Estudio de preferencias usuarios en

glassdoor

Abstracto:

Este trabajo explora el potencial de aprovechar los datos de Glassdoor mediante algoritmos de machine learning. A pesar de desafíos como la tendencia a reseñas negativas y la falta de verificación de la experiencia laboral, es un repositorio invaluable sobre las experiencias laborales en el mundo de trabajo. El uso de algoritmos y técnicas estadísticas avanzadas pueden analizar patrones en las opiniones de los empleados para predecir tendencias y evaluar la satisfacción laboral.

Utilizando algoritmos de aprendizaje supervisado, nos enfocaremos en técnicas de regresión para analizar patrones complejos en las opiniones de los empleados. Estos algoritmos nos permitirán identificar correlaciones entre variables clave, como los rangos salariales y las puntuaciones de reseñas, arrojando luz sobre cómo factores como el salario impactan directamente en la satisfacción laboral. Además, implementaremos algoritmos de procesamiento del lenguaje natural (NLP) para analizar las reseñas cualitativas de los empleados.

Al desglosar estos datos no estructurados, podremos extraer sentimientos y temas clave, transformando las experiencias subjetivas en información objetiva y cuantificable para nuestro análisis. Este enfoque integrado nos permitirá comprender no sólo las tendencias generales, sino también los matices sutiles en las opiniones de los empleados.

Audiencia:

Cualquier compañía que busca tener referencias y mejor información que con este modelo logrará tener una perspectiva sobre qué variables predicen mejor la tenencia de una opinión de la empresa. Este recurso es vital para las empresas al permitirles anticipar necesidades de fuerza laboral, fortalecer áreas clave y mejorar estrategias de atracción y retención de talento. Es una herramienta útil para anticiparse a las necesidades de tu fuerza de trabajo y ser una foco de mercadeo para la empresa.

Metadata:

El Dataset está compuesto por 18 columnas o variables. Tiene tres columnas de valores cualitativos nominales. Cinco columnas con valores cualitativos ordinales. Nueve columnas con variables numéricas de carácter discreto y una columna con data de tipo fecha, esta puede ser cualitativa o numérica dependiendo de la interpretación que se le de a la data, en nuestro caso es de será de tipo numérico. El número de registros o filas únicas sin contar los duplicados o eliminados son 286374.

Variables o features del Dataset

1. **Date_review (numerico)** : Fecha en la que se realizó la reseña.
2. **Job_title (nominal)**: Rol dentro de la empresa.
3. **current**: Estatus dentro de la empresa actual.
4. **location**: Lugar de la empresa (ciudad).
5. **overall_rating**: puntuación total de la empresa (entre 1-5).
6. **work_life_balance**: opinión del equilibrio entre la vida y del trabajo (1-5)
7. **culture_values**: opinión de la cultura empresarial (1-5)
8. **diversity_inclusion**: apreciación sobre el nivel de diversidad e inclusión (1-5)
9. **career_opp**: opinión sobre oportunidades dentro de la empresa (1-5)
10. ***comp_benefits**: opinión sobre compensación y beneficios (1-5)
11. **senior_mgmt:** apreciación sobre los directivos de la empresa (1-5).
12. **recommend:** apreciación sobre si recomendarías esta compañía ("mucho","normal", "malo").
13. **ceo_approv:** apreciación sobre el CEO o dueño de la compañía ("mucho","normal", "malo")..
14. **outlook:** apreciación sobre el futuro de la empresa ("mucho","normal", "malo").
15. **headline**: comentario a manera de título sobre la empresa.
16. **pros**: comentarios positivos sobre la empresa.
17. **cons:** comentarios negativos sobre la empresa.

Hipótesis

Hipótesis sobre Localización y Roles:

Hipótesis 1: Existe una correlación entre la localización y roles y las puntuaciones de reseñas en Glassdoor. Es decir, a medida que habrá localizaciones y roles ,donde la satisfacción laboral tiende a mejorar.

Hipótesis sobre Cultura Empresarial:

Hipótesis 2: Las opiniones sobre la cultura empresarial (culture_values) están directamente relacionadas con las puntuaciones totales de las empresas. Las empresas con altas calificaciones culturales tienden a tener mejores puntuaciones generales.

Hipótesis sobre Oportunidades de Carrera:

Hipótesis 3: La percepción de oportunidades de carrera (career_opp) influye significativamente en la satisfacción laboral general. Los empleados que perciben más oportunidades dentro de la empresa tienden a estar más satisfechos.

