Universidad Nacional Autónoma de México

Facultad de Estudios Superiores
Acatlán

Diplomado de Ciencia de Datos
Módulo 1

proyecto

Análisis de ofertas de empleo para Data Scientists

Módulo I
Profesora : Malerva Reséndiz Carla Paola
Alumno : Hernández González Ricardo Paramont
Fecha : sábado 20 de febrero de 2021
Ubicación del texto : https ://es.overleaf.com/read/wwgcvjfmtvsy

Índice

1. Presentación

1.1 Objetivo

En el presente proyecto, se presenta el procesamiento de una tabla de datos correspondiente a varias ofertas de empleo del área de ciencia de datos en los Estados Unidos, obtenidas por webscrapping del portal Glasddor. El objetivo de la transformación de los datos es llegar a una Tabla Analítica de Datos, lista para su uso en aprendizaje automático no supervisado, que pueda generar un modelo que a partir de características conocidas de la oferta de empleo, prediga un estimado del sueldo de la misma.

1.2 Diccionario de datos

Variable	Tipo	Descripción
Job Title	discreto	Nombre de la oferta de trabajo.
Salary Estimate	continuo	Rango del salario estimado por el portal Glassdoor, en miles de dólares
Job Description	texto	Descripción de distintos rubros de la empresa, puesto y solicitud.
Rating	discreta	Calificación por parte de usuarios de la empresa en cuestion. Valor entre 1.0 y 5.0
Company Name	Texto	Nombre de la compañía que ofrece la oferta de trabajo.
Location	discreto	Ciudad donde se ubica el trabajo ofertado, junto con el estado o país respectivo.
Headquarters	discreto	Ciudad y estado/país donde se ubica la sede de la empresa.
Size	discreto	Divide a las empresas según su número de empleados en rangos específicos.
Founded	discreto	Año de fundación de la empresa.
Type of Owners	discreto	Tipo de compañía según razón social o identidad.
Industry	discreto	Tipo de compañía según idustria en la que trabaja.
Sector	discreto	Sector en la que la compañía ofrece sus bienes o servicios.
Revenue	discreto	Utilidades de la compañía en dolares al año.
Competitors	discreto	Principal compañía competidora de la compañía que ofrece la oferta de trabajo.
Easy Apply	discreta	Etiqueta que indica si la aplicación por el puesto es sencilla.
Salary Minimum	continuo	Rango inferior del estimado del salario. Extraido de Salary Estimate.
Salary Maximum	continuo	Rango superior del estimado del salario. Extraido de Salary Estimate.

2. Calidad de datos

2.1 Etiquetado de variables

Se revisó el tipo de dato, valores únicos, y primeros cinco registros de cada variable para su clasificación dentro de los criterios de :

- Continua
- Categórica
- Texo
- Fecha

Los hayazgos llevaron al etiquetado de las variables de la siguiente forma :

Continuas

- c_salary_minimum
- c_salary_maximum

Categóricas

- v_job$_t itlev_salary_estimate$
- v_rating
- v_location
- v_headquarters
- v_size
- v_founded
- v_type_of_ownership
- v_industry
- v_sector
- v_revenue
- v_competitors
- v_easy_apply
- v_salary_estimate_source'

Texo

- t_company_name
- t_job_description

Fecha

No se encontró ninguna variable de fecha

Se observa que la gram mayoría de las variables en este problema son categóricas.

2.2 Duplicados

No se encontro ningún registro duplicado a lo largo de la tabla de datos.

2.3 Completitud

Se encontraron varias columnas con valores faltantes :

	columna	total	completitud
0	v_easy_apply	3745	4.195446
1	v_competitors	2760	29.393707
2	v_revenue	1392	64.389870
3	v_founded	977	75.006395
4	v_sector	546	86.032233
5	v_industry	546	86.032233
6	v_rating	409	89.536966
7	v_headquarters	240	93.860322
8	v_size	229	94.141724
9	v_type_of_ownership	229	94.141724
10	v_job_title	0	100.000000
11	v_location	0	100.000000
12	t_company_name	0	100.000000
13	t_job_description	0	100.000000
14	salary_estimate	0	100.000000
15	c_salary_minimum	0	100.000000
16	c_salary_maximum	0	100.000000

Las variables de v_easy_apply, v_competitors, v_revenue y v_founded fueron eliminadas por contar con más de 20% de valores faltantes, los cuales son demasiados para intentar imputarlos.
Los valores faltantes del resto de las variables son imputados en la sección de Valores Ausentes.

2.4 Limpieza de texto

Se comprende como limpieza de texto a la remoción de caracteres especiales (.,-_% etc.), incluyendo caracteres acentuados y la leta 'ñ', además de la transformación de todas las mayúsculas a minúsculas. Debido a la naturaleza de algunas, se conservaron algunos caracteres especiales, como en el caso de localidad para indicar la separación ciudad, estado.
Las siguientes variables sufrieron limpieza de texto :

- v_job_title
- v_location
- v_headquarters
- v_industry
- v_sector
- v_type_of_ownership
- v_size
- t_company_name
- t_job_description

2.5 Consistencia

- Se asegura que el valor del mínimo salario estimado es mayor al salario mínimo estadounidense. Para lo mismo se considera un salario mínimo de $7.5/h, una jornada laboral de medio tiempo de 4h/día y el calendario laboral estadounidense de 2019 que fue de 261 días. No se encuentra ninguna inconsistencia.
- Se asegura que en ninguno de los registros, el salario mínimo tenga un valor mayor al salario máximo. No se encuentra ninguna inconsistencia.
- Se asegura que ninguno de los registros de la calificación de la empresa (v_rating) esté por debajo de 1 y por arriba de 5. No se encontró ninguna inconsistencia.

2.6 Normalización

Normalización de v_job_title
Una exploración simple de la variable nos permite identificar al menos 1971 valores únicos que sin embargo, comparten grandes similitudes. Por ejemplo, existen muchos puestos que son en esencia de científico de datos pero con algún nombramiento adicional como trainee, junior, senior, manager, entre otros. Por lo que se procede a agrupar grandes categorías de empleos :

- data scientist
- data engineer
- data analyst
- machine learning professional
- business intelligence analyst
- analyst of other nature
- data architect

Adicionalmente, se encuentran varios títulos que tienen una sola ocurrencia y cuyas especificaciones parecen muy especiales. A dichos empleos con una sola ocurrencia que no fueron agrupados en las categorías anteriores, se les incluyó en la categoría de highly specific. A las categorías restantes, se les agrupó en la categoría de others, la cual posee el menor número de ocurrencias.

Normalización de v_location
La variable consta de la ciudad y estado o país donde se localiza la oferta de trabajo, por lo que se extraen de ellas las variables de :

- v_city : con 191 de categorías.
- v_state : con 10 categorías.
- v_big_city : variable dummy que indica si la oferta de trabajo se ubica en una de las 20 ciudades más pobladas de Estados Unidos.

