Andrea Valenzuela Ramírez

Stack Overflow Annual Developer Survey

En la práctica final de la asignatura de Visualización de Datos del Máster Universitario en Ciencia de Datos de la UOC, me gustaría trabajar con el *data set Stack Overflow Annual Developer Survey* de este año 2021. *Stack Overflow* se define como "Un sitio de preguntas y respuestas para programadores y profesionales de la informática". Es considerada una de las plataformas más grandes de la comunidad de desarrolladores. Millones de profesionales IT de cualquier área, y sobre todo desarrolladores, pueden fácilmente día tras día, buscar u ofrecer conocimientos y así ayudar, aprender, enseñar, colaborar y compartir con otros alrededor del mundo. Cada año se realiza una encuesta a sus usuarios desarrolladores para conocer las herramientas y tendencias que dan forma cada año al ámbito del desarrollo de software. Tanto a nivel personal como a nivel profesional, me resulta interesante analizar este *data set* para poder observar dichas tendencias y crear una infografía sobre ellas. Considero que presentar bien una información tan relevante podría ayudar a cualquier persona a entender el camino que el desarrollo de software está tomando y poder tomar decisiones personales en consecuencia, por ejemplo, a la hora de elegir qué lenguaje de programación le sería útil aprender próximamente si lo que quiere es que su salario aumente.

Este conjunto de datos se encuentra disponible online[1]. Se trata de datos actuales del año en curso que son importantes para la comunidad del desarrollo de software. En este caso, se ha tenido en cuenta no sólo una perspectiva de género, si no también de orientación sexual, etnia y salud mental. Así mismo, sería interesante analizar hasta qué punto estos campos pueden ser relevantes para dicho estudio.

Se trata de un conjunto de datos con 83.439 registros correspondientes al número de personas que han respondido el cuestionario. Cada registro cuenta con 48 variables de tipos variados (*int64*, *str* y *float64*) que corresponden a los siguientes campos:

- *ResponseId*: Identificador asociado a la respuesta.
- *MainBranch*: Descripción personal respecto al desarrollo de software.
- *Employment*: Situación laboral.
- *Country*: País de residencia.
- *US_State*: Estado de US en el que reside (en el caso de residir en US).
- *UK_Country*: País de UK en el que reside (en el caso de residir en UK).
- *EdLevel*: Nivel educativo.
- *Age1stCode*: Edad a la que escribió la primera línea de código.
- *LearnCode*: Medio por el que aprendió a programar.
- *YearsCode*: Años que lleva programando.
- *YearsCodePro*: Años que lleva programando como parte de su trabajo.

[1] Stack Overflow Insights - Developer Hiring, Marketing, and User Research

- *DevType*: Descripción de su trabajo actual respecto al desarrollo.
- *OrgSize*: Tamaño de la organización para la que trabaja.
- *Currency*: Moneda que usa diariamente.
- *CompTotal*: Compensación total.
- *CompFreq*: La compensación es semanal, mensual o anual.
- *LanguageHaveWorkedWith*: Lenguajes con los que se ha trabajado.
- *LanguageWantToWorkWith*: Lenguajes con los que se quiere trabajar en los próximos años.
- *DatabaseHaveWorkedWith*: Bases de datos con las que se ha trabajado.
- *DatabaseWantToWorkWith*: Bases de datos con las que se quiere trabajar en los próximos años.
- *PlatformHaveWorkedWith*: Plataformas con las que se ha trabajado.
- *PlatformWantToWorkWith*: Plataformas con las que se quiere trabajar en los próximos años.
- *WebframeHaveWorkedWith*: Web frames con las que se ha trabajado.
- *WebframeWantToWorkWith*: Web frames con los que se quiere trabajar en los próximos años.
- *MiscTechHaveWorkedWith*: Otros frameworks o librerías con las que se ha trabajado.
- *MiscTechWantToWorkWith*: Otros frameworks o librerías con las que se quiere trabajar en los próximos años.
- *ToolsTechHaveWorkedWith*: Otras herramientas con las que se ha trabajado.
- *ToolsTechWantToWorkWith*: Otras herramientas con las que se quiere trabajar en los próximos años.
- *NEWCollabToolsHaveWorkedWith*: Entornos de desarrollo con los que se ha trabajado.
- *NEWCollabToolsWantToWorkWith*: Entornos de desarrollo con los que se quiere trabajar en los próximos años.
- *OpSys*: Sistema Operativo con el que trabaja.
- *NEWStuck*: Qué hace cuando se queda bloqueado con algún problema.
- *NEWSOSites*: Páginas de *Stack Overflow* que ha consultado.
- *SOVisitFreq*: Frecuencia de visita a *Stack Overflow.*
- *SOAccount*: Tiene una cuenta de *Stack Overflow o no.*
- *SOPartFreq*: Frecuencia con la que participa en *Stack Overflow.*
- SOComm: Se considera miembro de *Stack Overflow.*
- *NEWOtherComms*: Nombre de las comunidades similares a *Stack Overflow* de las que es miembro.
- *Age*: Edad.
- *Gender*: Género.
- Trans: Transexualidad.
- Sexuality: Sexualidad.
- Ethnicity: Etnia.
- Accessibility: Accesibilidad.
- MentalHealth: Estado de salud mental.
- SurveyLength: Longitud del cuestionario.
- SurveyEase: Facilidad del cuestionario.