Hipótesis sobre Compensación y Beneficios:

Hipótesis 4: La opinión sobre compensación y beneficios (*comp_benefits) afecta directamente a la retención de talento. Empresas con altas calificaciones en este aspecto tienen una menor tasa de rotación de empleados.

Hipótesis sobre la Aprobación del CEO:

Hipótesis 5: La aprobación del CEO (ceo_approv) tiene un impacto en la percepción general de la empresa. Las empresas con altas calificaciones en la aprobación del CEO tienden a tener mejores puntuaciones generales.

Hipótesis sobre Comentarios Positivos y Negativos:

Hipótesis 6: La presencia de comentarios positivos (pros) y negativos (cons) en las reseñas afecta la puntuación general. Empresas con un equilibrio positivo tienden a tener mejores calificaciones.

Hipótesis sobre Diversidad e Inclusión:

Hipótesis 7: La percepción de diversidad e inclusión (diversity_inclusion) está relacionada con la satisfacción laboral. Empresas que son percibidas como inclusivas tienden a tener empleados más satisfechos.

Hipótesis sobre Balance Entre Trabajo y Vida:

Hipótesis 8: La opinión sobre el equilibrio entre trabajo y vida (work_life_balance) está asociada con la satisfacción laboral general. Un mejor equilibrio se correlaciona con mayores puntuaciones.

Hipótesis sobre la Gerencia :

Hipótesis 9: La opinión sobre el equipo de gerencia media y alta (senior_mgmt) está asociada con la satisfacción laboral general. Una mejor opinión se correlaciona con mayores puntuaciones.