*La variable v_location es eliminada por lo tanto.

Normalización de v_city
La variable consta de la ciudad y estado o país donde se localiza la sede de la empresa, por lo que se extraen de ellas las variables de :

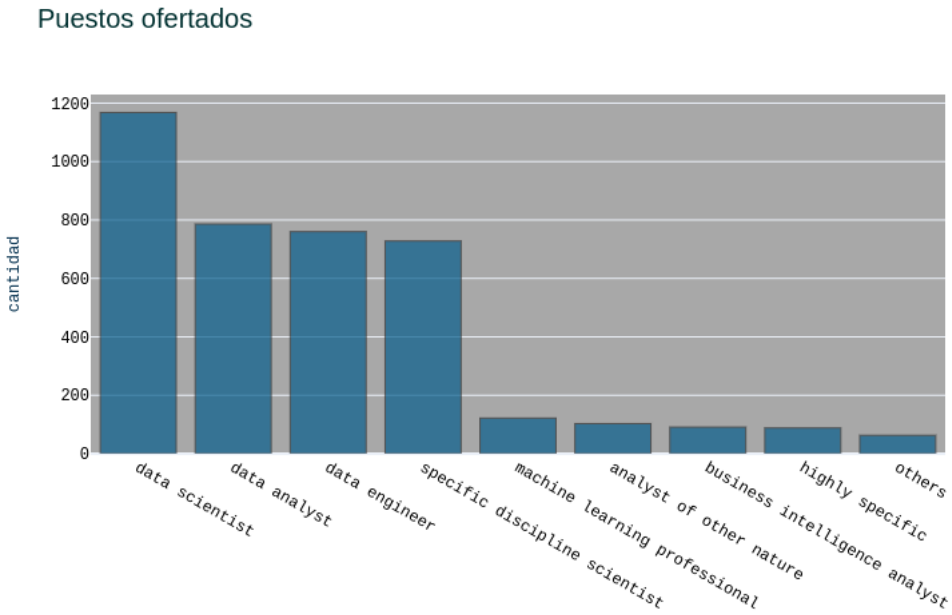
- v_headquarters_city : consta de 522 categorías
- v_headquarters_state : 50 categorías.

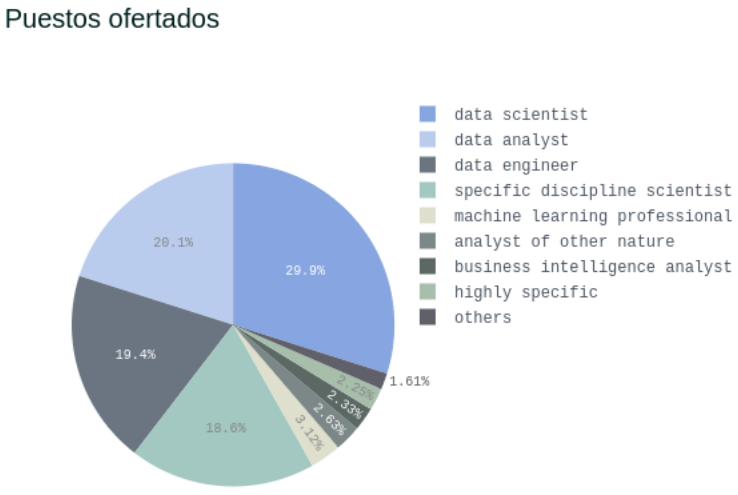
En la variable v_headquartes_state se identifica a aquellos registros de paises fuera de los Estados Unidos y se les agrupa en la categoría foreign countries.

Normalización de v_industry
Se agrupan todas las categorías con 4 o menos ocurrencias en la categoría others.

3. Análisis Exploratorio de Datos

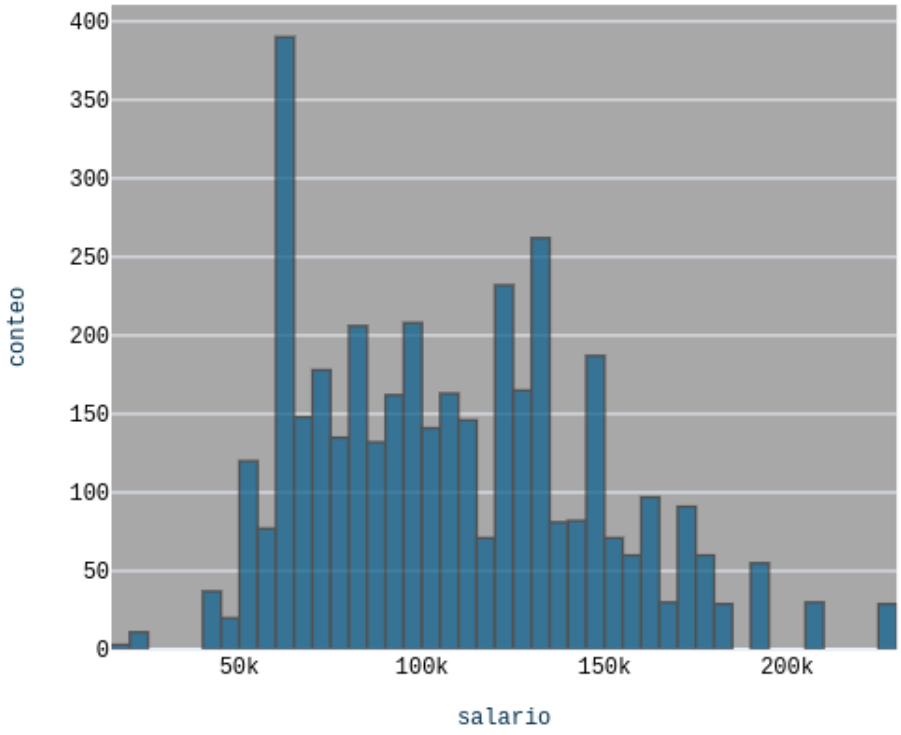
Distribución de puestos ofertados





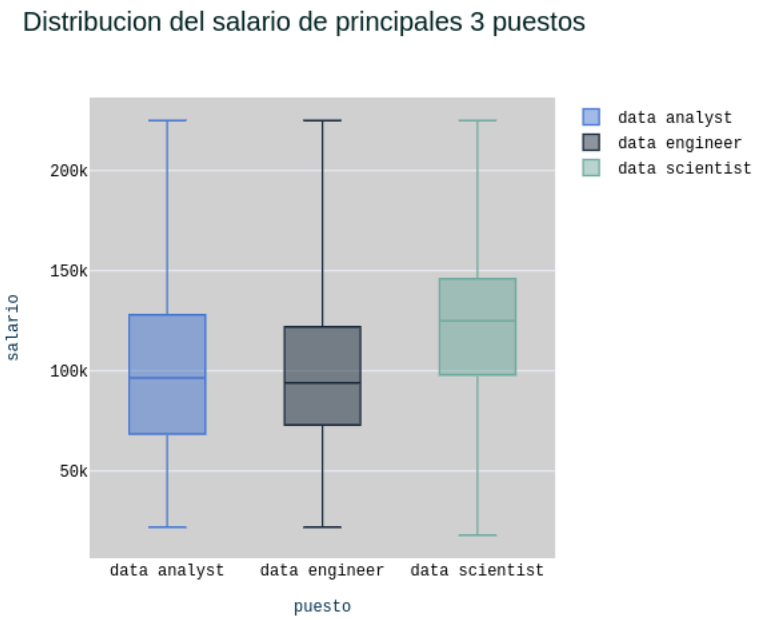
Casi uno de cada tres puestos ofertados tiene el título de Data Scientist. Los tres títulos de empleo más comunes son Data Scientist, Data Analyst y Data Engineer, entre los tres componen 70% de las ofertas.

