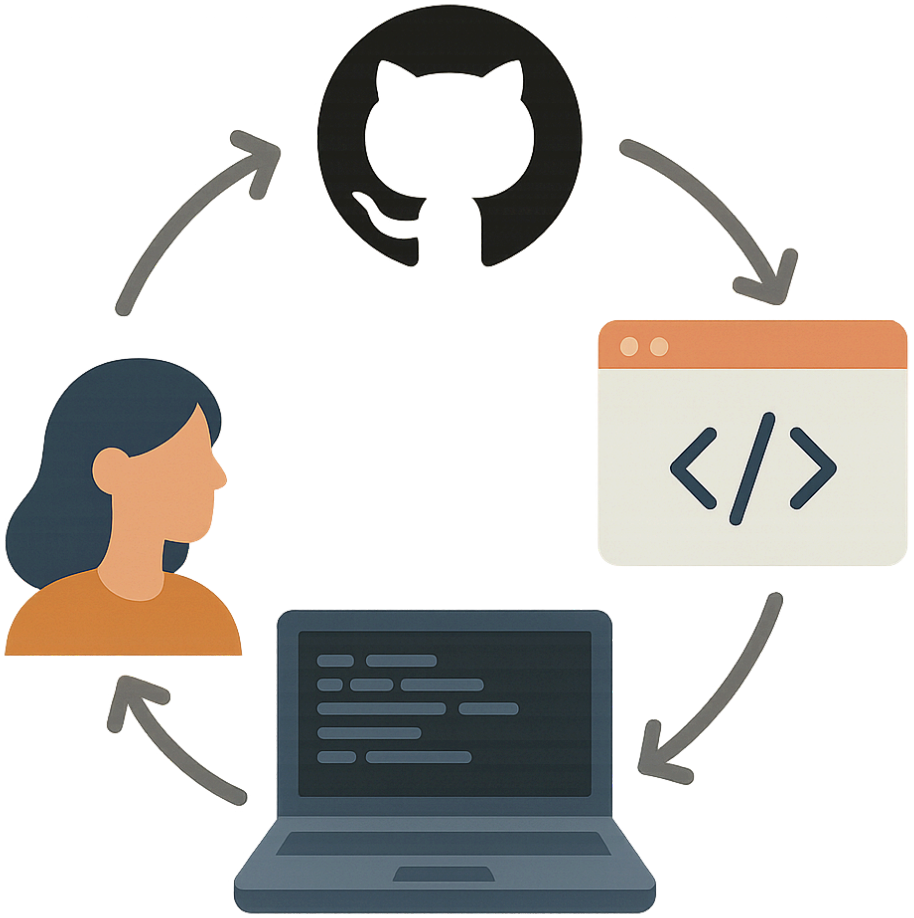
memoria de análisis

ecosistema de desarrolladores en Github



Mikel Guillén

abril 2025

1. BUSINESS CASE & DATA COLLECTION

El objetivo del EDA es conocer la situación del ecosistema de desarrolladores en el mundo. A la hora de adentrarse en una nueva materia para muchos de nosotros como es la ciencia de datos, es importante obtener una perspectiva global de la materia.

La obtención de datos la vamos a realizar desde Git Hub. GitHub es una plataforma en línea que se utiliza para el desarrollo de software y la gestión de proyectos. Es especialmente conocida por su integración con **Git**, un sistema de control de versiones que permite a los desarrolladores llevar un registro de los cambios en el código fuente a lo largo del tiempo. GitHub facilita la colaboración entre programadores al permitirles trabajar en proyectos de forma conjunta, compartir código y gestionar versiones del mismo

Según una encuesta realizada por JetBrains en 2021, GitHub fue utilizado por el 91% de los 32,000 desarrolladores encuestados, superando a competidores como GitLab (48%) y Bitbucket (30%) . Por lo tanto, la información de GitHub debería ser una muestra significativa del ecosistema global.

