FACULTAD DE INGENIERIA
Universidad de Buenos Aires

75.06 Organización de Datos
Primer cuatrimestre de 2018

Trabajo Práctico Nº1
Análisis Exploratorio

Apellido y Nombre :	Padrón :
Vergara, Ariel	97010
Hernandorena, Gonzalo	98022

Repositorio de Github

https://github.com/GonzaH/Detos/blob/master/AnalisisFinal.ipynb

Introducción

El trabajo práctico consiste en hacer un análisis exploratorio sobre los datos proporcionados por Navent. Las preguntas en las que basamos nuestro análisis y serán respondidas en las conclusiones son las siguiente:

- ¿Hay algún sexo que predomina en cantidad sobre el otro?
- ¿Los datos pertenecen a alguna zona específica?
- ¿Cómo es la distribución de usuarios por estudio?
    - ¿Es igual para los estudios en curso y para los graduados?
    - ¿Se postulan más los usuarios con determinado estudio?
- ¿Cumplen la cantidad de postulaciones por aviso con una ley de potencia?
- ¿Cuáles son los niveles más buscados?
    - ¿Cómo es relación con los estudios de los postulantes?
    - ¿Cómo es su relación con el tipo de trabajo?
- ¿Cómo se distribuyen las postulaciones en el tiempo?
    - ¿Hay algún horario donde haya más postulaciones?
    - ¿Y algún día?
    - ¿Y una combinación de ambas cosas?
- ¿Hay edades en las que sea más usado el sistema?

Herramientas utilizadas

Para realizar el trabajo práctico se utilizó Python como lenguaje de programación y la biblioteca pandas para el manejo de dataframes. Para las visualizaciones se utilizaron las biblioteca matplotlib y en el caso de los heatmaps seaborn.

Los sets de datos utilizados fueron:

- fiuba_1_postulantes_educacion.csv
- fiuba_2_postulantes_genero_y_edad.csv
- fiuba_3_vistas.csv
- fiuba_4_postulaciones.csv
- fiuba_5_avisos_online.csv
- fiuba_6_avisos_detalle.csv

Limpieza de los sets de datos

Luego de revisar los datos en cada uno de los csvs provistos, se realizó un análisis previo para determinar sobre qué dataframes se debía llevar a cabo una limpieza, con el fin de quedarnos únicamente con aquellos datos que nos aporten información y estén completos.

Comenzamos con el dataframe en donde se indica la fecha de nacimiento y el género de cada usuario:

```
Cantidad de valores nulos por columna:
idpostulante          0
fechanacimiento    4750
sexo                  0
dtype: int64

Cantidad total de usuarios: 200888
```

Podemos ver que existen 4750 de 200888 usuarios que no declaran su fecha de nacimiento (aproximadamente un 2,36%). Para que el set de datos esté completo podíamos rellenar estos datos faltantes realizando una regresión o sacando un promedio siguiendo algún tipo de criterio. Sin embargo, como la cantidad de usuarios en cuestión es baja, optamos por eliminarlos del set de datos. Para asegurarnos que esto no genere una pérdida de información considerable en los sets de visitas y postulaciones, se calculó cuántas postulaciones y visitas correspondían a estos usuarios y obtuvimos que tan solo el 1,31% de las postulaciones (44464 de 3401623) y el 2,37% de las visitas (22767 de 961897) pertenecen a estos usuarios. Por lo tanto, se los eliminó de los dataframes, quedando un total de 196138 usuarios.

Luego seguimos con el dataframe de educación. Este presenta una columna con el id del usuario (‘idpostulante’) y otras dos columnas con variables categóricas. La columna ‘nombre’ presenta los valores 'Otro', 'Secundario', 'Terciario/Técnico', 'Universitario', 'Master', 'Posgrado', 'Doctorado' mientras que la columna ‘estado’ puede adoptar los campos 'En Curso', 'Abandonado' o 'Graduado'. A su vez, encontramos que existían múltiples entradas de estudio por usuario e incluso, en algunos casos, aparecen múltiples entradas para un mismo tipo de estudio pero con distintos estados. Un ejemplo de esto se puede observar en el siguiente usuario:

	idpostulante	nombre	estado
373	YlMLGD	Terciario/Técnico	En Curso
374	YlMLGD	Otro	En Curso
375	YlMLGD	Universitario	En Curso
2206	YlMLGD	Terciario/Técnico	Graduado
2207	YlMLGD	Universitario	Graduado
2208	YlMLGD	Otro	Graduado
2209	YlMLGD	Posgrado	Graduado
3659	YlMLGD	Universitario	Abandonado
3660	YlMLGD	Terciario/Técnico	Abandonado

Analizando este usuario vemos que, por ejemplo, para el estudio ‘Universitario’ tiene 3 entradas, una con cada valor de estado. A su vez, tiene una entrada con el tipo de estudio Posgrado y en estado Graduado, siendo un Posgrado un nivel de estudio superior al Universitario. Asumimos que esto se puede deber a que las múltiples entradas de estudio corresponden a un historial del usuario. Se optó entonces por generar un nuevo dataframe, con un nuevo par de columnas referenciadas al estudio: estudio_graduado, en donde se indica el mayor nivel de estudio en el que se graduó el usuario, y estudio_en_curso, en donde se muestra el mayor nivel de estudio que tiene en curso el cual, a su vez, debe ser mayor al nivel que tiene en la columna graduado (sino no se cargará ningún estudio en curso para el mismo), de manera tal que cada usuario tenga una única entrada en el dataframe. Los niveles lo definimos de la siguiente manera de menor a mayor:

1. Otro
2. Secundario
3. Terciario/Técnico
4. Universitario
5. Master
6. Posgrado
7. Doctorado

La importancia de cada título la definimos arbitrariamente, lo cual nos introduce un bias en el análisis. Sin embargo, consideramos que este orden es el que mayor se acerca a la realidad ya que, en algunos casos, existen dependencias entre los títulos. Pensamos que ‘Otro’ hace referencia a talleres o cursos por lo que optamos que sea el de menor peso. En caso de no presentar estudios en curso o completos o que no cumpla con las condiciones impuestas para cada columna, se les cargará el valor ‘Ninguno’. A continuación veremos la entrada del usuario mostrado anteriormente en nuestro nuevo dataframe.

	idpostulante	estudio_en_curso	estudio_graduado
117605	YIMLGD	Ninguno	Posgrado

Como queremos tener usuarios que tengan tanto información sobre su educación como de su fecha de nacimiento (para poder calcular posteriormente su edad) y la cantidad de usuarios con fecha de nacimiento pero sin entradas de estudio no es alta (10260 de 196138, que es un 5,23%), optamos por eliminarlos.

Finalmente, analizamos si existen datos incompletos en el csv con detalles sobre los anuncios:

```
Cantidad de valores nulos por columna:
idaviso                    0
idpais                     0
titulo                     0
descripcion                0
nombre_zona                0
ciudad                 13487
mapacalle              12662
tipo_de_trabajo            0
nivel_laboral              0
nombre_area                0
denominacion_empresa       5
dtype: int64

Cantidad total de avisos: 13534
```

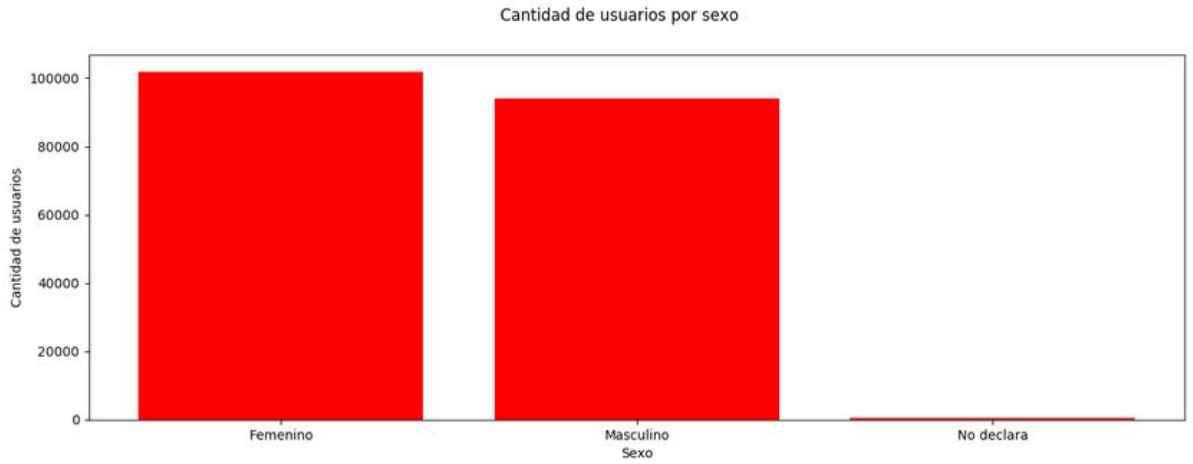
Como se puede ver en la imagen anterior, una cantidad ínfima de avisos presentan información en las columnas ‘mapacalle’ y ‘ciudad’, por lo que ambas fueron eliminadas del dataframe. Otra cuestión observada es que, salvo 5 avisos, todos tienen un valor en la columna ‘denominacion_empresa’. Antes de eliminar estas filas del dataframe, se analizó cuántas postulaciones y visitas de los sets de datos mencionados anteriormente se eliminarían en dicho caso y obtuvimos que solamente solamente 236 postulaciones y una visita estaban asociadas a estos 5 avisos (ninguno de estos estaba online por lo que no hacía falta eliminar ningún aviso del dataframe que contiene esta información). Por lo tanto, se eliminaron estos 5 avisos y las postulaciones y visitas asociadas.

