Proyecto de ETL - 1ra Parte

Kevin Santiago Artunduaga Vivas

Universidad Autónoma de Occidente

Cali, Colombia

Primeras aproximaciones

El conjunto de datos es un resultado de un web scraping de trabajos relacionados con los datos por medio de la plataforma de trabajos llamada “Glassdoor” este proceso se realizó en el año 2019 recuperando un aproximado de 165.000 ofertas de trabajo, incluyendo todos los países disponibles en los que se encontraban estas ofertas. Se recolectaron ofertas en 171 países en los cuales de acuerdo a los trabajos relacionado estaban; científico de datos, ingeniero de software, analista de datos, investigador científico, analista de negocios, gerente de producto, gerente de proyecto, ingeniero de datos, estadístico, administrador de base de datos, ingeniero de base de datos, ingeniero de aprendizaje automático.

Este dataset logró llamar mi atención debido a que conocer el panorama de cómo está posicionada nuestra carrera en el mundo laboral es algo importante al querer lanzarse a tomar esta como una profesión, ahora bien evaluar otros aspectos como salarios y condiciones de trabajo son características importantes para poder definir que tan bien está la carrera. Hay más puntos importantes en los cuales fijarse pero estos dos son los principales. Además del hecho de que tomar varios países como referencia es algo característico para tener en cuenta.

Dando a conocer estadísticas importantes para resumir la oferta laboral de los trabajos relacionados a los datos, desarrollando un seguimiento a la profesión y aún más sabiendo que este proceso fue realizado en un año reciente se infiere que la oferta y demanda de ello va creciendo a medida que se va dando a conocer esta carrera.

Por otra lado, al hacer un análisis al dataset como se trata de un proceso de web scraping hay fallas que se dan como resultado valores faltantes y nulos, que pueden ser redundantes en esta investigación, pero dando a su vez un trabajo de transformación de los datos necesario para su debido proceso de extracción, transformación y carga de ellos.

Acerca del dataset

Volumen de datos: Este dataset dispone de más de 165 mil filas y 165 columnas respecto a la tabla principal (glassdoor.csv).

Complejidad: Tiene un total de 15 tablas y distintos tipos de formatos de datos, volviéndolo en un proceso complejo y extenso para poder ver si toda la información que incluye es relevante o no.

Calidad: Desde una vista rápida a la base de datos se nota que hay errores en columnas con valores nulos, además de campos con información faltante, sin embargo al tratarse de un dataset bastante largo se pueden obviar estos inconvenientes y hacer uso de las filas y columnas que contienen la información correspondiente.

Rendimiento: Debido al tamaño del dataset se cargaran las filas completas, que a pesar de que sea una gran cantidad a la hora de manejar las transformaciones necesarias se garanticen consultas completas.

Información relevante: Al utilizar este conjunto de datos, se puede investigar cómo las tendencias económicas, la ubicación geográfica y otros factores influyen en la dinámica del mercado laboral. También explorar patrones de contratación en diferentes industrias y comparar las tasas de empleo en diferentes grupos demográficos.

Objetivo del proyecto

Se espera crear un modelo para predecir las habilidades más demandadas en el mercado laboral. Esto puede ayudar a los profesionales, estudiantes y la industria en general a tomar decisiones informadas sobre qué habilidades desarrollar y mejorar.

El objetivo principal del proyecto es analizar las ofertas laborales en el campo de los datos y con el propósito de identificar patrones y tendencias que revelen las habilidades más demandadas en el mercado laboral actual. Esta investigación puede ayudar a las personas interesadas en este ámbito o que quieran meterse a este, ayudando a determinar si es la predicción deseada como una decisión laboral.