2. DATA UNDERSTANDING

La información obtenida en Git Hub contiene (1) información pública (2) que se encuentre en GitHub (3) por economía (4) trimestralmente. La información disponible es desde Q1-2020 hasta Q3-2024. Por lo tanto esta información no nos valdrá para entender:

- Actividad privada
- Actividad fuera de GitHub
- Información geográfica más detallada que por economía
- Información temporal más detallada que trimestralmente y anterior a Q1-2020

Para más información en referencia a la información obtenida:

https://github.com/github/innovationgraph/blob/main/README.md
https://github.com/github/innovationgraph/blob/main/docs/datasheet.md

Los datasheet los vamos a conseguir directamente desde github en la fuente oficial:

https://github.com/github/innovationgraph/tree/main/data

En dicho enlace tenemos diferentes archivos en .cvs, para este EDA nos vamos a centrar en los siguientes datasets:

- **Repositorios:** el número de proyectos de software en una economía determinada, basado en la ubicación principal de todos los miembros del

repositorio con acceso de triage o superior. Consulta la documentación sobre Repositorios para más información. Ten en cuenta que este conteo incluye repositorios que pueden ya no ser activamente desarrollados o mantenidos.

- **Desarrolladores**: el número de cuentas de desarrollador ubicadas en una economía determinada, basado en la ubicación diaria principal. Este conteo excluye a los usuarios que son bots o que de otra manera están marcados como "spam" dentro de los sistemas internos. Consulta la documentación sobre cuentas personales para más información. Ten en cuenta que este conteo incluye cuentas de desarrollador que pueden ya no estar activas.
- **Lenguajes de programación**: el número de desarrolladores únicos en cada economía que realizaron al menos un *git push* a un repositorio con un determinado lenguaje de programación. Consulta la documentación sobre lenguajes de repositorios para más información sobre cómo detectamos los lenguajes de programación.
- **Temas**: el número de desarrolladores únicos que realizaron al menos un *git push* a un repositorio con un determinado tema. Consulta la documentación sobre Temas para más información sobre cómo los desarrolladores asignan temas a los repositorios.

Los documentos cuentan con la siguiente estructura y contaremos con las siguientes variables.

Schema for `repositories.csv`

repositories	iso2_code	year	quarter
integer	string	integer	integer

Schema for `developers.csv`

developers	iso2_code	year	quarter
integer	string	integer	integer

Schema for `languages.csv`

num_pushers	language	language_type	iso2_code	year	quarter
integer	string	string	string	integer	integer

Schema for `topics.csv`

num_pushers	topic	iso2_code	year	quarter
integer	string	string	integer	integer

3. DATA CLEANING

Al contar con información de la fuente oficial de Git Hub, no se ha requerido de mucha limpieza. Los archivos con los que hemos trabajado no contaban con valores nulos.

De todos modos, se han realizado las siguientes acciones para poder trabajar con la información de manera correcta:

- El código ISO2 (también conocido como código de país de dos letras) es una abreviatura de dos caracteres que representa a un país o territorio en particular según los estándares establecidos por la Organización Internacional de Normalización (ISO). En los dataset hemos observado que había una agrupación por países de la Unión Europea (EU), como un dato adicional. Por lo que se ha eliminado para no duplicar datos, ya que estos también estaban representados mediante los países independientes que lo componen.
- En el dataset de Temas (topics), el tema Hacktoberfest está escrito de diferentes maneras (Hactober, Hackoberfest2020, Hacktoberfest2021, Hackberfest…) por lo que se ha procedido a agrupar todos ellos con el nombre de Hactoberfest. Ya que el año ya lo tenemos representado en la columna “year”.
- En dataset de Temas (topics) se han agrupado lenguajes que tenían diferente nombre como “python” y “python3” o “react” y “reactjs”.
- En el dataset de Temas (topics), se ha procedido también a eliminar los temas relacionados con los lenguajes de programación, de esta manera pudiendo observar los temas más relevantes al margen de los lenguajes.

4. ANALYSIS

A la hora de realizar las gráficas se han realizado las siguientes acciones:

- Agrupar y sumar por la variable que queremos representar.
- Se ha decidido de representar los datos por Año-trimestre, ya que al no contar con información de años completos (por ejemplo 2024 solo tenemos hasta Q4), si representamos por año no obtendríamos datos uniformes y comparables. Para ello hemos creado una columna adicional year_quarter.
- Al contar con tanta información, en la mayoría de los casos se ha limitado el estudio a los 15 valores más frecuentes. Para ello se han ordenado los datasheets y se les ha asignado un índice para poder filtrar posteriormente según el número de valores que quisieramos visualizar.

