FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Organización de datos - 75.06/95.58

Trabajo Práctico 1

Análisis exploratorio de datos

1º cuatrimestre 2018

Grupo 8

Nombre	**Padrón**
Rodrigo Etchegaray Campisi	96856
Maximiliano Pagani	94754
Alejandro Nicolas Peña	98529
Fonzalida Miguel Angel	86125

1. Introducción general

Cuando se nos dió una introducción en clase sobre los temas que iba a tratar el TP1, una de las cosas a las que nos recomendaron apuntar en nuestro análisis era la búsqueda de la predicción de postulaciones, entre otros objetivos. Poder predecir a qué anuncio se va a postular cierta persona o conjunto de personas (agrupadas según un criterio), basándonos en el análisis de un gran volumen de datos anteriores.

Ahora bien, no es que nos interese realizar predicciones por el solo hecho de realizar futurología. A nuestro criterio, el objetivo fundamental en este tipo de empresas que ofrecen búsqueda de empleos online (como zonajobs.com), es acercar a la empresa ideal con el empleado ideal, en el menor tiempo posible. Para este fin entendemos que sirven las predicciones de postulaciones mencionadas anteriormente: para ofrecerle a la persona dichos anuncios (de mucha probabilidad de postulación) mediante una recomendación, antes de que se postule a los mismos por cuenta propia. De esta forma, lograríamos acortar el tiempo de conexión entre empresa ideal y empleado ideal.

Desde nuestro punto de vista, entendemos que para realizar estas predicciones hay, fundamentalmente, dos vías de análisis:

- La primera es mediante el estudio de los antecedentes de comportamiento de un usuario. Este estudio se focaliza más en el análisis individual de cada postulante, cosa que para el objeto del TP no nos sirve, además de que contamos con datos insuficientes para tal fin.
- La segunda vía sería el análisis a nivel macro de las tendencias generales de distintos conjuntos de usuarios o casos de estudios, agrupados por cierta característica. Luego, las recomendaciones a usuarios individuales estarían basadas en las tendencias de comportamiento dichos conjuntos.

Nosotros realizaremos nuestros análisis en base a la segunda vía. Trataremos de extraer conclusiones que nos ayuden a comprender las tendencias de comportamiento, para poder ofrecer mejores recomendaciones y en el momento más oportuno.

2. Procesamiento y análisis de los datos

2.0 Información general sobre los análisis realizados

2.0.1 Datos utilizados

Se analizaron los datos provistos por Navent sobre la búsqueda de empleos en el sitio https://www.zonajobs.com.ar/ . Para dicho análisis se contaba con 6 set de datos en formato csv con la siguiente información:

- **fiuba_1_postulantes_educacion.csv**:
  nivel educativo de los postulantes.
- **fiuba_2_postulantes_genero_y_edad.csv**:
  fecha de nacimiento y género de los postulantes.
- **fiuba_3_vistas.csv**:
  vistas de avisos online y offline, del 23 al 28 de Febrero de 2018.
- **fiuba_4_postulaciones.csv**:
  postulaciones del 15 de Enero al 28 de Febrero de 2018.
- **fiuba_5_avisos_online.csv**:
  avisos online al 8 de Marzo de 2018.
- **fiuba_6_avisos_detalle.csv**:
  detalle de avisos vistos y postulados tanto offline como online.

2.0.2 Lenguaje y librerías utilizadas para el análisis

Para el procesamiento de los datos se utilizó Python 3, junto a la librería Pandas.

Para todo el tema de graficación y visualización de los datos se utilizaron las librerías MatPlotLib y Seaborn.

Para la implementación y aplicación del algoritmo Apriori se utilizó la librería mlxtend (más adelante se explica todo este tema)

2.0.3 Repositorio de GitHub

Se utilizó un repositorio en github para la integración del trabajo. En dicho repositorio se pueden encontrar todos los notebooks con el código utilizado para realizar los análisis y visualizaciones.

Link: https://github.com/alepenaa94/TP1

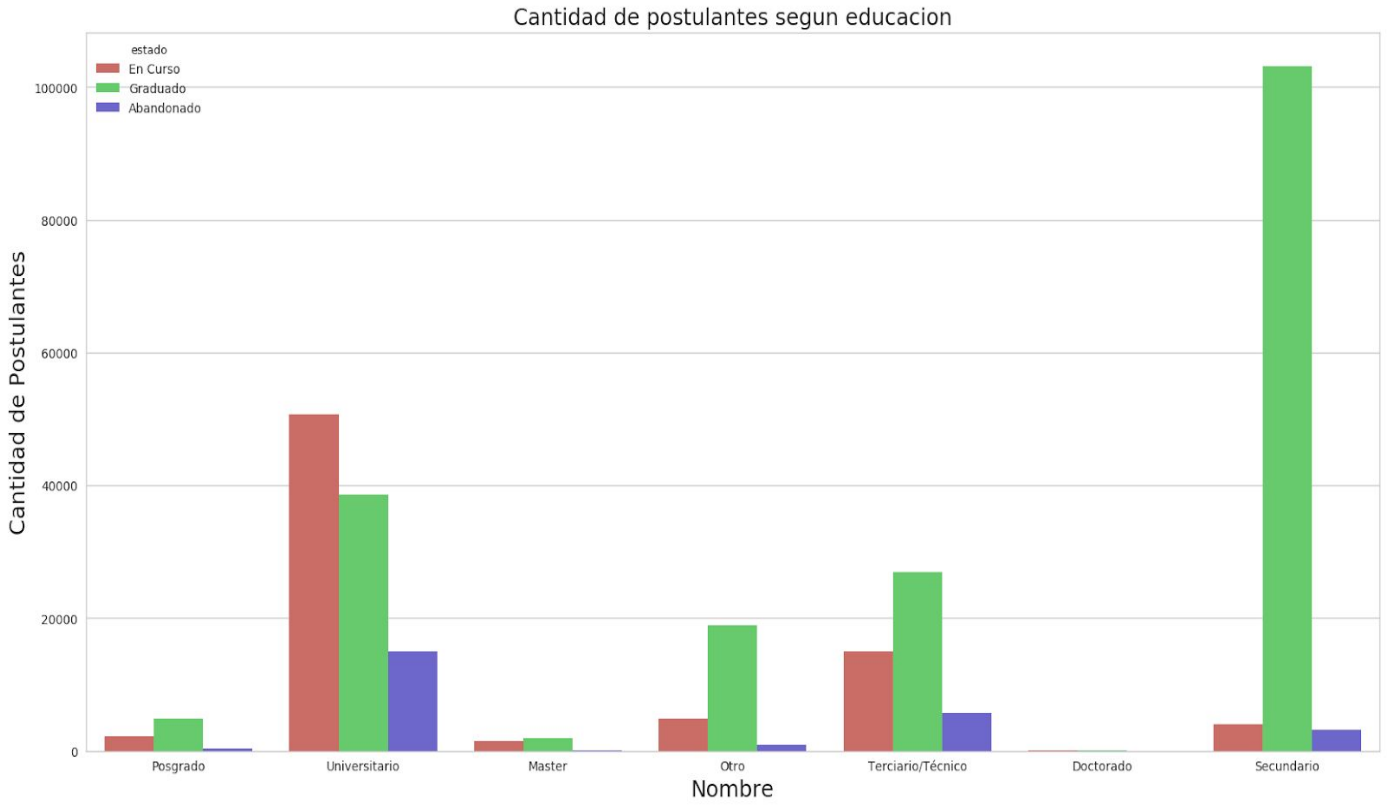
2.1 Postulantes según sus estudios y el estado de los mismos

2.1.1 Introducción

La información sobre la educación de los postulantes que teníamos disponible, constaba básicamente de un listado de estudios (Otros, Secundario, Terciario, Universitario, Posgrado, Máster, y Doctorado), y asociado a cada uno, un estado (completado, en curso, o abandonado). Luego, lo que queríamos era obtener un vistazo inicial general de las postulaciones de todo el set según su educación y el estado de cada una, para entender un poco cómo se distribuían los postulantes respecto a ese tema.

2.1.2 Cantidad de postulaciones según estudios y estado de los postulantes

Los datos que procesamos estaban en “buen estado”, no encontramos nulos ni datos extraños que arrojaran alguna dificultad o error durante el procesamiento y análisis de los mismos.



Como vemos, la mayoría de las postulaciones se encuentran claramente concentradas en dos combinaciones estudio-estado. Por un lado observamos un dominio muy marcado de postulaciones cuyos postulantes son graduados del secundario. Y por otro lado, observamos que hay mucha concentración de postulaciones en estudios universitario-en curso. Esto nos indican que la gran mayoría de los postulantes son estudiantes de universidad, con secundario completo (y que posiblemente estén buscando -o no- su primer trabajo). Más adelante veremos que la mayor cantidad de postulaciones se concentra en los jóvenes, lo que iría de la mano con esta conclusión.

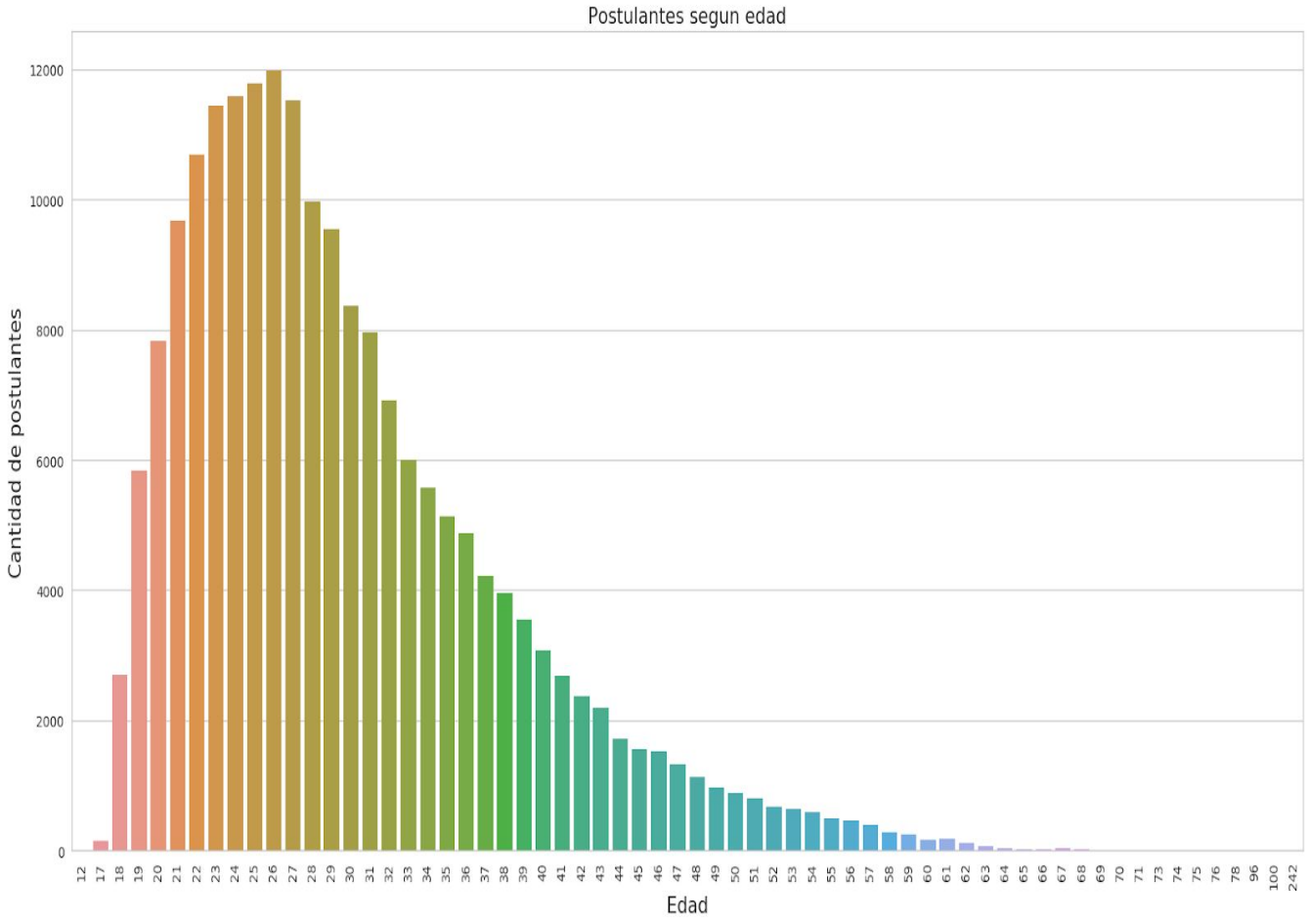
2.2 Edad de los postulantes

2.2.1 Introducción

Un análisis bastante esencial es ver cómo se distribuyen las postulación según la edad de los postulantes. Lo que se podría suponer en un principio, es que habrá dominio de los jóvenes, debido a que los adultos mayores podrían estar ya asentados laboralmente y con un trabajo estable. Veremos si eso ocurre.

2.2.2 Postulantes según la edad

Tras un procesamiento inicial de las fechas de nacimiento, y su posterior transformación a edades, obtenemos:



Con este gráfico se pudo observar una clara mayoría de postulantes en las edades entre 21 y 31 años. También con este gráfico se observa la persona de alrededor de 250 años, sumándole también postulantes con 12, 96 y 100 años los cuales podrían deberse a datos incorrectos y que son necesarios filtrar.

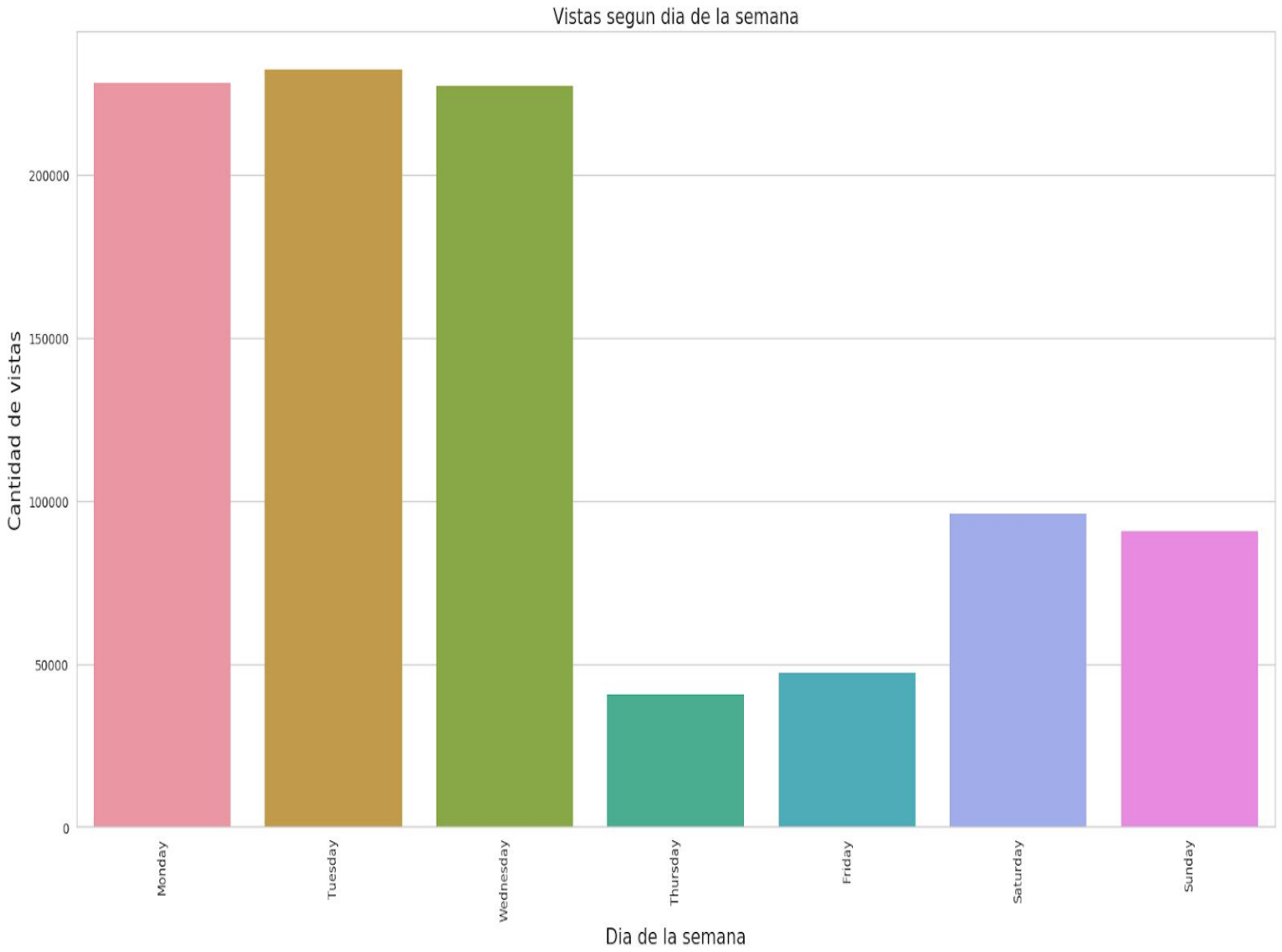
2.3 Vistas por día de la semana

2.3.1 Introducción

Queremos ver la distribución de las vistas a los anuncios según el día de la semana, para ver si podemos obtener alguna información sobre cómo son las tendencias en ese sentido. ¿Da lo mismo el día?

