TP: Análisis exploratorio

Lucas Risaro - Juan Pablo Capurro

94335 - 98194

Preparado de los archivos para el análisis

Antes de comenzar con el análisis de los datos, se procedió a adaptar los archivos proveídos por Navent, con el objetivo de filtrar posibles datos que no fueran útiles para el análisis en sí o que pudieran ocasionar inconsistencias en el mismo.

Análisis de presencia de elementos nulos:

Procedimos a revisar la presencia de elementos nulos en los archivos. Se encontraron datos nulos en los archivos ‘postulaciones_genero_edad’ y ‘avisos_detalle’.
En el caso de ‘postulaciones_genero_edad’ la columna que indicaba la fecha de nacimiento de los postulantes presentaba elementos nulos, para corregir esto se procedió a rellenar esas casillas nulas con el año actual así al querer trabajar con la edad a lo sumo la resta para determinar no surgiría algún error respecto al tipo de los datos.
El archivo ‘avisos_detalle’ se encontraron datos nulos en las columnas 'ciudad', 'mapacalle' y 'denominacion_empresa'. Este inconveniente se subsano eliminando esas columnas además de la columna ‘descripción’, tomando como criterio que la información que proveían las mismas no era relevante para el posterior análisis de los datos, quedando de la siguiente manera:

	idaviso	idpais	titulo	nombre_zona	tipo_de_trabajo	nivel_laboral	nombre_area
0	8725750	1	VENDEDOR/A PROVINCIA DE SANTA FE	Gran Buenos Aires	Full-time	Senior / Semi-Senior	Comercial
1	17903700	1	Enfermeras	Gran Buenos Aires	Full-time	Senior / Semi-Senior	Salud
2	1000150677	1	Chofer de taxi	Capital Federal	Full-time	Senior / Semi-Senior	Transporte
3	1000610287	1	CHOFER DE CAMIONETA BAHIA BLANCA - PUNTA ALTA	Gran Buenos Aires	Full-time	Senior / Semi-Senior	Transporte
4	1000872556	1	Operarios de Planta - Rubro Electrodomésticos	Gran Buenos Aires	Full-time	Senior / Semi-Senior	Producción

En los demás archivos no se encontraron elementos nulos.

Filtrado de filas:

Analizamos todos los archivos para ver si poseían filas duplicadas, lo que indicaría la presencia de datos de usuarios o avisos duplicados en el Data Frame, los cuales harían que el análisis no fuera certero.
En ningún archivo se encontraron filas con estas características. Si, en cambio, en el archivo ‘postulaciones_educacion’ se encontraron filas con ids de postulantes duplicados varias veces. Esto indica que hay usuarios con diferentes tipos de estados de su educación. Se decidió dejar esos datos tal y como están, debido a que los usuarios pueden tener

diferentes estados de educacion, carreras terminadas y otras en curso o abandonadas en sus perfiles, y se consideraron datos válidos en el dominio.

Creación de nuevas columnas:

Se crearon nuevas columnas en algunos dataframes para poder manejar más cómodamente algunos datos como fechas , horas y la edad de los postulantes.
Un ejemplo de esto es el archivo ‘postulaciones’ al cual se le agregó dos nuevas columnas ‘date’ y ‘time’ para tener por separado la fecha y la hora, quedando de la siguiente manera:

	idaviso	idpostulante	fechapostulacion	date	time
0	1112257047	NM5M	2018-01-15 16:22:34	2018-01-15	16:22:34
1	1111920714	NM5M	2018-02-06 09:04:50	2018-02-06	09:04:50
2	1112346945	NM5M	2018-02-22 09:04:47	2018-02-22	09:04:47
3	1112345547	NM5M	2018-02-22 09:04:59	2018-02-22	09:04:59
4	1112237522	5awk	2018-01-25 18:55:03	2018-01-25	18:55:03

Escala de los datos usados en los gráficos

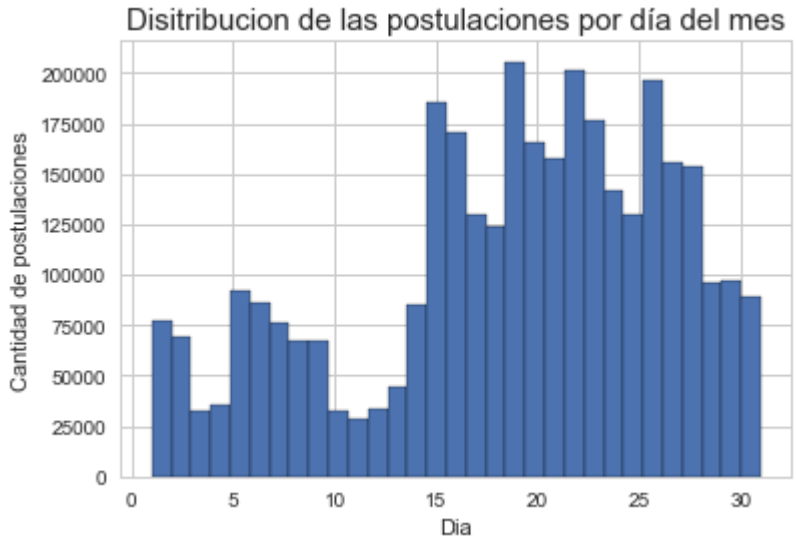
Algunos gráficos tienen una aclaración de (log) en el título, para indicar que se usa una escala logarítmica.

Análisis de las postulaciones

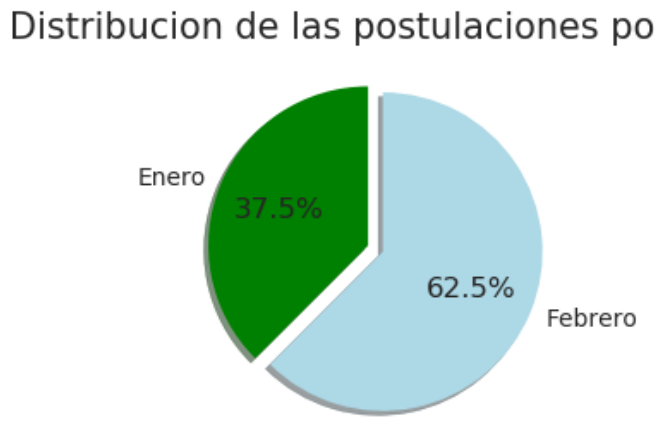
A continuación veremos cómo se distribuyen las postulaciones.

Postulaciones por mes:

A continuación se muestra la distribución de las postulaciones por día, tomando los dos meses juntos.



A continuación se presenta el porcentaje de las postulaciones de cada mes.