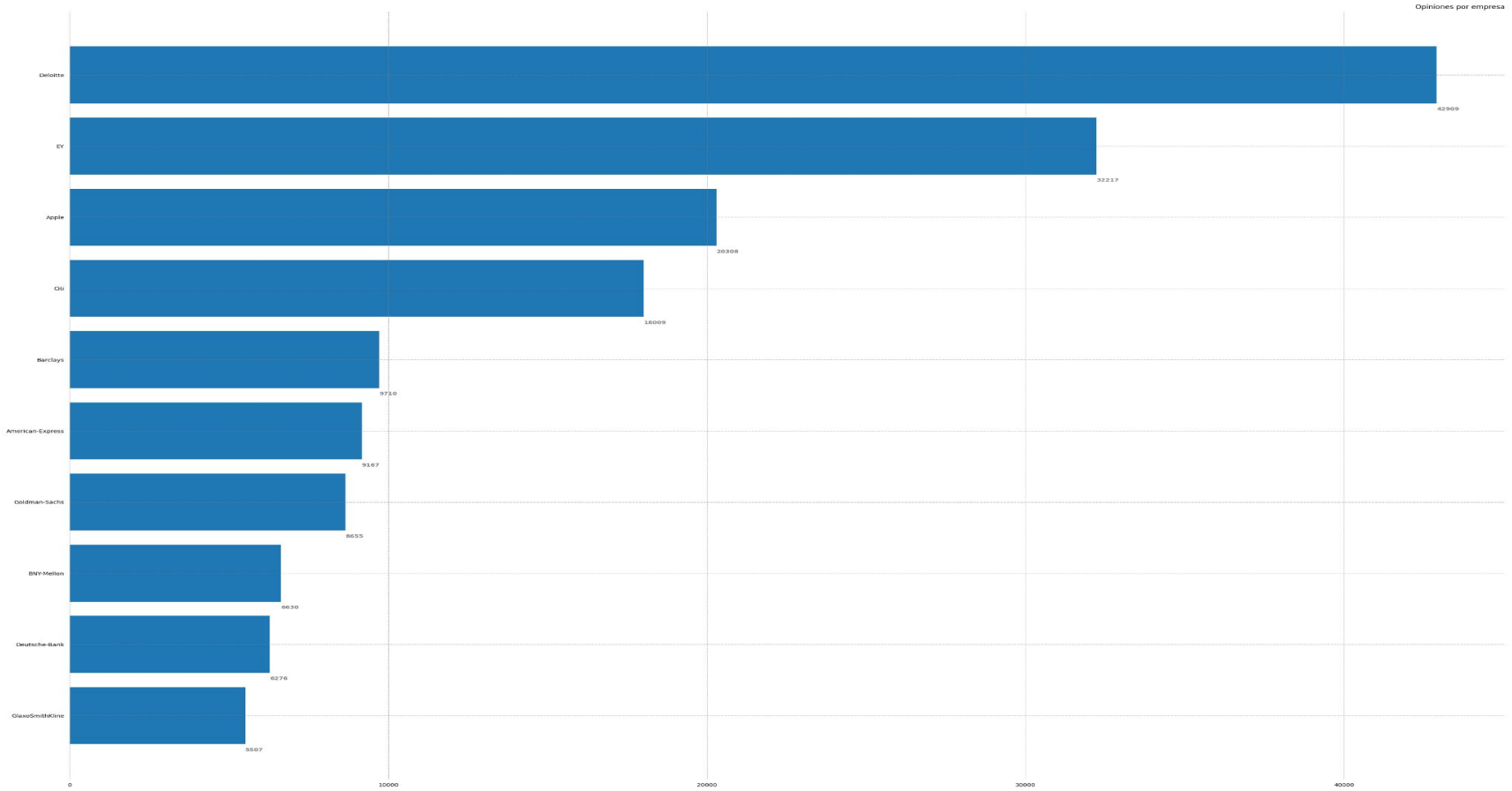
✖ Hipotesis 1: Localización y Roles

Algunas consideraciones:

Distribución de la muestra por:

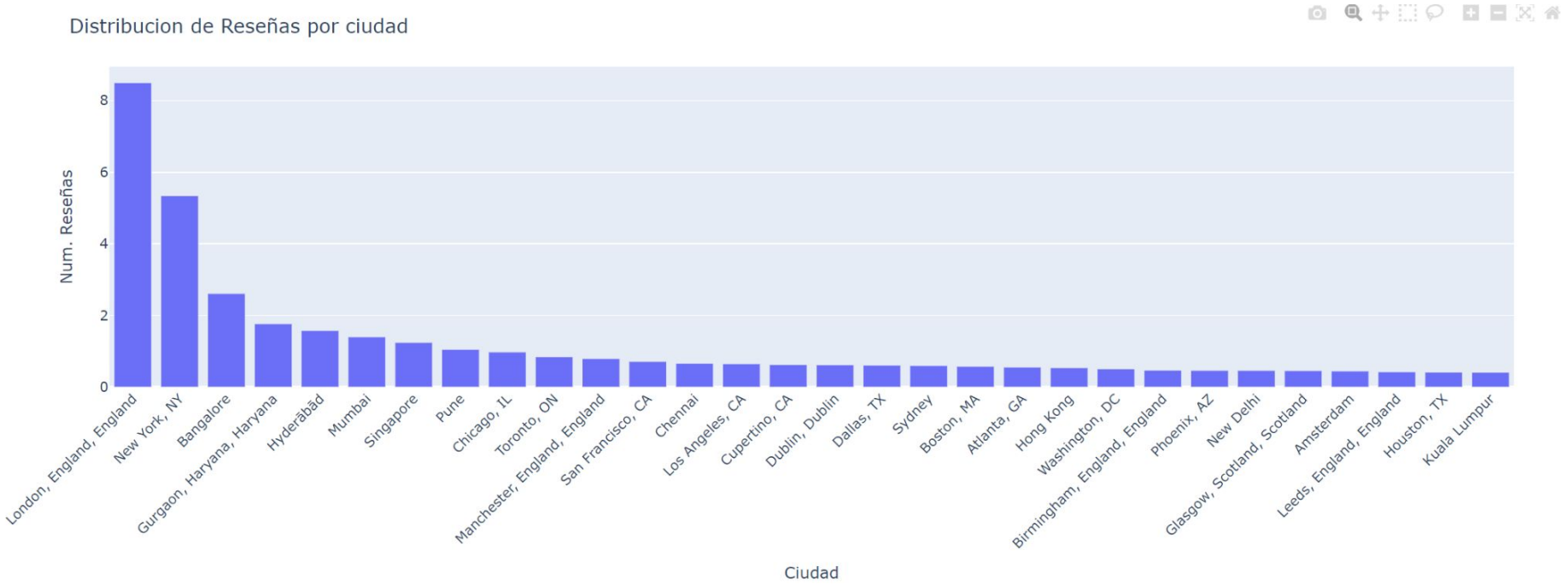
- Localización geográfica de las reseñas.
- Empresa
- Industria