2.3.2 Avisos visitados según el día de la semana

Luego de una transformación a datetime, filtrado de datos extraños o erróneos, y ordenamiento de los resultados, obtuvimos que las vistas se distribuían según el día semanal de la siguiente manera:



Notamos un pico sostenido entre Lunes, Martes y Miércoles, y a primera vista también vemos que, llamativamente, existe un descenso muy fuerte hacia los días Jueves y Viernes. Luego se observa un ascenso para los fines de semana, que igualmente se encuentran muy por debajo de los primeros tres días laborables.

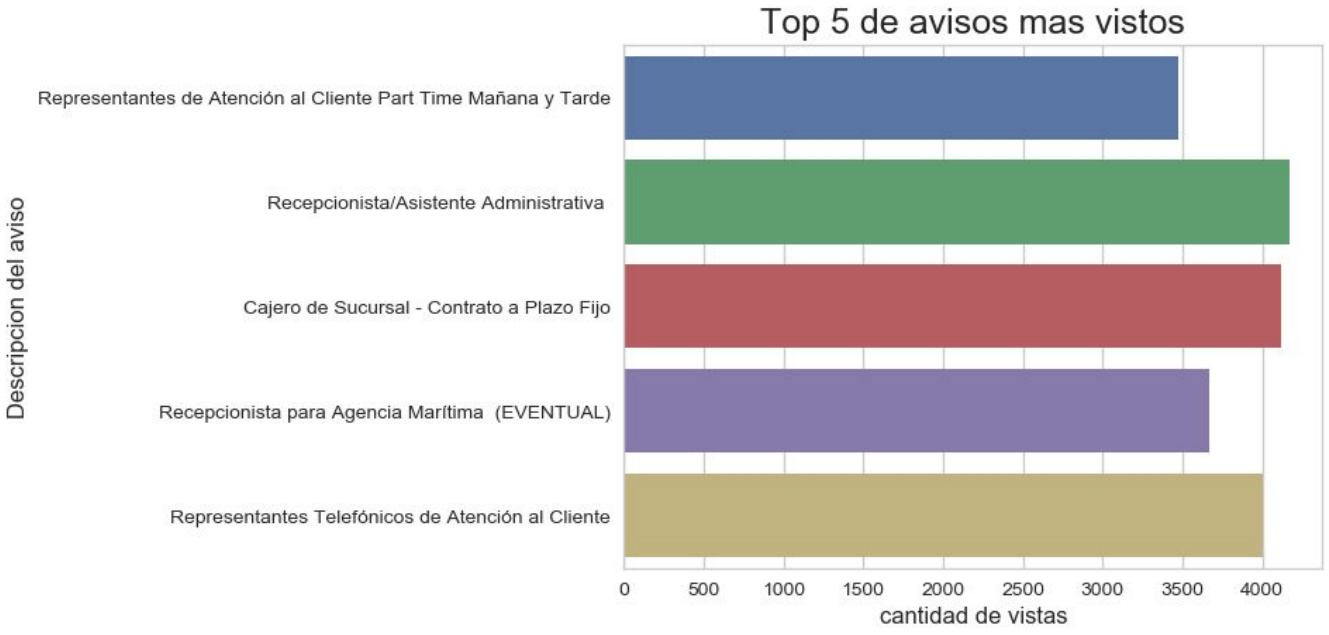
2.4 Avisos más visitados

2.4.1 Introducción

En cuanto a las vistas de los avisos, también nos resultó interesante poder encontrar cuales son los avisos más vistos, y para ello analizamos en un principio los datos que teníamos de las visitas. Luego de obtener el top 5, realizamos la integración con sus detalles para tener la información de su descripción y área.

2.4.2 Top 5 avisos con más vistas

Obtuvimos los siguientes resultados:



Llama la atención que llegaran a esos números de cantidad de vistas sin haber encontrado un empleado antes. Más allá de estos anuncios puntuales obtenidos en el top 5, podemos ver que existen dos avisos de área recepcionista, y dos de atención al cliente, un dato que nos podría estar indicando que dichas categorías son muy usuales (más tarde se comprobará).

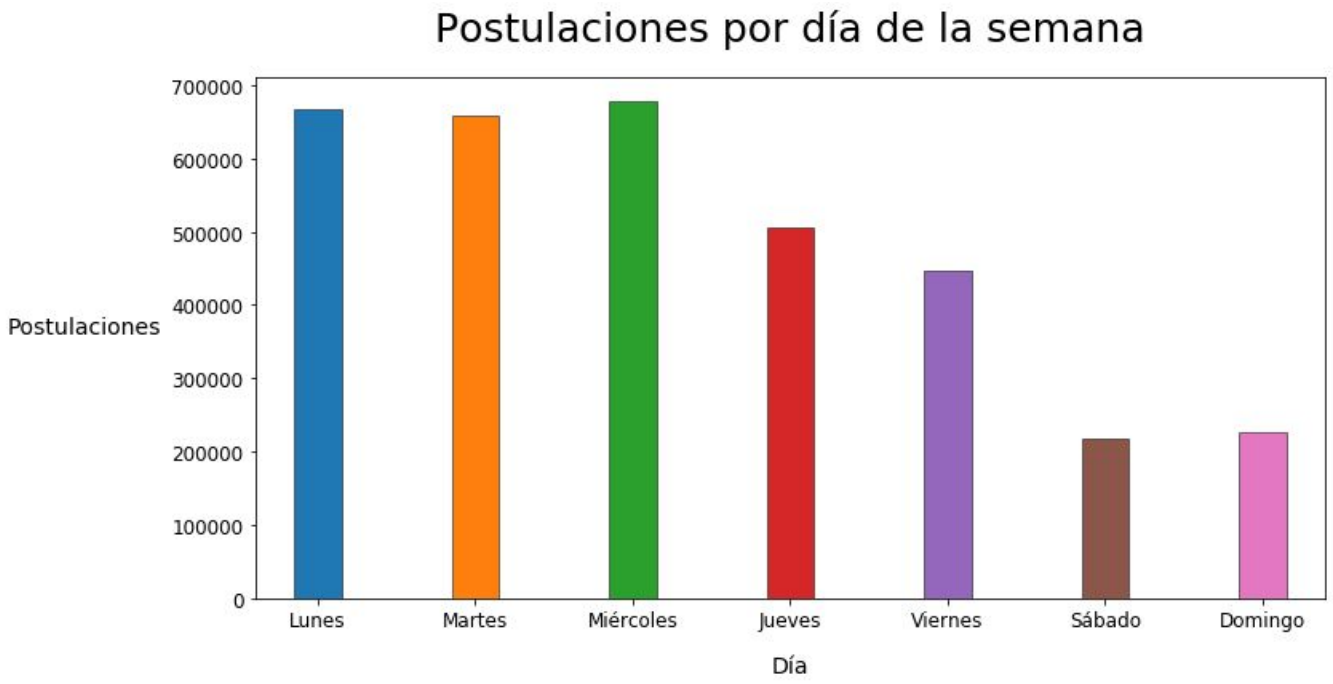
2.5 Postulaciones según el día

2.5.1 Introducción

Uno de los análisis que consideramos más útiles de hacer, es aquél que analiza la distribución de las postulaciones según el tiempo. Nuestro objetivo es identificar períodos de tiempo en los que se daba un alza en las postulaciones (ideales para recomendar avisos a los usuarios) y los períodos en los que había una baja (estudiar por qué). ¿Cualquier día es lo mismo? ¿Incide el hecho de estar en fin de semana o día laborable?

2.5.2 Postulaciones por día de la semana

Para este análisis se tomó la totalidad de los datos (al estar analizando por semana ,no nos interesaba el mes).

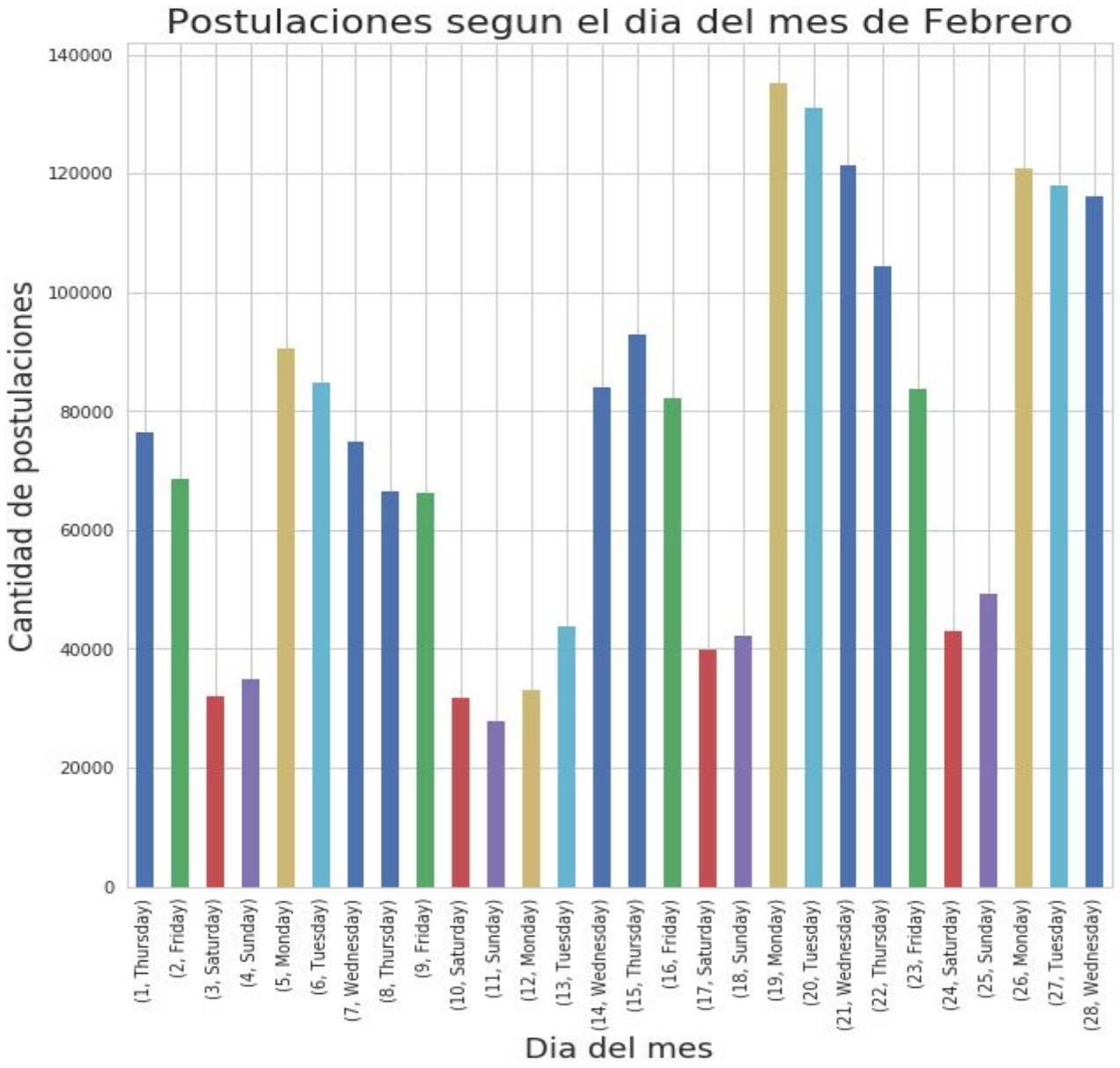


Claramente vemos que dominan los primeros días laborables, Lunes-Martes-Miercoles, muy parejos. Luego observamos un descenso marcado en los Jueves-Viernes, y nuevamente un descenso para el Sábado-Domingo, donde encontramos la menor cantidad de postulaciones. **Entonces, sería una buena práctica ubicar la mayor parte de nuestras comunicaciones de recomendación de avisos los primeros días de la semana.**

Vemos una diferencia clara con respecto a los avisos vistos por día de semana (que analizamos anteriormente). En dicho caso existía un descenso muy brusco y marcado los días Jueves y Viernes, que en el caso de las postulaciones no ocurrió.

2.5.3 Postulaciones por día del mes

Para este caso, tuvimos que acotar los datos procesados a aquellos que eran únicamente de un mismo mes. Como contábamos con mitad de Enero y la totalidad de Febrero, lógicamente utilizamos solamente los datos de Febrero. Luego, las postulaciones según el día para dicho mes nos quedaron distribuidas de la siguiente manera:



Como se puede observar en el gráfico anterior, la mayor concentración de postulaciones en los días de semana se produce generalmente los primeros tres días laborables, en concordancia con lo visto en el análisis anterior. En fines de semana se produce la menor concentración de postulaciones a lo largo de todo el mes. Como una conclusión de este análisis, podemos ver que hacia fin de mes se ve un incremento en la cantidad de postulaciones en todos los días de la semana.

2.6 Postulaciones y avisos visitados según la hora

2.6.1 Introducción

Una de las preguntas que nos surgieron se relacionaba con la hora del día en que los usuarios usan la plataforma: ¿Había alguna tendencia en las postulaciones y avisos visitados con respecto a la hora en las que se hacían?

Para tal fin se trabajó con los set de datos de postulaciones y de avisos visitados. Lo primero que se tuvo que hacer fue adaptar las fechas y horas en cada set a formato datetime, luego de un filtrado de datos nulos. Esta tarea fue sencilla en el caso de las postulaciones (ya se encontraban en un formato amigable para el tipo datetime), pero resultó de mayor dificultad en el caso de avisos vistos, ya que el formato pre-existente de las fechas era bastante inusual. Por esto mismo, hubo que realizar una transformación previa a un formato que el método .astype entendiese.