Cabe destacar que existen entradas en el dataframe de postulaciones cuyos id de aviso no matchean con ninguna entrada del set de datos en donde se indican los detalles de los mismos. Estas no serán tenidas en cuenta a la hora de realizar el análisis sobre cantidad de postulaciones por aviso.

Análisis de los sets de datos

Análisis de géneros

Lo primero que hicimos al analizar los datos fue graficar algunas cuestiones básicas para cerciorarnos que no hubiera nada extraño en esos campos. El género de los usuarios fue uno de ellos ya que, si bien habíamos limpiado los datos, podía suceder que hubiera una gran cantidad de integrantes de un género por sobre el otro.



En el gráfico se puede apreciar que no hay una diferencia significativa en cantidad entre los usuarios femeninos y masculinos, es decir que no se detecta ninguna anormalidad. Por otra parte, también se puede ver que los usuarios que no declaran el género son extremadamente pocos (445 de 196138), hasta el punto en que a duras penas se ve la barra correspondiente.

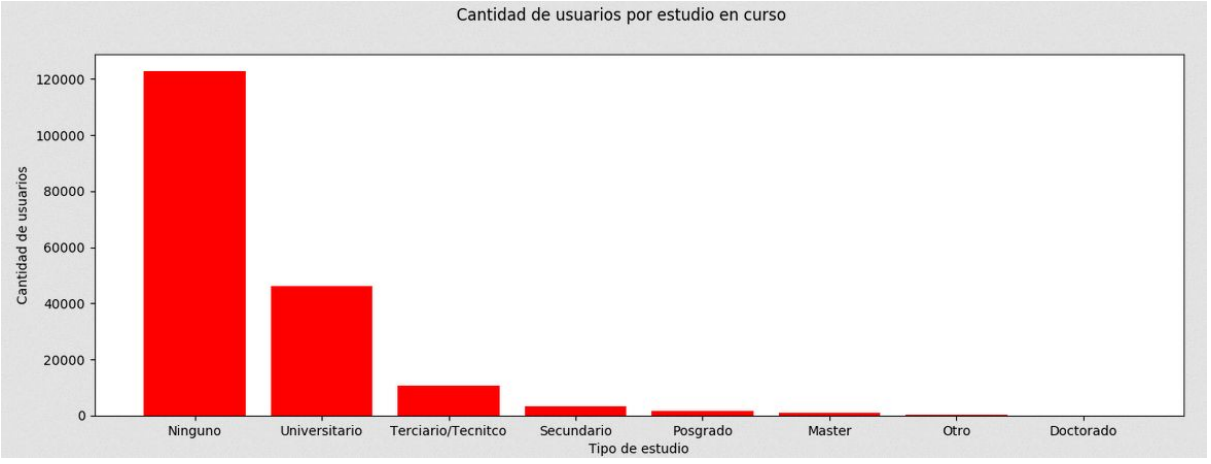
Análisis de zonas

Otro chequeo que hicimos fue la cantidad de avisos por zona. Previamente a hacer el análisis, descubrimos que la variable ‘nombre_zona’ adoptaba 4 valores, ‘Gran Buenos Aires’, ‘Capital Federal’, ‘Buenos Aires (fuera de GBA)’ y ‘Zona Oeste’. Como ‘Zona Oeste’ solamente contaba con dos entradas y pertenece al Gran Buenos Aires, decidimos anexar Zona Oeste al Gran Bs As. En el caso de ‘Buenos Aires (fuera de GBA)’, también había solo dos registros, pero como no pertenece a ninguno de las otras zonas no lo combinamos. Al graficar nos llevamos una sorpresa ya que esperábamos más avisos en CABA:

Con este gráfico podemos saber que la gran mayoría de las empresas que ponen sus anuncios en este sistema son de GBA, pero no sabemos si esto es así porque nos dieron datos sesgados o si es una cuestión del target de Navent. No podemos sacar mucha información significativa más que lo expresado anteriormente.

Análisis de estudios

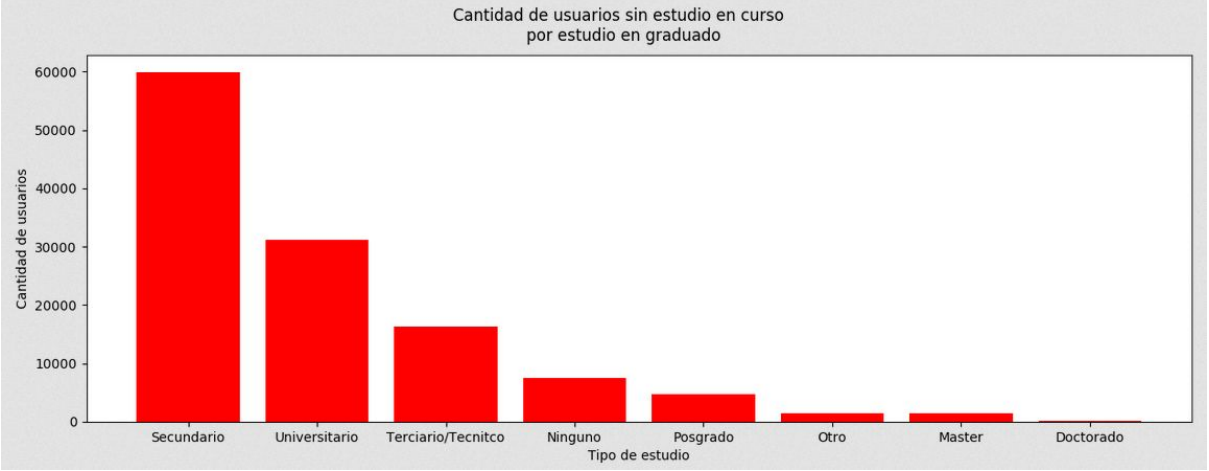
Una vez concluidos estos dos análisis básicos, pasamos a ver cuánta gente se había graduado y cuánta tenía en curso cada estudio. En el caso de los estudios en curso, nuestra hipótesis era que la mayoría habría cargado un estudio universitario, mientras que en los casos de posgrado, máster o doctorado iban a ser una cantidad muy poco significativa. Teniendo en cuenta que en la mayoría de los trabajos, y más cuando utilizan una página para conseguir empleados, se requieren estudios secundarios completos, consideramos que habrá muy pocos usuarios que hayan marcado que tienen el secundario en curso:



En el gráfico podemos ver que la mayoría de los usuarios no presenta un estudio en curso. Esto se puede deber a que pueden tener un estudio graduado y sus entradas de estudio

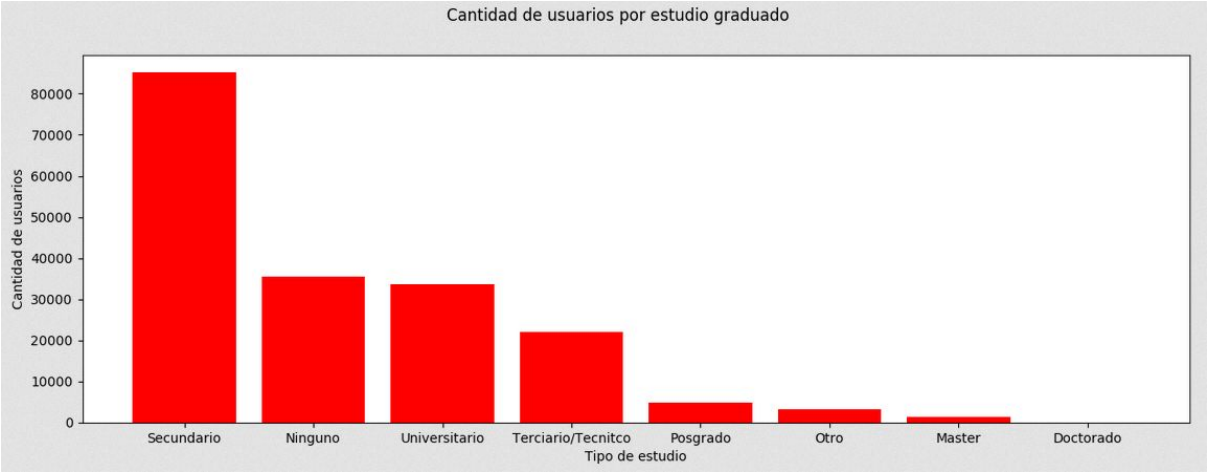
en curso en el dataframe de educación original tengan un menor nivel al mismo. Exceptuando esto, nuestras predicciones se cumplieron.

