Choosing a baby name using Python

Desarrollando un proyecto personal de Data Science

Victor Cuspinera

2021-11-04

Comentarios iniciales

⚠️ Presentation in Spanish

⭐ Preparada para el webinar de la Maestría en Ciencia de los Datos de La Universidad de Guadalajara

Sobre el autor

Nombre: Victor Cuspinera

Formación: Licenciatura en Actuaría (2007), y Maestrías en Economía (2014) y Ciencia de Datos (2020)

Experiencia: Analista e investigador en el Banco de México, en el área de dinero en efectivo

¿Por qué hacer proyectos? Para crear un portafolio de proyectos, ejercitar mis habilidades como *Data Scientist*, y seguir aprendiendo herramientas nuevas

Baby Names Baby Names Baby Names Baby Names Baby Names
Baby Names Baby Names Baby Names Baby Names Baby Names
Baby Names Baby Names Baby Names Baby Names Baby Names
Baby Names Baby Names Baby Names Baby Names Baby Names
Baby Names Baby Names Baby Names Baby Names Baby Names
Baby Names Baby Names Baby Names Baby Names Baby Names
Baby Names Baby Names Baby Names Baby Names Baby Names
Baby Names Baby Names Baby Names Baby Names Baby Names
Baby Names Baby Names Baby Names Baby Names Baby Names
Baby Names Baby Names

Choosing a baby name using Python

1ra parte. El proyecto

Intro

Motivación

Seleccionar el nombre de un bebé puede no ser un asunto trivial, o al menos no para mi esposa y para mí cuando supimos que íbamos a tener una bebé.

Buscamos nombres para bebés en todos lados: en libros y páginas web de nombres populares, nombres internacionales, recomendaciones de amigos, nombres de personajes de series y películas.

Para inicios de abril ya teníamos 7 nombres: los favoritos de mi esposa eran Elisa y Macarena, y los míos Aisha, Amanda, Carlina, Gina y Victoria.

Intro

Dudas

Sin embargo, empezaron a surgir varias preguntas:

- ¿hay alguna mejor manera de elegir un nombre para nuestra bebé?
- ¿cuáles son los nombres más populares?
- ¿existen tendencias de nombres?
- ¿hay algún nombre que suene mejor combinado con nuestros apellidos (Cuspinera Martínez)?
- ¿esto aplica también para nombres en países donde se habla inglés?

Bases de datos

Español

México 🇲🇽
- Registro Nacional de Población ❌

España 🇪🇸
- Instituto Nacional de Estadística ✅

Inglés

U.S.A. 🇺🇸
- Social Security Agency ✅

Canadá (provincia de British Columbia) 🇨🇦
- Government of B.C. ✅

Bases de datos

Ejemplo de base de datos original (Canadá)

```
##          Name  1920  1921  1922  1923  ...  2016  2017  2018  2019  Total
## 0      AADHYA     0     0     0     0  ...     0     0     0     5      5
## 1     AALIYAH     0     0     0     0  ...    22    26    22    19    421
## 2       AANYA     0     0     0     0  ...     0     0     0     0     12
## 3       AARYA     0     0     0     0  ...     0     0     0     5     16
## 4       ABBEY     0     0     0     0  ...     5     0     0     0    132
## ...       ...   ...   ...   ...   ...  ...   ...   ...   ...   ...    ...
## 2641    ZOFIA     0     0     0     0  ...     5     0     0     0      5
## 2642     ZOIE     0     0     0     0  ...     0     0     0     0     22
## 2643     ZOYA     0     0     0     0  ...     0    10    13     7     83
## 2644      ZO�     0     0     0     0  ...     0     0     8     5     19
## 2645      ZO�     0     0     0     0  ...    12     9     8    11    264
##
## [2646 rows x 102 columns]
```

Exploración de datos (EDA)