Nota: Para ver los gráficos acceder a la presentación adjunta.

Gráfico desarrolladores en el Mundo

Mediante la función *sns.barplot* de la librería Seaborn (una biblioteca de visualización de datos en Python) se crea un gráfico de barras. El objetivo es mostrar la evolución del número de desarrolladores totales en el mundo

Eje X: En el eje X se muestra el Año-Trimestre.
Eje Y: En el eje Y se muestra el número de desarrolladores en el mundo (en millones).
Cada barra representa el número de desarrolladores en un trimestre concreto.

Como se puede observar el crecimiento del número de desarrolladores en el mundo entre 2020 y 2024 tiene un valor constante de aproximadamente 27% YoY.

Gráfico desarrolladores por país - Ranking del 1 - 10

Mediante la función *sns.lineplot* de la librería Seaborn se crea una gráfica de líneas, en la cual tenemos:
Eje X: Representa la variable independiente, Año-Trimestre.
Eje Y: Representa la variable dependiente, desarrolladores (en millones)
Hue: Mediante este parámetro podemos añadir una tercera dimensión de datos, representada por diferentes líneas de colores. En este caso representaremos los diferentes países.
Es un desglose del número de desarrolladores en el mundo por país.

La mayor dificultad para realizar esta gráfica ha sido el hecho de ordenar la leyenda en orden con los últimos datos de 2024-Q3 y no con los datos de 2020-Q1.
El estudio se limita a que haya un ranking de 10 países por trimestre, si algún país no está en el siguiente trimestre, entrará otro. Es por ello que en la legenda tenemos 11 países, ya que Francia cae después de 2021-Q1

Gráfico repositorios totales en el mundo.

Lo creamos mediante la función *sns.barplot* de la librería Seaborn. El objetivo es mostrar la evolución de los repositorios en el mundo.

Eje X: En el eje X se muestra el Año-Trimestre.
Eje Y: En el eje Y se muestra el número de repositorios en el mundo (en millones).

Como se puede observar el crecimiento del número de repositorios en el mundo entre 2020 y 2024 tiene un valor constante de aproximadamente 25% YoY.

Gráfico repositorios - Por país

Gráfico similar a los desarrolladores por país, pero en este caso estamos contabilizando los repositorios creados.

Gráfico lenguajes de programación

Mediante la función *sns.lineplot* de la librería Seaborn se crea una gráfica de líneas, en la cual tenemos:
Eje X: Representa Año-Trimestre.
Eje Y: Representa Git push por desarrollador (en millones) de un lenguaje de programación específico. El comando git push permite subir los cambios de un repositorio local a un repositorio remoto. Esto permite que otros colaboradores puedan incorporar los cambios a sus propios repositorios.
Hue: Representaremos los diferentes lenguajes de programación.

La leyenda está ordenada de modo que muestra los lenguajes de programación más utilizados en 2024-Q3.

Gráfico Temática

Mediante la función *sns.lineplot* de la librería Seaborn se crea una gráfica de líneas, en la cual tenemos:
Eje X: Representa Año-Trimestre.
Eje Y: Cada punto de datos corresponde un tema basado en el número de desarrolladores únicos que subieron código a un repositorio etiquetado con ese tema durante un trimestre determinado
Hue: Representaremos los diferentes temas.

Gráfico Temática - Sin lenguajes de programación

Igual que la anterior pero quitando los temas sobre los lenguajes de programación.

5. RESULTADOS

Categoría	Tendencias
Desarrolladores y repositorios	Crecimiento de un 27% y 25 % anual respectivamente. Top 3 US, India y China.
Lenguajes dominantes	JavaScript, Python, Shell, TypeScript
Desarrollo Web	Área más fuerte. Herramientas como React, TailwindCSS, Next.js, JavaScript, TypeScript, Node.js, HTML, CSS, Node.js
Ciencia de Datos / IA	Machine Learning, Deep Learning, AI
DevOps & Contenedores	mongodb, Docker, Kubernetes
Colaboración abierta	Hacktoberfest marca picos fuertes cada octubre

Los resultados concluyen que el ecosistema de desarrolladores sigue siendo dominado por el desarrollo web. Sin embargo existe un segundo ámbito como es la Ciencia de Datos e Inteligencia Artificial que está entrando con fuerza.

