Universidad de Buenos Aires

Facultad de Ingeniería

(75.06) Organización de Datos

Cátedra: Luis Argerich

Análisis exploratorio del set de datos Navent
Repositorio: https://bitbucket.org/ignamiguel/7506-datos
1° Cuatrimestre 2018

Alumnos:
GARATE, Julián Matías (93043)
IGLESIAS, Ignacio (95050)
SHOKIDA, German (96172)
TULIPANI, Gaston (96570)

Materia: 75.06 Organización de Datos		Repo: https://bitbucket.org/ignamiguel/7506-datos
Grupo: Bitcoin	Cuatrimestre: 1C 2018	Informe: TP 1

Contenido

Materia: 75.06 Organización de Datos		Repo: https://bitbucket.org/ignamiguel/7506-datos
Grupo: Bitcoin	Cuatrimestre: 1C 2018	Informe: TP 1

Introducción

La empresa Navent tiene un servicio de búsqueda de empleo y vivienda.
Permite publicar avisos de empleos o propiedades y acercar los usuarios a las ofertas del mercado.
En este sentido el presente trabajo se basa en realizar un análisis exploratorio de la base de datos de la empresa de los avisos de enero y febrero 2018.
Link: http://www.navent.com/

Análisis del set de Datos

El set de datos está compuesto de los siguientes archivos:

#	Archivo	Estructura
1	fiuba_1_postulantes_educacion.csv	idpostulante | nombre | estado
2	fiuba_2_postulantes_genero_y_edad.csv	idpostulante | fechanacimiento | sexo
3	fiuba_3_vistas.csv	idAviso | timestamp | idpostulante
4	fiuba_4_postulaciones.csv	idaviso | idpostulante | fechapostulacion
5	fiuba_5_avisos_online.csv	idaviso
6	fiuba_6_avisos_detalle.csv	idaviso | idpais | titulo | descripcion | nombre_zona | ciudad | mapacalle | tipo_de_trabajo | nivel_laboral | nombre_area | denominacion_empresa

Lenguaje de Programación

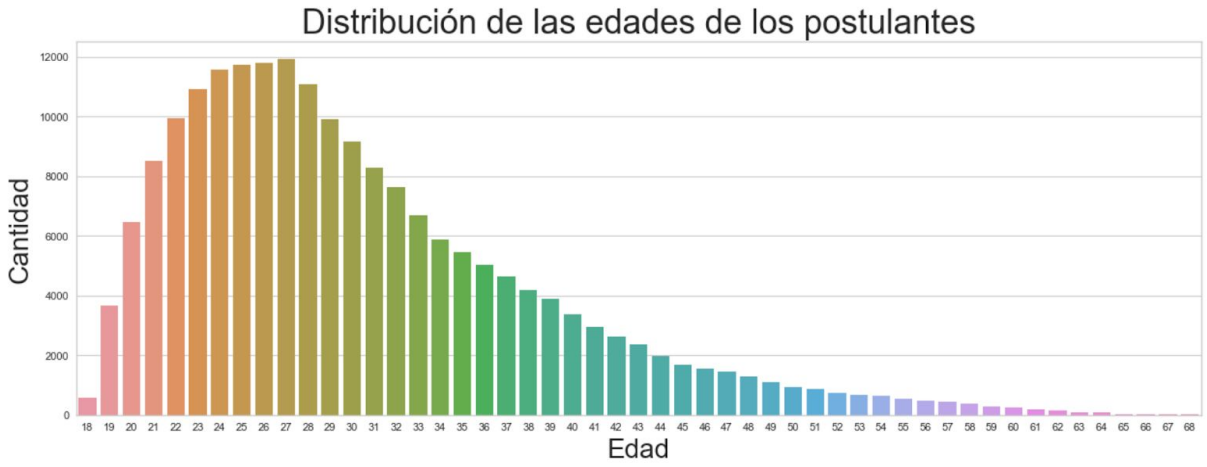
El análisis está hecho en Python.

Exploración de los Sets de Datos

Empezamos evaluando los postulantes.

Distribución de los postulantes por edad

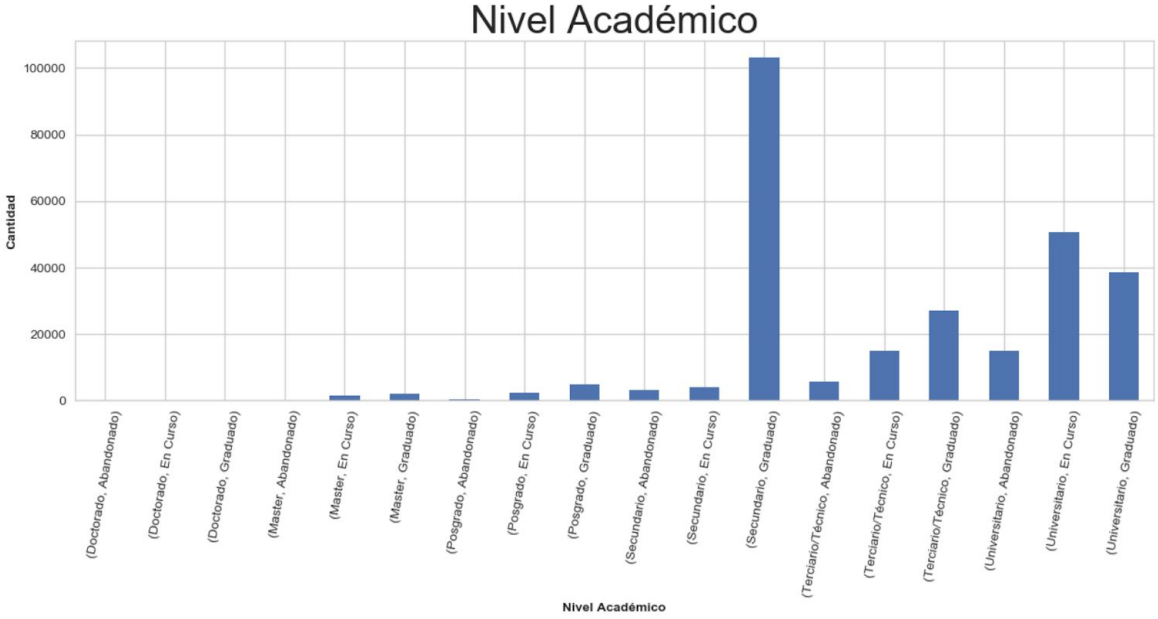
Vemos que la mayor concentración de postulantes del dataset se encuentra en el grupo de entre los 20 y 30 años, principalmente entre 25 a 28.



Queremos ver si este comportamiento está relacionado a que generalmente las personas completan la formación universitaria en ese rango de edad.