Empresas con el mayor número de reseñas



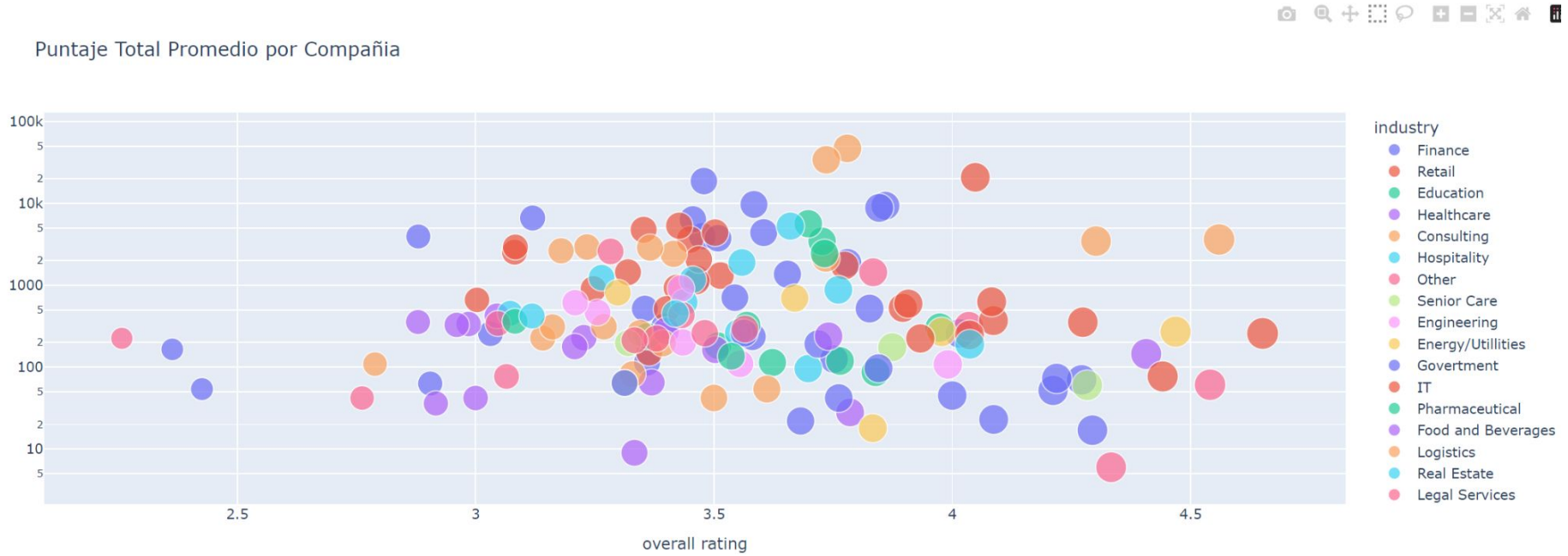
Podemos observar que hay una representación muy grande en la muestra de pocas empresas, siendo estas Deloitte (16.4%), Ernst & Young (12%), Apple (7%), Citi Bank (6,5%) y la entidad financiera Barclay's (3.39%). Con casi el 40% de los reviews.