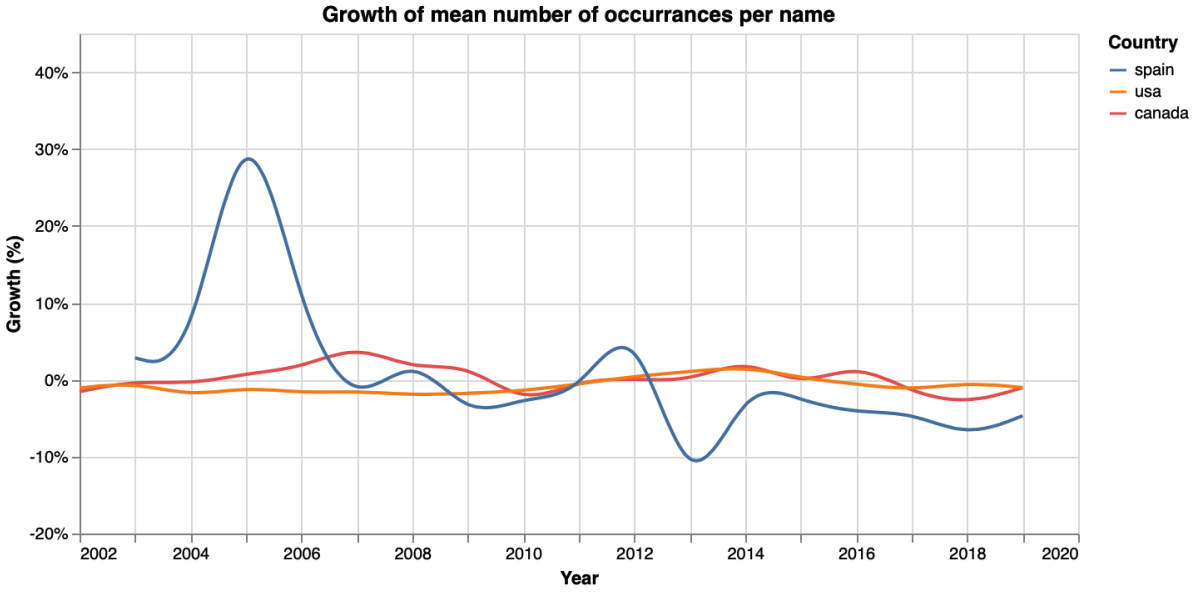
Aunque la información en las bases de datos es similar, al revisar con mayor detalle su estructura, temporalidad, formato y alcance son distintos:

País	No. de nombres distintos	Total de años	Periodo	Características de la BD	No. de archivos (formato)
España[1]	379	18	2002-2019	100 nombres más populares	18 (XLS)
Canadá (B.C.)	4,340	100	1920-2019	Nombres con 5 observaciones o más	2 (CSV)
USA	99,444	140	1880-2019	Nombres con 5 observaciones o más	140 (TXT)

1: Para España también se identificó una base de datos de los nombres con 20 observaciones o más en 2019, del INE.

Exploración de datos (EDA)

- El crecimiento anual en el número de nombres en las bases de datos de USA y Canadá es estable alrededor de cero.
- En la base de datos de España hay variaciones de hasta 30%, lo cual está relacionado con la estructura de la base de datos (sólo los 100 nombres más populares).



Limpieza de datos

- Abrir 160 archivos en distintos formatos y estructura, los cuales contienen información con las bases de datos de España, USA y Canadá.
- Estandarizar las columnas y consolidar la información en una única base de datos.
- Convertir los nombres a minúsculas.
- Revisar casos especiales (e.g. el nombre `Nan` equivalente a valores nulos).
- Conservar la información por país aunque algunos nombres se repitan en la base de datos.

Limpieza de datos

Base de datos consolidada.

```
##           name  number  year sex country language
## 0        maria    8838  2002   F   spain  spanish
## 1        lucia    7712  2002   F   spain  spanish
## 2        paula    5956  2002   F   spain  spanish
## 3        laura    5544  2002   F   spain  spanish
## 4        marta    4644  2002   F   spain  spanish
## ...        ...     ...   ...  ..     ...      ...
## 2457803    zoë      10  2015   F  canada  english
## 2457804    zoë      12  2016   F  canada  english
## 2457805    zoë       9  2017   F  canada  english
## 2457806    zoë       8  2018   F  canada  english
## 2457807    zoë      11  2019   F  canada  english
##
## [2457808 rows x 6 columns]
```