Para hacer un análisis más profundo sobre los usuarios que tienen el valor ‘Ninguno’ en la columna ‘estudio_en_curso’, decidimos filtrarlos para graficar la cantidad de usuarios sin estudio en curso por su valor en la columna ‘estudio_graduado’:



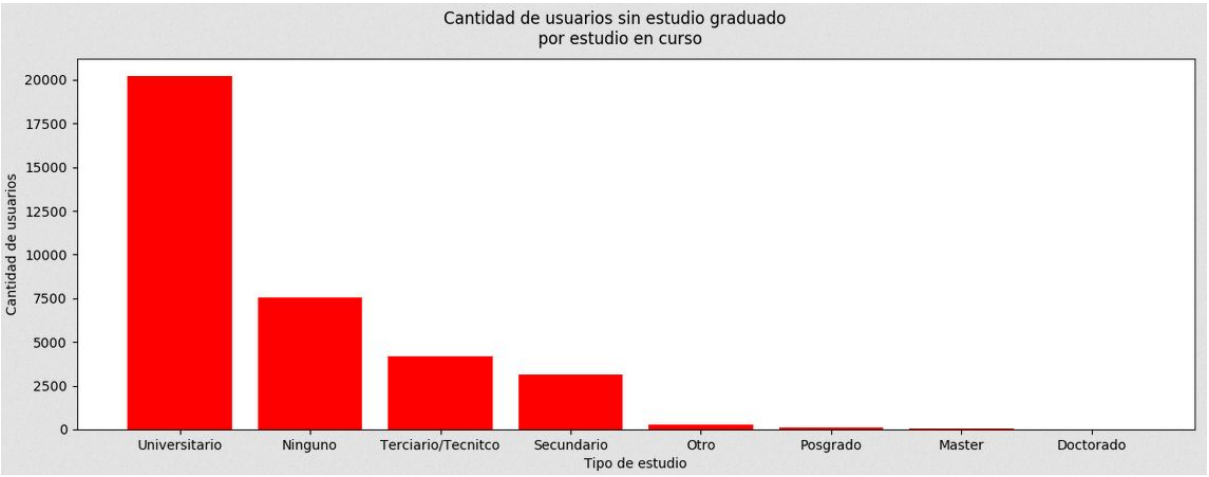
Al observar el gráfico encontramos algunos puntos interesantes. Para empezar, la mayoría de los usuarios que no ingresó un estudio en curso sólo tiene el secundario completo (de lo contrario lo habrían ingresado como dato ya que es algo importante). A su vez, los usuarios sin ningún estudio se encuentran en cuarto lugar en este gráfico y son aproximadamente 7500. Estos se corresponden con los usuarios que tenían solamente entradas de estudio con estado ‘Abandonado’ en el dataframe de educación original. Los estudios abandonados no fueron tenidos en cuenta dado que consideramos que no se podía realizar un análisis a partir de ellos sin hacer suposiciones. Un ejemplo de esto sería si un usuario abandonó un posgrado, él debería tener como mínimo un título de grado completo. La cantidad de estos usuarios superan solamente a la de las categorías relacionadas con estudios avanzados u otros, que no sabemos exactamente qué es (como se dijo anteriormente, suponemos que hace referencia a talleres o cursos extra). Habiendo visto esto, nos resultó necesario hacer un estudio equivalente sobre los graduados.

Con respecto a ellos, suponemos que la mayor parte de los usuarios tendrá como estudio graduado el secundario, dado que es el mínimo estudio para poder comenzar una carrera universitaria o un terciario; también, en la mayoría de los casos, es lo mínimo que se necesita para buscar trabajo. A continuación se muestra el gráfico de cantidad de usuarios por tipo de estudio finalizado:



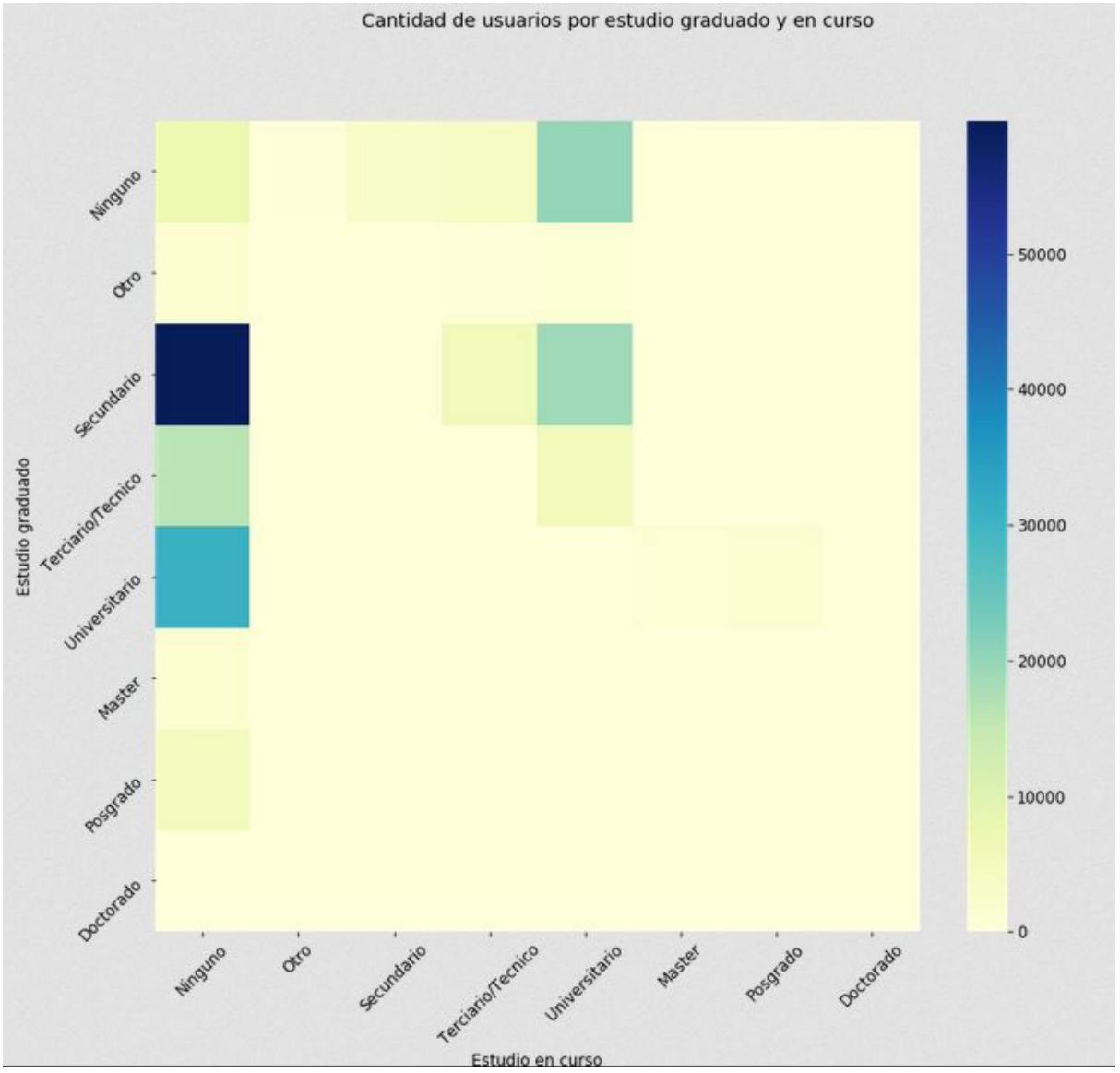
A diferencia del caso anterior, nuestra hipótesis se cumple con respecto a dónde está la mayor agrupación de usuarios. Podemos apreciar que los títulos avanzados como máster, posgrado y doctorado tienen una cantidad ínfima de gente (en el caso del doctorado ni siquiera se puede vislumbrar el color de la barra), lo que esperábamos, ya que es sabido que es muy poca gente decide continuar su formación con esos estudios.