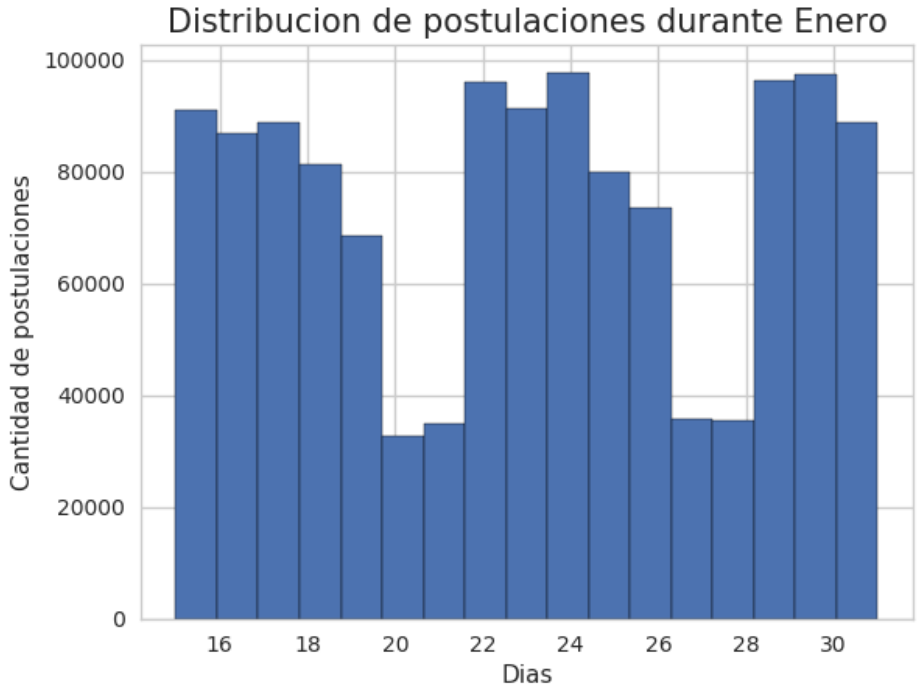


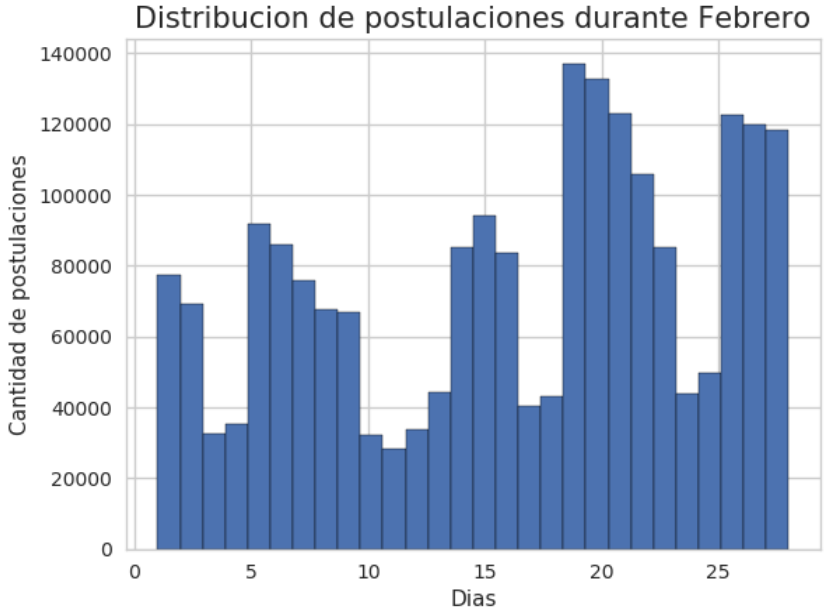
Visto en números la cantidad de postulaciones en Enero es de 1276198 y en Febrero 2125425.

En el primer gráfico a simple vista se podría observar que la cantidad de postulaciones es considerablemente mayor en la segunda mitad de los meses que en la primera, y que en el segundo Febrero supera en gran parte a Enero en cantidad de postulaciones por mes, pero es así realmente?

Los datos de los gráficos no son del todo ciertos debido a lo siguiente, solo se tienen datos de 17 días de Enero mientras que de Febrero tenemos los datos de todo el mes. Además esos 17 días de Enero corresponden a la segunda mitad del mes.
Es por esto que el primer gráfico muestra que en la segunda mitad de los meses se tiende a tener mayor postulaciones, y en el segundo gráfico se ven muchas más postulaciones en Febrero que en Enero.

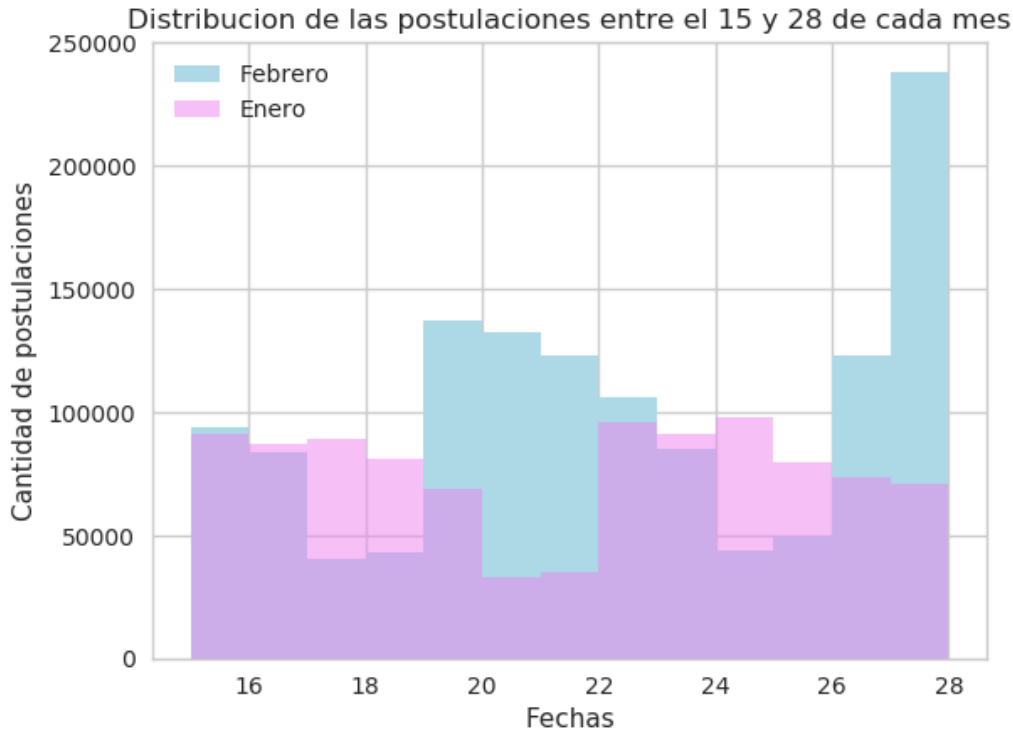
Ahora veamos la distribución de las postulaciones de cada mes por separado.