Análisis descriptivo

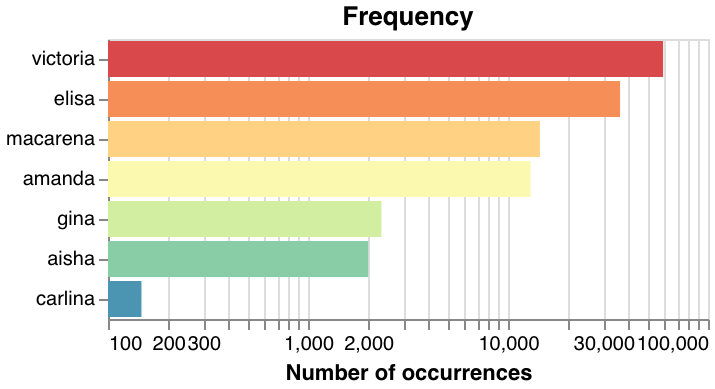
El análisis del resto del documento está relacionado sólo con los nombres favoritos para nuestra bebé: **aisha, amanda, carlina, elisa, gina, macarena** y **victoria**.

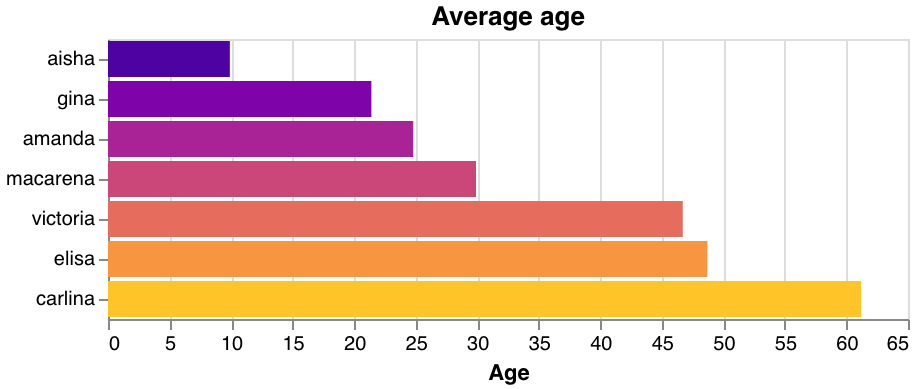
⭐ Para ver el análisis completo el cual contiene detalle sobre los nombres más populares y de género por país, revisar el jupyter notebook `analysis.ipynb` del repositorio de *Baby Names* en GitHub.

Análisis descriptivo: España 🇪🇸

De nuestros nombres favoritos en la base de datos de España, el nombre más popular es **victoria** con 59.6 mil registros, seguido de **elisa** con 36.3 mil, **macarena** con 14.4 mil, **amanda** con 12.9 mil, **gina** con 2.3 mil, **aisha** con 2 mil y **carlina** con 147 registros.