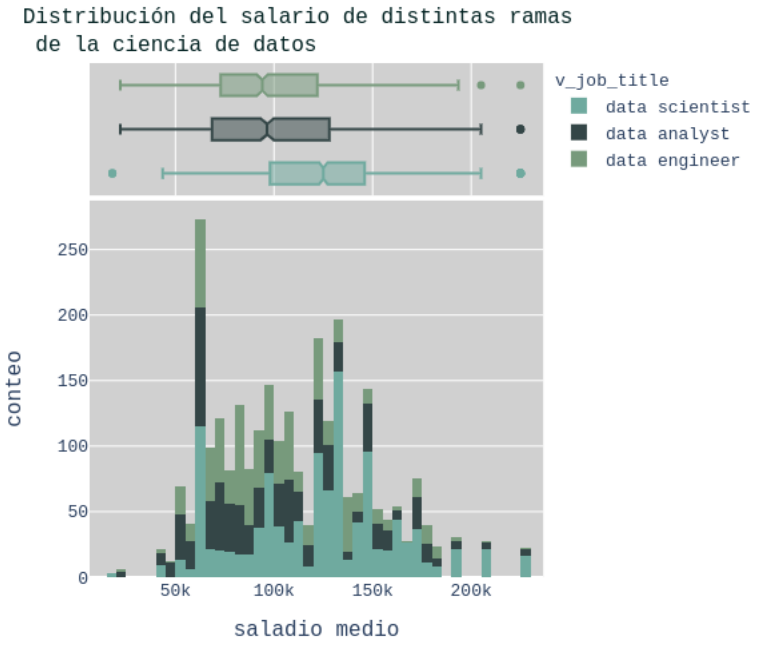
Distribución de la media de la estimación de salarios



Prácticamente todos los salarios sobrepasan los $50k anuales, la mayoría se encuentran alrededor de los $100k anuales, sin embargo, hay casos que llegan hasta los $225k anueales.

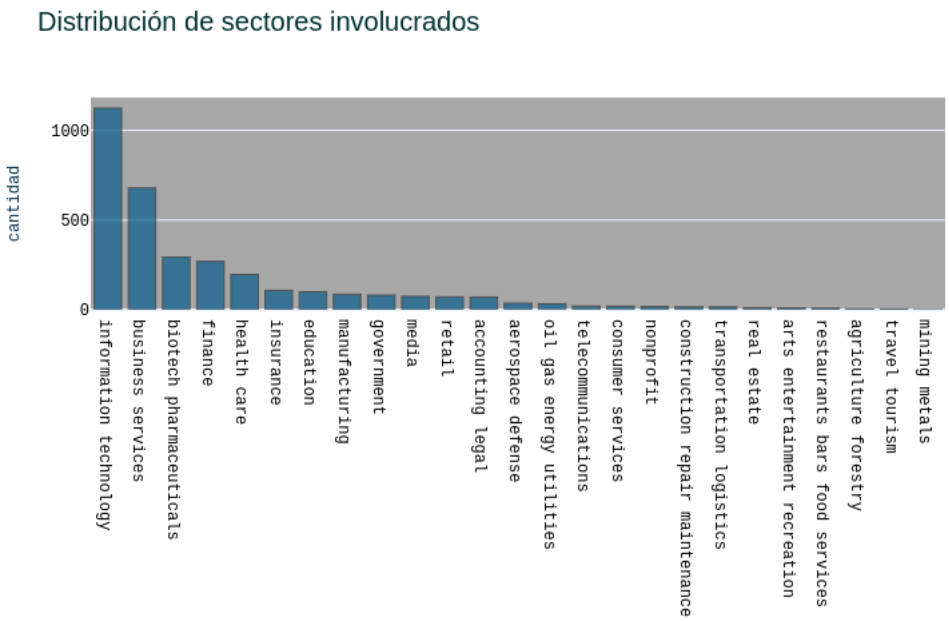
Distribución del salario de los primeros tres puestos ofertados

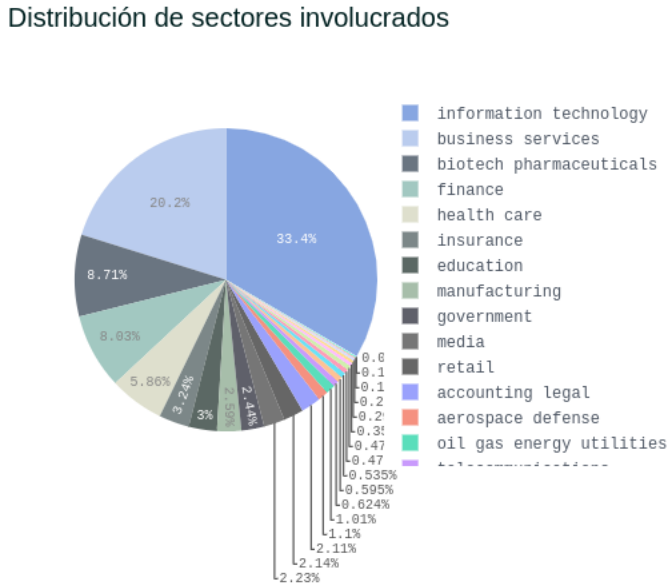




Se observa que los puestos de Data Scientist además de ser más comunes, son también mejor pagados, seguidos por los puestos de Data Analyst y Data Engineer.