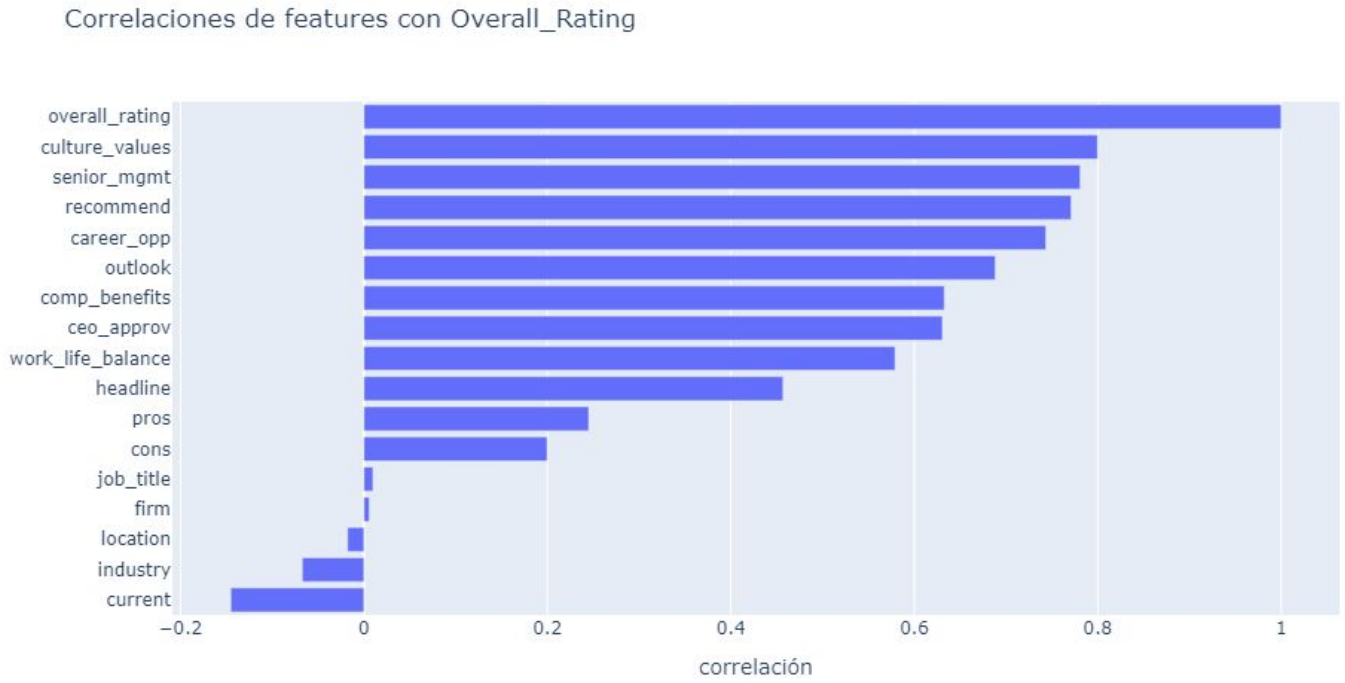
Distribución en ciudades con más de mil personas



En el gráfico están representadas todas las ciudades con más de 1000 personas con mucha representación de ciudades estadounidenses. Pero este total representa apenas el ***36.2%*** de todas las reseñas por lo que la muestra representa una diversidad de ciudades de todo el mundo. De una muestra de 277206 (*eliminando duplicados y nulos*) personas 14 % por ciento son de Londres y New York, siendo Londres la ciudad con el mayor número de reseñas (8,5%)

Distribución de promedio de reseñas de las compañía por tipo de industria





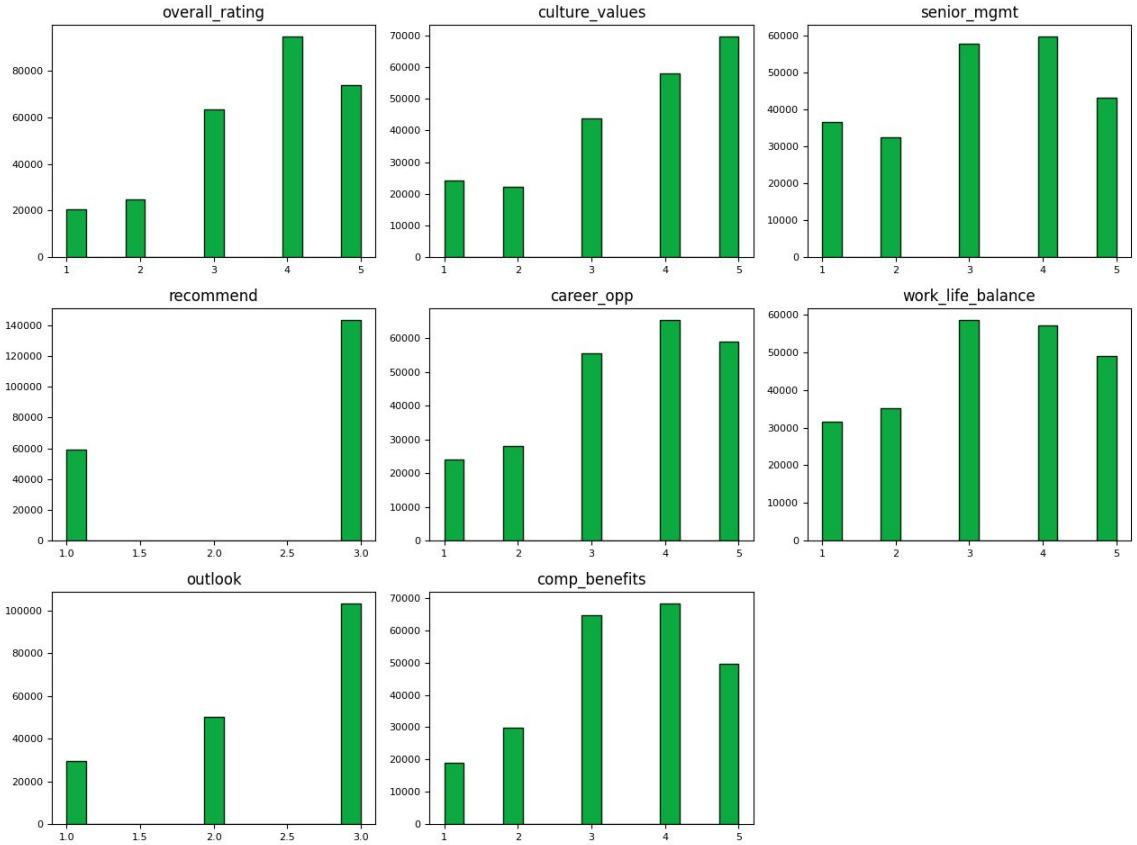
Sólo hay 6 ó 7 features con una correlación lo suficientemente fuerte para ser utilizadas en nuestro modelo. Más adelante podríamos evaluar relaciones no lineales.

