un arreglo.

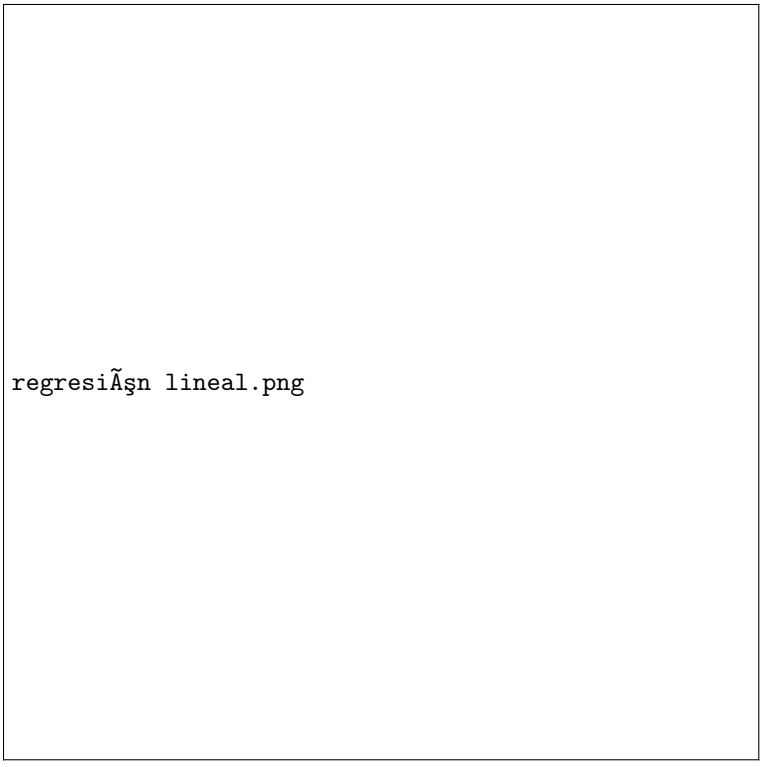

Figure 15: Regresión lineal por municipio.

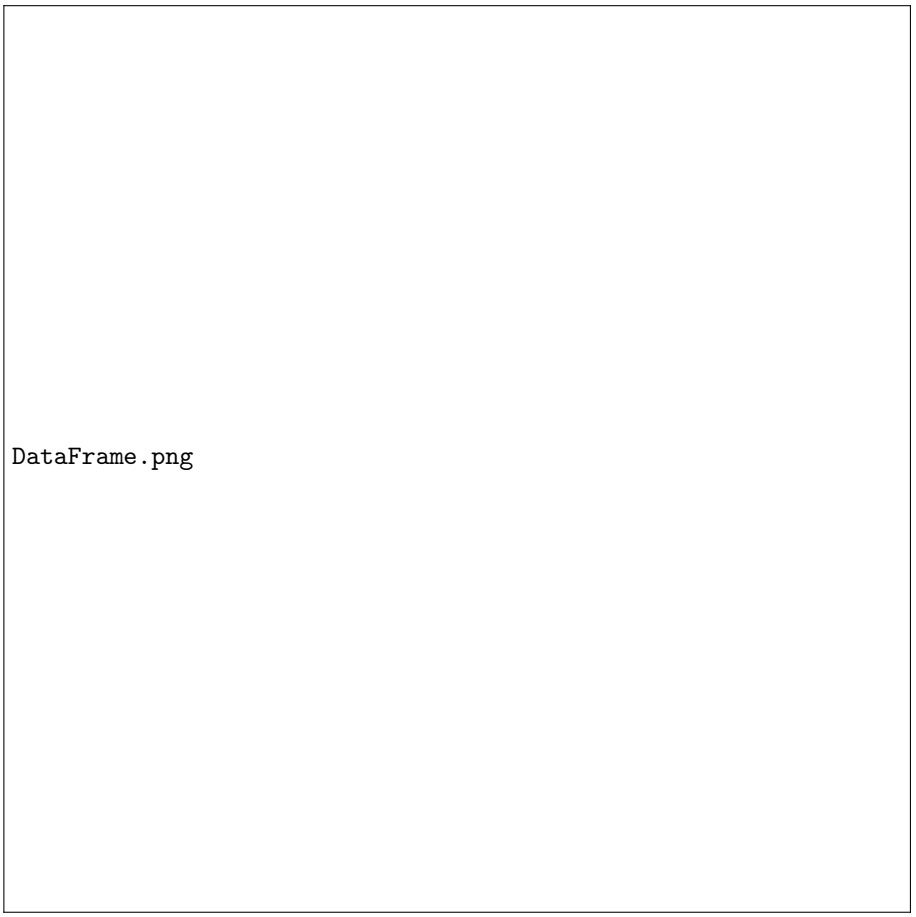

Figure 16: Añadimos la predicción al DataFrame

Figure 17: Suma de las predicciones.

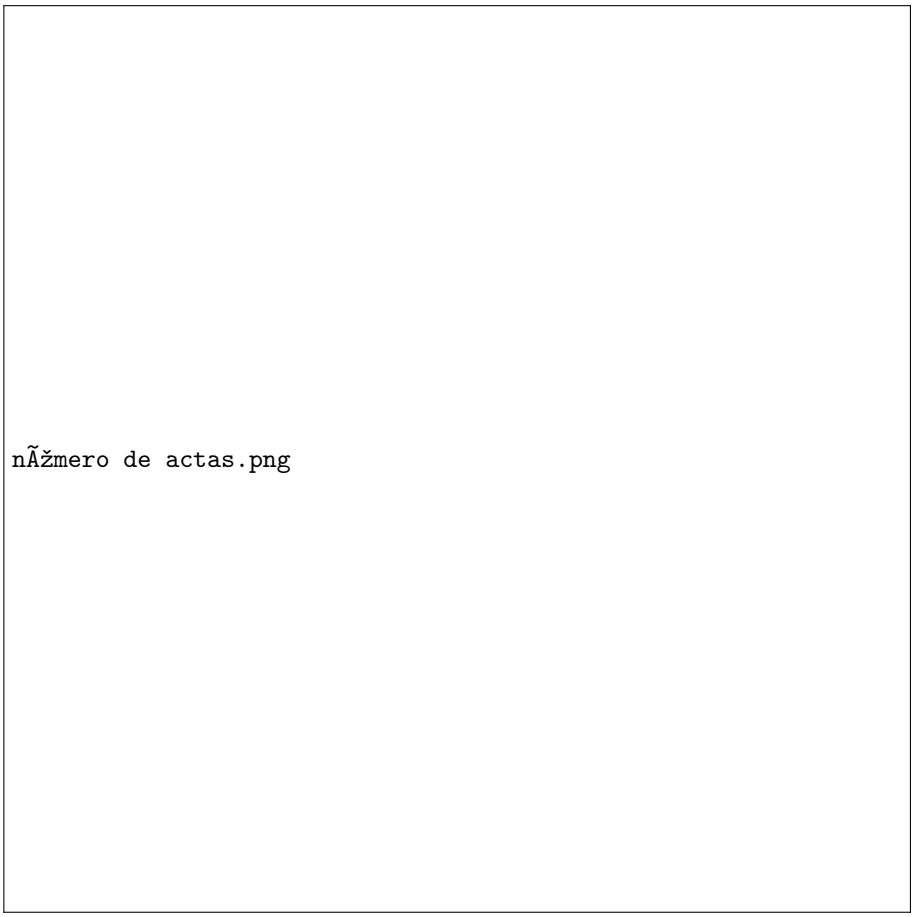

Figure 18: División para obtener el número de actas.