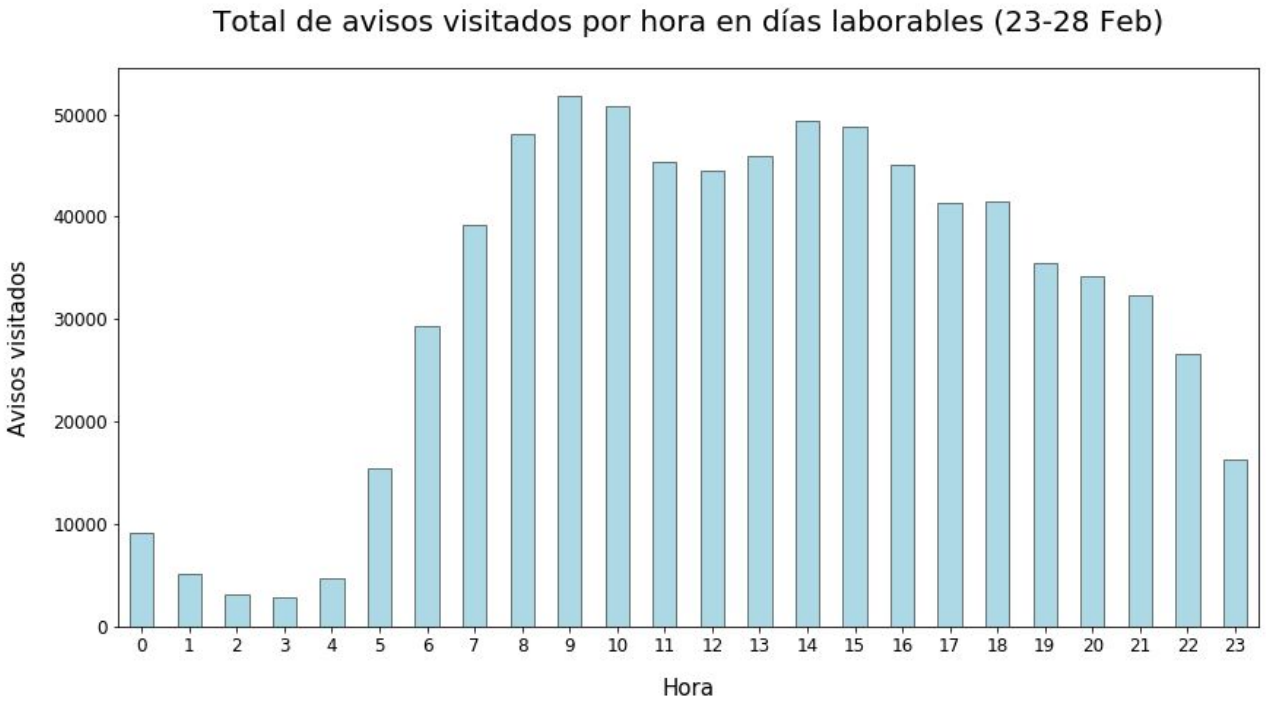
Con respecto a las postulaciones, se analizará todo el set de datos, que comprende datos desde el 15 de Ene hasta el 28 de Feb. No haremos distinción de meses, ya que los procesamos por separado y confirmamos que no hace falta; las tendencias se mantienen. Preferimos abarcar la mayor cantidad de datos posibles. Con respecto a las vistas, el rango de datos era más acotado: del 23 al 28 de Feb.

2.6.2 Avisos visitados

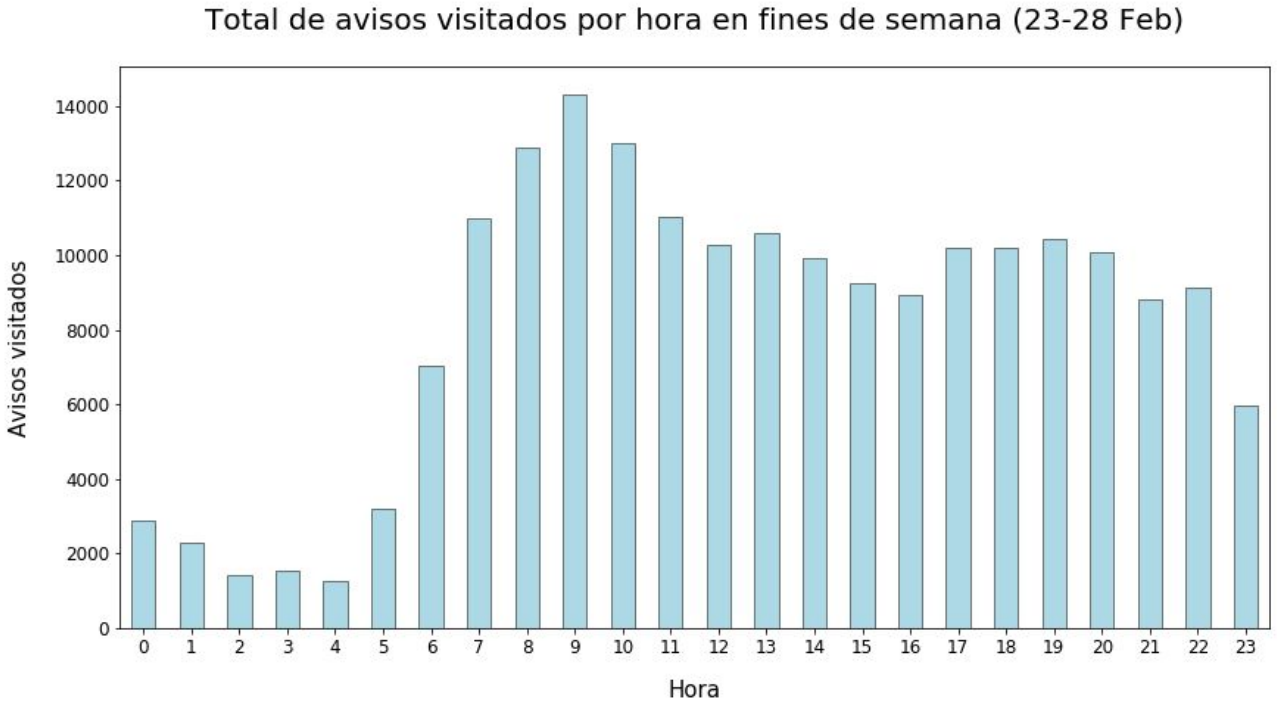
En un principio se realizó un análisis para todos los días de la semana, y obtuvimos algunas conclusiones interesantes:

Se puede ver que el pico máximo de avisos vistos se da a las 9 am, seguido de cerca por las 8 y 10 am. Luego, surgió la inquietud acerca de la diferencia entre día laborable y día de semana. ¿Existía alguna diferencia?

Luego, se procedió a realizar un filtrado extra, seguido de un análisis de cada caso:



Como podemos ver, para los días laborables la tendencia es muy parecida al caso anterior, tanto en forma, como en valores. Algo distinto sucedió para los fines de semana:

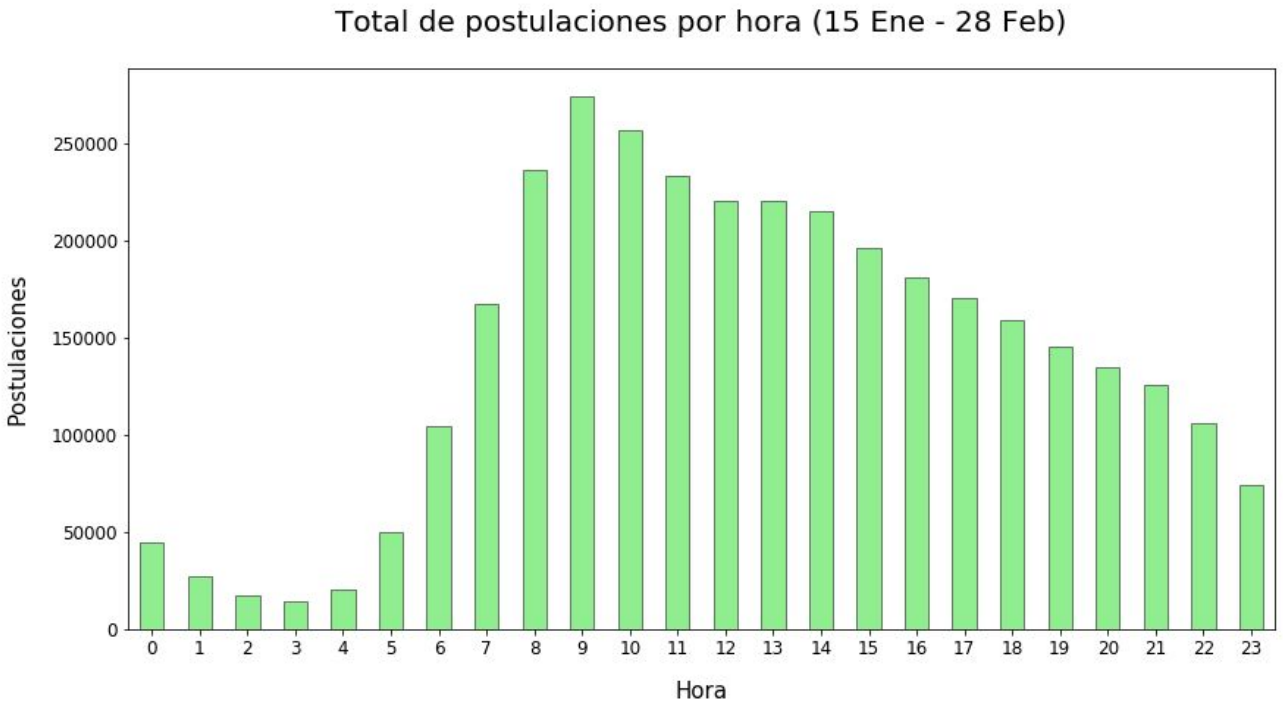


En este caso se mantienen los picos entre las 08 y las 10 hs, pero se observa un leve cambio en la tendencia entre las 12 y las 16 hs, existiendo un descenso. Además, podemos notar que la cantidad de avisos visitados fue significativamente menor a aquella de los días laborables.

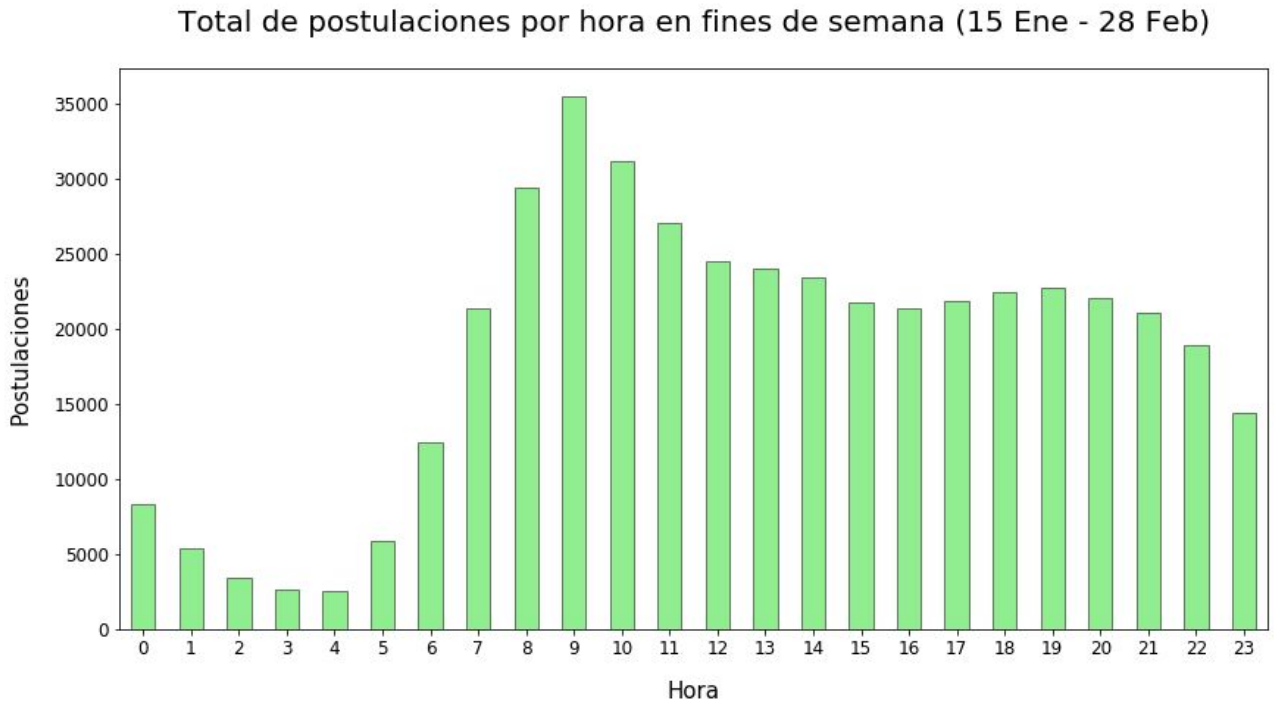
Tiene sentido entonces que el 1er caso analizado (todos los dias) sea prácticamente igual al caso de días laborables, ya que al tener los fines de semana cantidades bastante inferiores, estos no tienen tanta incidencia sobre el global como para transformar demasiado su tendencia.

2.6.3 Postulaciones

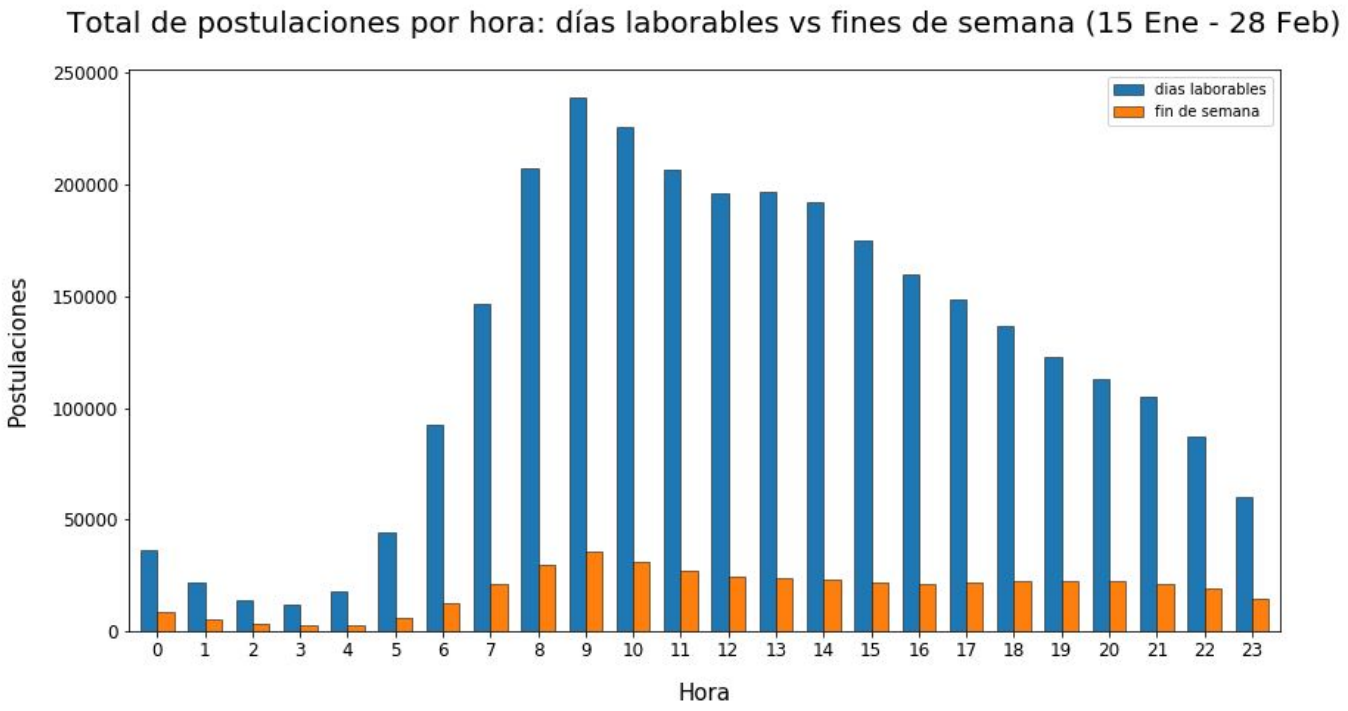
De la misma forma que se hizo con los avisos vistos, se analizaron las postulaciones (pero para otro período, como se mencionó anteriormente):



Al igual que con los avisos visitados, vemos que los picos máximos de postulaciones se dan de 09 a 10 hs. Pero si analizamos el descenso desde esos picos hacia las horas nocturnas, vemos que es un poco diferente al de las vistas, casi diríamos que es un descenso lineal.