Migración de datos

Para hacer la migración de los datos del csv ‘glassdor.csv’ hacia la base de datos PostgreSQL se hizo uso de python en conjunto de la librería psycopg2, en donde primero se realizó en el script la conexión a la base de datos llamada ETL, en conjunto de el host (use el local), el usuario y la contraseña, estos últimos dos puntos fueron guardados en un db_config.json para mantener estos datos ocultos. Luego de la conexión se creó un cursor para poder recorrer y manipular el conjunto de filas de la database. Para posterior a ello hacer la creación de la DB llamada “datajobs_glassdoor”.

Ahora bien antes de hacer el procesado de los datos tuve que guardar un csv con ayuda de pandas a partir del dataset entregado de kaggle, porque por alguna extraña razón entraba en conflicto al intentar cargar los datos a postgres. Hay que

tener en cuenta de que a pesar de que tuviera claro qué columnas eliminar a partir del EDA, procese todas las columnas completas para de esta manera hacer la comparación de como se ha ido evolucionando en la optimización y uso del dataset.

Columnas eliminadas

El dataset cuenta con demasiadas columnas que disponen de bastantes valores nulos, además del hecho de que hay bastantes que de acuerdo a el objetivo principal que se busca no son necesarias, por lo tanto determinó que columnas eliminar a partir de estos factores:

- ***breadcrumbs:*** Migas de pan, son una herramienta de navegación para las páginas web, por ende no es relevante.
- ***gaTrackerData_empSize, gaTrackerData_expired, gaTrackerData_industryId, gaTrackerData_jobId_long, gaTrackerData_jobId_int, gaTrackerData_locationId, gaTrackerData_locationType, gaTrackerData_pageRequestGuid_guid, gaTrackerData_pageRequestGuid_guidValid, gaTrackerData_pageRequestGuid_part1, gaTrackerData_pageRequestGuid_part2, gaTrackerData_sectorId, gaTrackerData_profileConversionTrackingParams_trackingCAT, gaTrackerData_profileConversionTrackingParams_trackingSRC, gaTrackerData_profileConversionTrackingParams_trackingXSP, gaTrackerData_jobViewTrackingResult_jobViewDisplayTimeMillis, gaTrackerData_jobViewTrackingResult_requiresTracking, gaTrackerData_jobViewTrackingResult_trackingUrl:*** Estas columnas parecen ser datos relacionados con el seguimiento y el análisis del sitio web en el que se publican las ofertas laborales. Aunque pueden ser útiles para rastrear el comportamiento de los usuarios en el sitio, no parecen estar directamente relacionadas con la demanda de habilidades laborales.
- ***header_adOrderId, header_advertiserType, header_applicationId, header_applyButtonDisabled, header_applyUrl, header_blur, header_coverPhoto, header_easyApply, header_employerId, header_expired, header_gocId, header_hideCEOInfo, header_locId, header_locationType, header_logo, header_logo2x, header_organic, header_overviewUrl, header_rating, header_saved, header_savedJobId, header_sgocId, header_sponsored, header_userAdmin, header_uxApplyType, header_featuredVideo, header_urgencyLabel, header_urgencyLabelForMessage, header_urgencyMessage, header_needsCommission, header_payHigh, header_payLow, header_payMed, header_payPeriod, header_salaryHigh, header_salaryLow, header_salarySource:*** Al igual que las columnas anteriores, estas parecen ser datos internos relacionados con la publicación y

administración de las ofertas laborales en el sitio web, hay algunas columnas que sí conserve referentes a la fecha que la oferta fue publicado, además del título de estas, pero gran parte de estas columnas son referentes a la navegación.

- ***job_eolHashCode, job_importConfigId, job_jobReqId_long, job_jobReqId_int:*** Estas columnas no disponían de una descripción, y al hacer la observación de los valores que habían en ellas eran número exagerados que no daban ninguna descripción sobre lo que podrían ser, además de los id que no cuentan ni siquiera con una tabla a la cual conectarse.

- ***job_jobTitleId, job_listingId_long, job_listingId_int:*** Al igual que los otros id a pesar de que tuvieran descripción son ID que no cuentan con la relación de las otras tablas disponibles del dataset.