A simple vista, el *data set* tiene la siguiente forma:

ResponseId	MainBranch	Employment	Country	US_State	UK_Country	EdLevel	Age1stCode	LearnCode	YearsCode	...	Age	Gender	Trans
1	I am a developer by profession	Independent contractor, freelancer, or self-em...	Slovakia	NaN	NaN	Secondary school (e.g. American high school, G...	18 - 24 years	Coding Bootcamp;Other online resources (ex: vi...	NaN	...	25-34 years old	Man	No
2	I am a student who is learning to code	Student, full-time	Netherlands	NaN	NaN	Bachelor's degree (B.A., B.S., B.Eng., etc.)	11 - 17 years	Other online resources (ex: videos, blogs, etc...	7	...	18-24 years old	Man	No
3	I am not primarily a developer, but I write co...	Student, full-time	Russian Federation	NaN	NaN	Bachelor's degree (B.A., B.S., B.Eng., etc.)	11 - 17 years	Other online resources (ex: videos, blogs, etc...	NaN	...	18-24 years old	Man	No
4	I am a developer by profession	Employed full-time	Austria	NaN	NaN	Master's degree (M.A., M.S., M.Eng., MBA, etc.)	11 - 17 years	NaN	NaN	...	35-44 years old	Man	No

La mayoría de variables son categóricas, aunque también se incluyen datos cuantitativos como el salario percibido por la persona.

En general, el propio *Stack Overflow* proporciona una infografía interactiva y bastante completa analizando los datos del cuestionario para este último estudio de 2021. Sin embargo, no se proporciona un análisis de los datos respecto de los años anteriores. Para este proyecto, me gustaría analizar los datos del cuestionario de 2021 focalizando en las respuestas de personas que residen en España (1.485 registros) y completando dicho análisis con la evolución de la tendencia en lenguajes, bases de datos, plataformas, etc durante los últimos 3 años (2019, 2020 y 2021) tanto en España como en el conjunto de los países presentes en el *data set*. No es posible ir más lejos en el tiempo ya que el tipo de preguntas cambia sustancialmente entre 2018 y 2019. Finalmente, los conjuntos de datos con los que pretendo completar el análisis también se pueden encontrar online en la misma dirección web.

Las preguntas que me gustaría resolver con esta visualización son:

- **[2021]** ¿Hay alguna relación entre el perfil social de la personas que participan en el cuestionario con los lenguajes, bases de datos, herramientas, etc, que usan o su categoría profesional y salario? En la infografía actual no se hace un análisis conjunto de las diferentes respuestas para caracterizar el perfil del participante.
- **[2021]** ¿Cuál es el perfil de los participantes en España? ¿Tenemos una representación heterogénea? ¿Y en el conjunto de datos total?
- **[2021]** ¿Cuál es la tendencia del desarrollo de software en España? ¿Sigue España la tendencia global?
- **[2019-2021]** ¿Muestran los cuestionarios consecutivos de 2019, 2020 y 2021 alguna tendencia global temporal? (Análisis temporal de las tendencias en los últimos tres años)

- **[2019-2021]** ¿Existen herramientas que muestran tendencias contrarias en años consecutivos? ¿Cuál es el motivo?