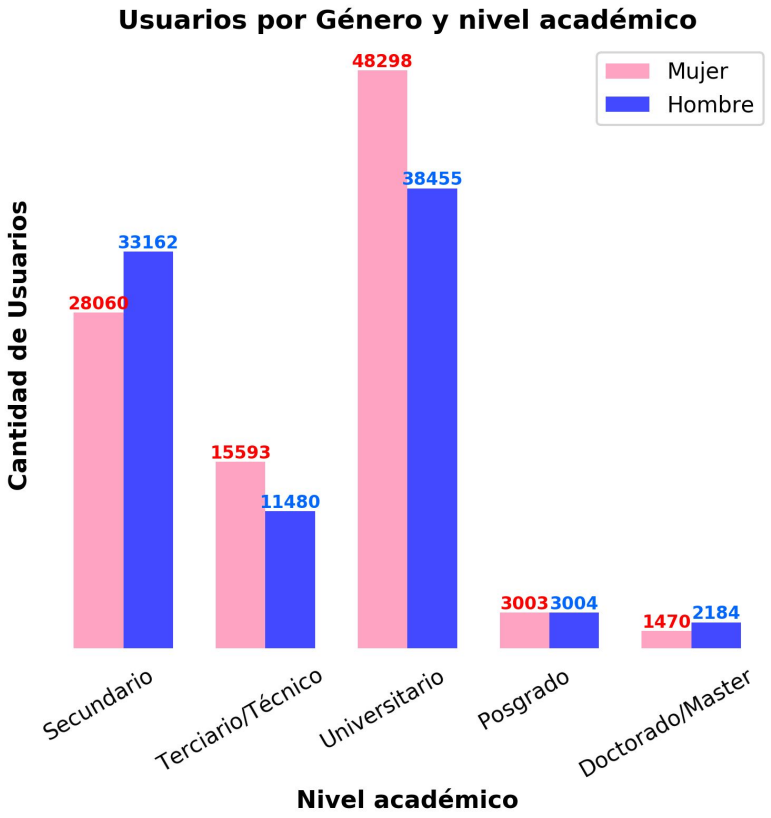
Nivel Académico

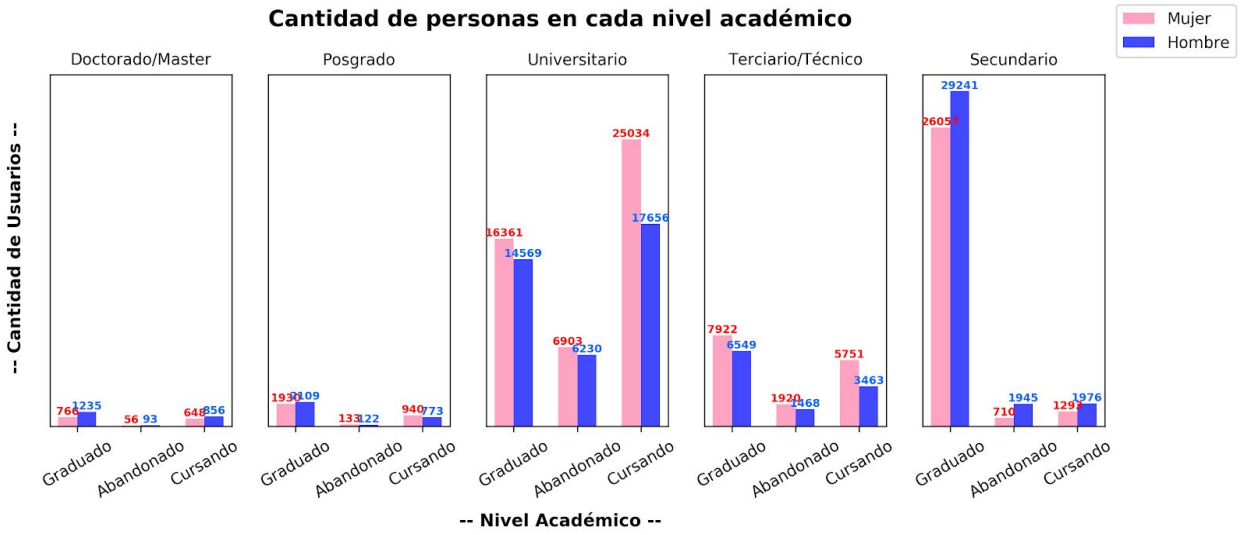
Agrupando por nivel académico y teniendo en cuenta el estado (En curso, Graduado, Abandonado, etc), veamos la distribución de los postulantes.



Nivel Académico por género

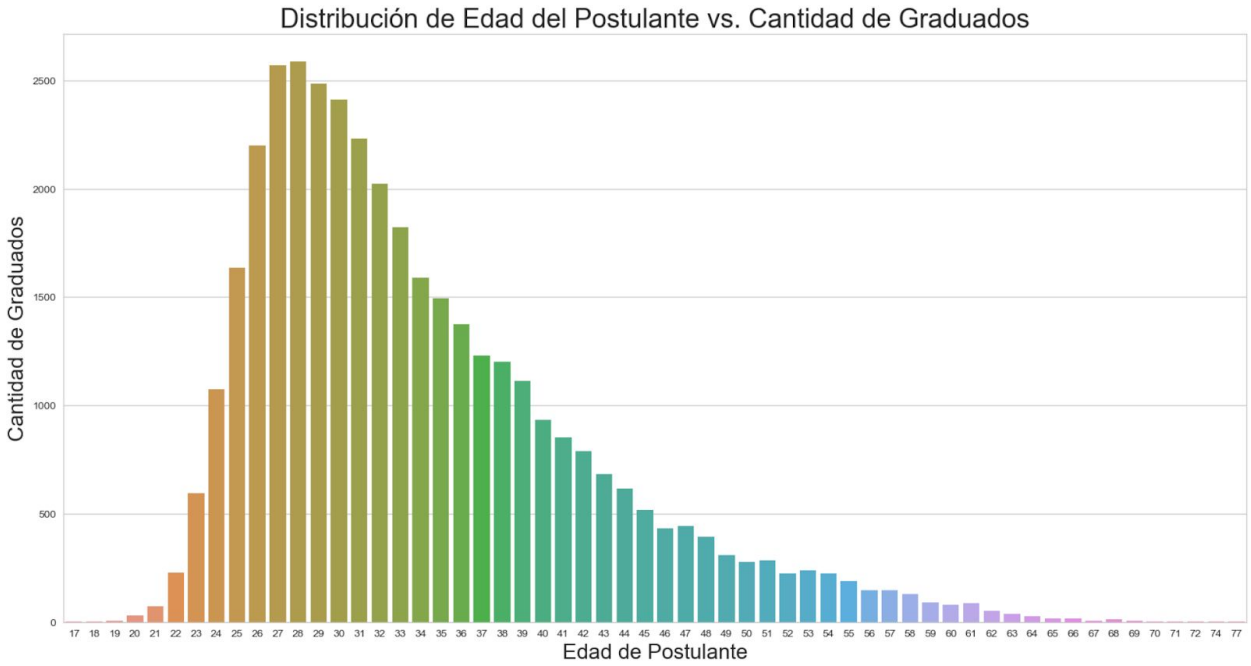
Si analizamos también por el género de los postulantes encontramos algunas diferencias entre hombre y mujeres. Las postulantes mujeres parecieran a primera vista que tienen un nivel académico superior al de los hombre.Como se puede ver en el gráfico siguiente luego de agrupar por nivel académico a los postulantes y discriminarlos por género, las mujeres aunque son menos que los postulantes hombres, pero tienen perfiles más profesionales.





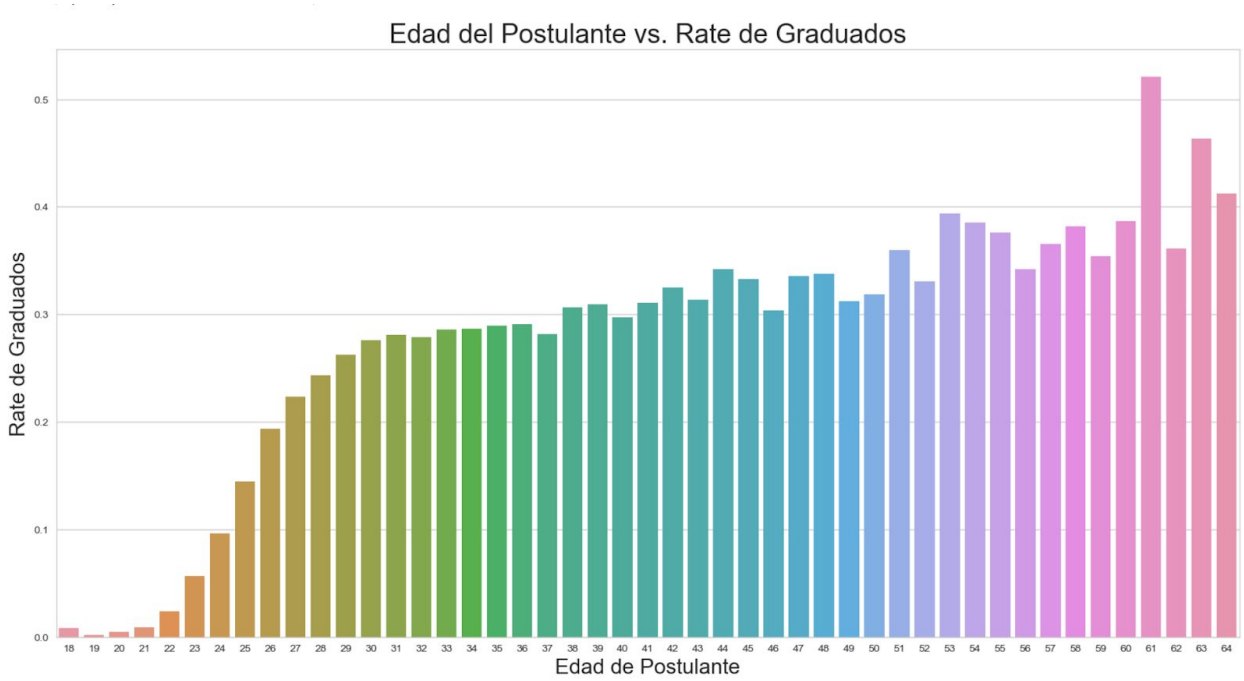
Distribución de Postulantes Graduados

Otro detalle interesante a analizar respecto al Nivel Académico de los postulantes, es la distribución de la cantidad de Graduados Universitarios de acuerdo a la edad:



Si bien no contamos con la información sobre la “**Fecha de graduación**”, podemos darnos una idea aproximada sobre la edad de graduación de los postulantes de la plataforma (entre los 26 y los 30 años). Sin embargo, no es casualidad que la forma del gráfico sea similar a la de la

distribución por edades, ya que tiene sentido que *a mayor cantidad de postulantes por edad*, *también mayor cantidad de graduados para esa misma edad*. Analicemos ahora, la cantidad de graduados por edad, pero tomando un valor relativo (respecto a la cantidad total de postulantes por edad):