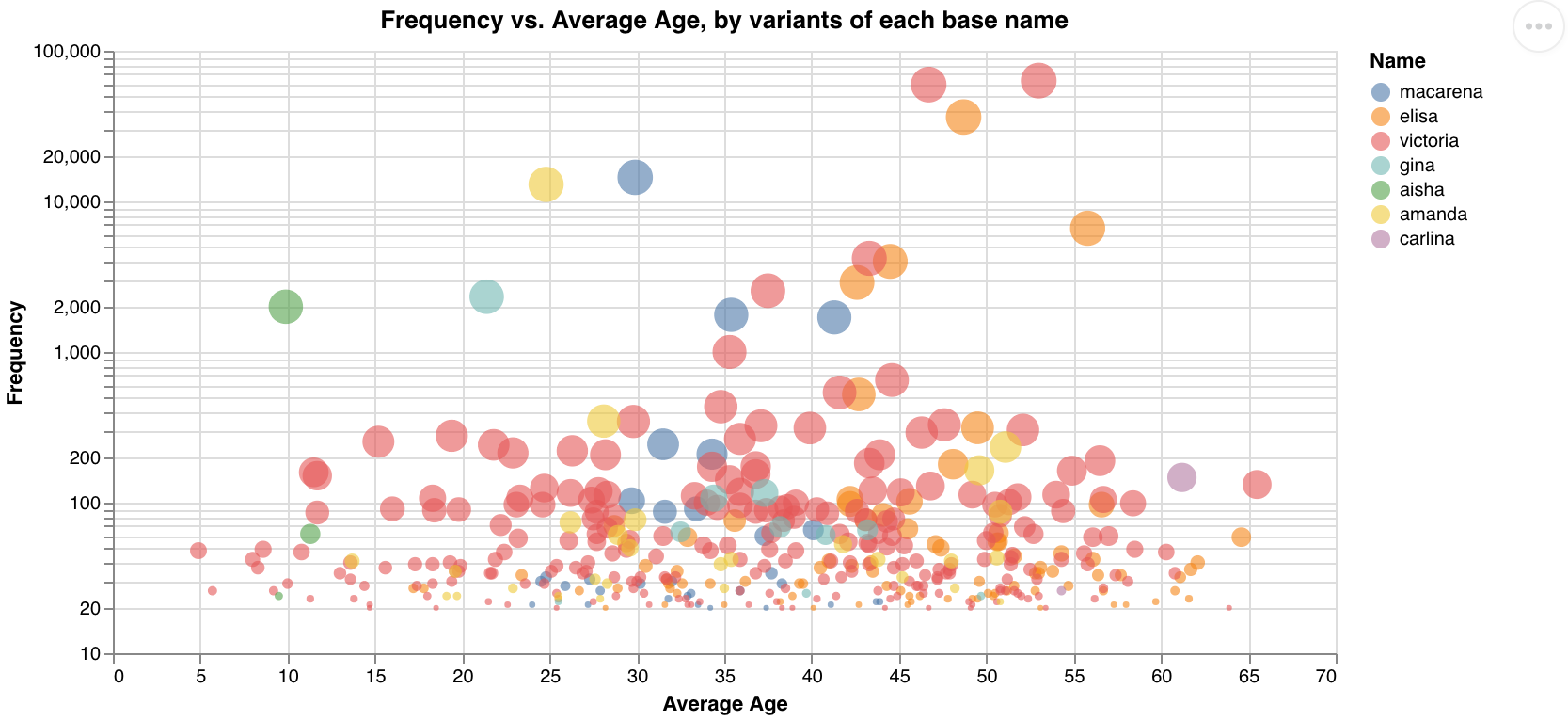
Respecto a la edad promedio, mientras **aisha** es el nombre para mujeres más jóvenes con 9.9 anõs promedio, en el otro extremo tenemos a **carlina** con 61.2 años.





Análisis descriptivo: España 🇪🇸

Frecuencia vs. edad promedio de nombres compuestos en la base de datos de 2019 de España.