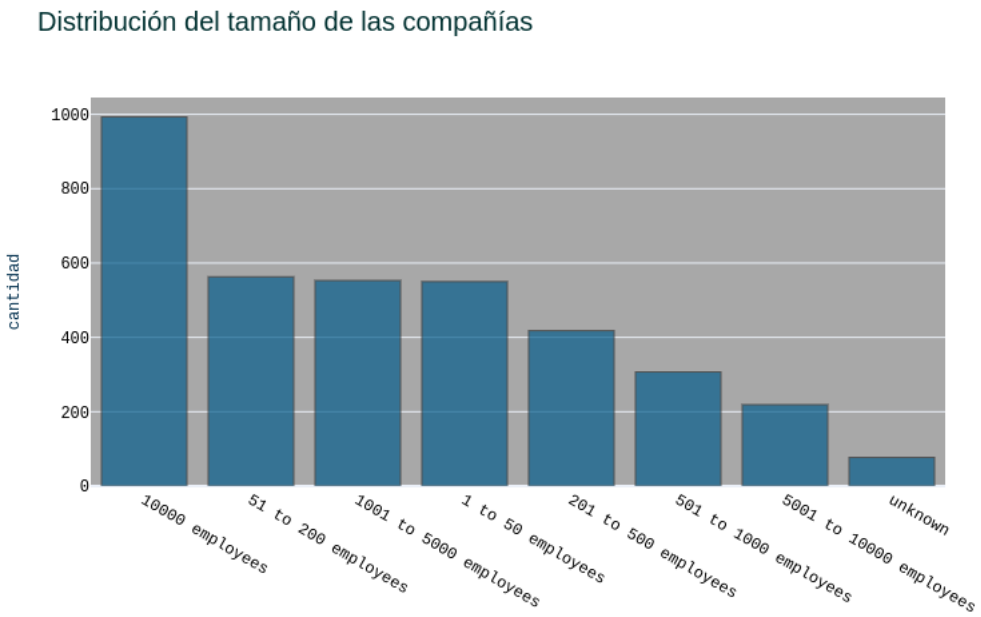
Principales sectores involucrados en las ofertas de trabajo

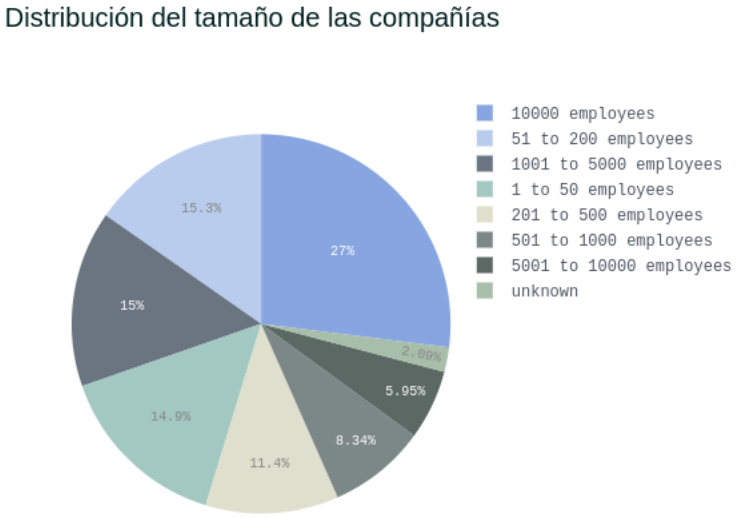




Las tecnologías de la información constituyen el principal sector involucrado en las ofertas de empleo. Uno de cada tres empleos es ofertado dentro del mismo. El sector de business services es el siguiente más común, la mitad de las ofertas de empleo son acaparadas entre estos dos.

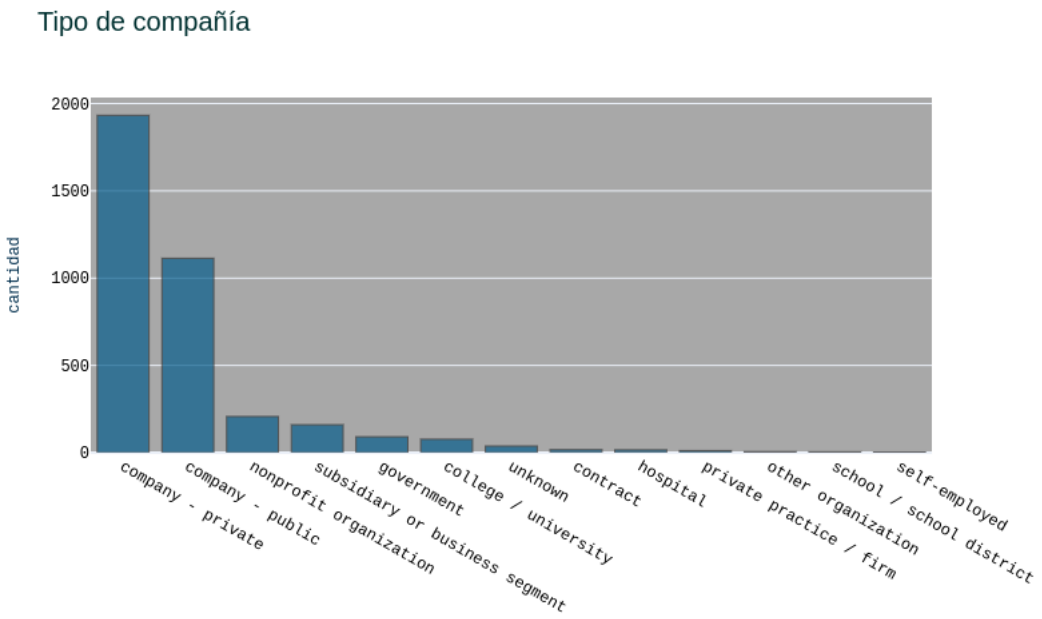
Distribución del tamaño de las compañías ofertando empleo

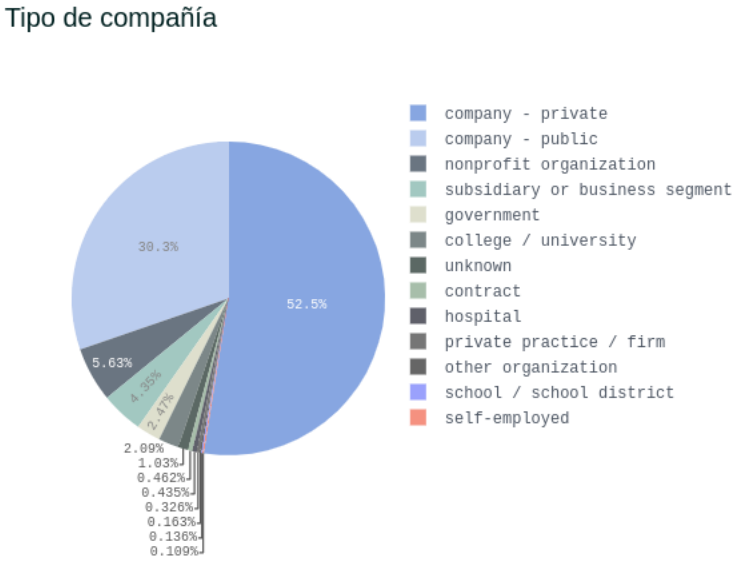




Las mayoría de los empleos son ofertados por las compañías más grandes dentro de la clasificación de la tabla (1000 o más emleados). Sin embargo, son empresas pequeñas (entre 51 y 200 empleados) las segundas en ofertar más empleos.