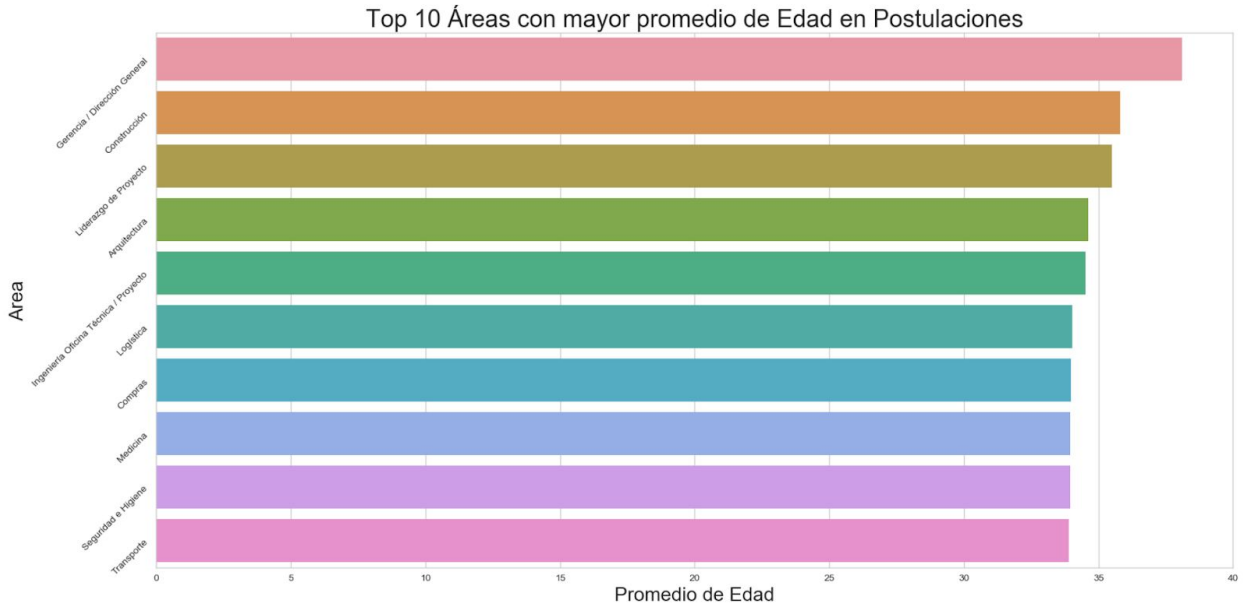


Como podemos ver ahora, la forma del gráfico ha cambiado, y la cifra relativa de graduados por edad se mantiene medianamente constante entre los 30 y los 40 años. Otro detalle interesante es el bajo **rate** de graduados en general, ya que podemos ver que en promedio este se encuentra en un valor del 30%.
Es útil contar con toda esta información al momento de crear un aviso en la plataforma, ya que tenemos una idea tangible sobre el nivel académico que podemos esperar (o aspirar) de acuerdo a la edad.

Segmentación del Mercado Laboral por Edad

Siguiendo con esta misma línea, nos interesaría saber si el Mercado Laboral se encuentra segmentado en distintas áreas de acuerdo a la edad. Es decir, si hay algunas áreas de trabajo en las que predominan los postulantes jóvenes, y otras en las que predominan postulantes más grandes. Para ello, podemos tomar las 10 Áreas con menor promedio de edad entre sus postulantes:

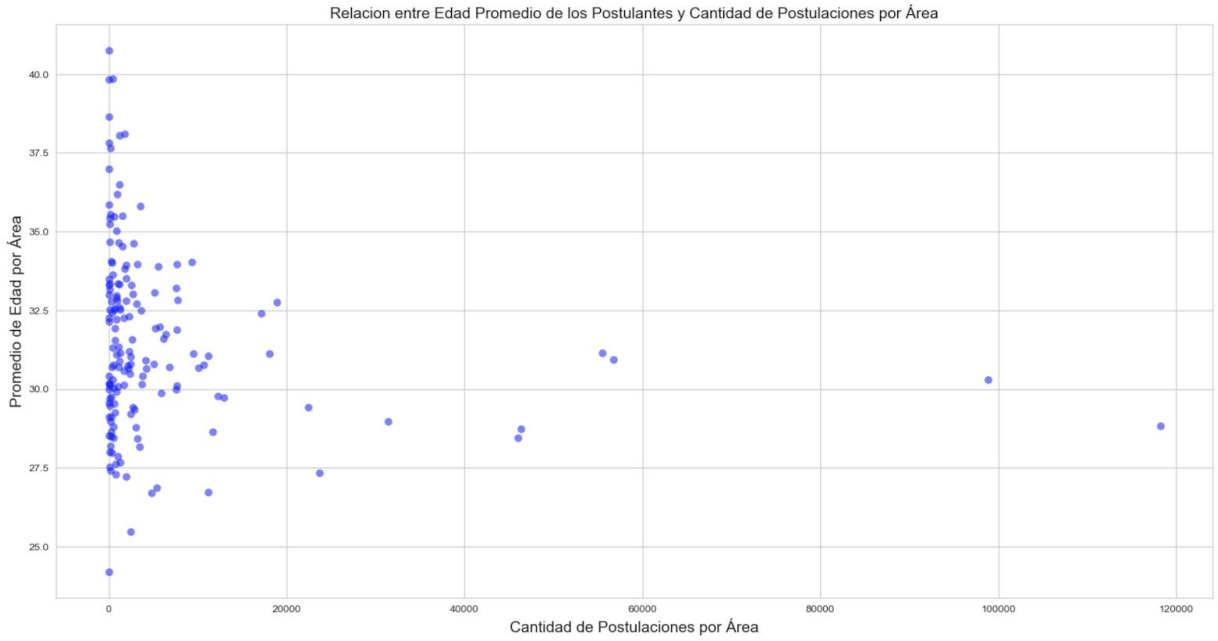
Podemos ver que algunas áreas tienen un promedio de edad de **~26 años**, cómo *Cadetería*, *Telemarketing*, *Data Entry o Atención al Cliente*, lo cual tiene mucho sentido, ya que son tareas para las que no se requiere mucha experiencia previa o conocimiento técnico. Esto puede ser útil para la plataforma, al momento de recomendar avisos de trabajo a perfiles jóvenes. Analicemos ahora el mismo escenario para perfiles de mayor edad:



Podemos ver que algunas áreas tiene un promedio de edad de **~35 años**, cómo *Gerencia*, *Arquitectura* o *Ingeniería*, lo cual tiene mucho sentido también, ya que son perfiles que requieren experiencia, manejo de equipo de personas y conocimiento técnico. Esta información también es útil para la plataforma, ya que por ejemplo no tendría sentido recomendar un aviso en estas áreas a un perfil recién graduado.

Relación Edad Promedio-Cantidad de Postulaciones por Área

Tratemos de analizar ahora si existe una relación entre la edad promedio de los postulantes para un área, y la cantidad de postulaciones a la misma a través de un **Scatter Plot**:

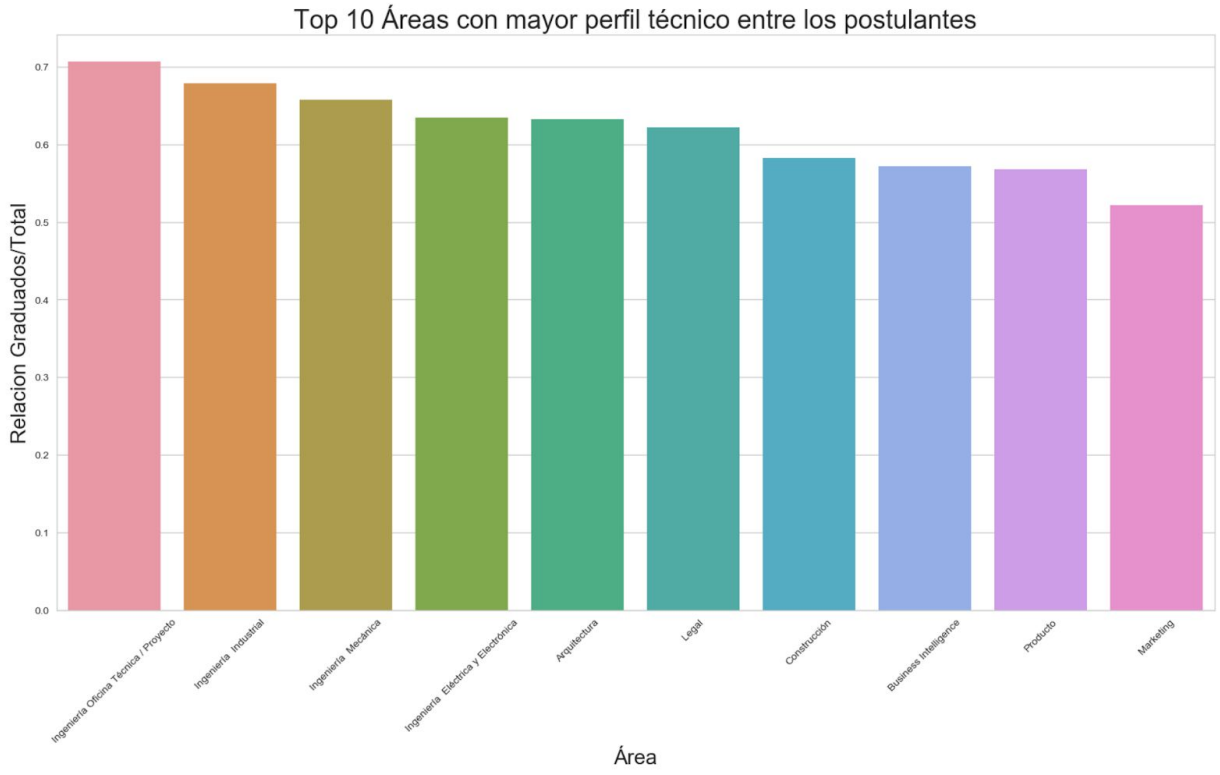


Hay mucha información interesante que se puede extraer del gráfico anterior:

- Las áreas con mayor edad promedio en sus postulantes, son las *menos frecuentes* en el sitio. Es decir que, por ejemplo, será complicado encontrar avisos en el sitio que busquen perfiles de más de 35 años.
- Las áreas con edad promedio entre 30-35 años son las *más frecuentes* en el sitio.
- Prácticamente no existen áreas con edad promedio menor a 25 años, lo cual representa una señal de la *poca búsqueda* - o la no búsqueda de trabajo - entre los jóvenes recién egresados de la Escuela Secundaria.