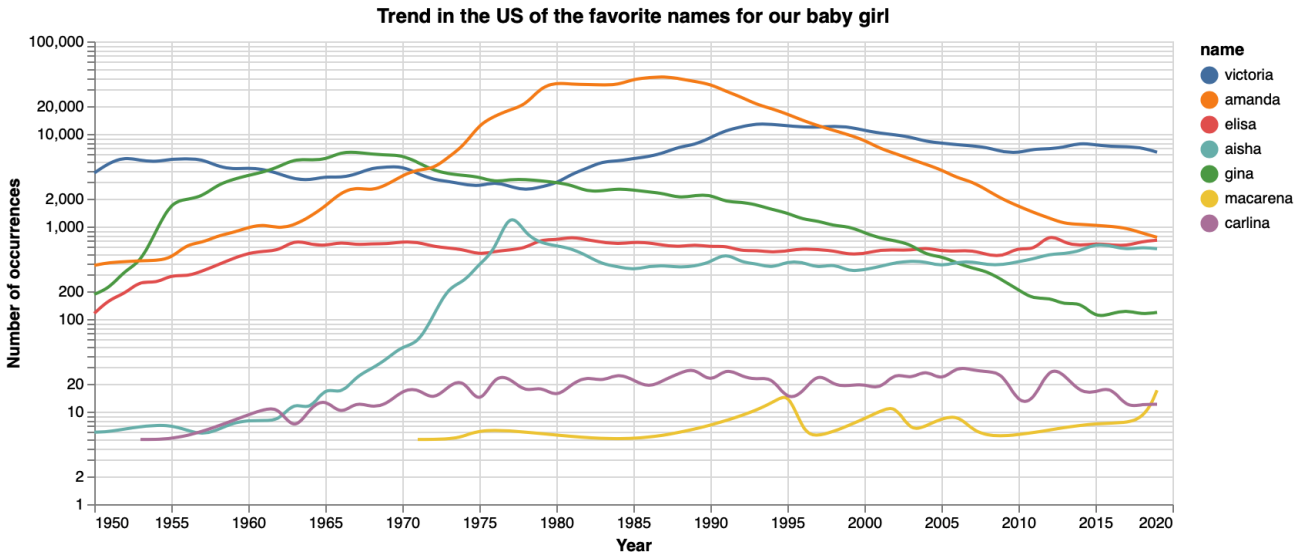
Análisis descriptivo: España 🇪🇸

Word cloud de nombres compuestos con 75 observaciones o más.

Los nodos representan los nombres, las flechas el orden y entre más obscuras las flechas más fuerte la conexión entre ambos nombres.

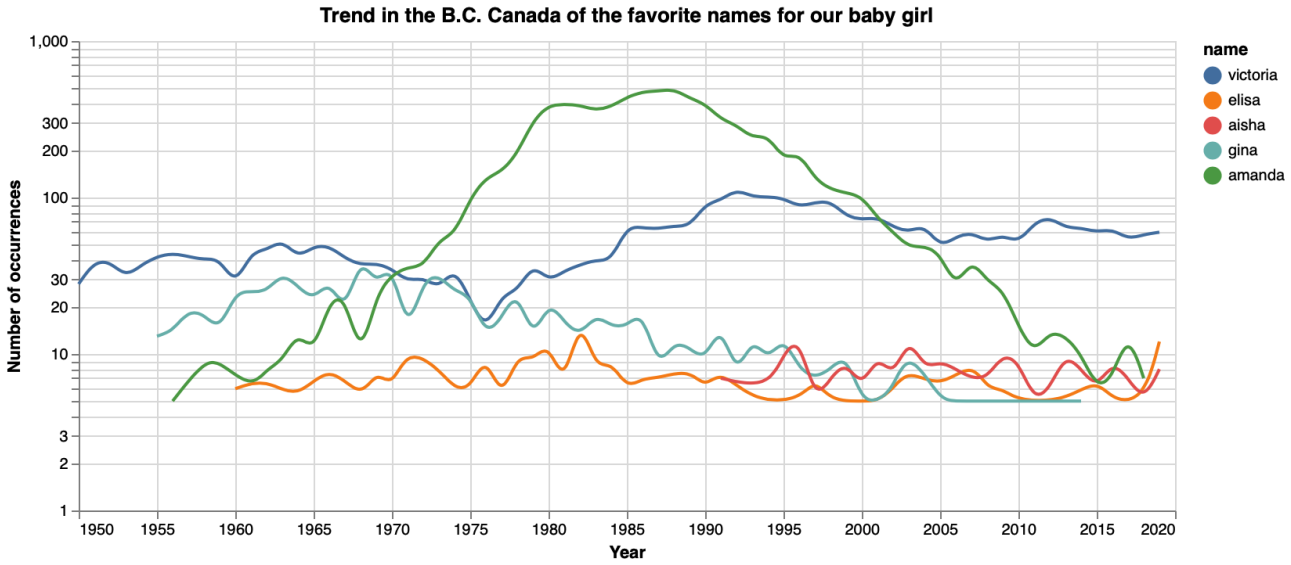
Análisis descriptivo: USA 🇺🇸

Gráfica de tendencias de nombres de 1950 a 2019, en USA. Se observa que **victoria** ha sido popular a lo largo del tiempo, mientras que **amanda** aumentó su popularidad entre los años 80's y 90's. Por otro lado, **carlina** y **macarena** tienen pocas observaciones.