Continuamos con el gráfico de los estudiantes que no completaron sus estudios graduados, lo que nos dió como resultado el gráfico que se muestra a continuación:

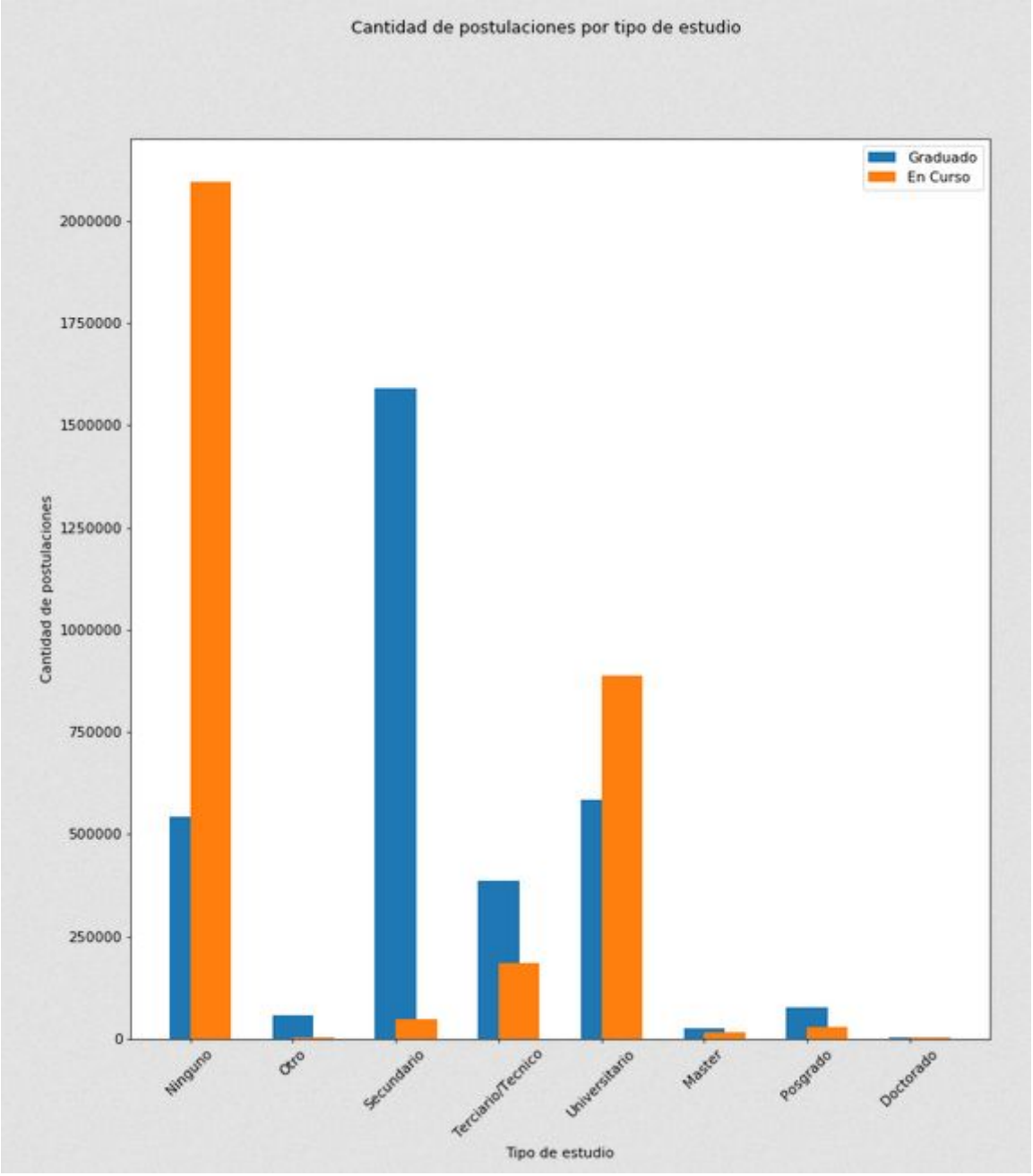


Lo primero que notamos fue que la columna del valor ‘Ninguno’ se corresponde con el gráfico análogo de gente sin estudio en curso (Cantidad de usuarios sin estudio en curso por estudio graduado). Por otra parte los estudiantes universitarios superan ampliamente al resto. De aquí podemos concluir que existe una gran cantidad de personas que no consideran necesario incluir un estudio finalizado previo teniendo uno de mayor valor en curso. Creemos que el caso universitario es el más popular porque la gran mayoría de los estudiantes universitarios vienen directamente del secundario por lo que se sabe que lo terminaron.

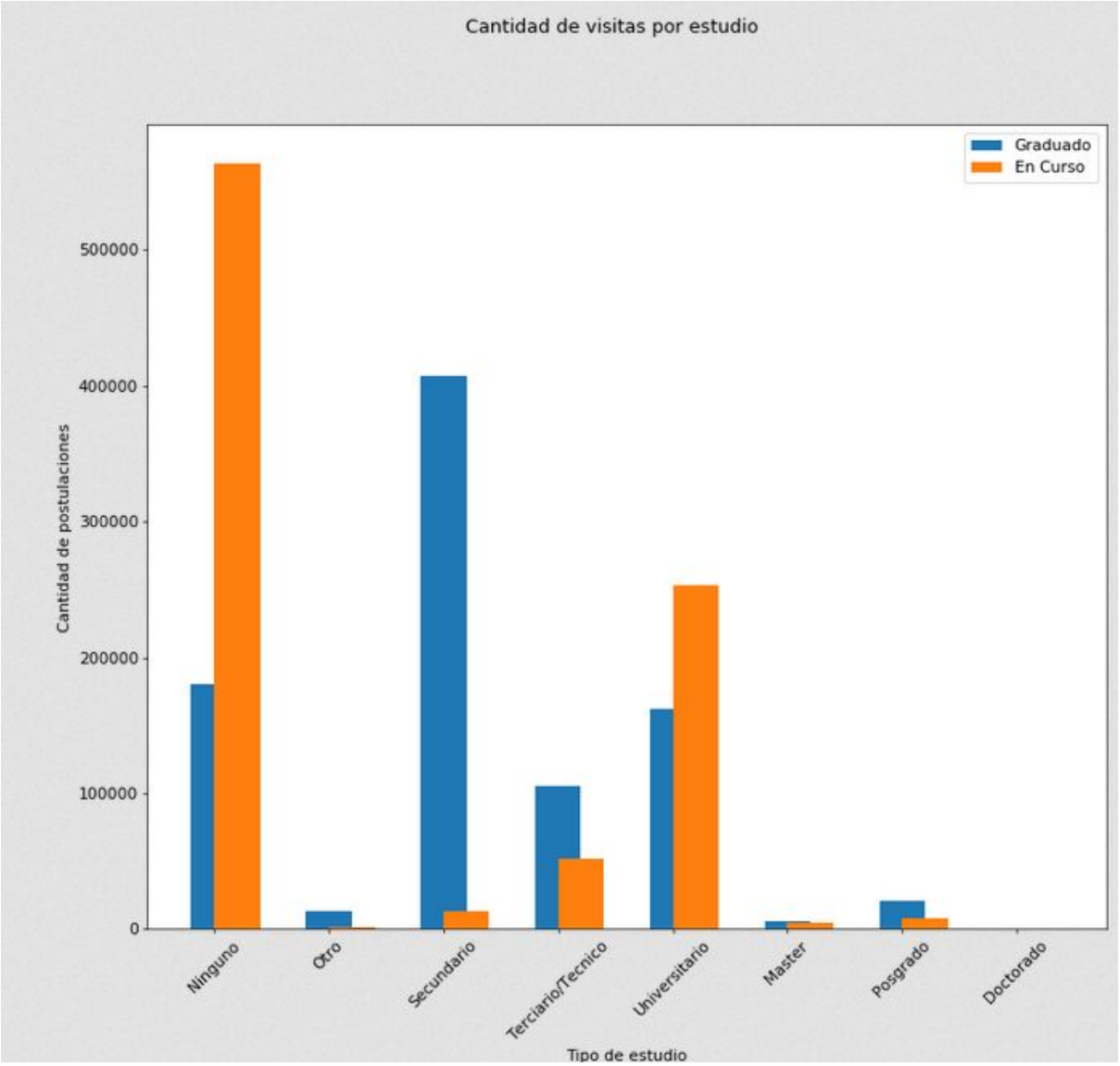
Para poder visualizar rápidamente el análisis realizado sobre las concentraciones de usuario tanto por estudio en curso como graduado, realizamos el siguiente heatmap:



Como estamos trabajando con los tipos de estudio, quisimos investigar qué categoría de estudio tiene más postulaciones. Para ello realizamos el siguiente gráfico de barras:



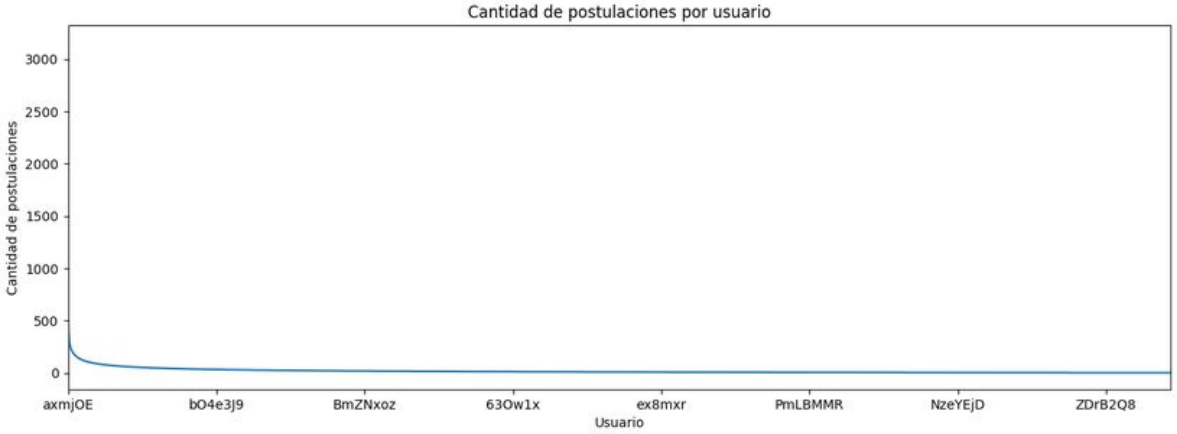
Podemos ver que con los estudios avanzados pasa lo mismo que en los anteriores análisis, los datos son tan pocos que son casi despreciables en comparación. El caso del secundario, es un resultado predecible ya que hay una gran cantidad de usuarios con éste estudio terminado, mientras que, como ya dijimos antes, los usuarios con el secundario en curso son muy pocas.