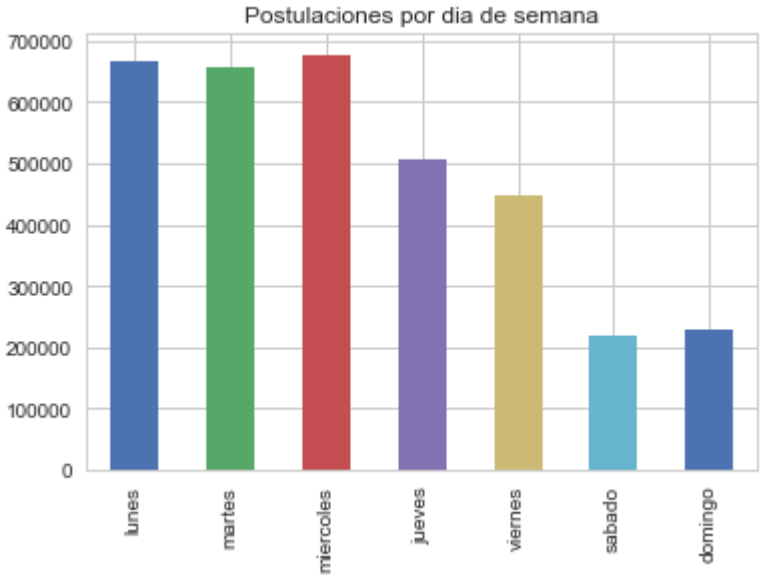


En estos gráficos se ve la diferencia de la cantidad de días de los cuales se tienen datos de cada mes.
Sin embargo aún no se puede hacer un análisis correcto de ambos meses juntos debido a que estamos teniendo en cuenta lapsos de tiempo diferentes.
Para hacer algo más cercano a la realidad y poder comparar los datos de ambos meses restringimos los datos de febrero solo a las fechas entre el 15 y el 28, inclusive, debido a que son las fechas de los días de Enero de los cuales se tiene información.

Veamos cómo queda entonces la distribución de postulaciones entre esas fechas para ambos meses:

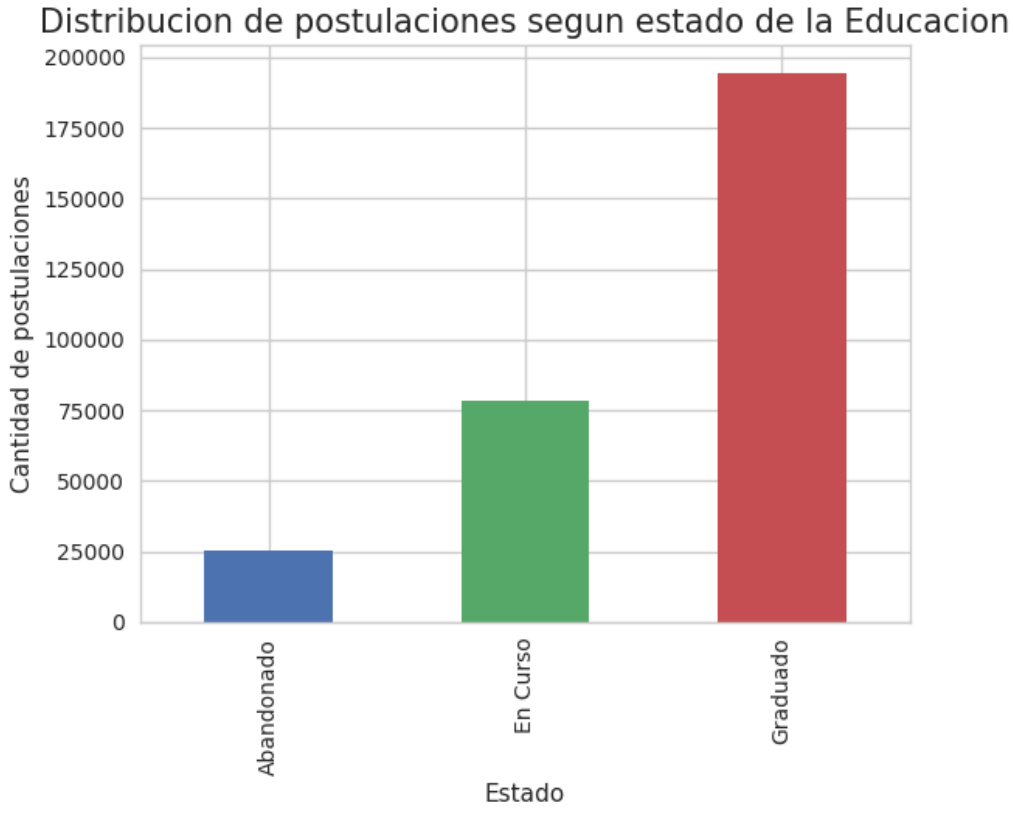


Ahora sí se puede tener una mejor comparación de la distribución de las postulaciones entre los dos meses, por lo menos entre el 15 y 28 de cada uno.
Se puede ver que predomina Febrero.
A modo de comentario esos 'baches' en los gráficos, corresponden a los fines de semana de cada mes. En febrero 17 y 18, y 24 y 25 fueron sábado y domingo respectivamente, lo mismo en el 20 y 21 de Enero. En el gráfico del mes de Febrero completo se aprecia bien este detalle.
También se aprecia que la pagina tiene bastante menos actividad durante los fines de semana:



Postulaciones según la educación de las personas:

A continuación veremos la distribución de postulaciones según estado de la educación y el tipo de carrera:

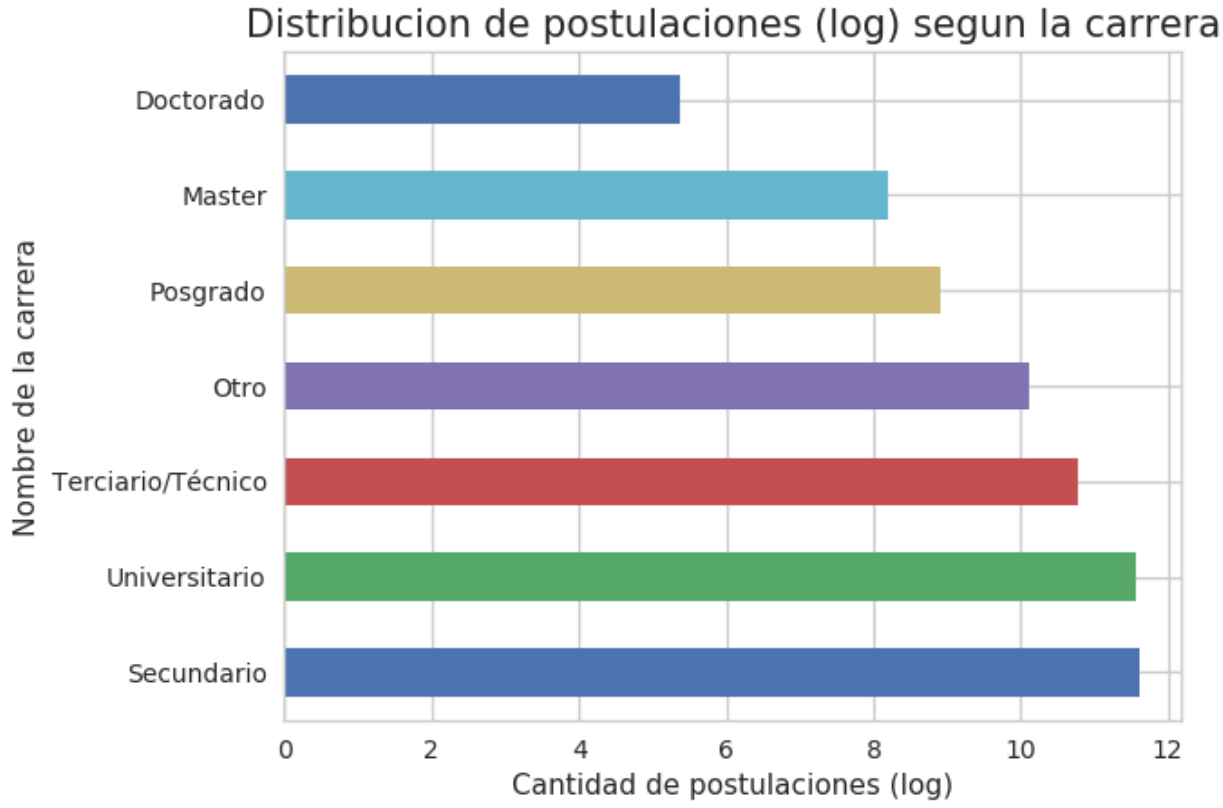


En números:

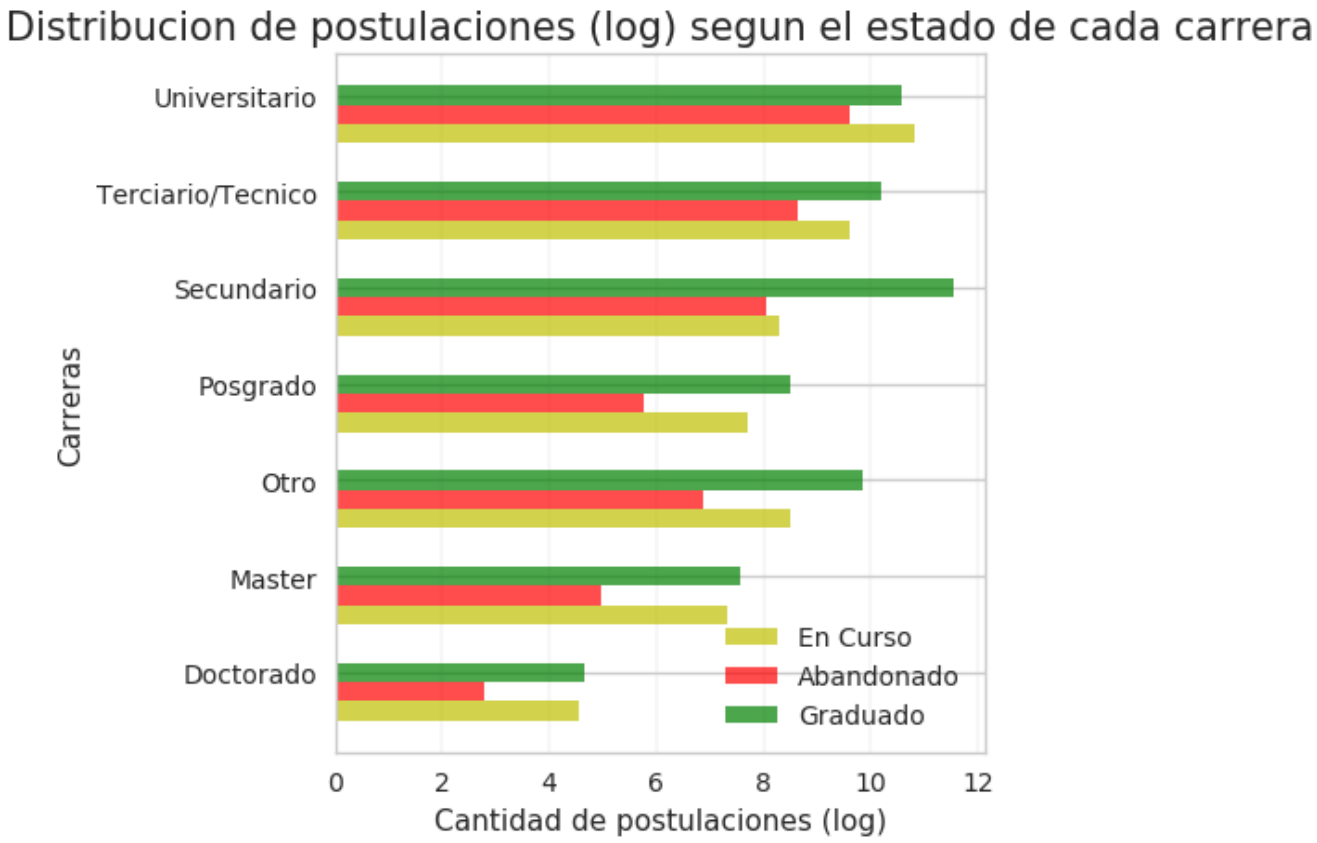
Estado	Cantidad
Abandonado	25226
En Curso	78531
Graduado	194474

Claramente se ve que las personas graduadas en general tienen más cantidad de postulaciones, lo cual es lógico debido a que tienen una mayor formación y su espectro de trabajos que pueden conseguir es más amplio que los de una persona que no terminó sus estudios o alguien que aún está en proceso de hacerlo.

El siguiente gráfico muestra las postulaciones según los niveles educativos de los postulantes:



Como era de esperar el grado con más postulaciones es el secundario.
Si suponemos que los postulantes ingresan todo su historial académico de las instancias que completaron no era posible que hubiera más postulaciones de universitarios, por ejemplo, o de alguna instancia superior, que de secundario, ya que para pasar a esas instancias primero deben terminar el secundario
Luego la diferencia entre las postulaciones de personas con secundario y con universidad puede deberse a que hayan decidido no seguir estudiando al terminarlo y empezar a trabajar. Esto lo podremos ver mejor en el siguiente gráfico.