Análisis descriptivo: Canadá 🇨🇦

Gráfica de tendencias de nombres de 1950 a 2019, en British Columbia, Canadá. Esta base de datos contiene información a nivel local, no nacional, por lo que no aparecen todos nuestros nombres favoritos (faltan **macarena** y **carlina**). Al igual que en USA, el nombre **victoria** ha sido popular a lo largo del tiempo, y **amanda** entre los años 80's y 90's.



Scores de nombres usando el IPA (Alfabeto Fonético Internacional)

El último punto de este proyecto fue crear un *scoring* para identificar qué tan bien combinan los nombres con nuestros apellidos.

Se utilizaron diccionarios de Python para traducir los nombres al Alfabeto Fonético Internacional (IPA) tanto en español como inglés.

Ejemplo:

```
##                    Nombre
## Texto original    Victoria
## IPA español       biktoɾja
## IPA inglés       vɪkˈtɔriə
```

Scores de nombres usando el IPA

Para obtener las métricas se identifican las rimas en ambos idiomas comparando uno a uno cada nombre con los otros nombres y apellidos.

En esta comparación, se identifica la rima asonante de la última vocal, rima asonante de las dos últimas vocales, rima consonante de la última sílaba, coincidencia de nombres en la primera sílaba, asiganando 1 en caso de coincidencia y 0 en caso contrario. Con estos resultados, y de forma ponderada, se obtiene un puntaje en español, otro en inglés y el puntaje total obtenido del promedio de los dos anteriores.

Ejemplo de fórmula para obtener el puntaje en español.

$$Score\ Spanish = \frac{\sum (last\_vowel_{SPA},\ 2 \times last\_2\_vowels_{SPA},\ 2 \times last\_syllable_{SPA},\ initial\_letters_{SPA})}{6}$$

Scores de nombres usando el IPA

Puntuación para nuestros nombres favoritos:

```
##                                Spanish  English  Total score
## aisha cuspinera martinez        0.0556   0.2222       0.1389
## amanda cuspinera martinez       0.0556   0.2222       0.1389
## carlina cuspinera martinez      0.0556   0.1667       0.1111
## elisa cuspinera martinez        0.0556   0.3334       0.1944
## gina cuspinera martinez         0.0556   0.3334       0.1944
## macarena cuspinera martinez     0.2222   0.3334       0.2778
## victoria cuspinera martinez     0.0556   0.3334       0.1944
```

macarena es el nombre con mayor puntaje de entre nuestros nombres favoritos, seguidos por **elisa, gina** y **victoria**.

Scores de nombres usando el IPA

⚠️ Cuando llegué a este punto, a mediados de abril 2021, mi esposa ya había elegido el primer nombre de nuestra bebé: **elisa**. Ahora estaba en mis manos el decidir si nuestra bebé tendría un único nombre o un nombre compuesto.

Al combinar todos los nombres de la base de datos consolidada con el nombre **elisa** y nuestros apellidos **cuspinera martinez**, los nombres con mayor puntaje fueron: aisa, akira, alisa, ariza, corisa, delisa, elfrida, elissa, elvina, elyria, elysia, erisa, isa, jazeera, lisa, liza, louisa, luisa, **macarena**, maeda, magdalena, makita, malina, malinda, malvina, marcelia, marcellina, marchita, margarita, marilda, marina, marquita, martina, martita, mathea, matthea, maurita, mayeda, miera, misa, raisa, riera, risa, shakira, viera.

Scores de nombres usando el IPA

El nombre de **macarena** se encuentra también entre los nombres compuestos que, combinados con **elisa**, tienen mayor puntaje.

```
##                                    Spanish  English  Total score
## elisa macarena cuspinera martinez   0.1667   0.4167       0.2917
```

Finalmente, decidimos que el nombre de nuestra bebé sería **elisa macarena cuspinera martinez**.