Esto mismo se puede concluir del análisis de los lenguajes de programación dominantes. Aunque JavaScript siga siendo líder y también contemos en el Top4 con TypeScript, ambos ligados al desarrollo web, observamos la llegada con fuerza de Python (Top2). Y si analizamos los Temas más relevantes de Github tenemos que Python sería la Top1. Esto se debe a que Python es muy utilizado en aprendizaje automático, ciencia de datos, computación científica y en ámbitos como los hobbies o la domótica.

Por otro lado, otra posible razón del incremento de la utilización de Python es debido a su gran uso en proyectos de colaboración abierta, como el Hacktoberst y la academia.

Viendo que los resultados muestran que la Ciencia de Datos se encuentra entre los ámbitos más importantes del ecosistema de desarrolladores, se decide analizar en más profundidad cuáles son las habilidades más demandadas en este ámbito. Para ello se realiza un estudio con un dataset de 12212 ofertas de trabajo de linkedin en 2024.

https://www.kaggle.com/datasets/asaniczka/data-science-job-postings-and-skills/data

Los resultados representados en la Gráfica Habilidades Data Science muestran que las habilidades más demandadas son Python, SQL y comunicación. Y en segundo lugar quedarían análisis de datos, visualización y machine learning.

6. GLOSARIO

TEMAS GITHUB

python: Lenguaje de programación popular por su simplicidad y versatilidad. Usado en web, ciencia de datos, automatización, etc.

javascript: Lenguaje de programación clave para el desarrollo web frontend y backend (usando Node.js).

react: Librería de JavaScript para construir interfaces de usuario interactivas, creada por Facebook.

typescript: Superset de JavaScript que añade tipado estático, facilitando el desarrollo a gran escala.

hacktoberfest: Evento anual que promueve la contribución al software libre durante octubre mediante pull requests.

java: Lenguaje de programación orientado a objetos, ampliamente usado en aplicaciones empresariales, Android y backend.

nodejs: Entorno de ejecución para JavaScript en el servidor, ideal para aplicaciones escalables y en tiempo real.

css: Lenguaje de estilos para definir la apariencia visual de páginas web.

reactjs: Es una biblioteca de JavaScript que se utiliza para crear interfaces de usuario (UI)

html: Lenguaje de marcado estándar para estructurar el contenido en la web.

docker: Plataforma de contenedores que permite empaquetar y desplegar aplicaciones con todas sus dependencias.

python3: Versión moderna de Python, con mejoras respecto a Python 2 en sintaxis y funcionalidades.

php: Lenguaje de programación del lado del servidor comúnmente usado para desarrollar sitios web dinámicos.

machine-learning: Campo de la inteligencia artificial que permite a las máquinas aprender a partir de datos.

nextjs: Framework para React que permite renderizado del lado del servidor (SSR) y generación de sitios estáticos.

tailwindcss: Framework de CSS que permite diseñar directamente en el HTML mediante clases utilitarias.

api: Interfaces que permiten la comunicación entre diferentes aplicaciones o servicios.

ai (Inteligencia Artificial): Campo que busca desarrollar sistemas capaces de realizar tareas que requieren inteligencia humana.

deep-learning: Subcampo del machine learning basado en redes neuronales profundas para tareas complejas como visión o lenguaje.

kubernetes: Sistema de orquestación para automatizar el despliegue, escalado y gestión de contenedores (como los de Docker).

mongodb: Base de datos NoSQL orientada a documentos, ideal para manejar datos no estructurados.

linux: Sistema operativo de código abierto, base de muchos servidores, sistemas embebidos y distribuciones para desarrolladores.

android: Sistema operativo móvil desarrollado por Google, basado en Linux.

website: Repositorios que contienen código o contenido para sitios web.

blog: Repositorios dedicados a blogs personales o técnicos, a menudo usando generadores estáticos como Jekyll o Hugo.