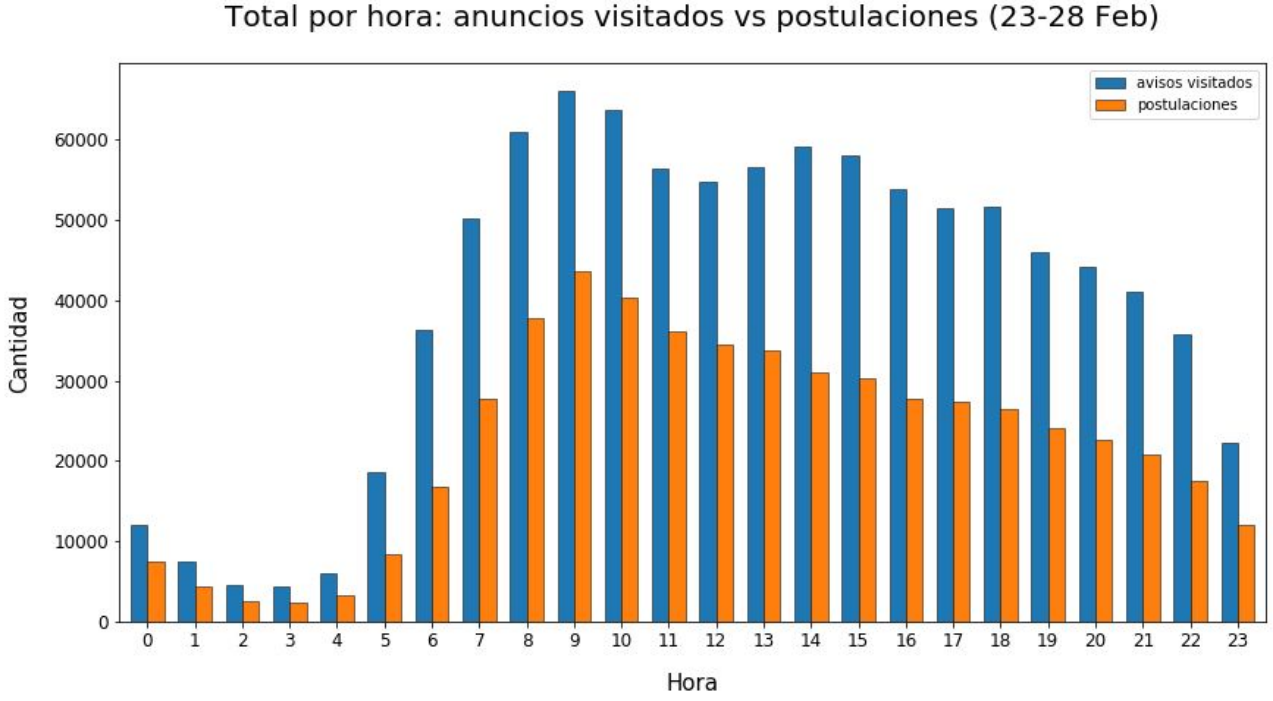


Nuevamente, observamos un leve cambio en la tendencia al ser fin de semana. Se nota un aumento entre las 18 y las 23 hs con respecto a los días laborables. Aún así, las cantidades de postulaciones son mucho menores en los fines de semana. Por ello, la tendencia global es dominada por los días de semana, deducible del siguiente gráfico:



2.6.4 Avisos visitados y postulaciones

De manera conjunta, analizamos el comportamiento de las visitas y las postulaciones, tomando en este caso el mismo intervalo, del 23 al 28 de Feb. Podemos ver el pequeño cambio de tendencias que notamos anteriormente:



Razonablemente, tenemos una mayor cantidad de avisos visitados con respecto a postulaciones. Si bien no tienen exactamente la misma forma de descenso, es muy parecida, y se puede afirmar en ambos casos que existe un pico entre las 08 y 10 hs, y que a partir de allí comienza un descenso sostenido hasta las 04 hs.

Esta distribución nos podría resultar de mucha utilidad para organizar nuestras comunicaciones de recomendación de avisos y concentrarnos en los picos observados (obviamente sin descuidar las otras franjas horarias).

2.7 Postulaciones según la hora del día y edad del postulante

2.7.1 Introducción

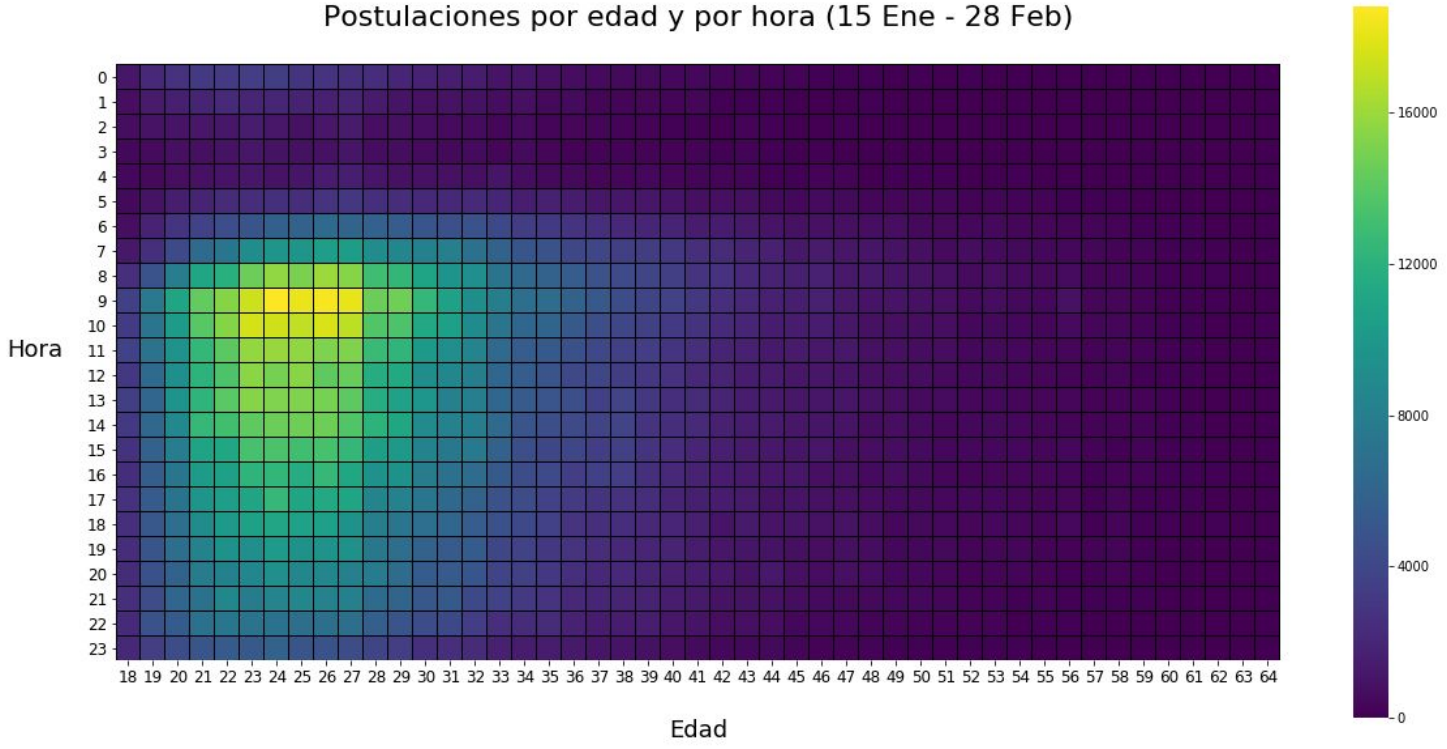
De la mano del análisis del punto anterior, profundizamos más el tema de horarios, basándonos en nuestra experiencia personal: ¿Tiene que ver la edad del postulante? ¿Un joven de 20 años se postula a la misma hora que una persona de 40 años? ¿Los jóvenes tienen tendencias a postularse en horarios nocturnos? Todas estas consultas pensamos que se podían visualizar a la vez mediante un heatmap.

Al igual que en las postulaciones por hora, se analiza el total de datos disponibles (del 15 de Enero al 28 de Febrero), ya que la tendencia se mantiene aún si se hace distinción entre meses (fue chequeado), y además porque preferimos tener la mayor cantidad de datos para poder realizar un análisis general lo más verídico posible (queremos ver la tendencia general, no si hubo mínimas variaciones por meses).

Previo a realizar algún análisis, tuvimos que filtrar fechas de nacimiento inválidas que existían en el set de datos. Luego, tras calcular las edades de cada postulante (edad al momento de referencia - 28 de Feb), realizamos un nuevo filtrado para trabajar únicamente con las personas que consideramos que tienen una “edad laborable estándar”: de los 18 a 64 años. Tras esta preparación previa, procedimos al análisis.

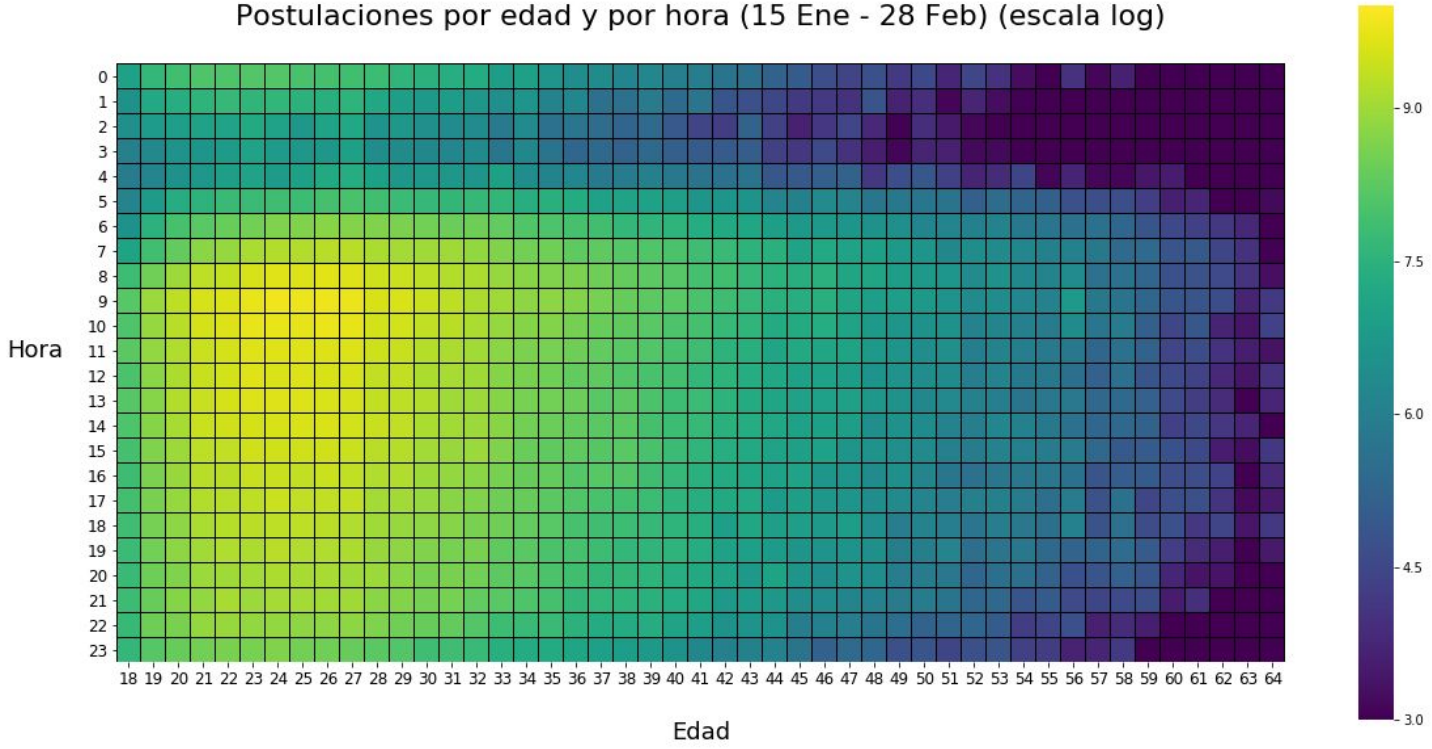
2.7.2 Postulaciones por edad y hora

Tras un acomodamiento en la estructura de nuestros datos integrados (postulante, edad, y hora de la postulación), pudimos obtener de manera gráfica la relación buscada:



En concordancia con lo analizado en el punto anterior, vemos que la mayor intensidad se da en el sector izquierdo medio del heatmap. Esto es, durante las horas de la mañana, específicamente entre las 08 y 10 hs. Pero hay algo nuevo que podemos observar: la mayor cantidad se da entre los 20 y 30 años aproximadamente, y vemos que este pico por la mañana que observamos, si bien sigue existiendo, ya no es tan marcado para los postulantes mayores a 30 años.

Pero esta visualización tiene un problema: al haber tanta diferencia en las cantidades de postulaciones entre los jóvenes y los más adultos, no se logra apreciar mejor las diferencias relativas entre cada edad. Por ello procedimos a utilizar una escala logarítmica, lo que nos permitió tener un mejor y más detallado panorama para las zonas más oscuras del heatmap:



Ahora sí, podemos apreciar mejor las diferencias relativas a lo largo y ancho de todo el heatmap. De esta manera, podemos concluir que en la franja de los 20 a 30 años, hay picos marcados en la mañana, pero también hay una gran intensidad de postulaciones durante la tarde y noche. Diferente es el caso con los mayores a 35 años, para los cuales no hay un pico tan marcado, y la intensidad a lo largo del día parece ser bastante más sostenida y pareja.