Perfil Técnico/Académico del Mercado Laboral

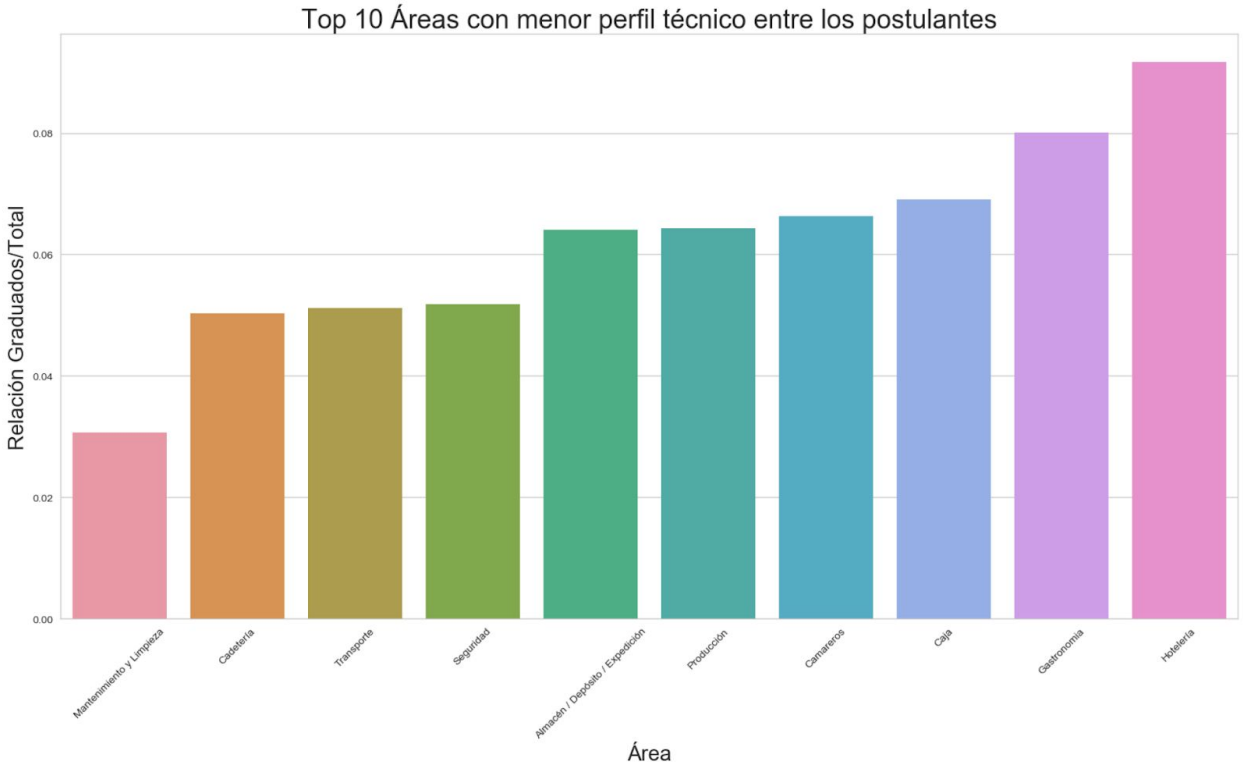
Nos interesa conocer ahora cuáles son las Áreas, cuyos postulantes tienen el mayor nivel académico. Para ello, nos proponemos a calcular nuevamente la cifra relativa de graduados respecto al total para cada área de trabajo:



Podemos sacar algunas conclusiones útiles sobre el gráfico anterior:

- En el área *Ingenieril*, *Legal* y *Construcción* existe una gran oferta de perfil técnico entre sus postulantes.
- Estos perfiles seguro se condicen con una mayor remuneración económica.
- Las empresas interesadas pueden tomar ventaja de este conocimiento y pedir un Título Universitario como condición obligatoria para sus búsquedas.
- Los empleados postulantes a estas áreas deben tener en cuenta que la falta de un Título Universitario puede significa una gran desventaja con respecto al resto

Analicemos ahora las áreas cuyos postulantes tienen el menor nivel académico:

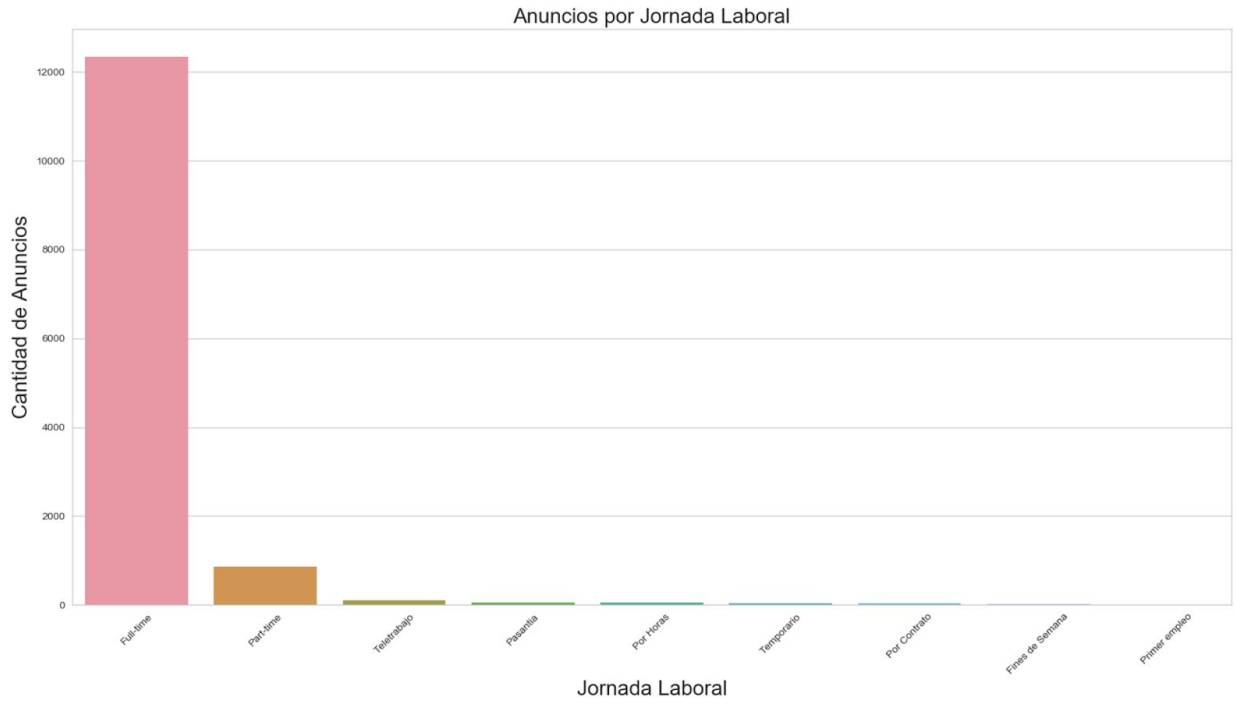


Nuevamente podemos sacar algunas conclusiones útiles sobre el gráfico anterior:

- En el área *Mantenimiento y Limpieza*, *Transporte* y *Seguridad* existe una baja oferta de perfil técnico entre sus postulantes. Por lo tanto, estos perfiles seguro se condicen con una menor remuneración económica.
- Estos perfiles seguro se condicen con una menor remuneración económica.