- ***map_address, map_postalCode:*** A pesar de que logren ser datos que tienen la posibilidad de servir, al hacer el análisis de los nulos ambas columnas cuentas con una cantidad de nulos de 151.108 y 151.317 respectivamente, un número bastante grande dejando aproximadamente un 91% de nulos.

- ***overview_allBenefitsLink, overview_allPhotosLink, overview_allReviewsLink, overview_allSalariesLink, overview_allVideosLink:*** Estos son variables referentes a los URL de fotos, videos, salarios elementos que no se van a analizar, ya que no se va a hacer un análisis respecto a las direcciones web, y no hablan más a partir de links.

- ***overview_industryId, overview_sectorId:*** Estos ID no disponen de una tabla a cuál corresponden, además con la vista de las gráficas de kaggle ambos tienen 47.000 filas con un valor de 0, es decir no tienen un ID al cual conectarse esta cantidad de filas.

- ***overview_stock:*** Esta variable al hacer el análisis de los nulos cuenta con un número de 118.933 una cantidad muy grande por ende se decidió eliminarla.

- ***overview_mission:*** A pesar de que la visión de las empresas sería algo chevere de poder ver dispone de 113.134 nulos.

- ***overview_competitors:*** Esta columna tiene 117.302 nulos, desperdiciando bastantes datos de los cuales se supone que es un ID que se conecta con otra tabla, pero a su vez dejando de lado demasiadas filas.

- ***overview_companyVideo:*** Son videos respecto a la compañía, sin embargo estos no son relevantes para el objetivo además dispone de 138.870 valores nulos.

- ➔ ***photos, rating_ceo_photo, rating_ceo_photo2x:*** Photos es un ID que conecta a una tabla de fotos, y los otros son URL de imágenes, elementos que no son relevantes para el análisis de la demanda laboral, además cuentan con un 52% de valores nulos.

- ➔ ***rating_ceo_ratingsCount, rating_ceoApproval, rating_recommendToFriend:*** Estas clasificaciones tienen una escala bastante rara debido a que hay hasta valores negativos, sin brindar tampoco una explicación a estos ratings.

- ➔ ***salary_country_cc3LetterISO, salary_country_ccISO, salary_country_continent_continentCode, salary_country_continent_continentName, salary_country_continent_id, salary_country_continent_new, salary_country_countryFIPS, salary_country_currency_currencyCode, salary_country_currency_defaultFractionDigits, salary_country_currency_displayName, salary_country_currency_id, salary_country_currency_name, salary_country_currency_negativeTemplate, salary_country_currency_new, salary_country_currency_positiveTemplate, salary_country_currency_symbol, salary_country_currencyCode, salary_country_defaultLocale, salary_country_defaultShortName, salary_country_employerSolutionsCountry, salary_country_id, salary_country_longName, salary_country_major, salary_country_name, salary_country_new, salary_country_population, 'salary_country_shortName, salary_country_tld, salary_country_type, salary_country_uniqueName, salary_country_usaCentricDisplayName, salary_currency_currencyCode, salary_currency_defaultFractionDigits, salary_currency_displayName, salary_currency_id, salary_currency_name, salary_currency_negativeTemplate, salary_currency_new, salary_currency_positiveTemplate, salary_currency_symbol, salary_lastSalaryDate:*** Todas estas columnas que mencione cuentan con una cantidad igual de 88.513 perteneciendo a el 54% de nulos un numero bastante grande descartando demasiada filas, de los cuales en cada una de las columnas estan solo especificadas los pesos especificados de Europa.

- ➔ ***wwfu:*** Esta columna es algo rara debido a que se supone que es un ID que está conectado a otra tabla del dataset pero esta solo tiene información referente a medias de imágenes, sin explicar valores importantes además contando con un número de nulos de 146.549.