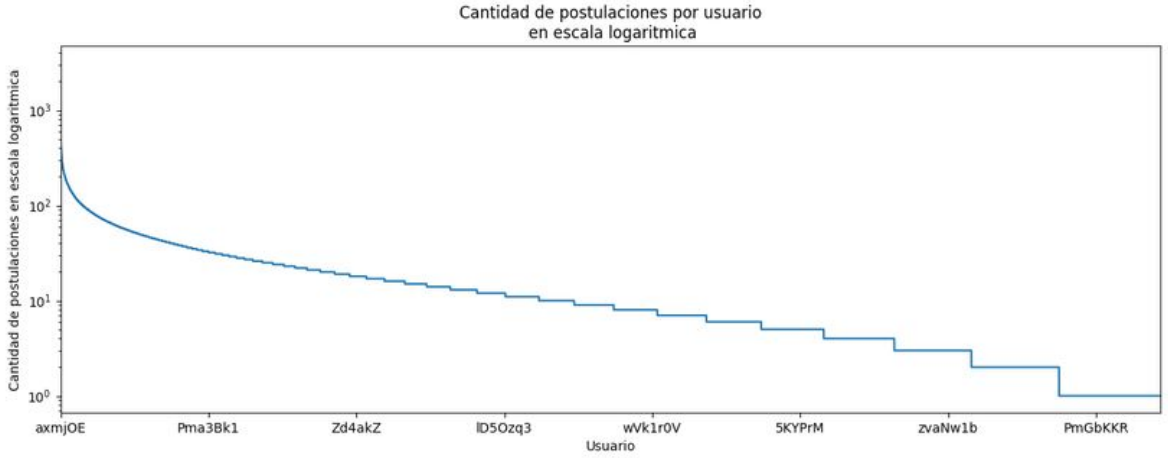
El gráfico es muy parecido al anterior (si bien las cantidades son mucho menores) por lo que asumimos que el comportamiento de los usuarios es similar tanto a la hora de mirar las publicaciones como de postularse.

Análisis de cantidad de postulaciones por usuario como ley de potencia

Procedimos a analizar si la cantidad de postulaciones por usuario responden a una ley de potencias. Esto es porque hay algunos usuarios que hicieron una cantidad inexplicable de postulaciones, mientras que la mayoría sólo hacían una pequeño número. El siguiente gráfico plasma esta relación:



En el caso de la cantidad de postulantes por aviso, a simple vista no se puede apreciar el principio de la curva. Decidimos hacer la gráfica en escala logarítmica para ver si el resultado era aproximadamente lineal:



Se nota que el principio sigue siendo curvo y se va escalonando a medida que se acerca a los avisos con menor cantidad de postulaciones (llegando hasta los que no tienen ninguno). Como hay muchos usuarios con una o dos postulaciones, los valores 1 y 2 estarán repetidos muchas veces en el gráfico. Por otra parte, por cómo se construye la escala del eje y los valores menores tienen un mayor espacio entre ellos que los de mayor orden. En consecuencia el salto del 1 al 2 es notorio y, debido a la gran cantidad de repeticiones de estos valores, se observan los escalones en el gráfico.

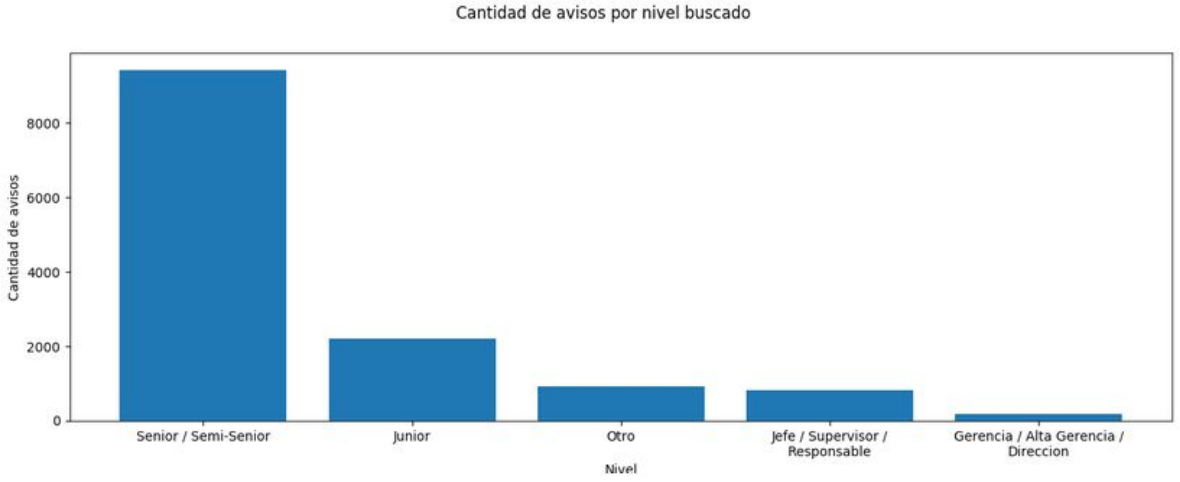
Para que las magnitudes del gráfico anterior cumplan con una ley de potencias, el 20% de los usuarios con más postulaciones debe englobar aproximadamente el 80% de las postulaciones. Al realizar los cálculos obtuvimos los siguientes resultados:

```
Total postulaciones: 3267818
El 20% de los usuarios con más postulaciones tiene un total de 2096161 postulaciones
El 20% de los usuarios con más postulaciones tiene el 64.145586% de las postulaciones
```

Como podemos ver, solamente abarcan el 64,15% por lo que no cumple con una ley de potencias. Esto puede deberse a que si bien los avisos pueden ser interesantes, solamente lo será para el grupo específico de gente al que apunta.

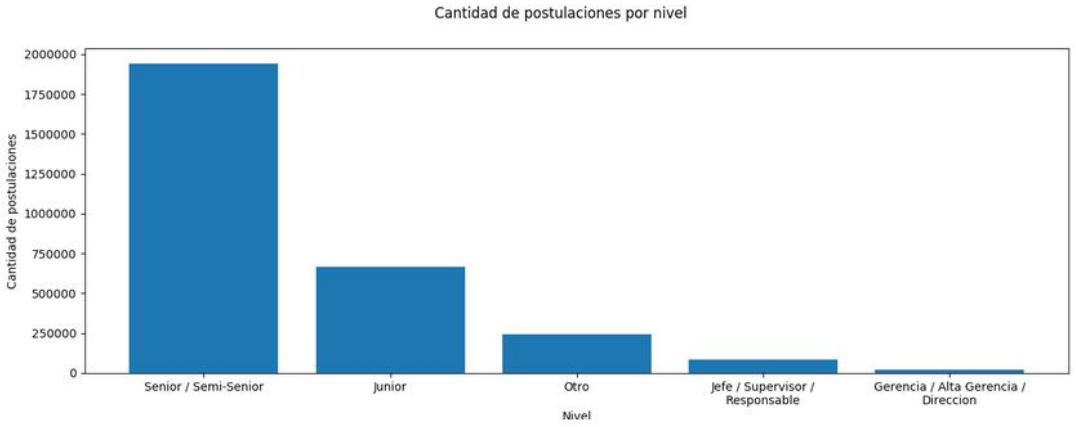
Análisis de nivel laboral y tipo de trabajo

Una cuestión que llamó nuestra atención fue la variable correspondiente al nivel laboral. Ésta tiene cinco valores categóricos, por lo que resulta fácil de graficar lo relacionado con ella, ya sea la cantidad o relación con otra columna. Para dar un pantallazo inicial a esta columna de datos, consideramos que lo mejor es ver el nivel más buscado. Realizamos el gráfico esperando pocos jefes y gerentes y una gran demanda de senior, semi senior y juniors, a su vez no sabemos nada del nivel 'Otro'. Creemos que lo más normal en los cargos más elevados es que no suelan ser buscados con sistemas de esta índole, sino que se lo asignan a gente que ya tiene experiencia dentro de la empresa o referencias externas.