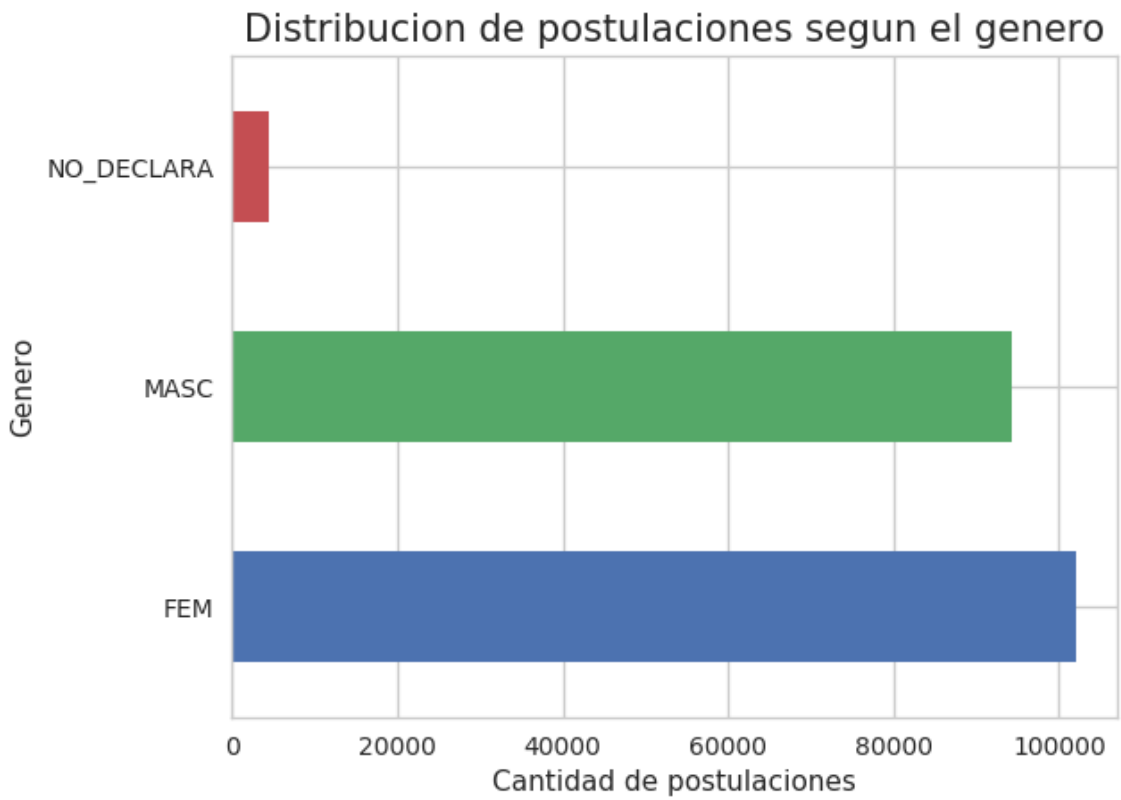
De este gráfico se aprecia que:

- El secundario es el único nivel de estudios en el que la cantidad de postulantes 'en curso' no es de aproximadamente el mismo orden de magnitud que los graduados. Esto refuerza la idea de que el título secundario es necesario para insertarse en el mercado laboral.

Postulaciones según el género y la edad de los postulantes:

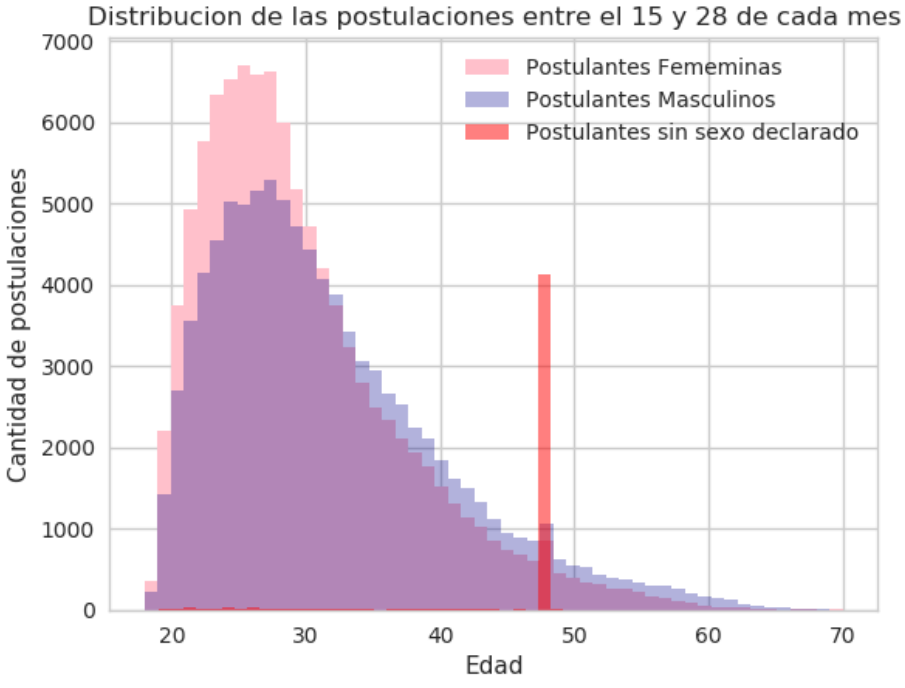
Para este análisis se filtró el archivo 'postulaciones_genero_y_edad'. En el mismo había años de nacimiento que iban desde el 1700 hasta el 2006. Decidimos dejar solamente los datos de postulantes de entre 18 y 70 años, pensamos que alguien de mayor edad no buscaría trabajo a través de una página de internet.
Dejamos hasta 70 años inclusive debido a que esa es la edad jubilatoria, se supuso que pasada esa edad no se buscaría trabajo, en blanco por lo menos.

Una vez aclarado lo anterior, pasemos a ver la distribución de postulaciones según el género de las personas.



Se ve a simple vista que el género femenino realizó más postulaciones a empleos que los demás.

Veamos a continuación cómo se distribuyen a partir de la edad y el sexo:



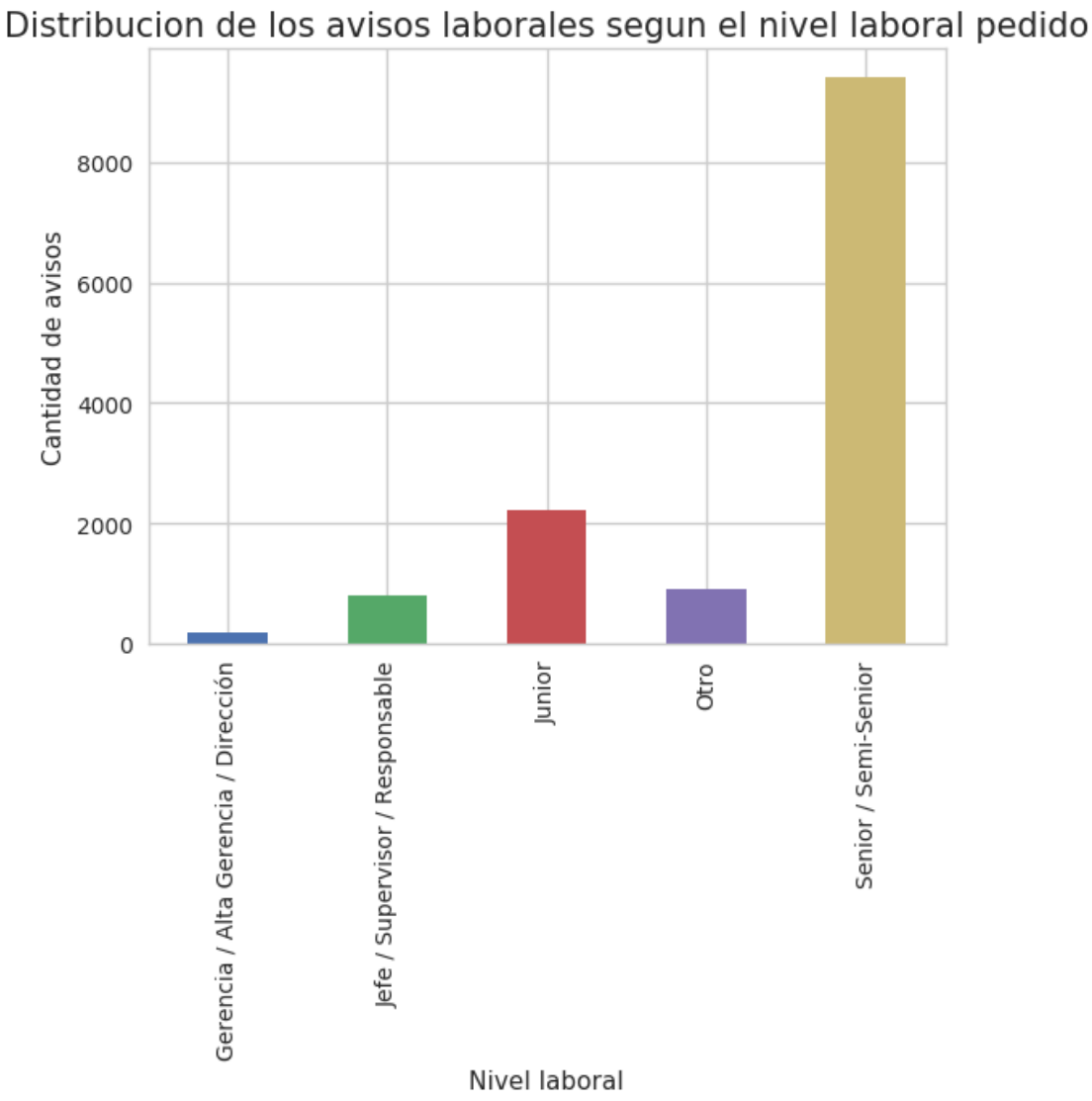
Puede verse que la mayor cantidad de postulaciones son de personas de entre 25 y 30 años y que la mayoría de ellos son mujeres lo que se condice con el gráfico anterior, además a partir de esa edad la cantidad de postulaciones comienza a descender drásticamente, hasta que por algún motivo entre los 45 y 48/49 años de pronto se da un pico de postulaciones.
Un dato curioso de ese pico de postulaciones es que casi toda esa gente prefirió no declarar su sexo. Si vemos con atención el gráfico vemos más casos de postulaciones de personas en esta condición pero entre los 47 a 49 años aproximadamente esa cantidad aumenta drásticamente.
Haciendo un análisis de esta situación en el archivo se pudo observar que efectivamente ese pico se produce a los 48 años exactamente.
Esto es seguramente consecuencia de datos faltantes, ya que el epoch Unix fue hace 48 años, por lo que a un conjunto de postulantes deben de tener al mismo tiempo la edad y el sexo como faltantes, *defaulteando* a tener nacimiento en el epoch Unix y sexo sin declarar.

Análisis de los avisos de empleo:

Ahora pasaremos a analizar un poco las características de los avisos de empleo de la página web. Veremos cómo se distribuyen según el tipo de empleo, el tiempo de trabajo, la zona donde se encuentra, entre otras.

Nivel laboral:

A continuación veremos cómo se distribuyen los avisos según el nivel que solicitan los empleadores.



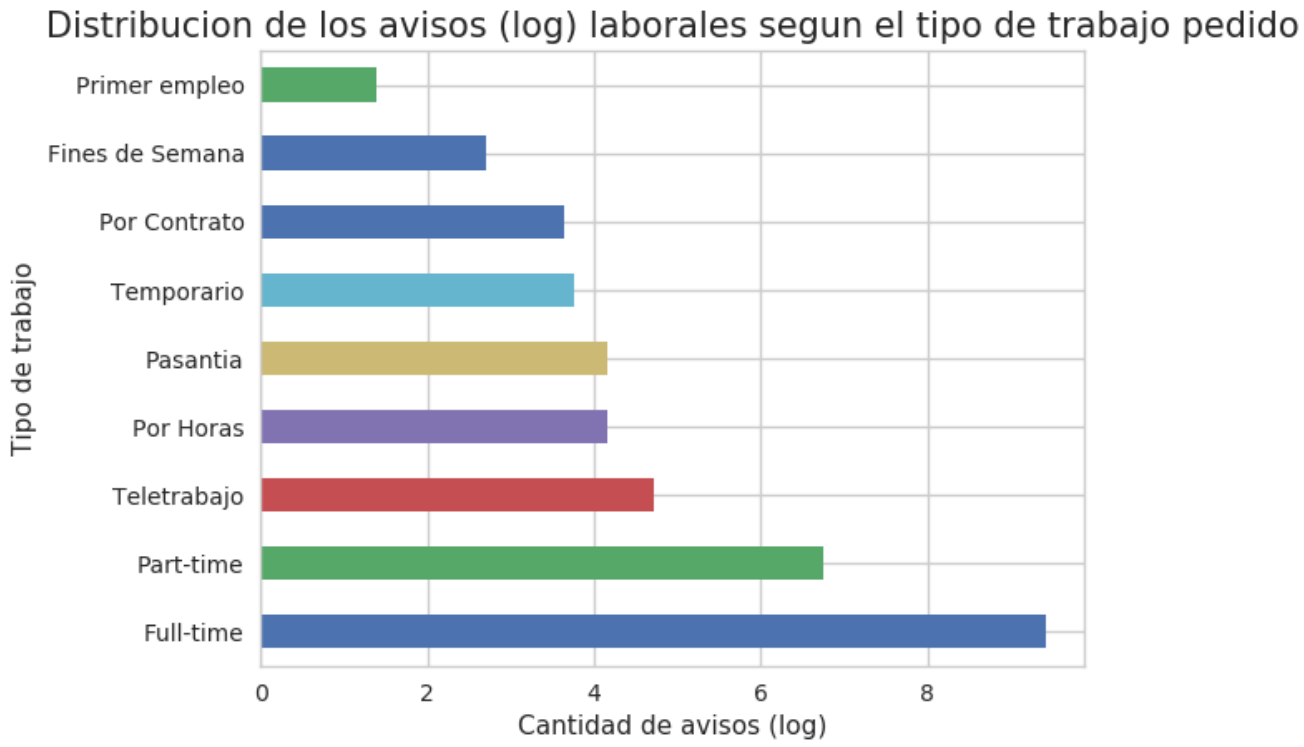
En números:

Nivel	Número de avisos
Senior / Semi-Senior	9407
Junior	2216
Otro	921
Jefe / Supervisor / Responsable	809
Gerencia / Alta Gerencia / Dirección	181

Se aprecia una forma aproximadamente de campana en el expertise requerido por las ofertas laborales, con cumbre en ‘Senior/SemiSenior’.

Tipo de trabajo solicitado:

Veamos ahora la distribución de los avisos según el tipo de los empleos.