✖ Hipotesis 1: Localización y Roles
✓ Hipótesis 2: sobre Cultura Empresarial
✓ Hipótesis 3: sobre Oportunidades de Carrera
✓ Hipótesis 4: sobre Compensación y Beneficios
✓ Hipótesis 5: sobre la Aprobación del CEO
✖ Hipótesis 6: sobre Comentarios Positivos y Negativos
✖ Hipótesis 7: sobre Diversidad e Inclusión
✖ Hipótesis 8: sobre Balance Entre Trabajo y Vida
✓ Hipótesis 9: sobre la Gerencia
✖ Hipótesis 10: sobre la Industria

Las features con mayor correlación con el promedio general (overall_rating) son, por orden de correlación:

1. Cultura Empresarial (culture_values) : 0.79
2. Opinión de Gerencia (senior_mgmt): 0.78
3. Recomendarias esta empresa (recommend_num): 0.77
4. Oportunidades de carrera (career_opp): 0.74
5. Futuro de la Empresa (outlook): 0.688
6. Compensación y Beneficios (comp_benefits): 0.63
7. Balance Entre Trabajo y Vida (work_life_balance): 0.579

Distribución de reseñas en las variables claves



Observaciones:

- Tanto las variables como Recommend (recomendarías a esta empresa) y Outlook (futuro de la empresa) tienen son variables que parecen tener un estándar muy alto con ninguna valoración en cantidades significativas en 4 ó 5.
- La única variable que tiene mayor cantidades de valoraciones altas que el resto es valores culturales. Todas las demás variables su valoración más alta es 4.
- En todos los casos los valores bajos (1,2) suelen ser más bajos que el resto de los valores. En la mayoría de los casos son aproximadamente la mitad del resto de los valores. Con excepción de Opinión de Gerencia y Balance Entre Trabajo y Vida que tienen una presencia importante de 1 y 2 en sus reseñas.

Empresas
10 Top y 10 Bottom

Las diez mejores empresas (sin ponderar) son:

firm	industry	overall_rating	
		mean	count
CarShop-UK	Retail	4.651163	258
Bain-and-Company	Consulting	4.559757	3623
Gateley	Legal Services	4.525424	59
Anglian-Water	Energy/Utillities	4.468635	271
CityFibre	IT	4.441558	77
Abcam	Healthcare	4.406897	145
Achieving-for-Children	Other	4.333333	6
Boston-Consulting-Group	Consulting	4.301789	3466
Engineering-and-Physical-Sciences-Research-Cou...	Govertment	4.294118	17
ActionCOACH	Senior Care	4.283333	60