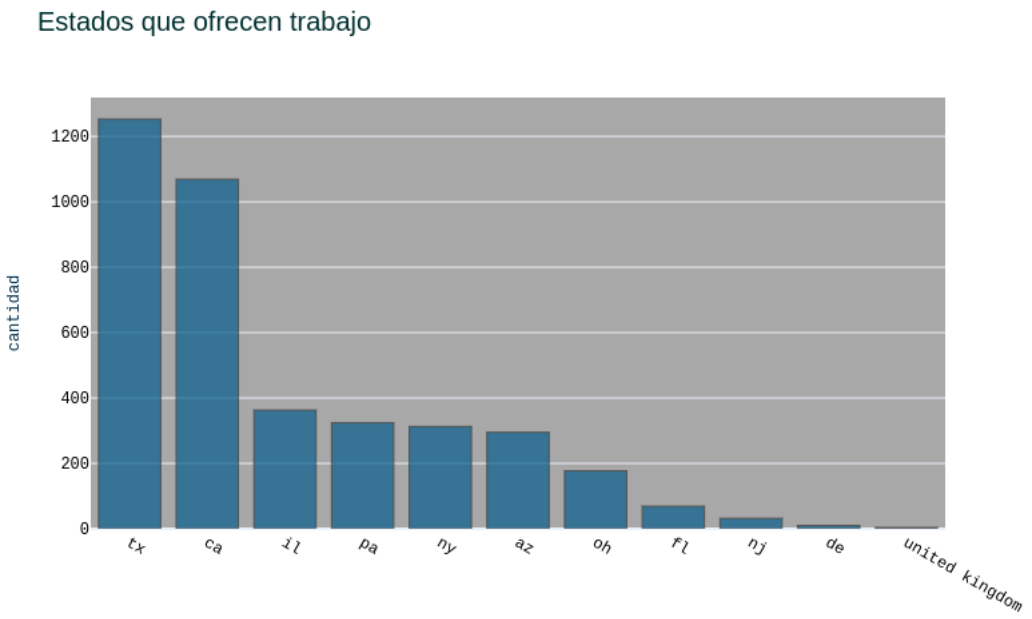
Distribución del tipo de propiedad de las compañías

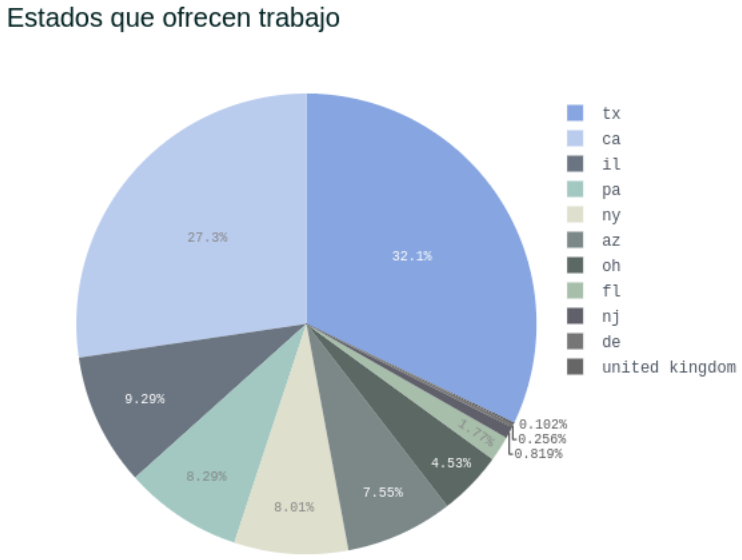




Más de la mitad de las compañías son privadas y aproximadamente un tercio son públicas. Apenas 0.1% de los trabajos ofertados se identifican dentro del autoempleo.

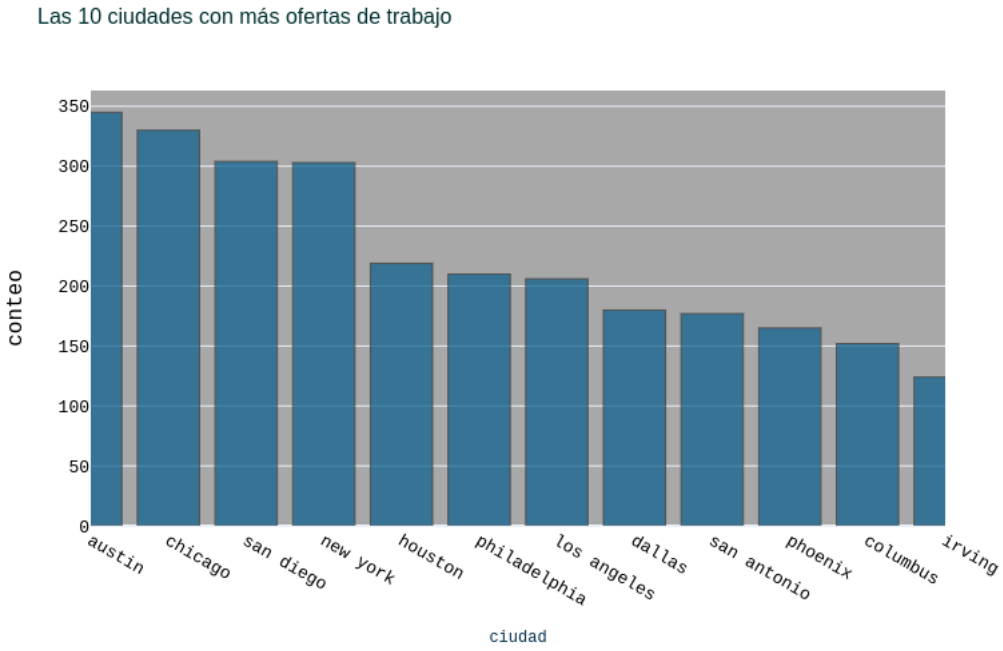
Estados donde se ubican los empleos ofertados