Modalidad de Trabajo en el Mercado Laboral

Un detalle que nos puede interesar sobre los avisos, es el tipo de Jornada Laboral al que las empresas aspiran para sus trabajadores. Podemos tener una idea de esto clasificando cada uno de los avisos de acuerdo al tipo de trabajo requerido:

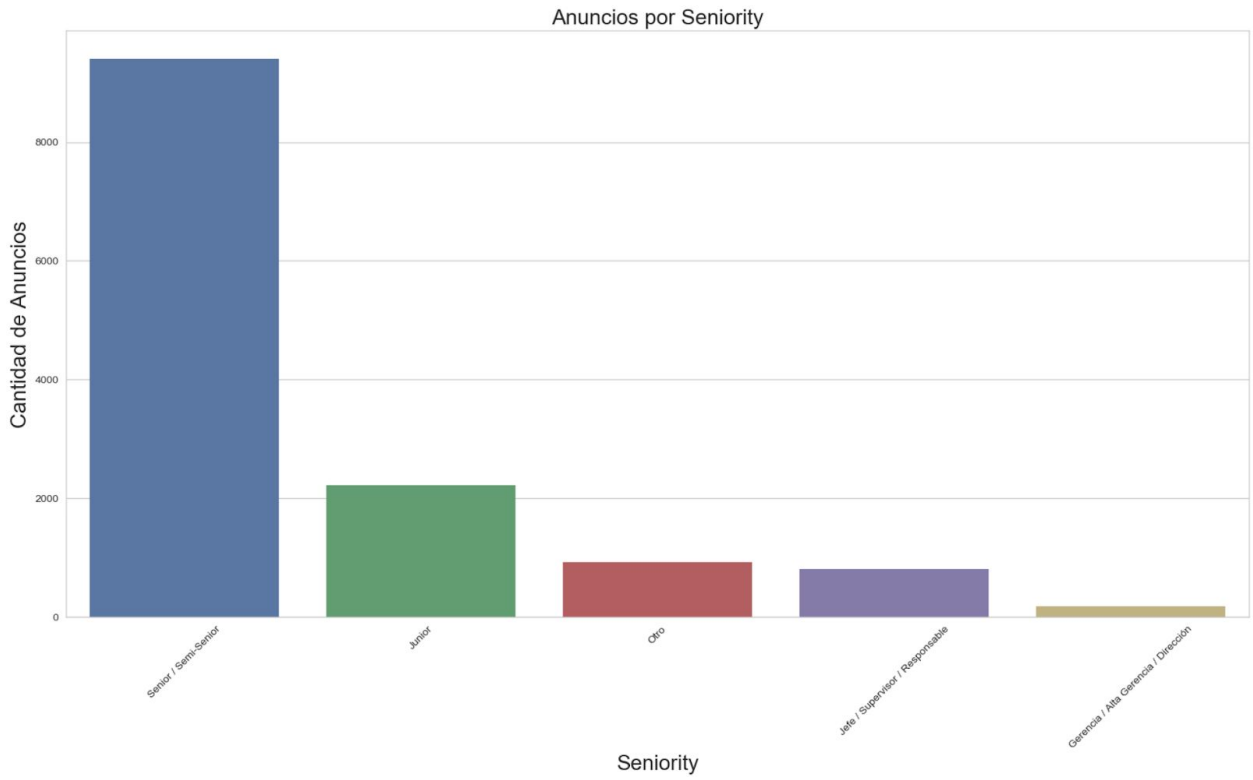


Podemos sacar algunas conclusiones de este gráfico:

- La inmensa mayoría de los anuncios del sitio son para trabajos de modalidad **Full-time**.
- Si bien hay una gran cantidad de ofertas para trabajos **Part-time** *(~1000 anuncios*), es claro que la jornada laboral *menor a 8 horas diarias* sigue siendo una utopía para aquellos estudiantes que deseen sumar experiencia previo a recibirse.

Seniority buscada en el Mercado Laboral

Otro detalle que nos puede interesar sobre los avisos, es el nivel de Seniority que las empresas buscan para llenar sus posiciones. Podemos tener una idea de esto clasificando cada uno de los avisos de acuerdo al nivel de experiencia deseada:

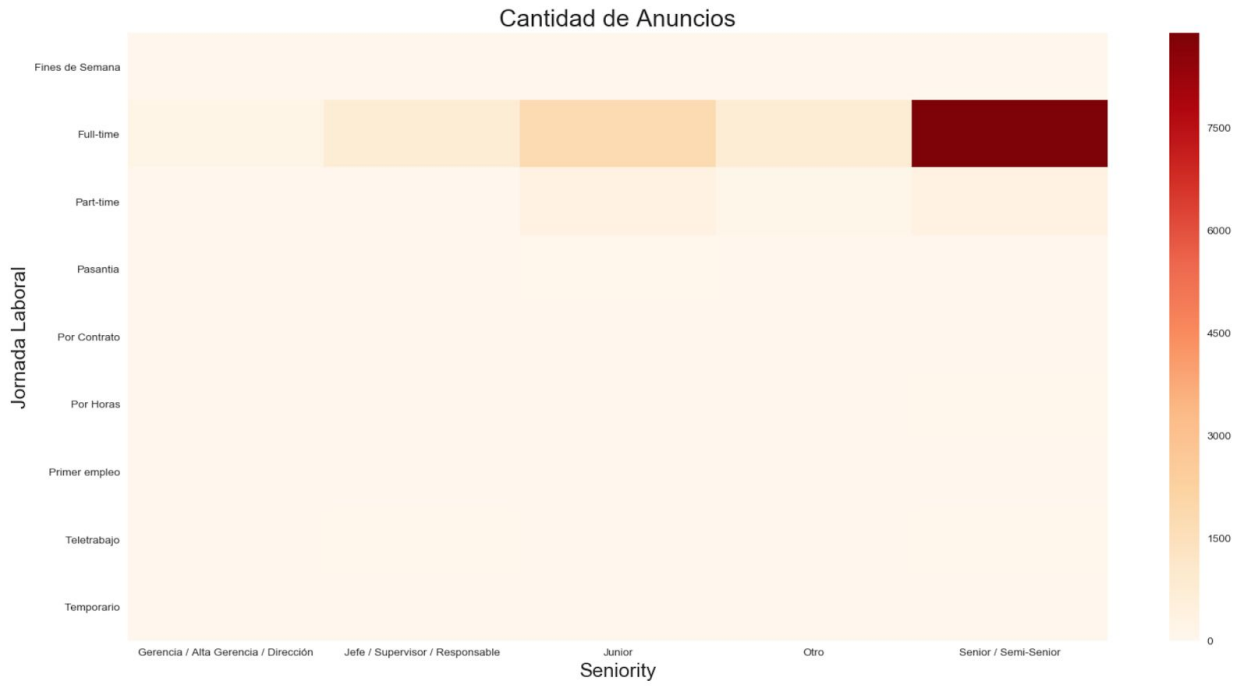


Podemos sacar múltiples conclusiones de este gráfico:

- La mayoría de los anuncios del sitio buscan un perfil **Senior / Semi-Senior**.
- Si bien hay una gran cantidad de ofertas para un nivel de experiencia inicial (**Junior**), es claro que sigue siendo muy complicado encontrar *el primer trabajo*, ya que la mayoría de las empresas parecen buscar perfiles con mucha más experiencia.

Relación Jornada Laboral-Seniority

Luego de analizar los gráficos anteriores, vale la pena ver la relación que existe entre la experiencia buscada en el Mercado Laboral y el tipo de jornada. Podemos observar esta relación a través de un **Heat Map**:

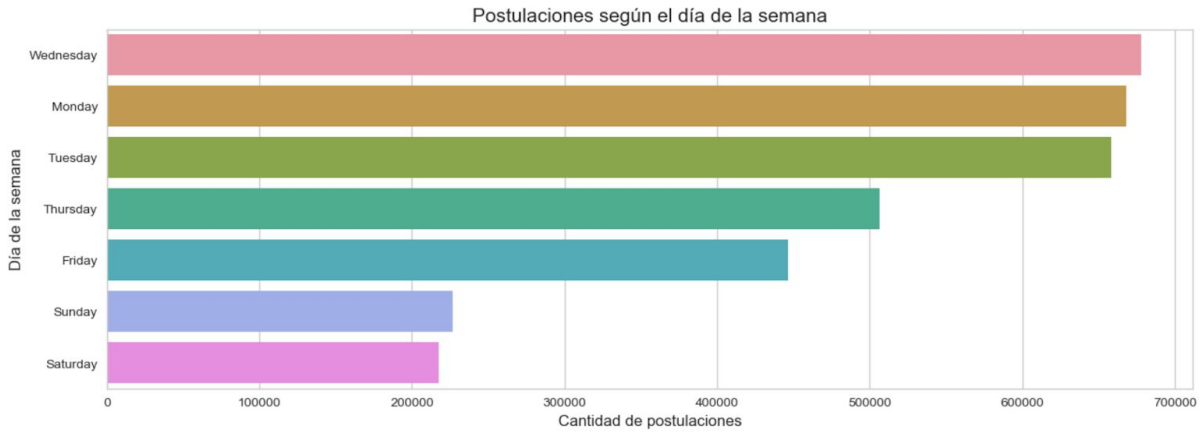


Algunas conclusiones visibles del gráfico:

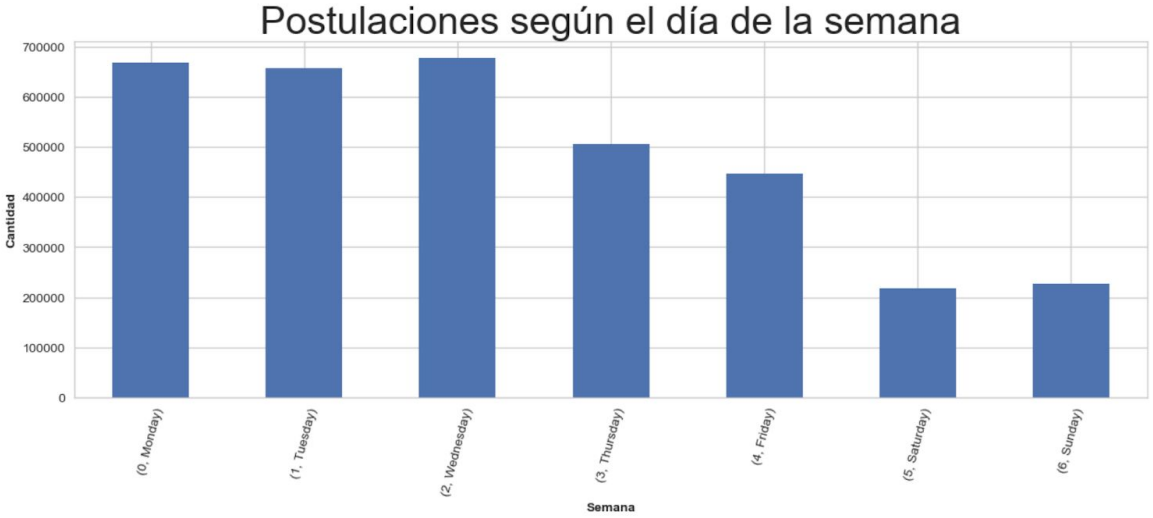
- Claramente la **gran mayoría** de los anuncios laborales persiguen tanto una jornada *Full-time* como un perfil *Senior / Semi-Senior*.
- La búsqueda de otro tipo de combinación de Jornada Laboral - Experiencia resulta muy complicada para los postulantes.

Publicaciones según el día de la semana

Si analizamos las publicaciones por día de la semana, vemos que los días de mayor postulaciones son los Miércoles, Lunes y Martes, por orden de cantidad.



Si analizamos la evolución de la semana, vemos que la cantidad de publicaciones disminuye a medida que nos acercamos al fin de semana.

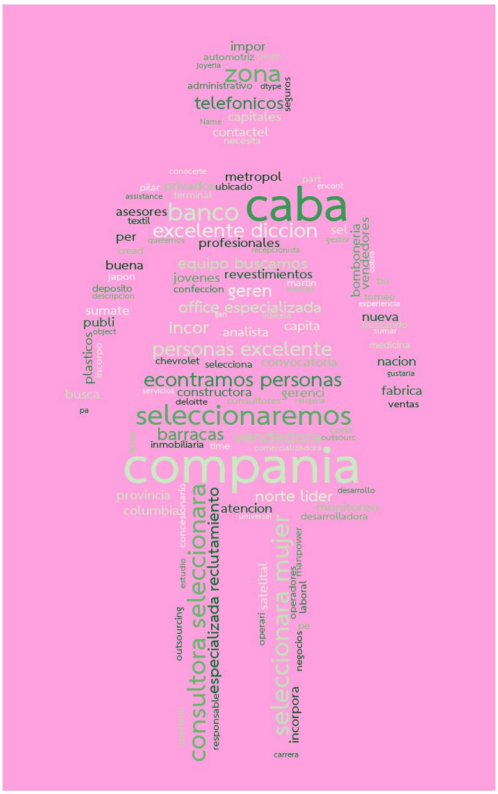


Análisis de las descripciones

Los textos de las descripciones tienen información explícita, de las búsquedas laborales para cada empresa y también tendencias del conjunto.
En el wordcloud siguiente se pueden apreciar palabras como, servicios, ventas, líderes, programadores, analista, son las que poseen mayor frecuencia de aparición, dándonos a conocer un poco de información de cómo se comporta el mercado laboral.

Las descripciones si se analizan junto a las postulaciones y las vistas que recibe cada publicación, se puede ver como dependiendo del género de los usuarios hay diferentes tendencias de palabras predominantes.

Palabras a las que, las mujeres, más ven.



Palabras a las que, las mujeres, más se postulan.

Como para las mujeres también se hace el mismo análisis para los hombres.
Las apreciaciones a primera vista, dan conocer ciertas palabras que entusiasman más a las mujeres que los hombres y viceversa.
Equipo, **compania**, **cliente**, **banco**; son palabras que tienen mucha frecuencia en las publicaciones elegidas por las mujeres.
Proyecto, **tecnología**, son palabras que son más comunes entre las búsquedas de los hombres.

Palabras a las que, los hombres, más ven.

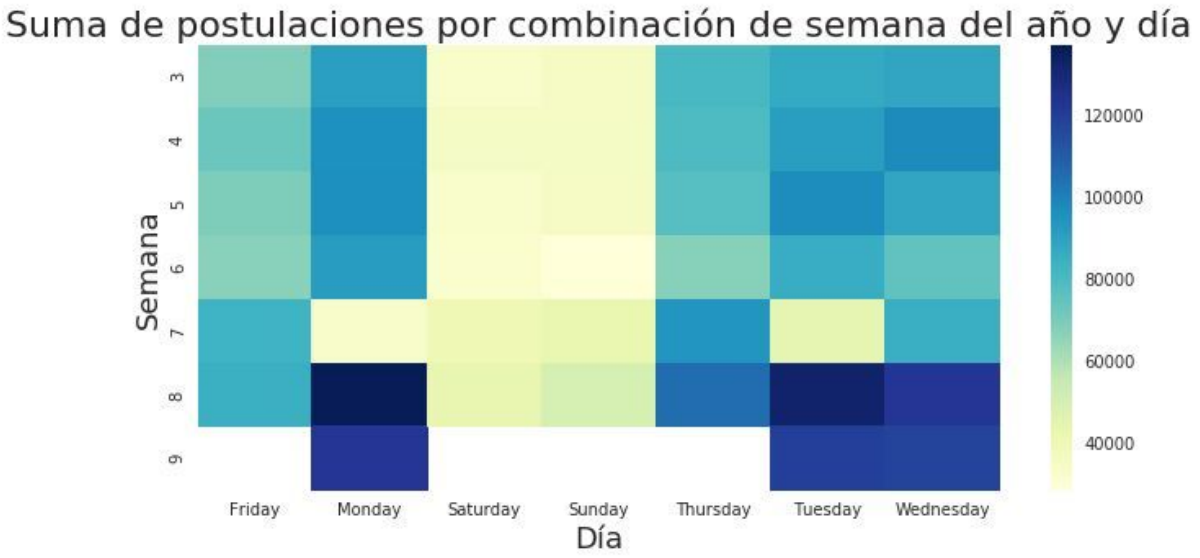
Palabras a las que, los hombres, más se postulan.

Localización

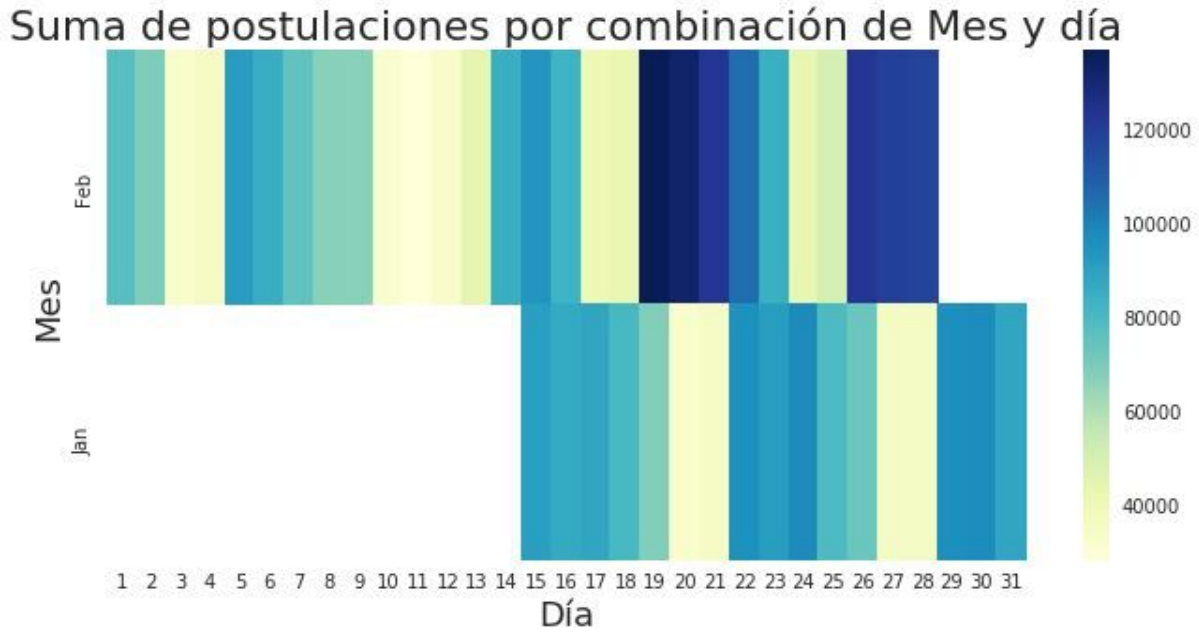
En las publicaciones están algunas de las direcciones de las empresas que publican (o de las consultoras de RRHH). De estas se puede observar la tendencia a acumularse a la zona centro. Más allá de que lamentablemente pocos avisos estaban con los datos en formato correcto, imposibilitando un análisis más significativo.