Columnas seleccionadas

- ***benefits.benefitRatingDecimal, benefits.comments, benefits.highlights, benefits.numRatings, benefits.employerSummary:*** Estas columnas son referentes a los beneficios tanto de comentarios como de calificaciones, a su vez está el id que es “benefits.highlights” el cual se conecta con la tabla de ‘glassdoor_benefits_highlights.csv’.

- ***gaTrackerData.empName, gaTrackerData.industry, gaTrackerData.jobTitle, gaTrackerData.location, gaTrackerData.sector:*** Está el nombre de la empresa, la industria, titulo del trabajo, localización y el sector de la compañía.

- ***header.employerName, header.jobTitle, header.location, header.posted, header.normalizedJobTitle:*** Título del trabajo el cual es parecido a “gaTrackerData.jobTitle”, también está el nombre de la empresa, la ubicación, la fecha en la que se publicó la oferta laboral y una pequeña normalización del título del trabajo.

- ***job.description, job.discoverDate, job.jobSource:*** Descripcion del trabajo, fecha en la que se descubrio y la fuente de trabajo.

- ***map.country, map.employerName, map.lat, map.lng, map.location:*** País en donde está la oferta laboral, nombre de la compañía, coordenadas de latitud y longitud, también la localización usualmente la ciudad.

- ***overview.foundedYear, overview.hq, overview.industry, overview.revenue, overview.sector, overview.size, overview.type, overview.description, overview.website:*** Año en el que se fundó la compañía, sede de la empresa, industria de la compañía, ingresos de la empresa, sector de la empresa, cantidad de empleados de la organización, tipo de empresa, descripción de la empresa y el website de la compañía.

- ***rating.ceo.name, rating.starRating:*** Nombre del ceo (director ejecutivo) y calificacion de estrellas sobre la oferta laboral.

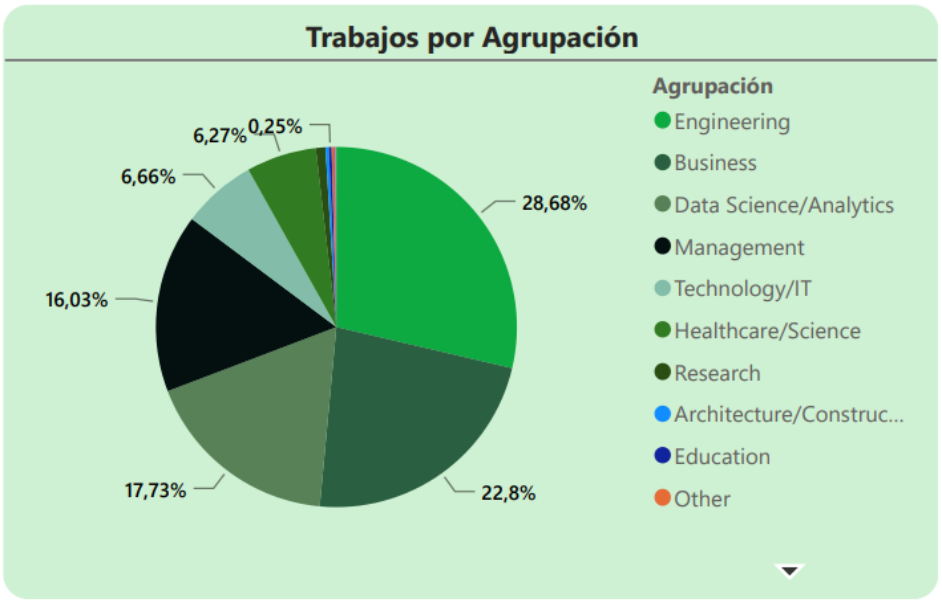
- ***reviews:*** Cantidad de opiniones.