Las diez mejores empresas (al ponderar) son :

firm	industry	mean	count	mean_weighted
Deloitte	Consulting	3.779091	42909	0.584969
EY	Consulting	3.733309	32217	0.433887
Apple	IT	4.048257	20308	0.296574
Citi	Finance	3.479705	18009	0.226063
American-Express	Finance	3.862223	9167	0.127721
Barclays	Finance	3.584037	9710	0.125542
Goldman-Sachs	Finance	3.847487	8655	0.120127
Deutsche-Bank	Finance	3.456182	6276	0.078249
BNY-Mellon	Finance	3.119608	6630	0.074612
GlaxoSmithKline	Pharmaceutical	3.702560	5507	0.073555

Observaciones:

- El promedio de las empresas ponderadas es de 3.7 en cambio si no se toma la cantidad de reseñas las empresas top tienen promedios alrededor de 4.43.
- En las empresas no ponderadas hay una presencia importante de instituciones financieras, en cambio las empresas sin ponderar son más diversas, aunque vale acotar que quienes tienen más de mil reseñas están en el área de consultoría.
- Apple es la única empresa que mantiene un promedio de 4 + a pesar de tener una gran cantidad de reseñas.

Las diez peores empresas (sin ponderar) son:

firm	industry	mean	count
Creative-Support	Other	2.257778	225
Diligenta	Finance	2.363636	165
Curtis-Banks	Finance	2.425926	54
ENABLE-Scotland	Other	2.761905	42
Angard-Staffing	Consulting	2.788991	109
Barchester-Healthcare	Healthcare	2.879213	356
Capita	Finance	2.879848	3945
AFH-Wealth-Management	Finance	2.904762	63
Colosseum-Dental	Healthcare	2.916667	36
Four-Seasons-Health-Care	Healthcare	2.960366	328

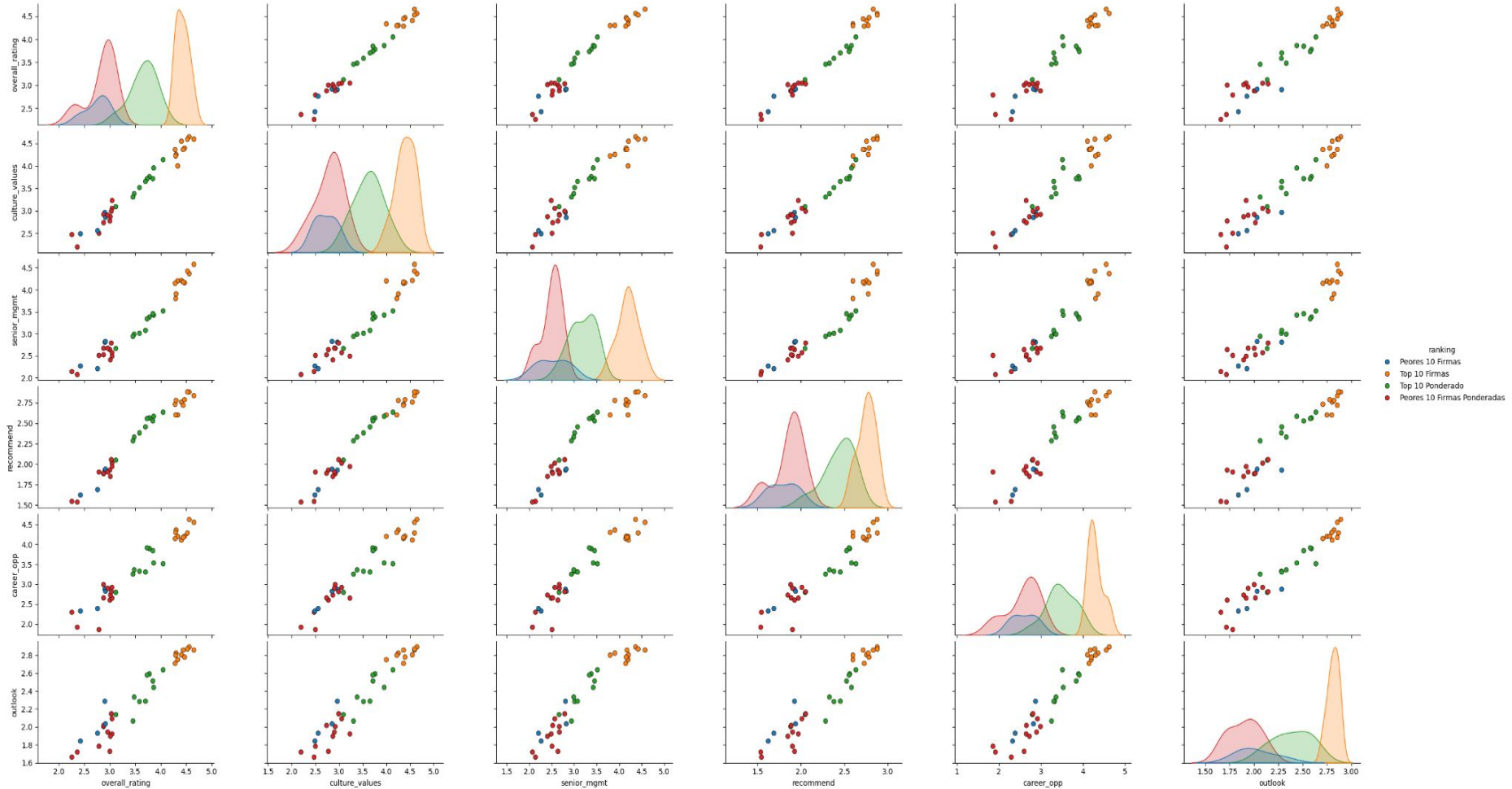
Las diez peores empresas (al ponderar) son :