En números:

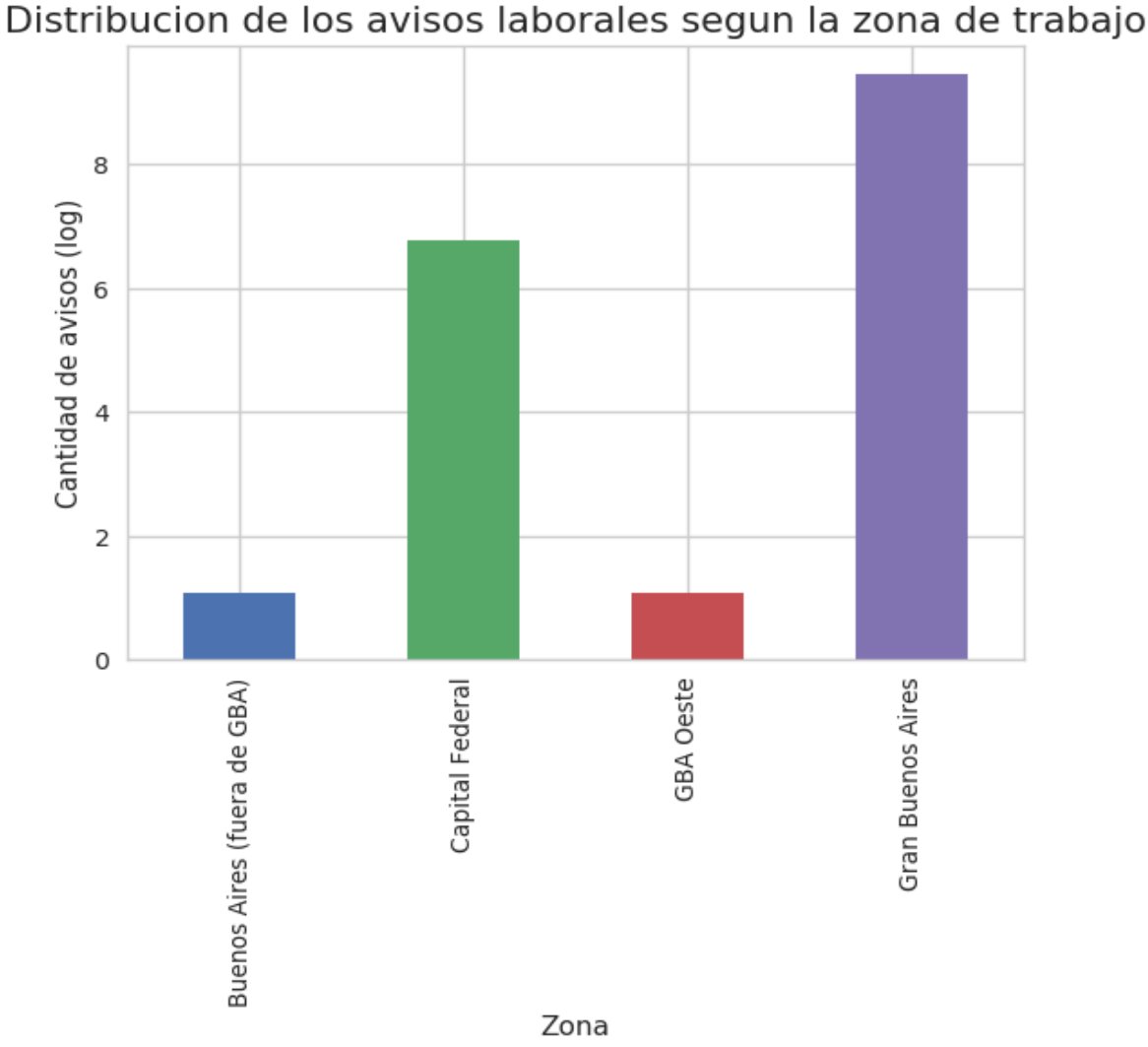
Nivel	Número de avisos
Full-time	12339
Part-time	863
Teletrabajo	110
Por Horas	63
Pasantia	63
Temporario	42

Por Contrato	37
Fines de Semana	14
Primer empleo	3

Es claro que los empleadores buscan contratar personal fijo, ya que el 97% de los anuncios son para empleos full-time o part-time.

Zona de trabajo:

Veamos cómo se distribuyen los avisos según la zona donde se encuentra el empleo.



En números:

Nivel	Número de avisos
Gran Buenos Aires	12654

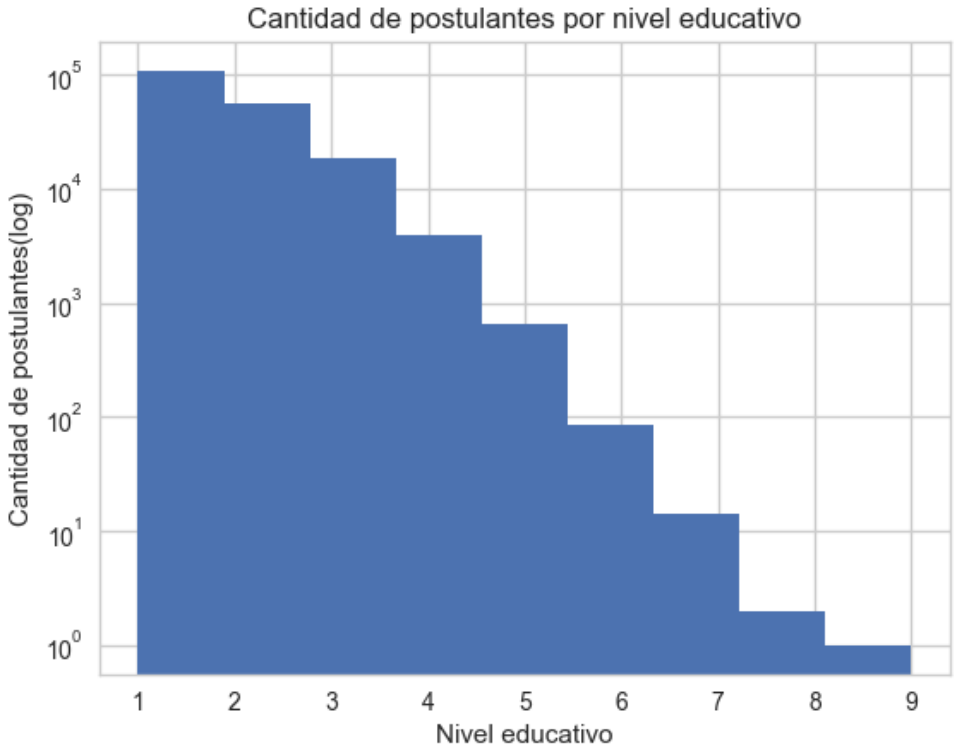
Capital Federal	876
GBA Oeste	2
Buenos Aires (fuera de GBA)	2

Por la poca cantidad de avisos y el hecho de que GBA Oeste es parte de GBA, correspondería descartar los datos correspondientes a ‘fuera de GBA’ y ‘GBA zona oeste’. Por otro lado, la cantidad de avisos por habitante no parece crecer linealmente, siendo el GBA más denso en ofertas laborales que CABA (estas áreas cuentan con 13 millones y 3 millones de habitantes respectivamente, según el último censo).