Postulaciones y vistas por fechas.

Podemos observar que se realizan más postulaciones los días lunes, martes y miércoles.
Pareciera que los datos de enero están incompletos, aunque mantienen la misma tendencia que los de febrero.
También queda claro que los fines de semana es donde menos se postula la gente.
Continuando con las postulaciones vamos a ver qué sucede si separamos las postulaciones por semana:



Dada la información separada por semanas, durante el mes de enero las postulaciones fueron bastante distribuidas en los días laborales. Llegando a fines de febrero hubo más postulaciones que en enero.
Se puede ver que hay un lunes y un martes que tienen muy pocas postulaciones, ahora lo veremos en el siguiente gráfico de forma más detallado.



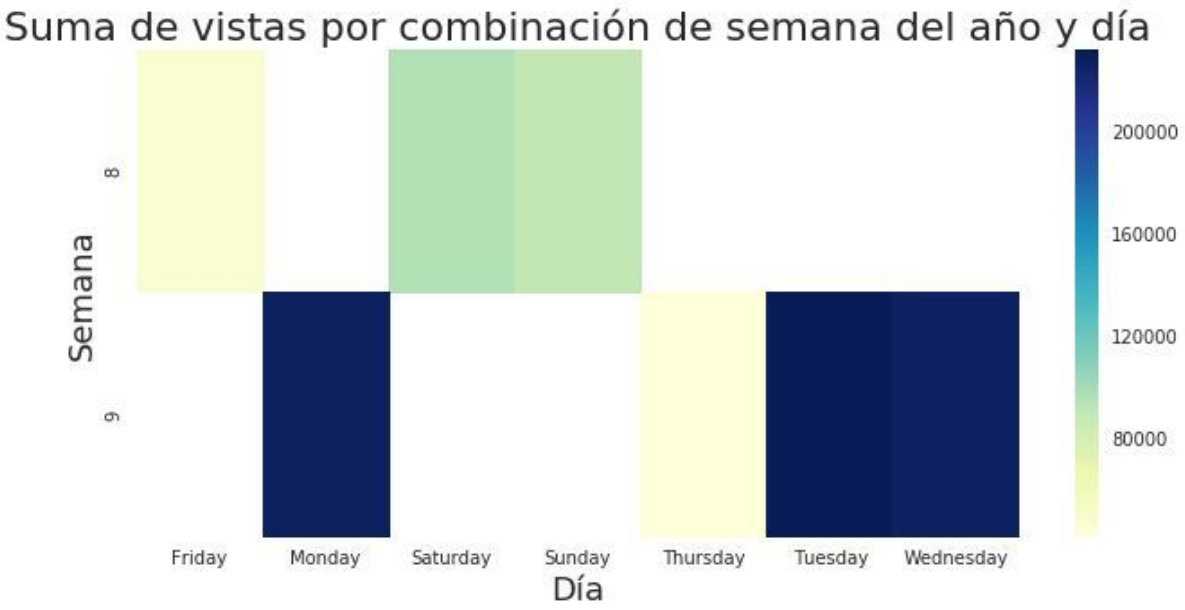
Aquí se aclara porque en la 7ma semana del año hubo pocas publicaciones, corresponden a los días 12 y 13 de febrero, los cuales fueron feriados por Carnaval (por lo visto la gente no se postula los días feriado).

Y confirmamos que tenemos los datos de la mitad del mes de enero.

Una pequeña conclusión que podemos sacar es que en la segunda quincena de febrero, posiblemente cuando vuelven de las vacaciones o luego de un feriado largo, suelen postularse más.

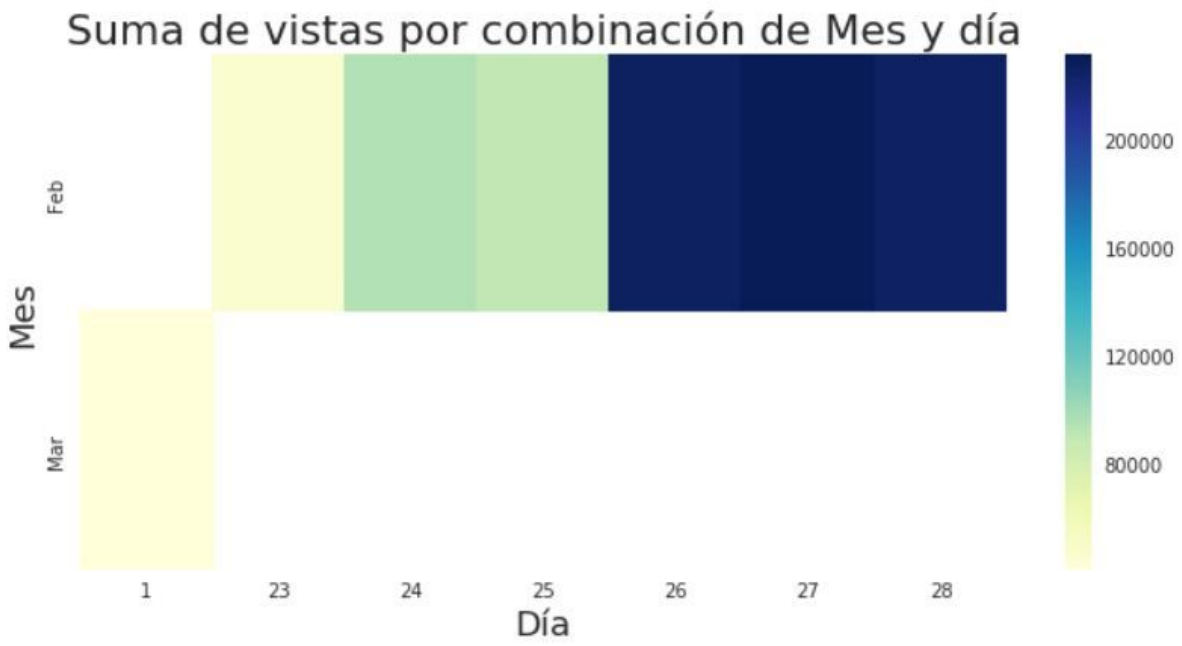
Vistas

Ya que revisamos la información sobre las postulaciones veamos los datos que tenemos de las vistas de los avisos:



En comparación al set de datos de postulaciones, claramente no contiene la información en el mismo rango de fechas, pareciera que tiene solo una semana cargada, y coincide en que los avisos se ven más los días lunes, martes y miércoles, con la diferencia que parece que los sábados y los domingos también se ven avisos.

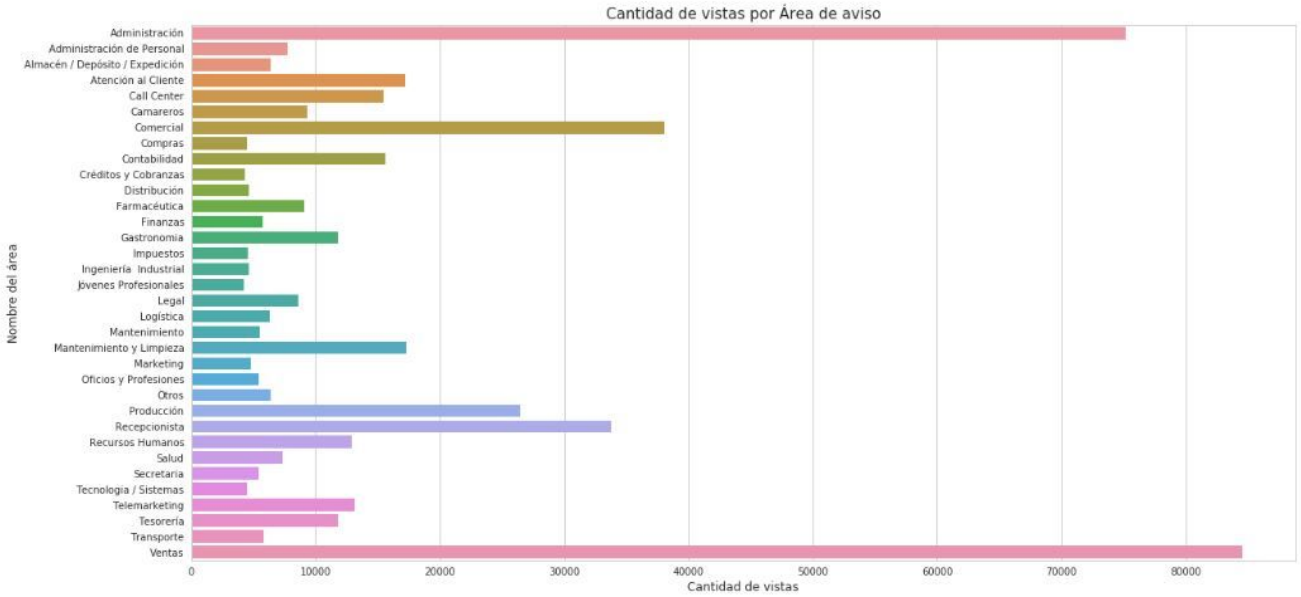
Veamos más en detalle separado por días:



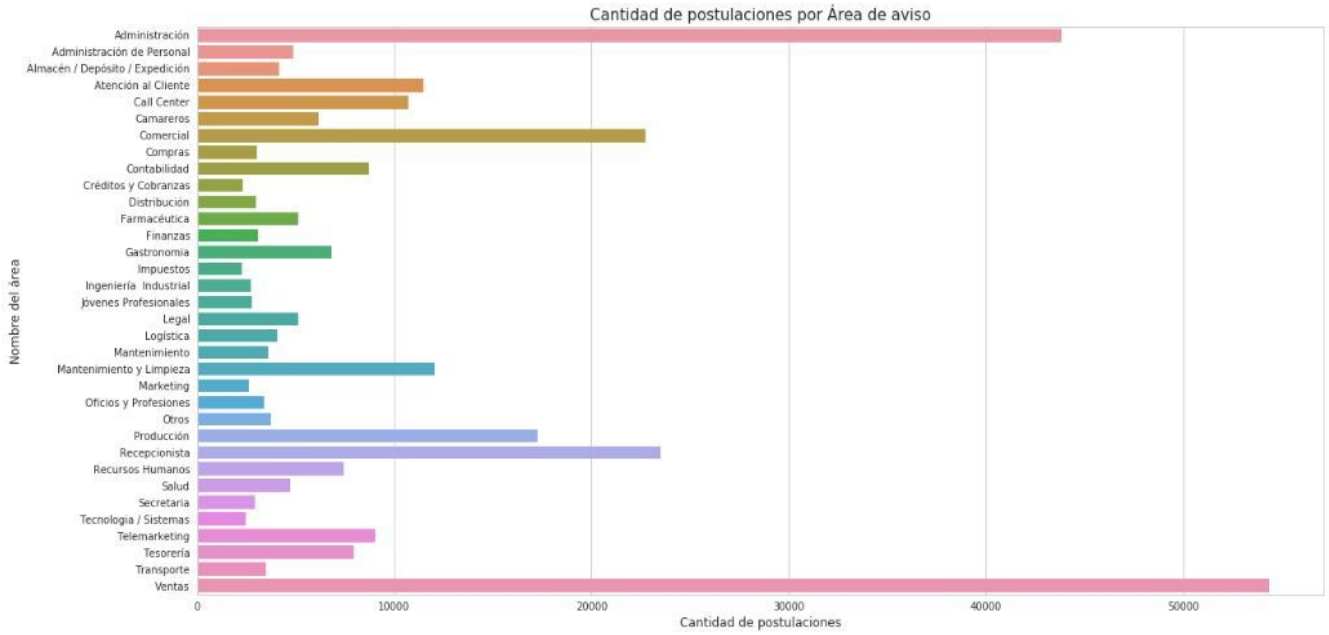
Como suponíamos arriba, tiene una semana completa cargada y observamos los mismo que en el ítem anterior.

Publicaciones y vistas

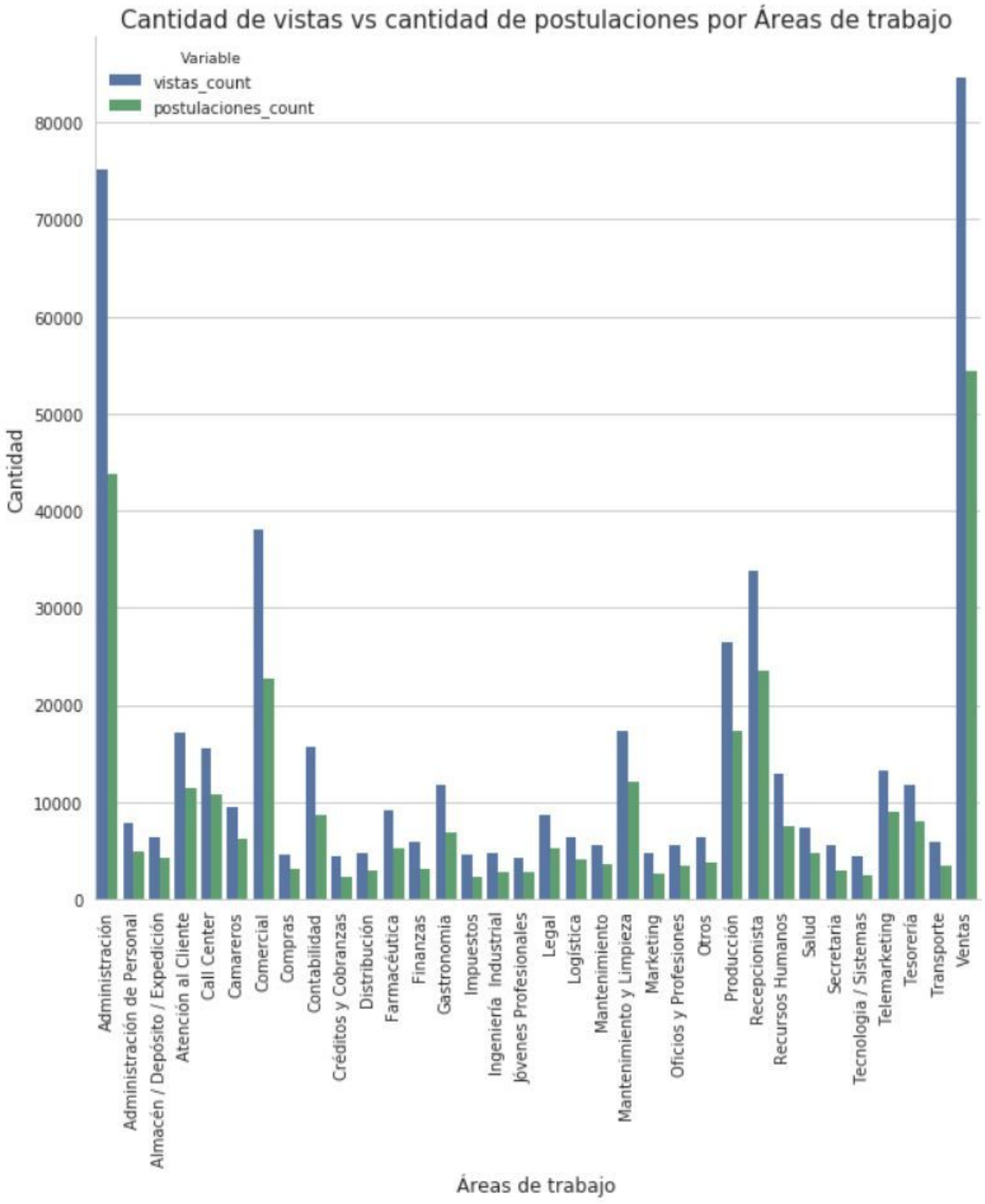
Vamos a ver la cantidad de vistas que hay sobre los avisos agrupados en sus respectivas áreas. Con la restricción de que al menos tengan 4000 vistas.



Aquí le agregamos un restricción más, de rango de fechas, para estar en las mismas condiciones que las vistas, observemos la cantidad de postulaciones:



Juntemos esos dos gráficos para entenderlos un poco mejor:



Observando el gráfico podemos concluir que al menos en estas áreas de trabajo hay mucha cantidad de postulaciones con respecto a las vistas, aunque todo esto queda definido solo para la última semana de febrero, que es donde tenemos los datos. También hay que tener en cuenta que las cantidades de vistas totales con las de postulaciones totales es diferente.

Conclusiones

A continuación listamos algunas conclusiones a las que llegamos luego del análisis de los datos:

#	Tema	Conclusion	Referencia
1	Edad de los postulantes	Entre 20 y 30 años, principalmente entre 25 a 28.	Distribución de los postulantes por edad
2	Jornada laboral	Principalmente jornada completa, seguido en menor medida de part-time.	Modalidad de Trabajo en el Mercado Laboral
3	Nivel Académico	El dataset se caracteriza por el nivel secundario graduado seguido luego por Universitario en curso.	Nivel Académico
4	Género	Las mujeres tienen perfiles más profesionales que los hombres, también buscan diferentes cosas cuando leen las descripciones	Nivel Académico por género
5	Localización	Los datos de las calles están incompletos, pero se puede ver como en la zona centro hay más trabajos o direcciones para hacer entrevistas.	Localización
6	Días de la semana	Las personas tienden a buscar trabajo los días hábiles, principalmente al principio de la semana.	Publicaciones según el día de la semana

Bibliografía

http://pandas.pydata.org/
https://docs.python.org/2/library/re.html
https://bokeh.pydata.org/en/latest/
https://github.com/amueller/word_cloud
https://amueller.github.io/word_cloud/generated/wordcloud.WordCloud.html
https://geopy.readthedocs.io/en/stable/
https://jakevdp.github.io/PythonDataScienceHandbook/
https://matplotlib.org/users/index.html