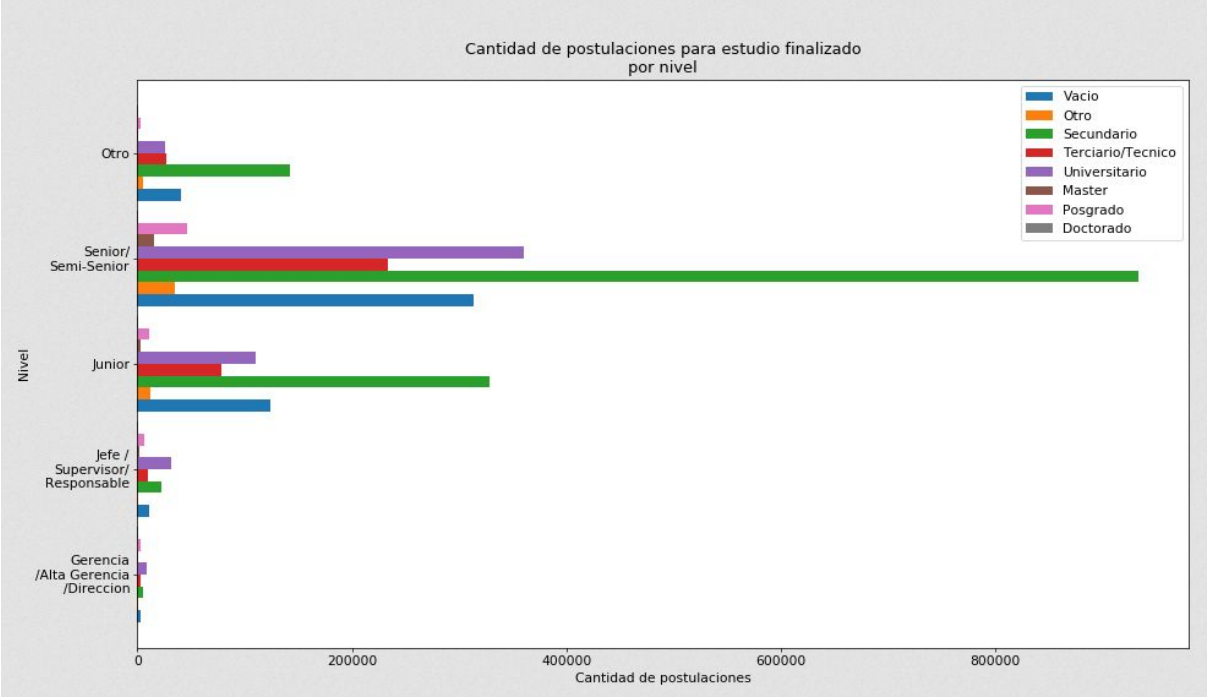
El gráfico confirma nuestra hipótesis de Senior/Semi-Senior, y los cargos altos, pero desacredita completamente la parte de Junior. Esto puede deberse a que la mayoría de las empresas suele buscar gente con experiencia.

Una idea que nos interesó fue comparar los resultados obtenidos con un gráfico análogo pero cambiando la cantidad de avisos por cantidad de postulaciones:



Las conclusiones que sacamos del gráfico no son muchas, porque es muy parecido al anterior. La gente se postula más para los niveles que más se buscan, lo que es completamente lógico.

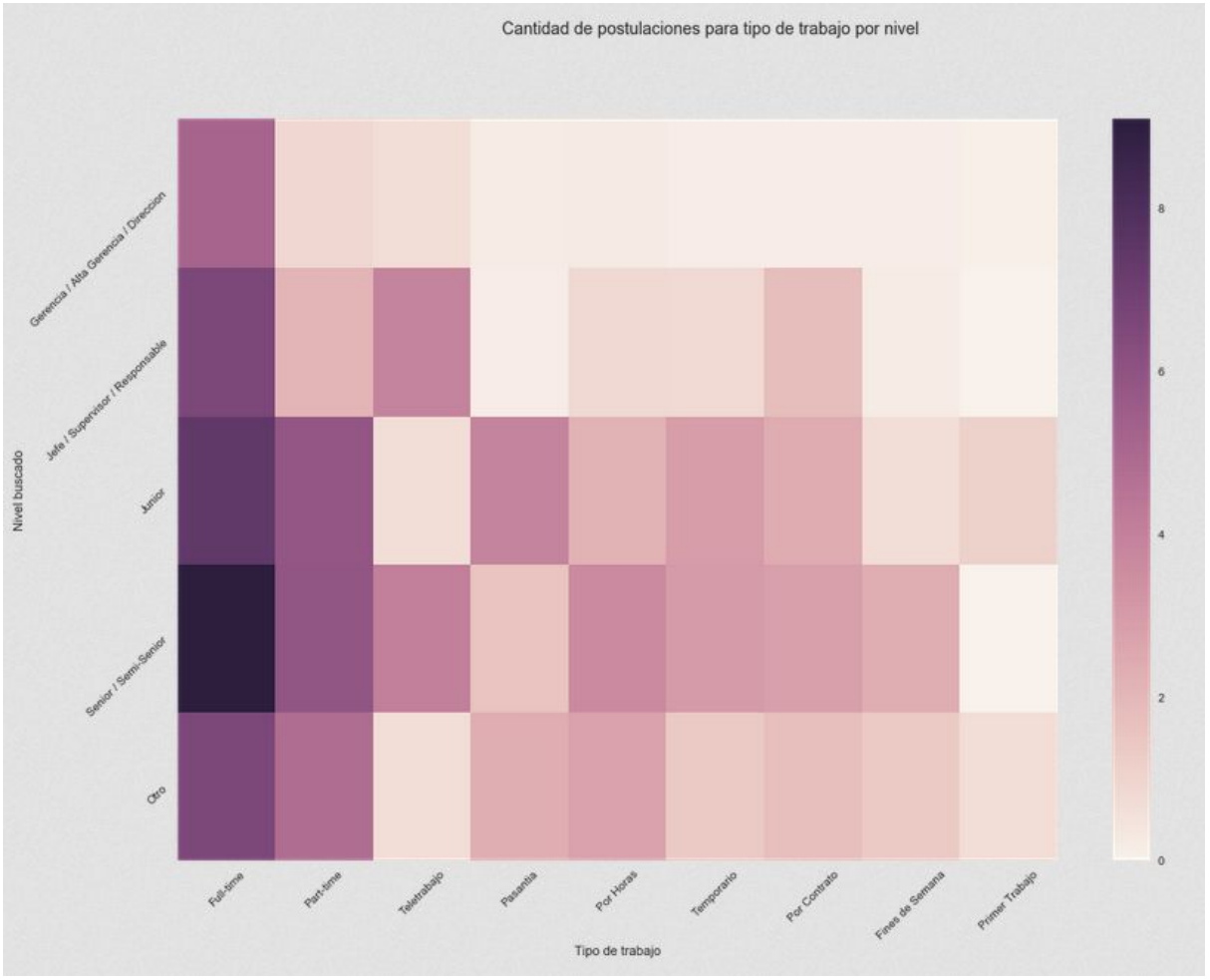
Posteriormente investigamos la relación entre el nivel laboral y los estudios de los postulantes a los mismos. Para ello graficamos un multibar plot con la cantidad de usuarios por estudio finalizado que se postula a un nivel específico:



Este gráfico nos permite visualizar claramente que el nivel al que más gente se postula es “Senior/Semi-Senior”, lo que es lógico ya que era el nivel más pedido por una amplia diferencia. También se refleja la cantidad de empleos que piden Gerentes y Jefes, ya que al ser niveles poco buscados hay pocas postulaciones. Se observa una situación análoga para los estudios en el color del Doctorado, que no se ve. Como el gráfico es de usuarios graduados

no es de sorprender que el color de postulantes con secundario completo sea la que más postulaciones acumula.

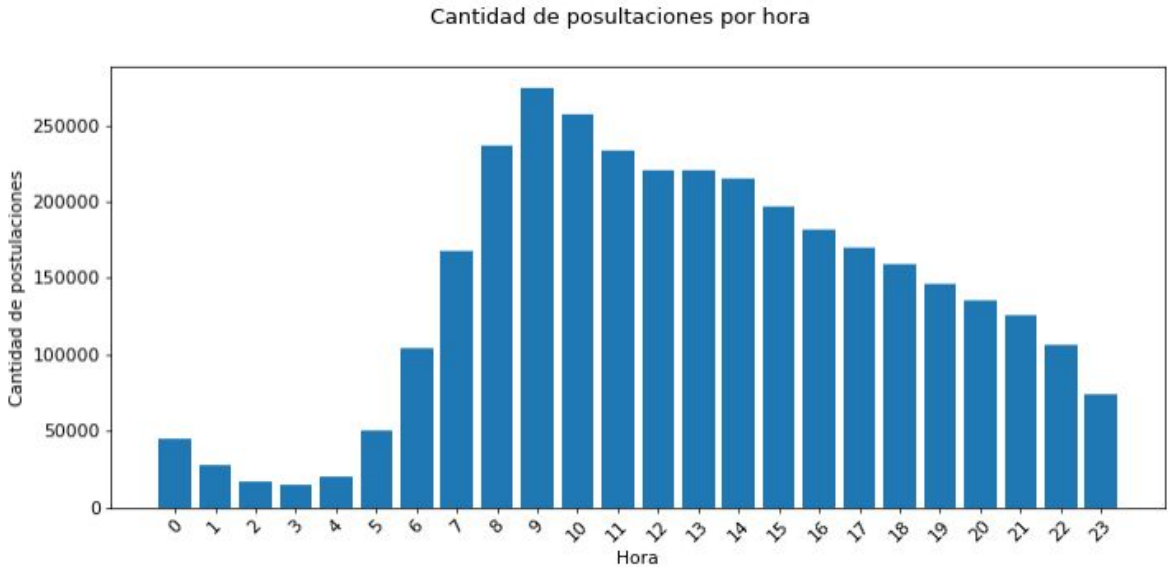
Viendo el gráfico anterior nos topamos con la duda de cómo sería la relación entre el nivel y el tipo de trabajo (Full time, Pasantías, Primer trabajo, etc), por lo que decidimos plasmarla en un heatmap:



Utilizamos una escala logarítmica a la hora de realizar el heatmap (en el repositorio se encuentra la versión con la escala real) para poder apreciar la diferencia de colores para los casilleros correspondientes a una menor cantidad de postulaciones, dado que solamente podíamos distinguir colores para la columna de Full-Time (especialmente para el casillero con Senior/Semi-Senior). Claramente se puede ver que los trabajos Full-time son los más buscados, seguidos por los Part-time. Como el nivel Senior/Semi-Senior es el predominante en los avisos, la combinación de las búsquedas de ambos es la celda que más postulaciones concentra. Por otra parte se puede ver que la proporción de Full-time sobre los tipos de trabajo es mucho mayor que la de Senior/Semi-Senior sobre su respectiva categoría, dado que la columna de Full-time acumula colores mucho más oscuros. Un dato curioso es que existen postulaciones para jefaturas que no son Full time.

Análisis de variables temporales

Para continuar con el análisis del set de datos, decidimos concentrarnos en las variables temporales. Comenzamos por la hora en las que se realizan las postulaciones. Nuestra suposición era que la gente iba a postularse en las horas que corresponden a su viaje al trabajo y luego de volver a su casa, y que las horas en las que se suele dormir iban a tener una caída drástica. Los resultados obtenidos se pueden ver en el siguiente gráfico:

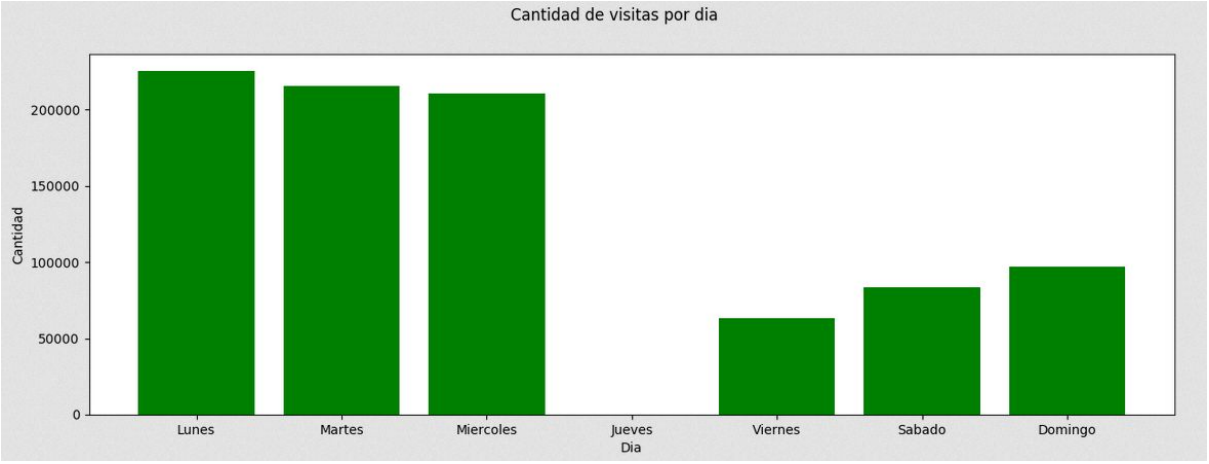


Podemos apreciar que desde la una de la 1 hasta las 5 de la mañana la cantidad de postulaciones es menor a los 50000 y luego comienza a crecer hasta llegar a su pico a las 9. Las 6 a.m. es un horario muy común en el que se levanta la gente, pero no se cumple la hipótesis de que a esas horas se acumula la mayor cantidad de postulaciones. Lo que sí sucede es que, a partir de que se empieza a levantar la gente, comienza a aumentar la cantidad de postulaciones lo que le da fuerza a la hipótesis de que la gente usa el viaje hacia el trabajo para revisar las ofertas y postularse. Las 9 de la mañana es un horario muy común para entrar a trabajar, por lo que es normal que decrezca la cantidad de postulaciones a partir de esta, ya que no pueden usar libremente internet para usar el sistema. Por otra parte, como la cantidad de gente despierta es mayor, aunque baje la cantidad de postulaciones no llega a los niveles anteriores de las 5 a.m. cuando la mayoría de los usuarios están dormidos. A medida que avanza el día, disminuye la cantidad de postulaciones. Esto podría deberse a que, como la gente ya se postuló a la mañana considera que ya vió las posibilidades y lo revisará otro día. Las postulaciones decrecen progresivamente hasta llegar a las 23 cuando caen excesivamente porque mucha gente ya está durmiendo.

Otro estudio con respecto a variables temporales que consideramos importante es la relación de postulaciones con días de la semana. Pensamos que la gente se postula durante la semana porque es cuando más se siente el stress del trabajo o cuando surgen crisis laborales.

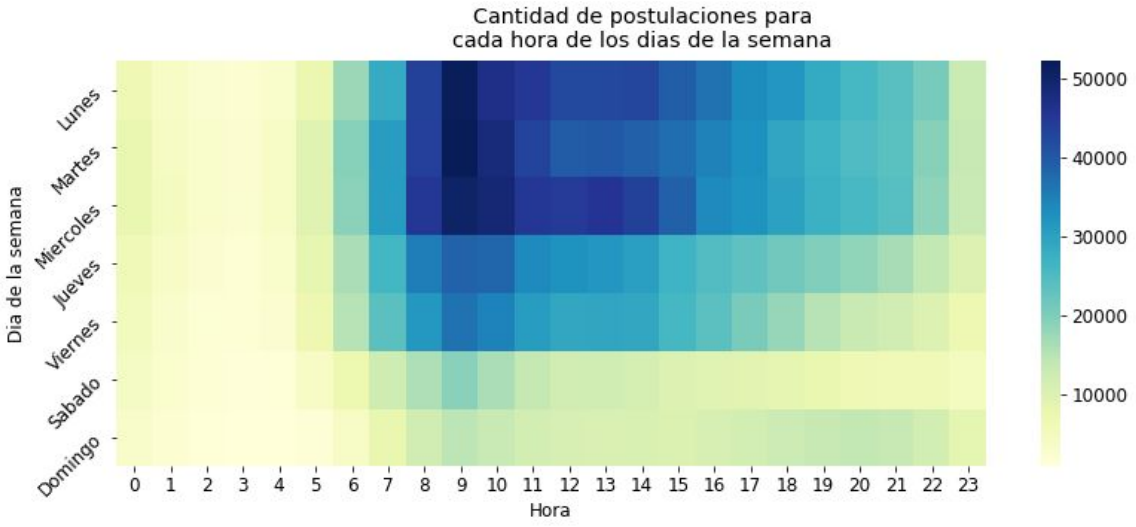
Por lo tanto, esperamos que el fin de semana tengo menor cantidad de postulaciones que el resto de los días.

El gráfico anterior parece alabar nuestra hipótesis, además de mostrar que en los jueves y viernes la cantidad de postulaciones decrece significativamente. Veremos si sucede lo mismo con las visitas:



En este caso nos topamos con algo muy extraño y es que no existen visitas para ningún día jueves del set de datos, tratamos de buscar este día de varias maneras pero siempre recibimos el mismo resultado. Desconfiamos de este gráfico y optamos por no tenerlo en cuenta ya que no sabemos qué está sucediendo con estos datos.

Como contamos con las visitas por día y por hora, se nos ocurrió que podíamos graficar esa relación, para ver si se obtenía alguna información extra, o se conseguía un nuevo punto de vista:



Este gráfico resume rápidamente lo que vimos en los gráficos anteriores sobre postulaciones. La gente no suele postularse en la noche ni durante los fines de semana. Por otra parte, se nota mucho la caída en la cantidad de postulaciones el día jueves, que sigue hasta el domingo, donde se eleva un poco nuevamente. También se puede apreciar que las 9 de la mañana es la hora pico de postulaciones a lo largo de toda la semana.