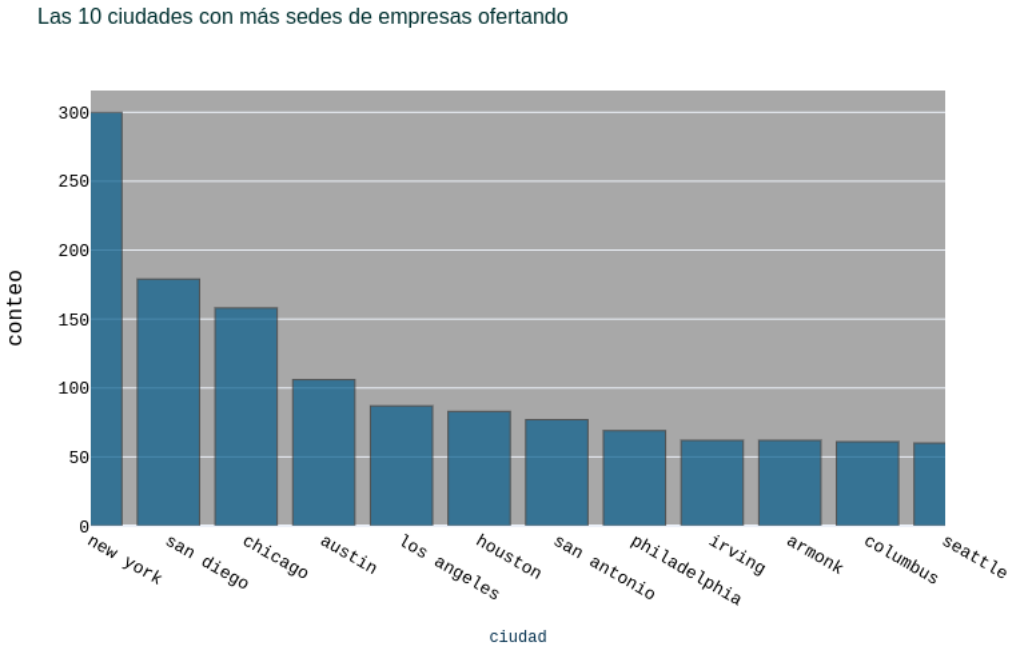
La mayoría de los empleos ofrecidos se ubican en Texas, fruto de la transición de la economía del estado del sector energético al sector de las tecnologías de la información. No muy lejos, sigue California, que representa el estado lider en inovaciones de tecnologías digitales y el que más aporta al PIB nacional. Curiosamente, se observa que algunos de los empleos ofertados son en el Reino Unido.

Las 10 ciudades con más ofertas de empleo



La ciudad que más empleos oferta es Austin, Tx, la cual fue calificada como la segunda con mayor crecimiento económico en Estados Unidos en 2020, superada sólo por Denver, Tx. La seguda ciudad que más empleos oferta es Chicago, la cual representa la economía más grande fuera de las costas estadounidenses.

Las 10 ciudades con más sedes (headquarters) de compañías que ofertan empleos

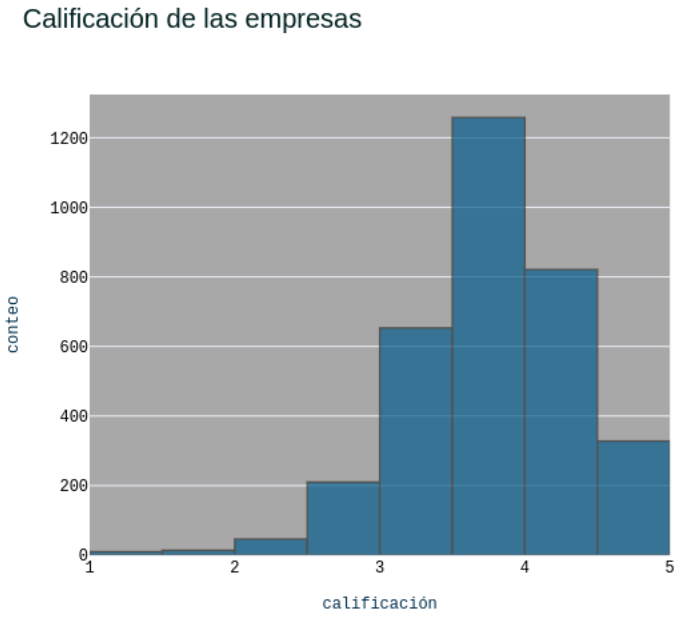


La ciudad que aloja más sedes de las compañías involucradas es Nueva York, superando a San Diego, el segundo lugar, por casi el duble. Curiosamente, la única ciudad de california en el top 10 es Los Angeles, apenas en el puesto 5, a pesar de que California es el estado lider en tecnologías digitales.

Los 10 estados con más sedes de empresas ofertando

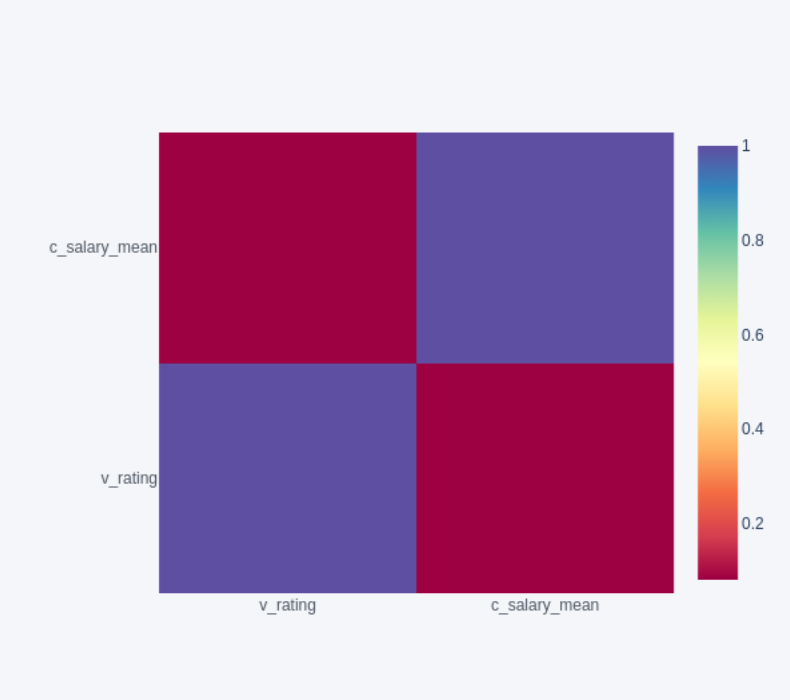
A pesar de que sólo una ciudad de California se encuentra dentro del top 10 de ciudades con más sedes, California concentra la mayoría de sedes a lo largo de sus ciudades.

Distribución de la calificación de las empresas



Se nota que la tendencia es que las empresas sean calificadas con una calificación de 3.5 estrellas en una escala de 1 estrella a 5 estrellas.

Correlación entre calificación de las empresas y salarios que ofrecen



Se aprecia que no existe una correlación significativa entre la calificación de la empresa y los salarios que ofertan. Lo que indica que los empleados no toman en cuenta el salario entre los criterior principales para calificar su entorno de trabajo.