- ***salary.country.currency.name, salary.country.defaultName, salary.salaries:*** Nombre de la moneda del país, nombre del peso predeterminado y el ID que se vincula a la tabla ‘glassdoor_salary_salaries.csv’

Gráficos

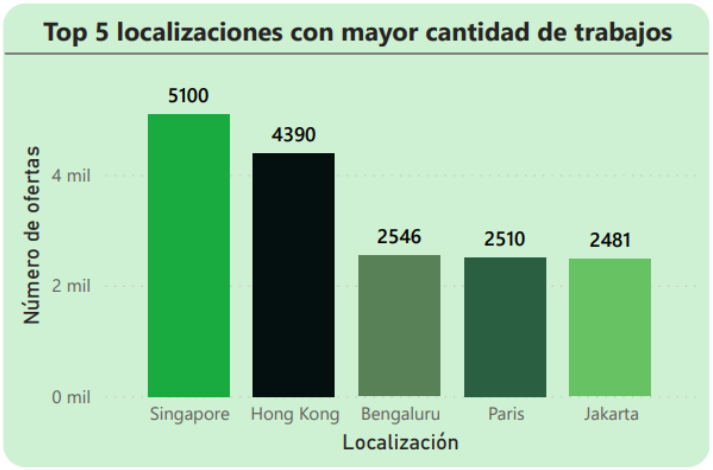
Para los gráficos me base en 3 por ahora, debido a que aun necesito hacer transformaciones a algunas de las columnas, así como la conexión entre las dimensiones que dispongo.

1) El primero es un gráfico circular que es ofertas de empleo por agrupación, para las agrupaciones realice un proceso de normalización de datos bastante extenso de los cuales saqué 19 categorías.



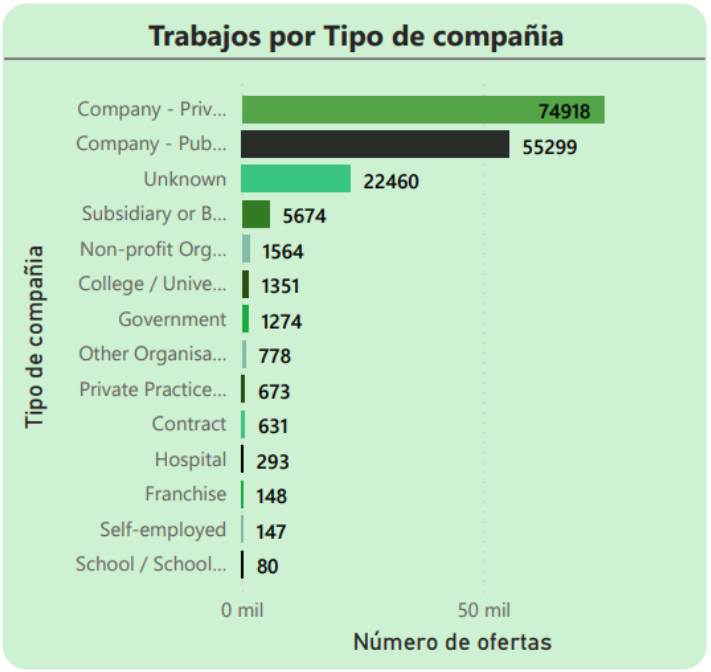
Como se puede observar gran porcentaje de las categorías se encuentra en ingeniería, negocios y data science / analytics; estas serían las agrupaciones que más predominan en los trabajos.

2) El segundo gráfico es uno de barras verticales que es un top 5 de localizaciones con más cantidad de ofertas laborales



De primero esta singapore con 5100, de segundo está Hong Kong con 4390, luego Bengaluru(ciudad de india) con 2546, de cuarto está París con 2510 y de quinto esta Jakarta (capital de indonesia) con 2481 ofertas.

3) Esta gráfica es de barras horizontales la cual es la cantidad de ofertas laborales por tipo de compañía, de los cuales hay 14 categorías.



Predominan las compañías privadas con 74.918 mientras que de últimas están las que se enfocan en las escuelas con nada más que 80 ofertas laborales.