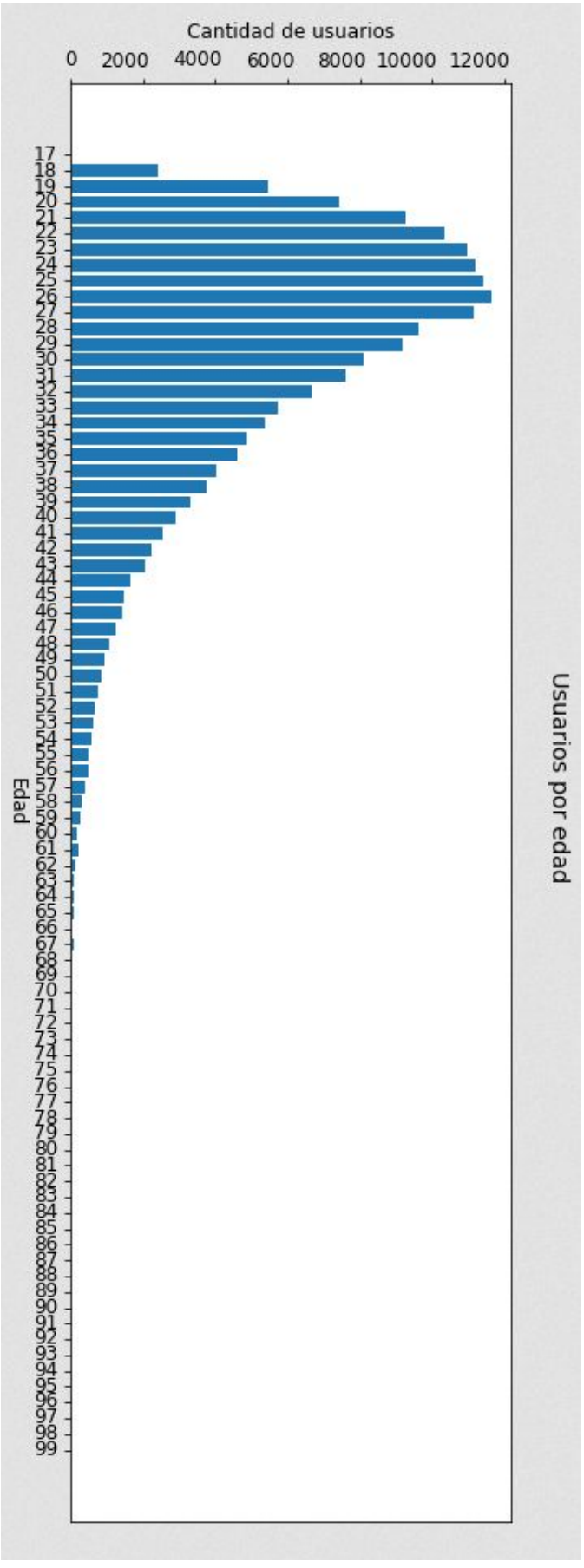
Nuestro último estudio temporal es sobre las edades de los usuarios. En este caso nos encontramos con algunos outliers. Por una parte un usuario tenía 12 años, que filtramos por ser demasiado joven, por el otro nos encontramos con los siguientes datos en el extremo opuesto:

	idpostulante	fechanacimiento	sexo	edades
56206	xkPwXwY	0031-12-11	FEM	1987
71458	LN85Y3b	0029-05-11	MASC	1989
130846	8M2R6pz	0024-02-09	FEM	1994
141832	A36Npjj	0033-09-14	FEM	1985
145683	dYjV0rb	0012-11-04	NO_DECLARA	2006
148638	GNZOvAv	0004-07-19	MASC	2014
149653	1QPQ8QL	0011-03-08	MASC	2007
154559	xkdvwm0	1775-07-09	MASC	243
164618	96X1loa	1917-07-08	MASC	101

En el caso del hombre de 101 años dudamos si dejarlo o no, pero decidimos arbitrariamente que el límite sea de 100 años. Esto es porque la gente tan mayor es poco probable que busque trabajo y que realmente hayan ingresado ellos a la página ya que no suelen ser muy hábiles con la tecnología. La gente de entre 90 y 99 años es el límite superior. Estos outliers probablemente se deban a que la gente ingresó su fecha de nacimiento en un

formato incorrecto, por ejemplo, la fecha de nacimiento de un usuario es 0031-12-11 (realmente no comprendemos que quiso ingresar dado que su edad sería de 7 u 8 años). El formato utilizado en el csv de género y edad corresponde con '%Y-%m-%d'.

Finalmente, analizaremos la cantidad de usuarios por edad. Esperamos que la mayoría esté entre los 18 y 30 años, porque es el grupo de personas que busca activamente trabajo y no está completamente ligada a los métodos tradicionales de búsqueda de empleo. Para graficar esto hicimos un histograma con la cantidad de usuarios por edad (de 18 a 99 años):


Cantidad de usuarios
0
2000
4000
6000
8000
10000
12000
Edad
Usuarios por edad

En las edades superiores a los 67 años no se puede apreciar una barra dado que la cantidad de usuarios para estas edades es muy pequeña en comparación a las demás (por ejemplo, no hay usuarios entre 80 y 90 años). Nuestra predicción sobre el rango de edades con mayor volúmen de usuarios se cumplió y, a su vez, podemos observar cómo crece la cantidad de usuarios hasta 26 años, alcanzando su pico, para luego decrecer en forma progresiva.

Conclusiones

De acuerdo al análisis realizado, podemos responder las preguntas que nos planteamos antes de llevar a cabo el análisis exploratorio de los sets de datos.

La primer pregunta planteada fue si existía un sexo predominante entre los usuarios de Navent. Pudimos determinar que se distribuyen de forma pareja, si bien hay más usuarios de género femenino que masculino (una cantidad ínfima de usuarios no marcaron su sexo).

Nuestra segunda pregunta fue con respecto a la zona de los avisos y pudimos determinar que Gran Buenos Aires abarca la gran mayoría de los avisos, en contraposición a nuestra hipótesis en donde considerábamos que habría una mayor cantidad de anuncios en CABA.

En cuanto a la investigación sobre la distribución de los estudios de los usuarios, observamos que la gran mayoría de los usuarios se concentra, en primer lugar, en gente con el secundario completo y, en segundo lugar, en personas con el universitario completo, ambos sin ningún estudio en curso. Son muy pocos los usuarios que ingresaron como tipo de estudio un Doctorado, Posgrado y Máster, ya sea habiéndose graduado o estando en curso. Como se dijo en el análisis correspondiente, esto era de esperarse dado que una cantidad baja de gente persigue estos títulos. Sin embargo la distribución de las otras columnas cambia drásticamente al analizar la columna ‘estudio_en_curso’.

Analizando las postulaciones por estudio descubrimos que la gran mayoría pertenece a usuarios que no tienen nada en curso. Consideramos a que esto se debe a que muchos usuarios caen dentro de esa categoría. Lo mismo ocurre con los usuarios graduados del secundario con respecto a los demás graduados.

Otra de nuestras dudas iniciales era si la cantidad de postulaciones por aviso respondía a una ley de potencias dado que, después de una mirada inicial del set de postulaciones, observamos que existían unos pocos usuarios con muchas postulaciones y muchos con una cantidad baja. Al analizar en detalle, pudimos concluir que dicha distribución no se condice con una ley de potencias dado que el 20% de los usuarios con más postulaciones abarcaba el 65% de las mismas, y no el 80% como se debe cumplir en esta ley.

La quinta pregunta planteada hace referencia a los niveles laborales que se muestran en el archivo de detalles de los avisos. El nivel más buscado es Senior/Semi-Senior, esto trae como consecuencia que la mayoría de los postulantes se decante por esa opción. Viendo los gráficos de cantidad de avisos y postulaciones por nivel pudimos descubrir que los niveles están en el mismo orden y proporción en ambos gráficos, por lo que concluimos que las postulaciones por nivel dependen, si bien no linealmente, de la cantidad de avisos de cada nivel. Al ver la relación del nivel buscado con los estudios completos de los postulantes, vimos que tanto el nivel como el estudio que más postulantes acumulaban, formaban la combinación de estudio/nivel con mayor volumen (Senior/Semi-Senior con Secundario). Algo análogo vimos al buscar la relación entre el nivel buscado con el tipo de trabajo, donde, Senior/Semi-Senior en nivel, y Full time en tipo, acumulaban la mayor cantidad de postulantes.

En cuanto a cómo se distribuyen las postulaciones en el tiempo, observamos que se reparten de forma similar durante los días de semana (con una recaída los días jueves y viernes) y disminuyen considerablemente en los fines de semana. Respecto a los horarios en los que se postula la gente, estas comienzan a crecer de forma muy acelerada a partir de las 5 de la mañana, que es cuando la gente se despierta (anteriormente el uso es extremadamente bajo porque la mayoría de las personas suele estar durmiendo). El pico se alcanza a las 9 de la mañana y a partir de ahí comienzan a caer lentamente. A partir de observar el heatmap de cantidad de postulaciones por día y hora, pudimos concluir que este comportamiento se mantiene durante toda la semana y es proporcional a la cantidad de postulaciones por día (mayor concentración los días Lunes, Martes y Miércoles, una reducción los Jueves y Viernes y finalmente una caída brusca los fines de semana).

Para concluir, al analizar las edades de los usuarios que tienen registros en el sistema, determinamos que la gran mayoría de estos se encuentran concentrados entre los 20 y los 30 años. La cantidad de usuarios crece progresivamente desde los 18 hasta los 26 años y a partir de esta edad comienza a decrecer año a año. Como era de esperarse, la cantidad de usuarios en edades superiores a los 50 años es muy baja.