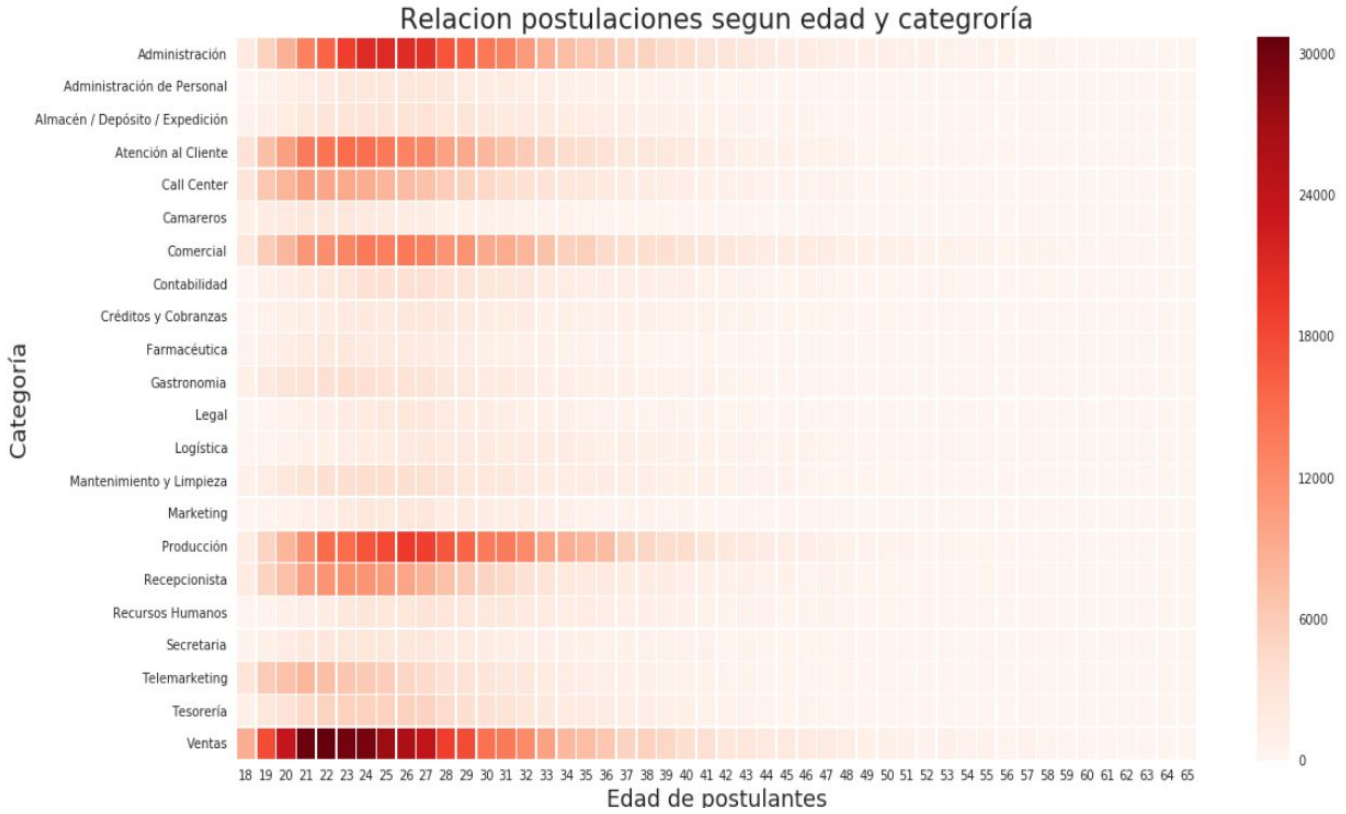
2.8 Postulaciones según edad y categoría

2.8.1 Introducción

De lo analizado previamente, se quiso profundizar en si existe algún tipo de relación entre la categoría y las edades: ¿Que categoría es más “popular” según la edad? ¿Cual es la tendencia del área de trabajo según la edad?

2.8.2 Postulaciones según edad del postulante y categoría del anuncio

Para contestar estas preguntas, se analizaron los datos de las postulaciones, filtrando por edad y vinculando este set de datos con los datos de los avisos, para poder así saber a qué categoría pertenece. Claramente como son muchas categorías el siguiente gráfico fue “recortado” a las categorías que tengan más de 1000 postulaciones.



Como se puede observar en el heatmap, la búsqueda laboral es muy frecuente para edades entre los 18 y 32 años en la que se suele finalizar el secundario, la universidad o se busca una estabilidad laboral. En este rango de edades podemos observar qué categorías son muy demandadas según la edad, pero vemos que en general se da que son: Ventas, Producción, Comercial, Atención al Cliente y Administración.

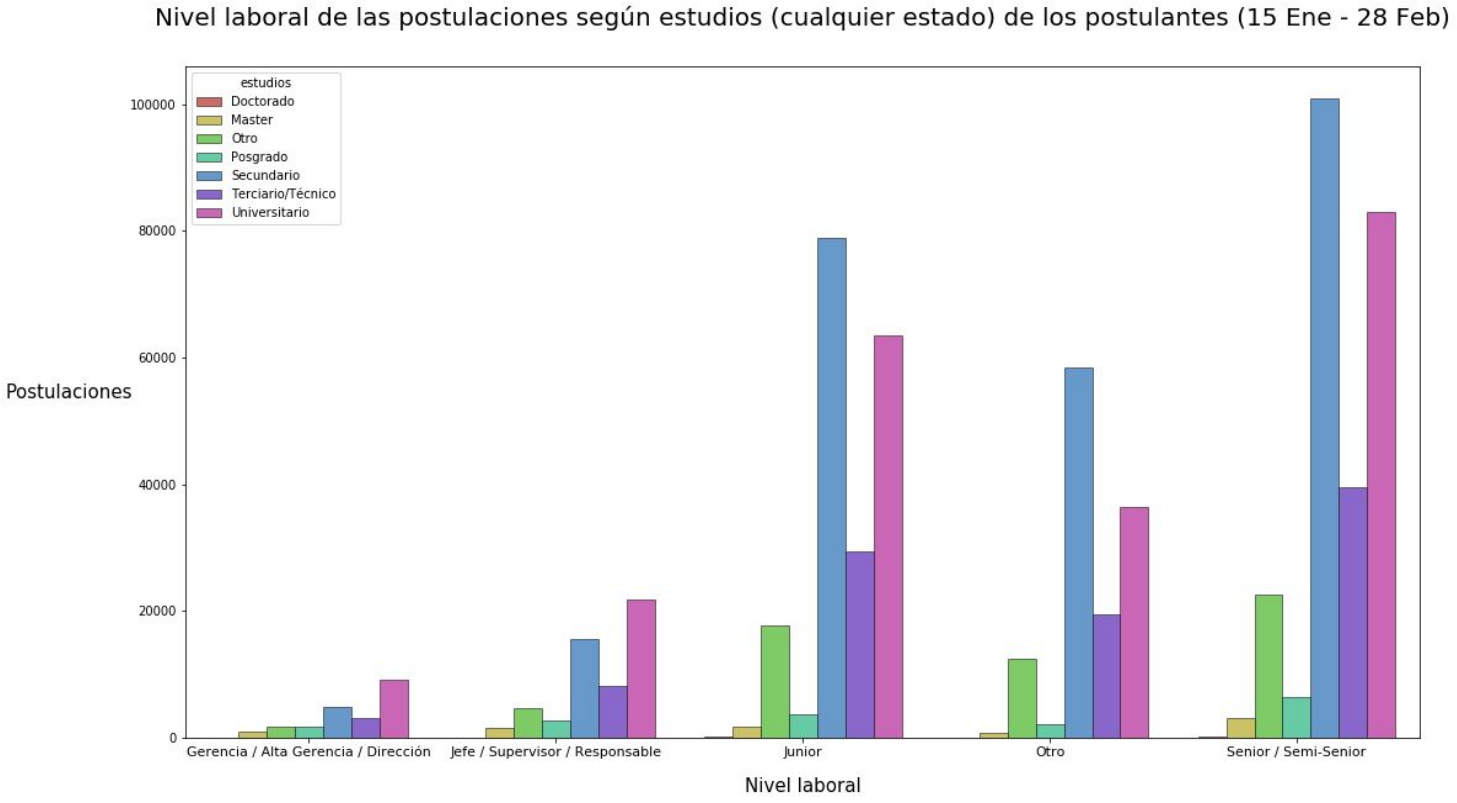
2.9 Nivel laboral de las postulaciones según la educación de los postulantes

2.9.1 Introducción

Una pregunta que surgió fue si había relación entre el nivel educativo de los postulantes y los niveles laborales a los que postulaban. Esta inquietud surgió básicamente de pensar que los niveles más altos eran ocupados por gente más preparada. Si bien este análisis no va a ser exactamente directo en ese sentido (uno no puede inferir que educación tienen las personas que pertenecen a altos cargos de empresas solamente por el análisis postulaciones, es decir, de intenciones de pertenecer a dichos cargos), veremos si podemos observar este comportamiento esperado, donde a mayor nivel laboral de las postulaciones suponemos habrá mayor cantidad de postulantes con educaciones de alto nivel.

2.9.2 Análisis sobre los datos brutos (raw data)

En principio analizamos los datos en bruto. Esto quiere decir, que dentro de un mismo nivel de estudios incluimos tanto estudios en curso como completados (no fueron incluidos estudios abandonados, filtrados en la carga inicial del set de datos), para tener un vistazo general inicial de lo que estábamos buscando. Luego de estructurar nuestros datos de la manera adecuada, obtuvimos que:



Podemos apreciar a simple vista una bajísima cantidad de postulantes con estudios de Posgrado/Máster/Doctorado con respecto a los estudios dominantes Secundario/Universitario/Terciario. Esta característica la podemos ver replicada en todos los niveles laborales, lo que a priori haría descartar

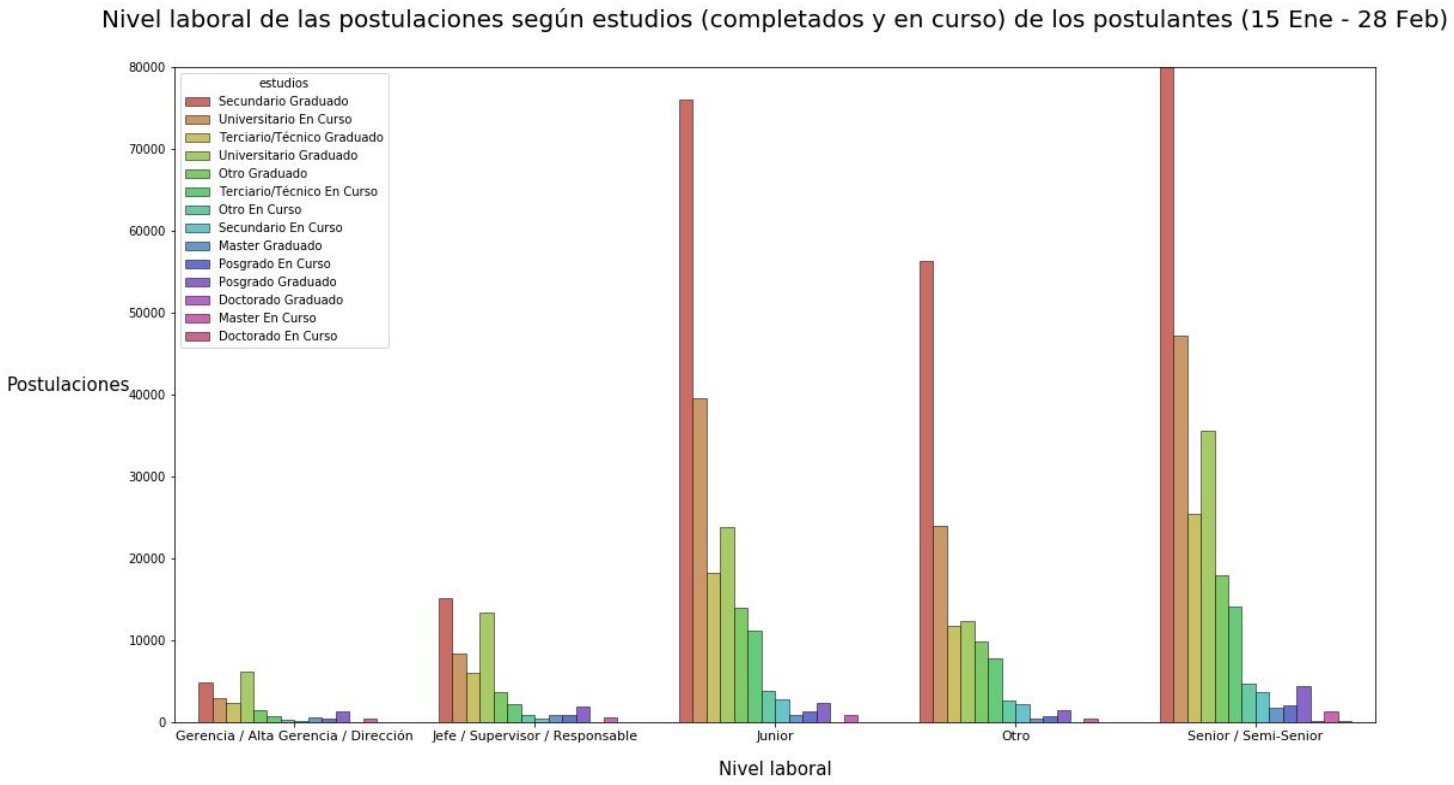
nuestra hipótesis inicial sobre el predominio de altos niveles de estudios para los niveles laborales superiores. Pero, si miramos bien, podemos distinguir ciertas cosas:

- A diferencia de los niveles laborales Junior/Otro/Sr-Ssr, en los cuales dominan los estudios secundarios seguidos por los universitarios, en los niveles superiores Jefaturas/Gerencias observamos un cambio de orden en ese sentido. Primero se ubican los estudios universitarios, seguidos por los secundarios. Es decir, que aumentó el nivel educativo en estos dos niveles respecto a los primeros tres, lo que podría ir de la mano con nuestra hipótesis inicial.

- Mas aún, podemos ver que, proporcionalmente, la diferencia entre Secundario/Universitario vs Posgrado/Máster/Doctorado se achicó significativamente si la comparamos con la de los niveles laborales más bajos, lo que supone mayor proporción de postulantes con estudios avanzados en los dos niveles superiores.