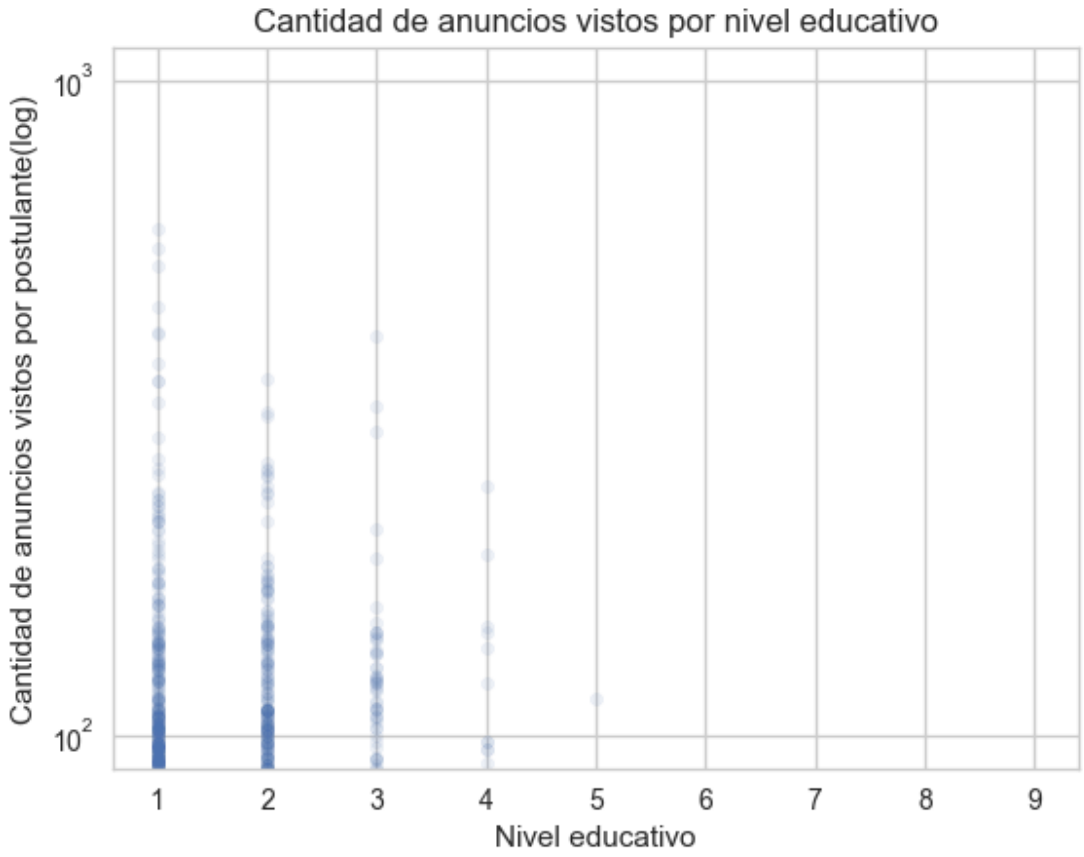
¿Hay relación entre el nivel de estudios de una persona y cuántos anuncios mira?

Para responder esta pregunta definimos la métrica de ‘nivel educativo’, que es una burda simplificación de qué tan avanzado en educación formal está un postulante. Consta de la suma de todos los estudios que alguna vez intentó.

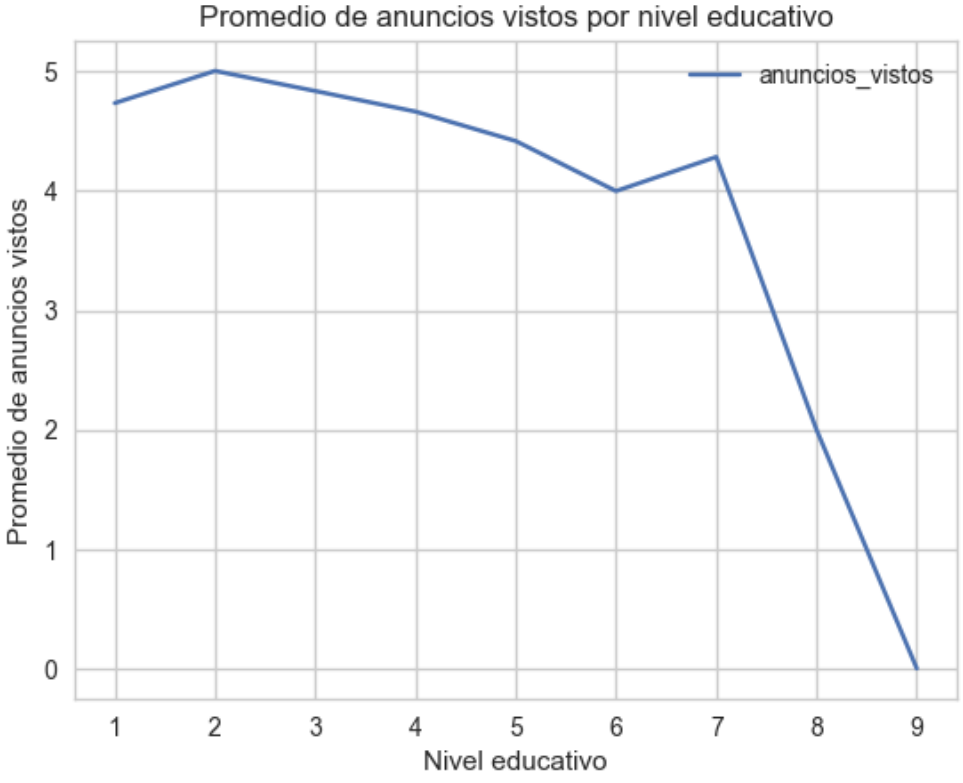
Por un lado, se ve que la mayoria de las personas reporta haber estudiado una sola cosa, y decrecen exponencialmente los postulantes con más estudios.



Esto a su vez, hace menos relevante al hallazgo de que cuanto menor el nivel educativo, más anuncios visita:



Para clarificar esto, se puede plantear qué cantidad de anuncios ven en promedio los postulantes según su nivel educativo:



El promedio para los valores que tiene sentido mirar (hay muy pocos postulantes con nivel educativo 9) nos dice que no hay correlación entre la cantidad de anuncios visitados y el nivel educativo como lo definimos.
Quizás con una métrica distinta mejor modelada de nivel de estudio podríamos llegar a conclusiones distintas.

Resumen de insights:

- El dataset está bastante limpio, habiéndose encontrado sólo unos pocos datos faltantes.
- Hay menos postulaciones en los fines de semana.
- El caso más usual es que un postulante sólo cuente con educación secundaria.
- Las distribución de ofertas laborales según nivel de expertise requerido toma una forma de campana, con el pico en Senior/SemiSenior.
- GBA es más denso en postulaciones que CABA.
- Aparentemente no hay relación entre qué tantos estudios tiene una persona y cuantos anuncios visita.

Código fuente:

Disponible en github.
Kernel de kaggle.