Palabras más mencionadas en la descripción de las ofertas de empleo

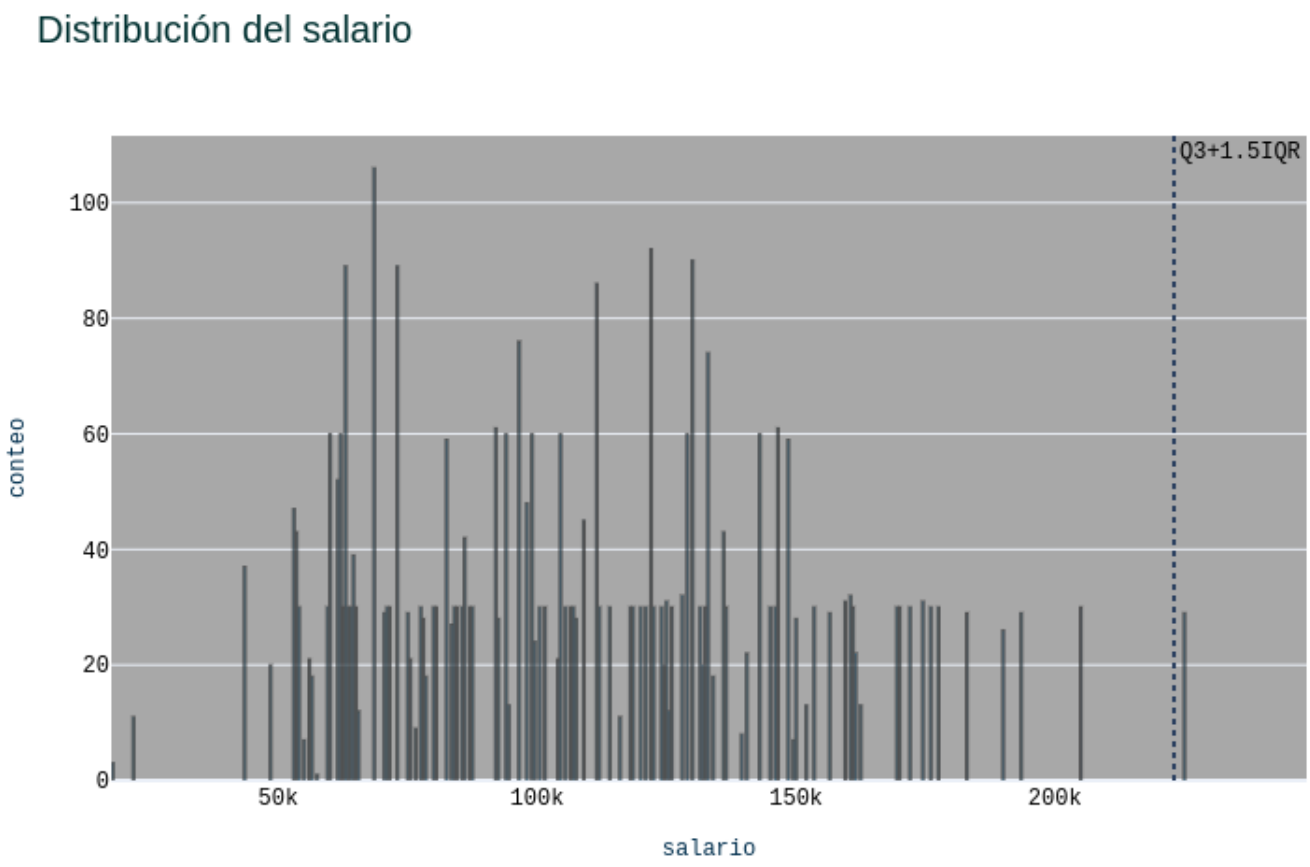
La palabra más común es 'Data', lo que resulta obvio. Las siguientes más mencionadas según su orden son experiencia, negocios y habilidades. Si nos enfocamos en lenguajes de programación, los más mencionados son python, seguido por R y después java.

Palabras más comunes en el nombre de las empresas

Se alcanza a apreciar que los empleos ofertados por Apple, Amazon, IBM, Facebook e incluso la Universidad de Texas son suficientes para resaltar en el gráfico.

4. Valores anómalos

Se decide hacer un analisis de valores atípicos para la variable objetivo de salario. Se encuentra que existen 29 valores atípicos demasiado altos (0.74% de los registros totales), que sobrepasan el límite heurístico de distribución de Q3 + 1.5IQR. Sin emabrgo, observando dichos valores, se descubre que estos sobrepasan el límite heurístico de distribución por apenas 1.85% el valor de la media del salario.



Este hecho, junto con que dichos salarios tan altos son de gran interés para su inclusión en el modelo predictivo hacen tomar la decisión de conservarlos.

5. Valores ausentes

Se cuenta con 7 variables que contienen valores ausentes a imputar :

	columna	total	completitud
0	v_industry	418	86.632555
1	v_sector	418	86.632555
2	v_rating	319	89.798529
3	v_headquarters_state	183	94.147745
4	v_headquarters_city	181	94.211705
5	v_size	174	94.435561
6	v_type_of_ownership	174	94.435561

La variable de calificación (v_rating) se imputa como si fuera continua, debido a que aunque es categorica, también es ordinal. Realizando la prueba KS, se descubre que le mejor valor a imputar es la mediana, la cual es un valor de 3.8.

Como todas las demás variables son de tipo categorico no ordinal, se imputa la moda para cada una. Se comprueba que no hay cambios significativos en la distribución por medio de la prueba de chi cuadrada para las variables de :

- v_type_of_ownership
- v_type_of_ownership
- v_headquarters_state

Sin embargo, las variables de :

- v_size
- v_industry
- v_sector

no pasan la prueba, sin embargo, son igualmente imputadas con la moda por no contar con el conocimiento para imputar dichas variables con el uso de un modelo.

6. Ingeniería de variables

Ya que la estimación de salario se compone de un rango en nuestra tabla de datos, con la variable de salario estimado mínimo (c_salary_minimum) y salario estimado máximo (c_salary_maximum), se decide asimilar ambas en una sola variable que consiste en la media aritmética entre las dos variables. Dicha variable se llama tgt_salary_media, con la etiqueta tgt (target) ya que consiste en el objetivo a predecir del podelo que se ajustará con la tabla final.

Con el objetivo de poder normalizar más sencillamente las variables de ubicación que son v_location y v_headquarters, se decide separar estas en ciudad y estado, ya que la catidad de categorías por estado es significativamente menor. Sin embargo, se conserva la información de ubicación por ciudad en caso de que alguna resulte ser relevante para la predicción del objetivo. Si una ciudad resulta o no ser relevante para el modelo, lo decidirán los filtros en la sección de Reducción de Dimensiones.
Ninguna transformación se realizará a las variables de v_raiting y tgt_salary_mean por considerar que su estructura en este punto del proyecto puede ser directamente incluida en el entrenaminto de un modelo.