		overall_rating	
firm	industry	mean	count
Angard-Staffing	Consulting	2.788991	109
Arcadia	Retail	3.003017	663
Babylon-Health	Healthcare	3.044811	424
Barchester-Healthcare	Healthcare	2.879213	356
Barnardo-s	Other	3.046921	341
Boots-Opticians	Healthcare	3.012945	309
Capita	Finance	2.879848	3945
Creative-Support	Other	2.257778	225
Diligenta	Finance	2.363636	165
Equiniti	Finance	3.031250	256
Four-Seasons-Health-Care	Healthcare	2.960366	328

Observaciones:

- El promedio de las empresas considerando el # de reseñas es de 2.83 en cambio si no se toma la cantidad de reseñas las peores empresas tienen promedios alrededor de 2.7.
- Los sectores con más presencia son los de salud y finanzas.
- La peor compañia, considerando la cantidad de reseñas, es *Capita*. Un fondo de inversión con un puntaje promedio de 2.87 y 3945 reseñas.

Atributos de Compañias por su Overall Ranking



Observaciones I

- Es de esperarse que al tomar en cuenta la cantidad de reseñas, la distribución mejores empresas se normalizará. Las empresas peores siguen una distribución con un sesgo negativo importante, pero la diferencia entre aquellas que se consideró la cantidad de empresas y las peores nominalmente, es notable.
- Tanto los mejores y peores empresas no ponderadas tienen una distribución con sesgos. Sin embargo las top empresas no ponderadas son más alargadas y con puntuaciones claramente entre 4 y 5 (leptocúrticas).Mientras las peores empresas son chatas (Mesocúrticas).

Observaciones II

- Por lo general la gran mayoría de las empresas mientras mejor salen en una categoría muestran una correlación positiva entre las variables.
- La agrupación de empresas por el overall_rating y la cantidad de reseñas parece ser un buen predictor de en qué grupo estará la empresa. Con muy pocas excepciones hay empresas en el top ten ponderado por debajo de las empresas peores, en regresiones que miden la relación entre las features. Sería interesante ver si un algoritmo de clasificación tomaría los mismos grupos para compararlos.
- Todas las features mantienen una correlación positiva, es decir que hay una colinealidad importante y habría quizás que reducir el número de features

Conclusiones:

1. El location y la industria no parece afectar mucho el promedio de evaluación de una empresa,14% de las reseñas son de Londres y Nueva York, el resto son ciudades de todo el mundo.
2. Las técnicas de sentiment analysis para medir la relación de los comentarios escritos con la variable target no fueron muy eficientes en este caso. Habría que intentar otro modelo o enfoque.
3. Las relaciones importantes con la variable target fueron todas de carácter positivo.
4. Las variables más importantes para el modelo fueron los valores culturales de la empresa, la apreciación sobre la gerencia, la acción de recomendar o no la empresa para trabajar, las oportunidades de crecimiento y el futuro de la empresa.
5. En menor medida los beneficios y compensación de la empresa, la opinión del ceo y el balance entre la vida y el trabajo también tuvieron una correlación significativa con la valorización.
6. Es posible que una reducción de estos factores sea de utilidad porque quizas hay algunas áreas redundantes que ayudarían a que el modelo sea más simple.