2.9.3 Análisis con distinción del estado de los estudios

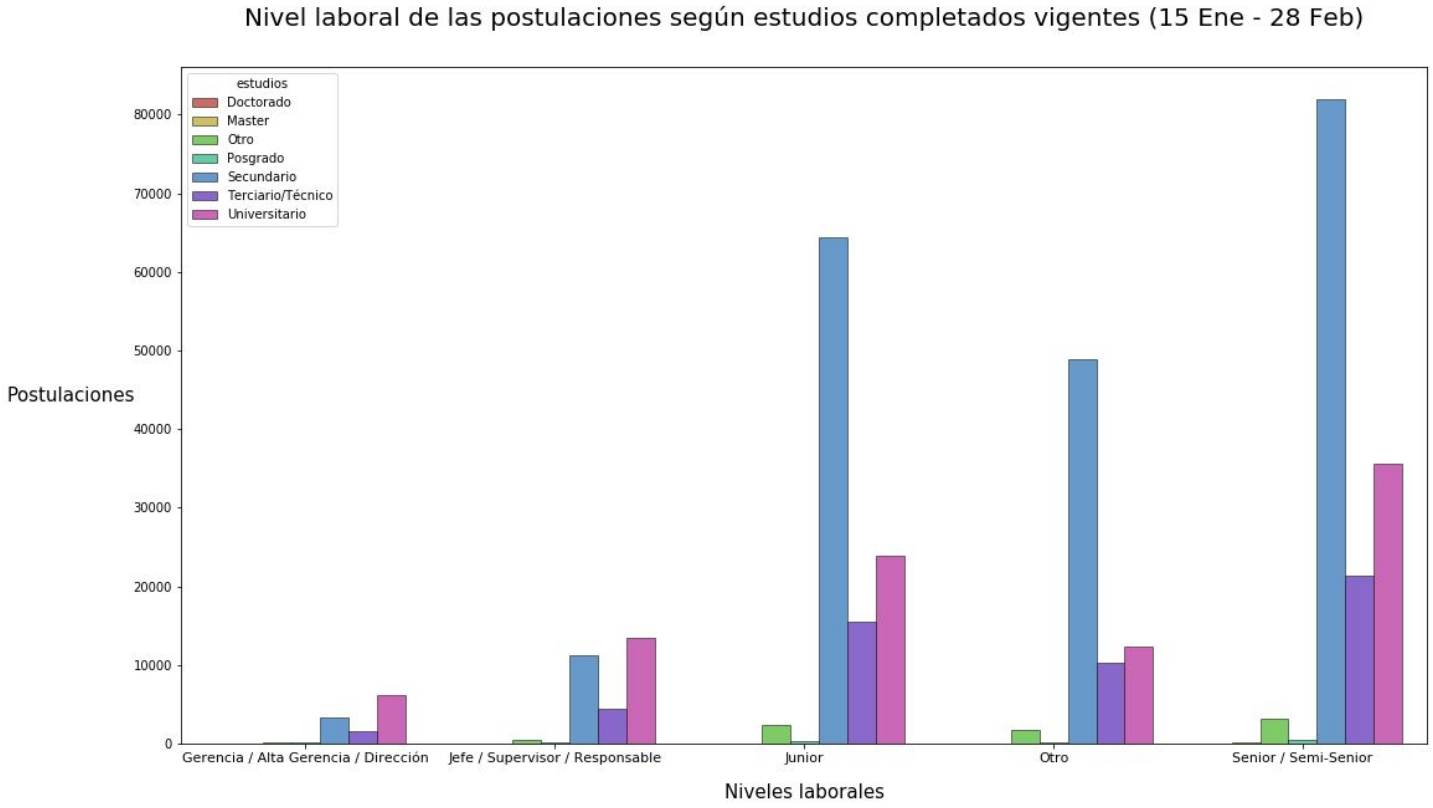
Ahora bien, como mencionamos antes, estábamos trabajando sobre los datos en bruto. Si bien ya pudimos sacar conclusiones, refinamos nuestros datos y realizaremos la distinción de cada nivel de estudio según su estado, es decir, completado o en curso.



Lo que podemos observar es una gran cantidad de estudiantes universitarios en curso, que en el anterior análisis estábamos englobando dentro de Universitarios y nos ensuciaba en cierta manera los datos, ya que nos hacía pensar que la diferencia entre Secundario y Universitario no era tan grande, cosa que evidentemente no es así, ya que ahora podemos ver a Secundario Graduado dominando muy por encima de Universitario en los primeros tres niveles laborales.

2.9.4 Nivel laboral según estudios completados

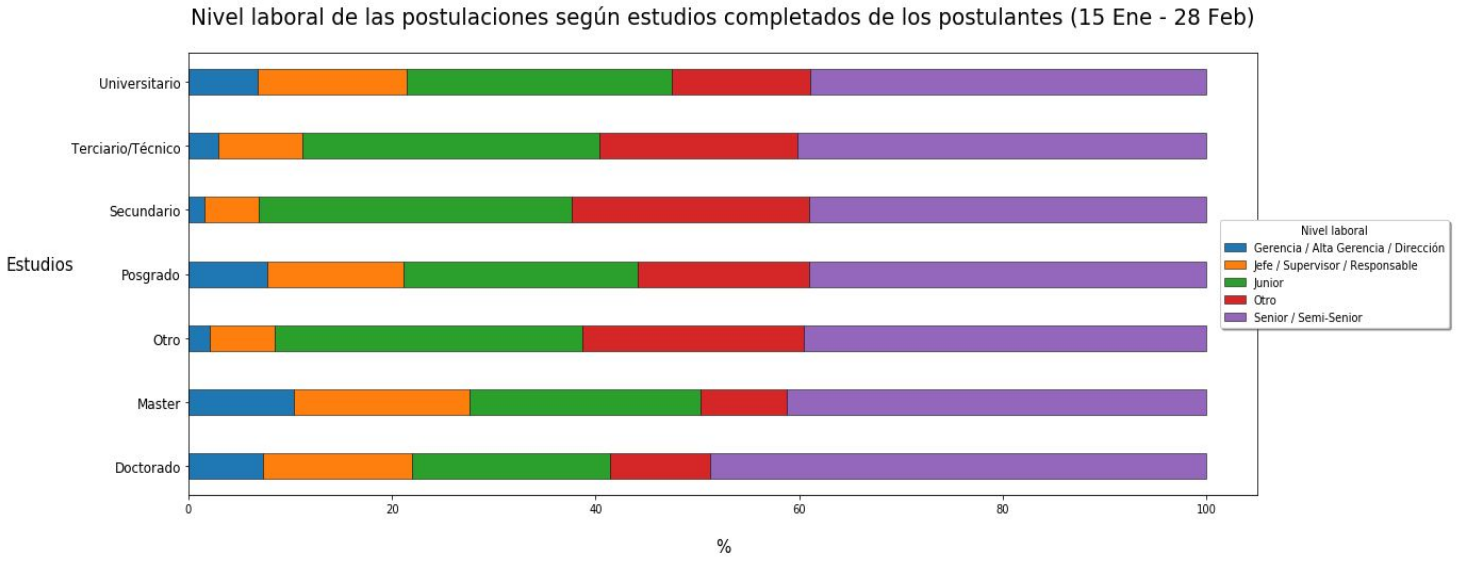
Para hilar aún más fino, en este punto vamos a seguir refinando los datos y basaremos el análisis en lo que a nuestro criterio consideraremos como estudios “válidos”. Esto es, estudios completados a la fecha, descartando así los estudios en curso:



Ahora si, podemos ver que se amplió considerablemente la diferencia entre Secundario y Universitario para los tres niveles laborales más bajos. Además, vemos que ya prácticamente no figuran los niveles de estudio más avanzados -Posgrado/Master/Doctorado- lo que nos hace deducir que la gran mayoría de los que figuraron en el 1er análisis eran estudios avanzados en curso.

2.9.5 Composición según nivel laboral de la educación de los postulantes

En esta ocasión daremos “vuelta” el enfoque para los mismos datos (seguimos con estudios únicamente completados), ya que miraremos la proporción de distintos niveles laborables con la que se componen las postulaciones de los usuarios con un nivel de estudio determinado:



En este caso resulta extraño el alto porcentaje de aplicaciones a niveles Junior proviniendo de estudios avanzados -Posgrado/Master/Doctorado-, aunque un poco más esperable era el aumento porcentual significativo de postulaciones a Gerencia en estudios avanzados con respecto a los estudios -Secundario/Otro/Terciario-.

2.10 Nivel de “cercanía” entre las distintas categorías. Algoritmo apriori.

2.10.1 Introducción

Tratando de buscar patrones generales de comportamiento de los usuarios al momento de postularse, surgió automáticamente la necesidad de buscar categorías muy dependientes una con respecto a otra. ¿A qué nos referimos con dependientes? Básicamente saber, a nivel macro, que categorías comparten mayor cantidad de postulantes. O dicho de otra forma, si A postuló a Ventas, y A postuló a Comercial, entonces Ventas y Comercial estarían en cierta forma relacionadas. Y si esto lo aplicamos a grandes volúmenes de datos, **podríamos encontrar duplas de categorías muy cercanas y basar nuestras recomendaciones en estas cercanías.**

Un caso muy común de esto ocurre en los supermercados, donde se estudian las n-uplas de productos más frecuentes comprados en una misma orden/transacción. De esta forma, se sabe que por ejemplo, es muy común que se compren papas fritas y gaseosa en una misma orden. Con esta información, se pueden armar estrategias de marketing, ubicar los productos físicamente cerca, etc.

Esta metodología se engloba dentro de lo que se conoce como Aprendizaje de Reglas de Asociación, útil para descubrir relaciones interesantes entre variables en grandes volúmenes de datos.

En nuestro caso, esta información la vamos a obtener mediante el uso del algoritmo apriori (https://en.wikipedia.org/wiki/Apriori_algorithm), un algoritmo diseñado para detectar y obtener estas n-uplas más frecuentes que mencionamos anteriormente. Para poder aplicar dicho algoritmo, hacemos uso de la librería “mlxtend” (https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/) que contiene una implementación para Python del mismo.

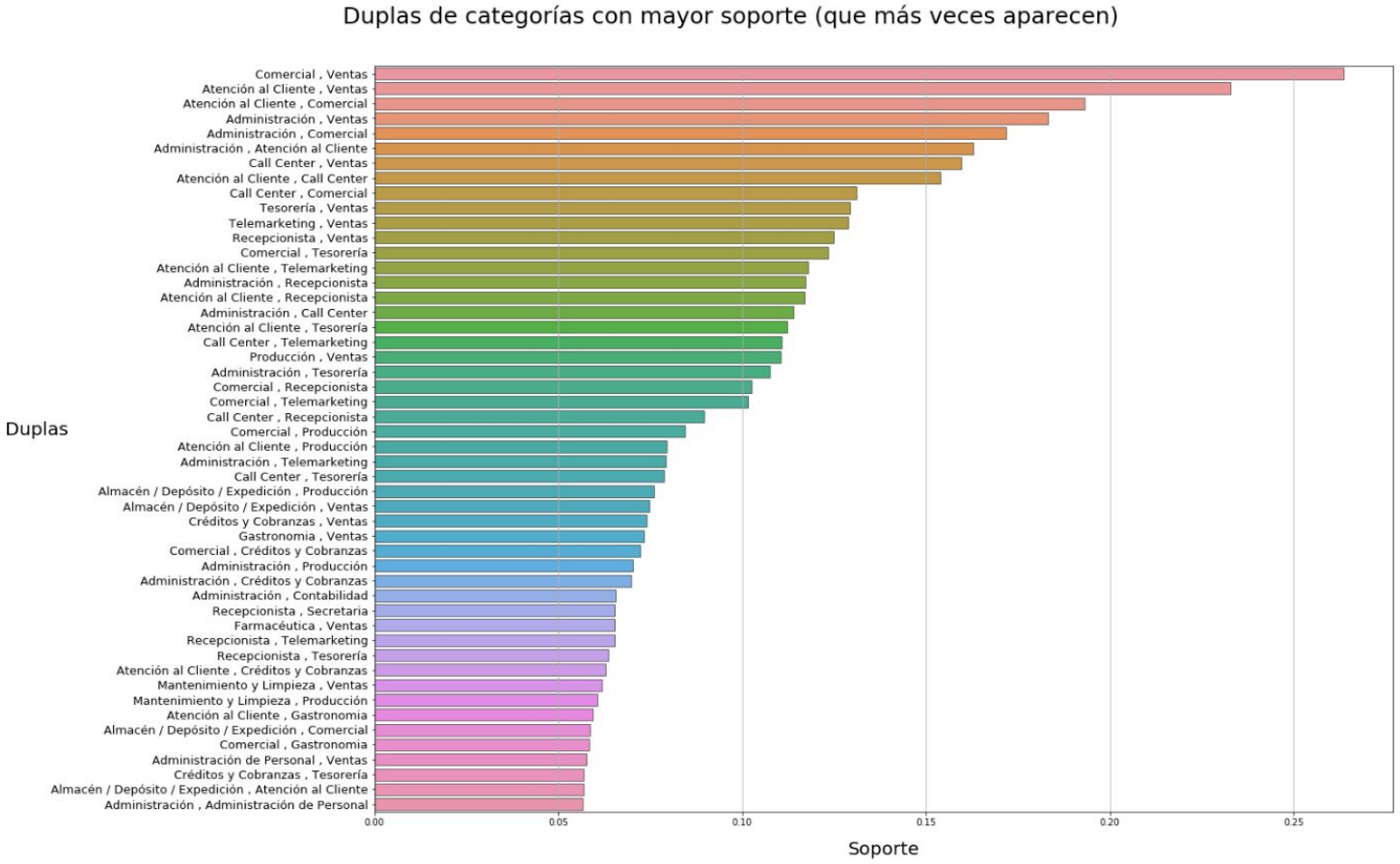
Previo a aplicar el algoritmo, debimos transformar nuestra estructura de datos a la requerida por la librería: a una matriz binaria de postulantes x categorías, donde se indica con 1 o 0 para cada postulante si se postuló o no a cada categoría.

2.10.2 Duplas de categoría con mayor soporte

Antes de avanzar con el análisis, tenemos que mencionar el concepto de soporte y de qué forma aplicamos el algoritmo. El soporte es una medida indicadora de que tan frecuentemente esa n-upla aparece en el set de datos (soporte * 100 = porcentaje de aparición en el set de datos). Nos va a interesar porque es un filtro que le pasaremos al algoritmo por parámetro: min_support, con el cual le estaremos diciendo que devuelva n-uplas que tengan como mínimo ese soporte.

Otro filtro importante que aplicaremos, elegido para no hacer muy extensivo el análisis, es la longitud de las n-uplas repetidas a buscar. Usaremos longitud 2, es decir, que solo busque duplas de categorías que se repitan para un mismo postulante.

Con estas consideraciones, aplicamos el algoritmo, y obtenemos los soportes de todas las duplas existentes (también de las categorías individuales, cuentan como 1-upla). Luego de realizar ciertos filtrados, y un ordenamiento, obtuvimos las duplas más habituales, es decir, las de mayor soporte:



A primera vista vemos que la dupla con mayor soporte es la de Comercial y Ventas, siendo un poco mayor a 0,25. Esto nos dice que aproximadamente más del 25% de los postulantes analizados aplicaron a anuncios de Comercial y también de Ventas.

Podemos ver que el área de Ventas está presente en muchas duplas, lo que nos dice que está muy relacionada con gran cantidad de otras áreas, y potencialmente puede ser un área muy buena para recomendar a los usuarios (ampliaremos mejor con el estudio de la confianza, en el punto de abajo).