6.1 One-Hot encoding (Dummies)

Se decide codificar a las variables categóricas con 15 categorías o menos con la creación de variables dummies. Dichas columnas son :

- v_job_title
- v_sector
- v_size
- v_state
- v_type_of_ownership

6.2 Target-Encoding

Para las variables con más de 15 categorías se decide codificarlas con el uso de la técnica target encoding. De esta forma, se intenta controlar la explosión de variables. Dichas columnas son :

- v_industry
- v_city
- v_headquarters_city
- v_headquarters_state

6.3 Vectorización de texto

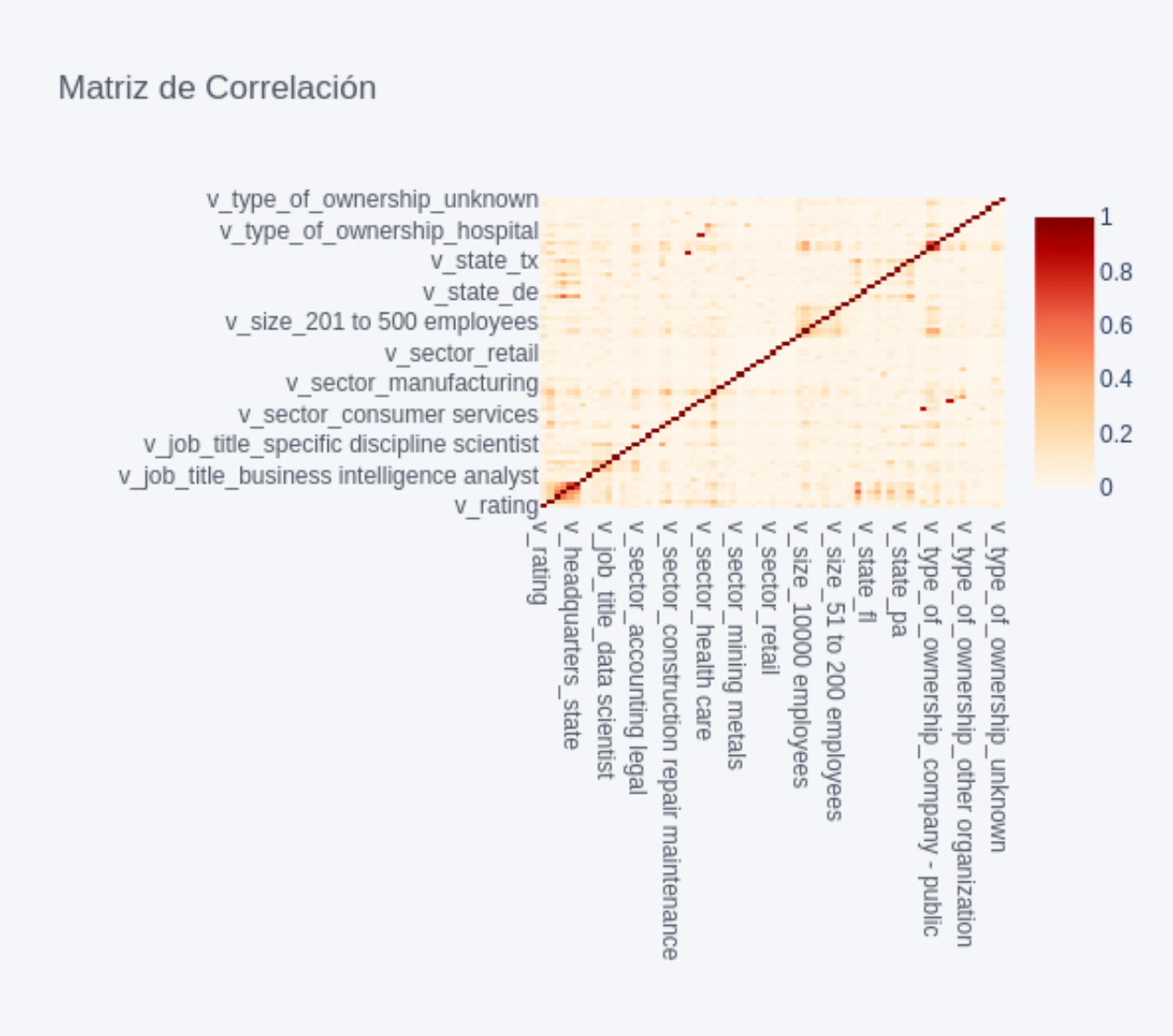
A las variables de texto (t_job_description y t_company_name) se les aplica una vectorización con la técnica de count vectorizer. Para esto, se debió eliminar stop words, eliminar hapaxes, tokenizar y lematizar previamente. La vectorización se aplicó especificandole al modelo incluir únicamente palabras con 15% de ocurrencias en el total del texto procesado. No se pudo realizar la remoción de hapaxes para la variable t_job_description, ya que la lista de hapaxes constaba de 43 mil palabras y la operación de buscar por cada palabra del texto en la misma se alargaba demasiado. Se decidió omitir dicho paso, ya que al fin y al cabo, el modelo sólo toma en cuenta las palabras con una alta frecuencia para vectorizar.

Al final se observa una gran explosión de variables, pasando de 16 columnas antes de la ingeniería de variables a 360.

7. Reducción de dimensiones

7.1 Filtro de alta correlación

Con el objetivo de eliminar variables de entrada, o independientes, que pueden ser descritas por otras variables de entrada, se explora el coeficiente de spearman entre cada una de las mismas :



Sólo una pequeña muestra de los nombres de la variable se muestran en el gráfico, pues en realidad se tienen 360 columnas en total

Se encuentras que las siguientes variables están altamente correlacionadas :

v_headquarters_city y v_headquarters_state se conserva v_headquarters_state
v_sector_education y v_type_of_ownership_college / university se conserva v_type_of_ownership_college / university
v_sector_government y v_type_of_ownership_government se conserva v_type_of_ownership_government
v_type_of_ownership_company - private y v_type_of_ownership_company - public se conserva v_type_of_ownership_company - public

7.2 Correlación con el objetivo

Se buscan variables con una correlación demasiado baja con el objetivo y se eliminan bajo el supuesto de que una baja correlación de spearman representa muy poca capacidad predictiva de la variable objetivo. Se encuentran 330 variables cuyo coeficiente de correlación de spearman está por debajo de 0.1, en su mayoría variables de palabras obtenidas por la vectorización del texto. Se procede a eliminar todas esas variables.

7.3 Multicolinealidad

Se estudia la multicolinealidad entre las variables independientes y se decide eliminar a las altamente relacionadas con el criterio de que su VIF (variable inflation factor) es mayor a 10. Se encuentran las siguientes variables con una alta multicolinealidad :

	variables	VIF
0	v_industry	86.566996
1	v_city	60.621267
2	v_headquarters_city	90.085193
3	v_headquarters_state	111.057827

Se procede a eliminar las variables presentadas en la tabla.

8. Tabla final

Al final se obtienen dos tablas analíticas de datos con 21 columnas. Una para el entrenamiento de un modelo de aprendizaje no supervisado con 3127 registros y otra para la validación del modelo con 782 registros. Un extracto de las tablas se presenta a continuación.

index	tgt_salary_mean	v_job_title_data scientist	v_state_ca	v_state_fl	v_state_il	v_state_ny	v_state_pa	v_state_tx
1	146000.0	1	0	0	0	1	0	0
4	146000.0	1	0	0	0	1	0	0
5	146000.0	1	0	0	0	1	0	0
10	146000.0	1	0	0	0	1	0	0
14	146000.0	1	0	0	0	1	0	0