Comentarios finales del proyecto

Escoger un nombre para un bebé puede no ser tan sencillo. Durante este proyecto, en el proceso de selección se desarrolló una herramienta que mide como suena un nombre completo asignando una puntuación para hacer más fácil la comparación entre los posibles nombres para un bebé. Esta herramienta es una propuesta inicial y tiene limitaciones, por lo que podría mejorarse considerando otros factores fonéticos (ritmo, acentuación, secuencia de nombres, entre otros), cambiando las ponderaciones en la función para obtener el *scoring*, o considerando otros idiomas.

Como una herramienta, este modelo es sólo una parte del rompecabezas que podría (y debería) ser complementado con otras variables cualitativas como el significado de los nombres, tradiciones regionales y/o familiares, tendencias, u otras ideas de los futuros padres, durante la selección del nombre del bebé.

Dónde encontrar este proyecto

Este proyecto se ha buscado compartir a través de distintos espacios:

1. Página personal
2. GitHub
3. LinkedIn
4. Towards DataScience
   (revista electrónica especializada)

2da Parte. Comentarios y recomendaciones

Recomendaciones

Lenguajes

...programación
- Python
- R language

...lengua
- Inglés (recomendado)

Herramientas:
- Entornos: Jupyter lab/notebook, RStudio
- Control de versiones: Git y GitHub
- Visualización: Matplotlib, Altair, ggplot2, Shiny, Tableau
- Bases de datos: SQL, MongoDB
- Servicios en la nube: AWS, Google

Otras habilidades:
- Data Wrangling
- Webscraping
- Trabajo colaborativo
- Reproducibilidad
- NLP (análisis de texto)
- Pronósticos (análisis temporal)
- Análisis espacial

Ideas para proyectos personales

¿Qué te interesa, gusta, o apasiona?

Favor de contestar la encuesta

Entra a www.menti.com

Ingresa el código **6230 9621**

Escribe tus ideas

Ideas para proyectos personales

Algunas de las ideas compartidas:

- Logística en transporte
- Tráfico aéreo
- Desempeño de equipos de futbol mexicano
- Selección de posgrado
- Fluctuación de criptomonedas
- Análisis de precios y ofertas
- Compra de inmuebles
- Portadas de películas
- Análisis de población
- Experiencia de usuario
- *Google lens* para personas con discapacidad
- Gestión de tiempo en redes sociales
- Análisis espacial en salud pública

Muchas gracias por sus respuestas, el compendio completo de ideas se encuentra en este link.

Otros proyectos personales interesantes

- Generador de Pokemmones por Andres Pitta
- Observación de ardillas en NYC Central Park por Cari Gostlic
- Análisis de sentimiento de elecciones en Canadá por Sam Edwardes
- Análisis de Sentimiento de respuesta de Canadá a Covid-19 por Leopoldo y Victor Cuspinera

Desarrolla tu portafolio de proyectos

Proyectos personales

su inspiración puede venir de hobbies, deportes, música, familia, pláticas, experiencia laboral o académica

Ventaja:
administrar su propio tiempo, desarrollar proyectos de su completo interés

Desventaja:
promover su proyecto por ustedes mismos

Desarrolla tu portafolio de proyectos

Voluntariado

participar en voluntariados de ciencia de datos con Omdena, United Nations Volunteers, Data Science for Social Good, entre otras instituciones

Ventaja:
existen múltiples opciones de proyectos, aprender nuevas herramientas y habilidades, ampliar su red de contactos

Desventaja:
consume mucho tiempo (15-20 hr por semana, de 4 a 8 semanas), no ingreso monetario

Desarrolla tu portafolio de proyectos

Hackathones y competencias

otra manera de aprender y desarrollar un portafolio es participando en competencias y hackathones, entre los más famosos están los de Kaggle y Driven Data. Algunas empresas lanzan hackathones para encontrar talentos, Grupo Modelo lanzó el Breweing Data Cup

Ventaja:
aprender mucho en corto tiempo

Desventaja:
los hackathones pueden ser maratónicos y desgastantes

¿Preguntas o comentarios?