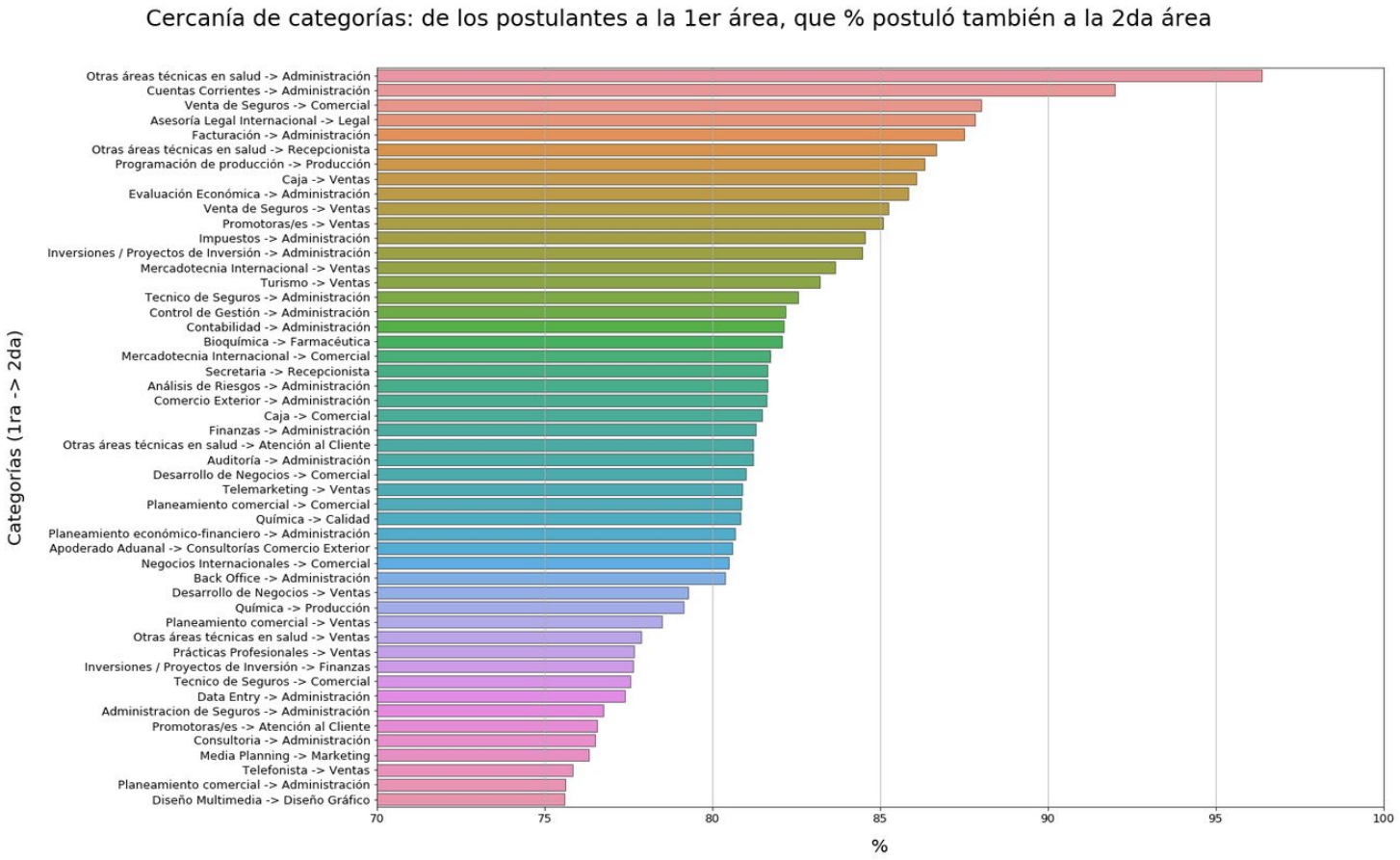
2.10.3 Duplas de categoría con mayor confianza

Antes que nada debemos definir el concepto de confianza: es un indicador de qué tan seguido la regla de asociación en cuestión se cumple. En nuestro contexto, sería la proporción de aparición de una dupla respecto a una categoría individual. Mucho más fácil es entenderlo con un ejemplo:

- Si calculamos el soporte de la dupla Comercial , Ventas sobre el soporte individual de Ventas (0,44) obtenemos la Confianza de esa dupla respecto a Ventas: 0,26 / 0,44 = 0,59
  Es decir, que de todas las personas que se postularon para al menos un anuncio de Ventas, el 59% también se postuló a uno o más anuncios de Comercial.

- También se podría calcular a la inversa, la confianza respecto a la categoría Comercial (soporte individual 0,36): 0,26 / 0,36 = 0,72
  Luego, el 72% de las personas que postularon a Comercial, también postularon a Ventas. **Estos datos resultan muy útiles para ver que otras categorias de anuncios se podrian recomendar a un usuario con antecedentes de postulaciones a cierta otra categoría.**

Tenemos que mencionar también que fue necesario un filtrado de resultados. Al haber duplas de soporte extremadamente bajo (muy baja ocurrencia en el set de datos), nos quedaba una confianza muy alta, y no era representativo de la realidad. Por ejemplo: una sola persona se postuló a Traducción, y esa misma persona también aplicó a otro trabajo del área de Caja (el único caso de postulante aplicando a ambas áreas). Esto nos daría una confianza de 1, es decir, que el 100% de las personas que se postulan a Traducción se postulan también a Caja, y obviamente este tipo de conclusiones no nos sirven. Por ello filtramos las duplas y nos quedamos solo con aquellas de soporte mayor a 0.001 (en nuestro set de datos de 200.000 postulantes aprox, serían unas 200 ocurrencias, número que a nuestro criterio consideramos apropiado para el análisis que estamos efectuando).



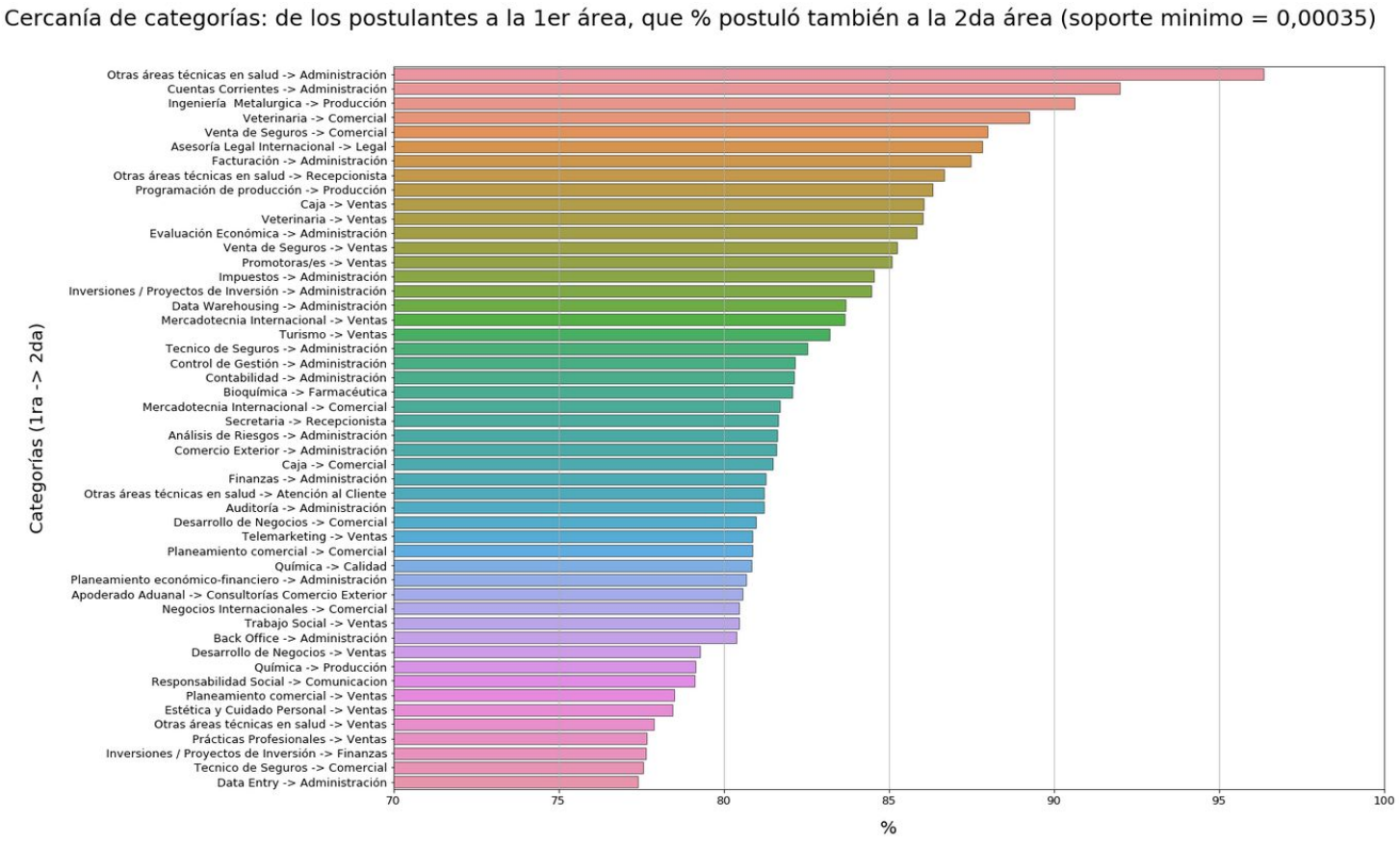
Nota: si bien es una práctica recomendable iniciar los gráficos de barras en 0, en este caso para visualizar de mejor forma la diferencia entre cada dupla se configuró el inicio en 70%. De lo contrario, obtendremos un gráfico de ocupación masiva con muy poca variación entre cada barra de las duplas.

Lo primero que podemos ver en este gráfico nos llama la atención. De todas las personas que postularon a trabajos de área “Otras áreas técnicas de salud”, aproximadamente el 96% aplicó también a trabajos de Administración. Es un dato bastante curioso, ya que no es algo que uno pueda suponer naturalmente, como

quizás pase con el dato de que el 88% de los postulantes a “Asesoría Legal Internacional” postularon también a “Legal”.

Este listado podría ser muy útil, ya que fijando un número arbitrario de confianza mínima deseada (confianza * 100 = %), se podrían realizar recomendaciones desde un área “X” a un área “Y”, si la confianza de (X,Y) respecto a (X) es mayor a la elegida.

Otro posibilidad es hilar más fino en el filtrado de los resultados por soporte. Al reducir el soporte mínimo, podremos observar nuevas duplas que quizás antes no aparecían, impensadas, pero corremos riesgo de realizar conclusiones erróneas debido al ser resultados provenientes de un sector muy reducido de datos. Si por ejemplo tomamos un filtrado más pequeño, de soporte mínimo 0,00035 (aproximadamente son 60 ocurrencias en todo nuestro set de datos de 200.000 postulantes aprox), podríamos obtener nuevas duplas:

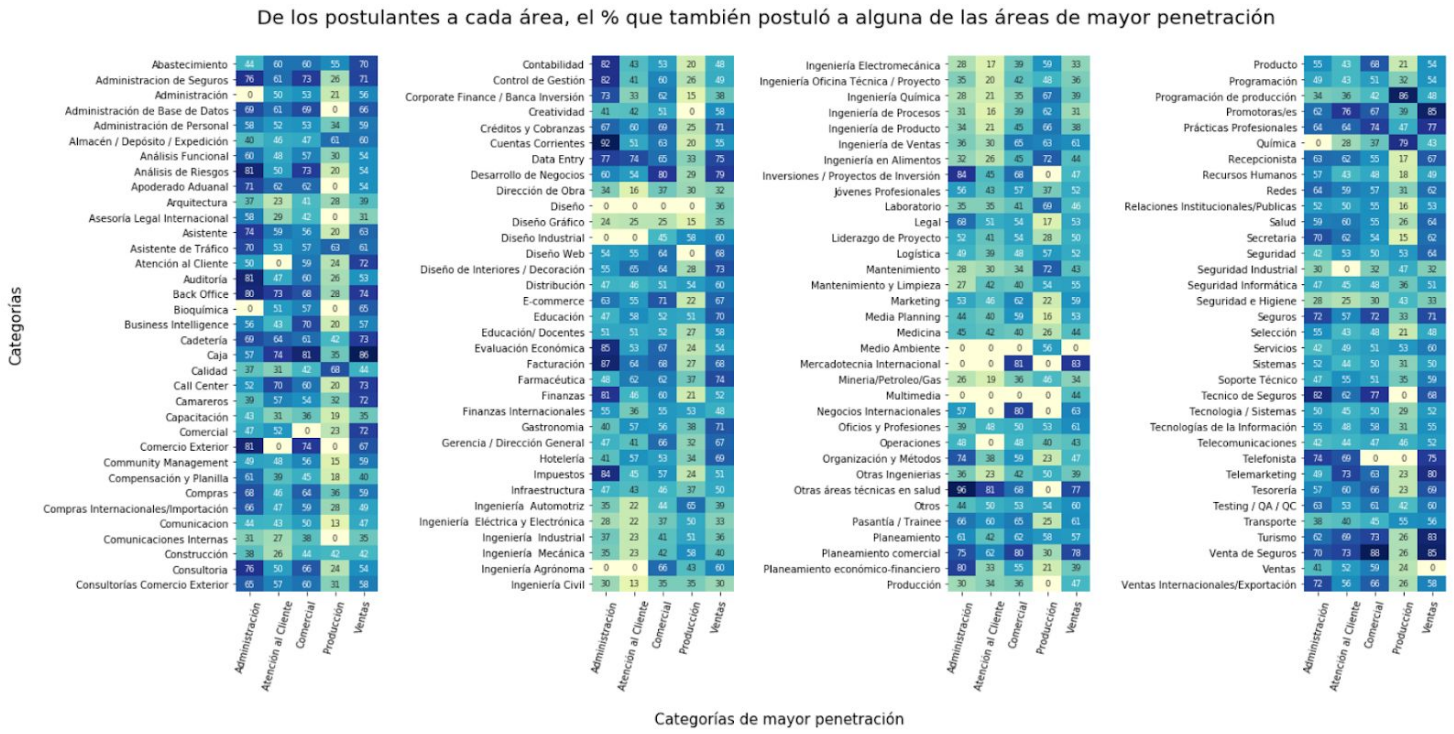


Inmediatamente vemos nuevas duplas surgidas, como la de [Veterinaria, Comercial] (de los postulantes a Veterinaria, el 88% postuló también a Comercial), o la de [Data Warehousing, Administración], o [Turismo, Ventas].

Qué soporte y confianza mínima elegir para aplicar a los algoritmos de recomendación no son únicos, dependerá de una combinación de cuánta seguridad al recomendar quiera, y también de cuánta diversidad de duplas quiera abarcar.

2.10.4 Top 5 áreas de mayor penetración: su relación con las demás categorías

En este caso, analizamos la confianza de todas las áreas respecto a las 5 de mayor penetración (las 5 de mayor soporte individual), para entender su alta tasa de ocurrencias:



En efecto, Administración, Atención al Cliente, Comercial, Producción, y Ventas son las de mayor penetración por la alta cantidad de personas que postularon a otras áreas y aplicaron a éstas. La diversidad de áreas “conectadas” con Administración, Atención al Cliente, Comercial, y Ventas es sorprendente. Prácticamente todas las demás áreas han postulado a éstas 4. Para el caso de Producción no vemos tanta diversidad, pero entendemos porqué se posicionó como una de las 5 de mayor penetración al ver mucho impacto desde las distintas áreas de ingeniería (2do heatmap).

Por otro lado, las celdas más oscuras nos indicarían una buena oportunidad de recomendarle a alguien que postuló a la categoría ‘y’, algunos avisos de la categoría ‘x’ en cuestión.

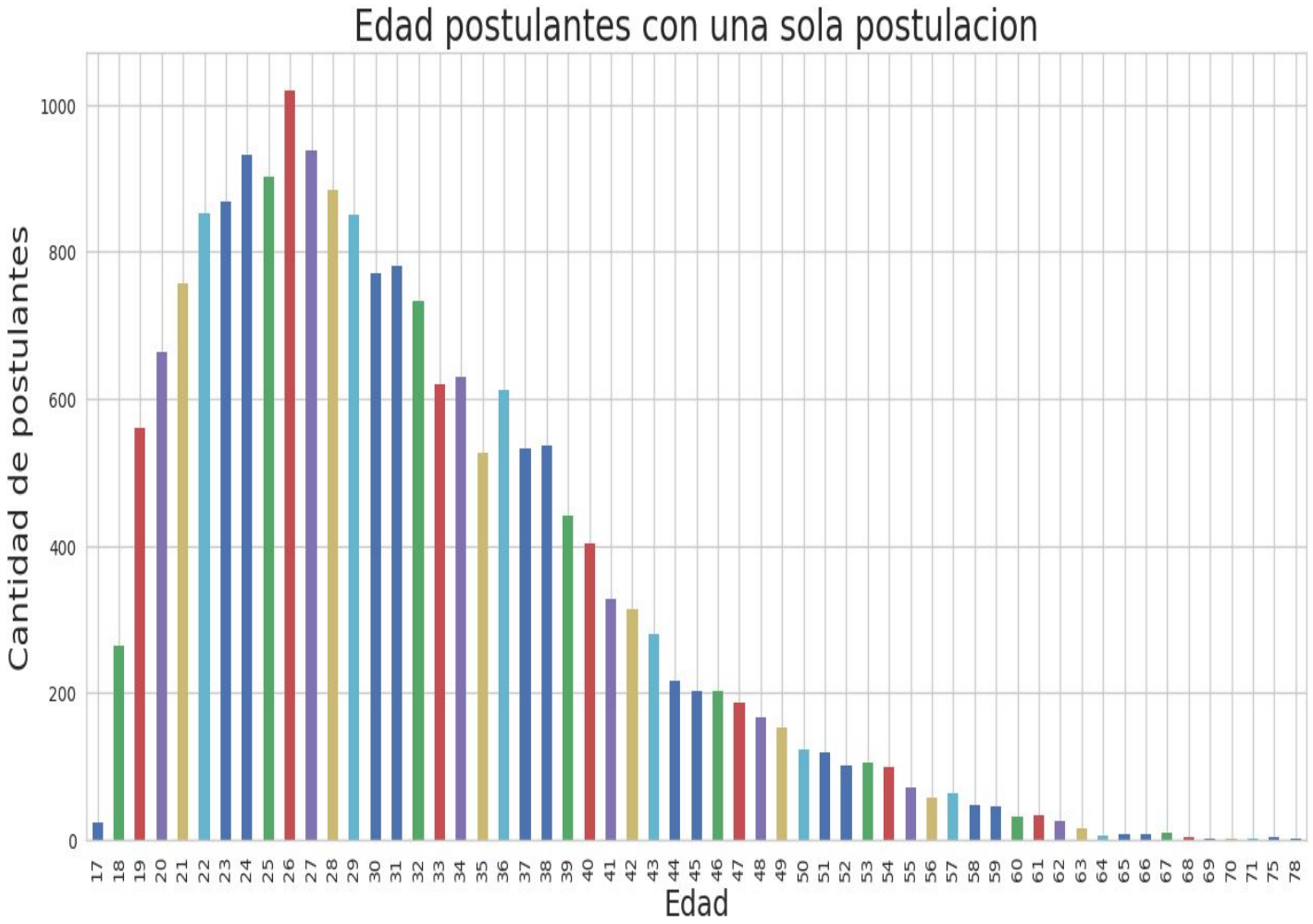
2.11 Tendencias de usuarios con una sola postulación

2.11.1 Introducción

Observando la información sobre las postulaciones se notó que de 200888 postulantes había 19947 postulantes con una sola postulación que es el mínimo que se maneja. Es decir que casi el 10% de los postulantes sólo se postularon una vez en mes y medio. Esto llevó a investigar si esas personas que se postularon “poco” tenían algunas características en particular.

2.11.2 Usuarios con una sola postulación según su edad

Se hizo una análisis acerca de las edades de los postulantes con pocas postulaciones. Se obtuvo el siguiente gráfico:

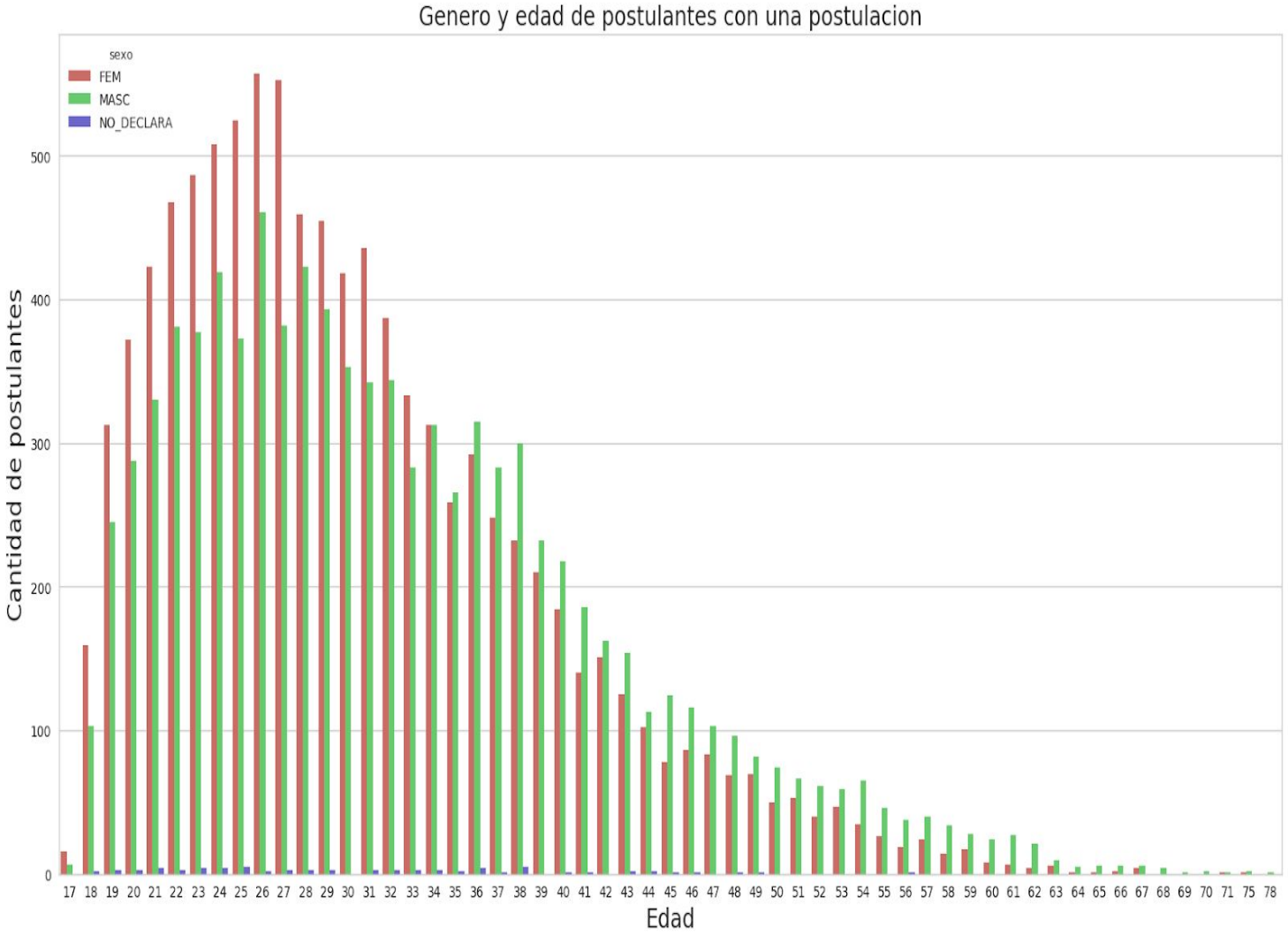


Aquí se ve que sigue dominando la edad de 26 años, esto es porque efectivamente hay muchos postulantes en el set de datos con dicha edad y por lo tanto hay gente con varias postulaciones como gente con pocas postulaciones.

Donde sí puede verse la diferencia es con la línea de tendencia hacia arriba hasta los 26 y luego hacia abajo que había en el gráfico de los postulantes según la edad en la sección 2.2. Ahí puede observarse que por ejemplo las personas de 36 años suelen tener más de una postulación a diferencia de las personas con 35.

2.11.3 Usuarios con una sola postulación según su edad y género

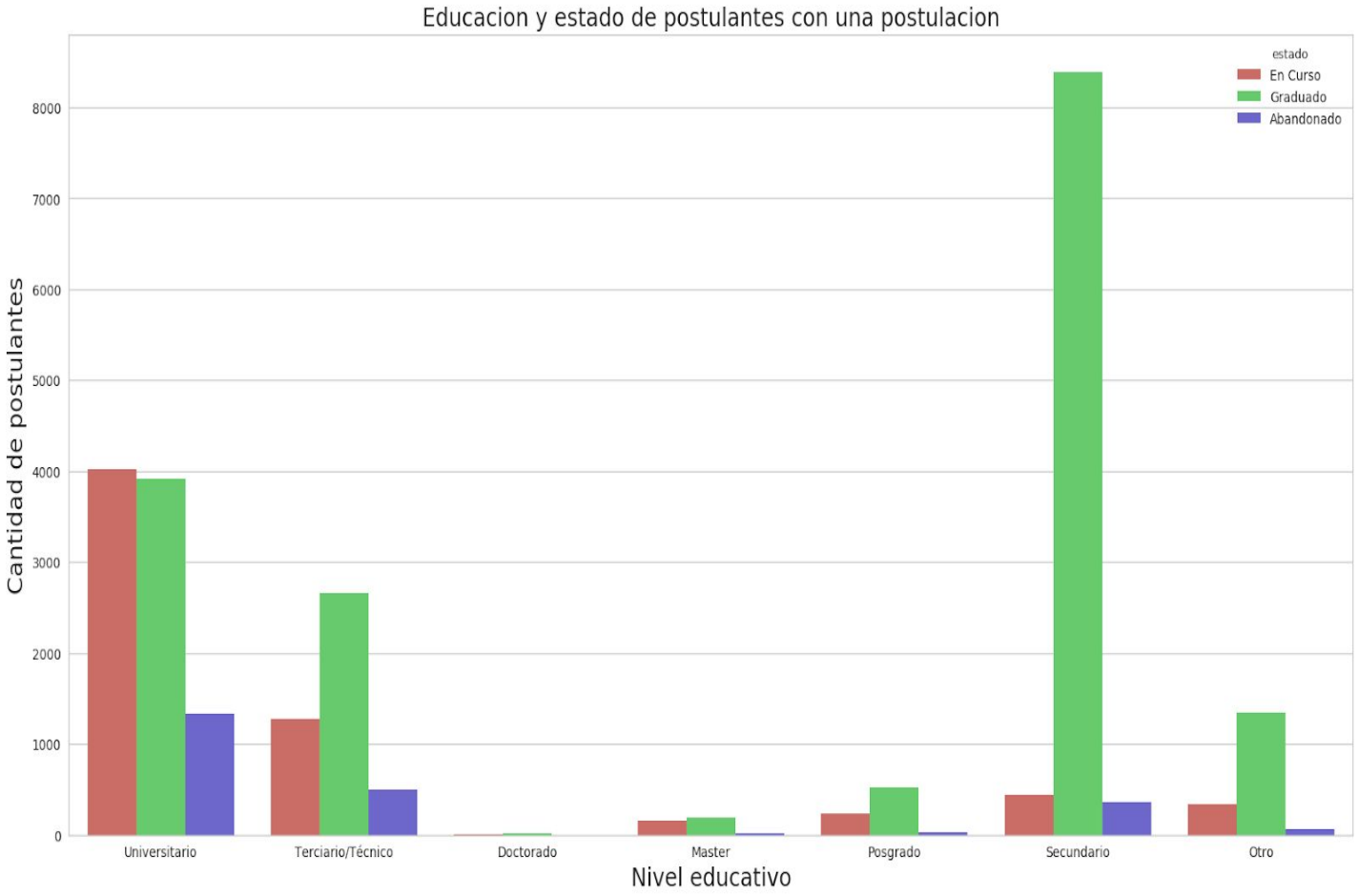
La información anterior también se integró con el género de las personas obteniendo un gráfico múltiple:



Aquí se pudo observar un dominio del género femenino hasta los 34 años y a partir de allí un dominio del género masculino, dando a entender que las mujeres menores de 34 años se postulan poco y los hombres mayores de 34 años también.

2.11.4 Nivel y estado de estudios para usuarios con una sola postulación

Luego de analizar el género y edad de las personas con pocas postulaciones se analizó la educación. Es decir, el tipo de educación que tenían y su estado. Se obtuvo el siguiente gráfico:



Como sucedió en el caso de las edades vuelve a haber un dominio del secundario completo porque hay muchos postulantes con esa situación. Comparando con el gráfico de la sección 2.1 se puede observar una notable diferencia en la educación universitaria. Por un lado los niveles de las personas graduadas se mantienen prácticamente iguales, pero el nivel en curso baja considerablemente. Esto arroja que las personas graduadas de la universidad se postulan poco y que hay una cantidad considerable de personas que están cursando la universidad suelen postularse más de una vez.

3. Avisos online a recomendar

Se realizó un algoritmo de recomendación en el cual se hizo uso de los set de datos de postulaciones y avisos_online para poder, en base a una postulación, sugerir avisos posibles. Lo que se realizó fue a una cierta postulación ‘A’ por el postulante ‘a’, se obtuvieron todos los otros postulantes a este mismo avios. A cada uno de ellos se buscó otras postulaciones realizadas (sin repetidos) para luego filtrar por el área del aviso ‘A’ y por último quedarme únicamente con los avisos que se encuentren online según indica el csv. A su vez a estas recomendaciones luego se las puede filtrar por tipo de puesto de trabajo, por jornada laboral u otra característica que distinga a los avisos. Dicho algoritmo se puede observar en el repositorio de GitHub.

4. Conclusiones generales

Aunque en cada punto de los distintos análisis se realizaron apreciaciones y conclusiones, queríamos remarcar algunos puntos (algunos de los que entendemos como más importantes fueron remarcados en negrita) que podrían tener una utilidad real, porque al fin y al cabo ese sería el objetivo de este análisis. Si bien tiene un objetivo académico de aprendizaje de herramientas y conceptos, también es importante poder rescatar cierta utilidad práctica de todo esto.

Entonces, de utilidad práctica, ¿qué comportamientos o conclusiones pudimos inferir de los análisis realizados?

- Sería una práctica recomendable distribuir nuestras comunicaciones con recomendaciones (esto sería por ejemplo los e-mail con avisos recomendados que envía la plataforma a un usuario) en mayor cantidad los primeros días de la semana (Lun, Ma, Mie), bajando la cantidad según vimos los días Jue, Vie, y aún menor número los fines de semana. Además, para cada uno de estos días, distribuir las comunicaciones según la hora, con un mayor número durante la mañana, de 08 a 10 hs, y a partir de allí descendiendo sostenidamente hasta las 04 hs del día siguiente.

- Recordar también que, como vimos, en los últimos 10 días del mes, se produce un incremento de las postulaciones, por lo que sería quizás buena idea aumentar el caudal de comunicaciones por esas fechas.

- También como vimos, al incrementar el nivel laboral al que postula un usuario, incrementaron en cierta medida los estudios de ese usuario, por lo que otra buena práctica podría ser recomendar anuncios más avanzados, como de Gerencia o Jefatura, a personas que tengan estudios más altos, es decir, a partir de Universitario Graduado.

- Como vimos en en el punto 2.10, pudimos detectar las duplas más frecuentes que se dan, y de allí detectar mediante la confianza que categorías son muy “cercanas”. Si fijamos un soporte mínimo y una confianza mínima, tendríamos categorías muy probables (por el análisis estadístico de la tendencia) a los que una persona se podría postular, y podríamos recomendar anuncios de dichas

categorías. O continuar e integrar estas recomendaciones con otros parámetros, y refinarlas aún más, para mejorar la probabilidad de que se postule.

También se pueden repetir muchas más conclusiones, que como mencionamos anteriormente se encuentran dispersadas en los distintos puntos de análisis, pero aquí citamos a aquellas a las cuales se les puede observar una utilidad práctica muy directa.