MEMORIA TÉCNICA: PROYECTO GEO-ARBITRAGE

Asignatura: Visualización (VIS)

Grado: Ciencia de Datos - 3º Curso

Autores: Daniel Herrán Gómez-Senent, Batu Senyucel, Iván Navarro Martínez y Marc Gómez Cidad

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf
Escola Tècnica Superior d'Enginyeria Informàtica

ÍNDICE

Introducción

El objetivo de este proyecto es desarrollar una herramienta interactiva para "Nómadas Digitales", permitiendo la exploración de destinos basada en el concepto de Geo-Arbitraje (ganar en una moneda fuerte y gastar en una economía de menor coste). A continuación, se detalla el flujo de desarrollo siguiendo el modelo de pipeline de visualización de Ben Fry, justificando las decisiones técnicas y de diseño implementadas en el código.

1. Adquisición (Acquire)

La adquisición de datos se realiza mediante la carga de archivos estáticos en formato CSV, una elección estándar para el intercambio de datos estructurados.

- **Fuente Principal:** Se utiliza el archivo data_master_clean.csv, que actúa como el dataset maestro conteniendo información a nivel de país (Índices de coste, salarios, velocidad de internet, felicidad), que ha sido creado combinando 5 datasets distintos pero de la misma índole.
- **Fuente Secundaria:** Se integra el archivo datasets/livable_cities.csv para obtener granularidad a nivel de ciudad, permitiendo un análisis más detallado dentro de los Estados Unidos mediante la función augment_us_data.

Se ha optado por la carga local mediante pandas para garantizar la latencia mínima en la aplicación. Al no depender de una API en tiempo real, se asegura la estabilidad del dashboard. La combinación de dos fuentes permite enriquecer la visualización, resolviendo el problema de que los datos a nivel de país suelen ocultar las variaciones internas de naciones grandes como EE. UU.

2. Formateado (Parse)

Los datos crudos requieren una estructura semántica para ser útiles. En esta fase se han implementado transformaciones clave en el código:

- **Categorización Geográfica:** Se creó la función get_continent. Los datos originales contenían regiones ambiguas. El código parsea estas regiones y países específicos para agruparlos en 7 continentes estrictos (América del Norte, Sur, Europa, Asia, África, Oceanía)
- **Segmentación de Costes:** Se transformó la variable numérica continua Índice de Coste en una variable categórica ordinal (Bajo, Medio, Alto) mediante pd.cut.

Según los principios de percepción, el cerebro humano procesa mejor los grupos discretos. Convertir regiones complejas en "continentes" reduce la carga cognitiva del usuario. Asimismo, la segmentación del coste permite al usuario filtrar mentalmente de forma rápida sin tener que evaluar números decimales específicos.

3. Filtrado (Filter)

Dada la magnitud del dataset (más de 100 países y ciudades), la visualización completa resultaría ilegible (el problema del overplotting). Se implementó un filtrado interactivo en el sidebar:

- **Filtro Categórico:** Selección de continentes mediante multiselect.
- **Filtro Cuantitativo:** El sistema filtra visualmente qué países son viables económicamente mediante el cálculo del Nomad Score.

El código genera un filtered_df y filtered_cities sobre los cuales se renderizan todos los gráficos posteriores.

Permitir al usuario eliminar continentes enteros (por ejemplo, si no desea vivir en Asia) limpia el área de dibujo y permite que la escala de los ejes se readapte automáticamente a los datos relevantes, mejorando la legibilidad de las diferencias sutiles entre los puntos restantes.

4. Minado (Mine)

Esta es la fase de mayor carga computacional, donde se generan nuevas métricas que no existían en el dataset original:

- **Métricas Derivadas:** Se calcula la Puntuación Nómada dividiendo el salario imputado por el índice de coste.
- **Algoritmo de Convex Hull:** Para el gráfico de dispersión. Este algoritmo matemático calcula el polígono mínimo que envuelve al conjunto de ciudades de un continente.
- **Normalización y Ponderación:** Para el mapa final, se aplicó una normalización Min-Max a las variables de Coste, WiFi y Felicidad, permitiendo crear un Personal_Score único basado en pesos definidos por el usuario

El "minado" transforma datos en información. El Convex Hull no es solo estético; define matemáticamente el "espacio de posibilidades" de un continente (su aura), permitiendo ver visualmente si Europa tiene un rango de calidad de vida más amplio que Asia. La fórmula de ponderación personalizada permite un análisis multicriterio que una simple tabla no podría ofrecer.

5. Representación (Represent)

Se han seleccionado modelos visuales específicos para cada tipo de dato:

Distribución (Violin Plot): Se usó px.violin para el Coste por Continente. A diferencia de un boxplot, muestra la densidad de probabilidad de los datos.

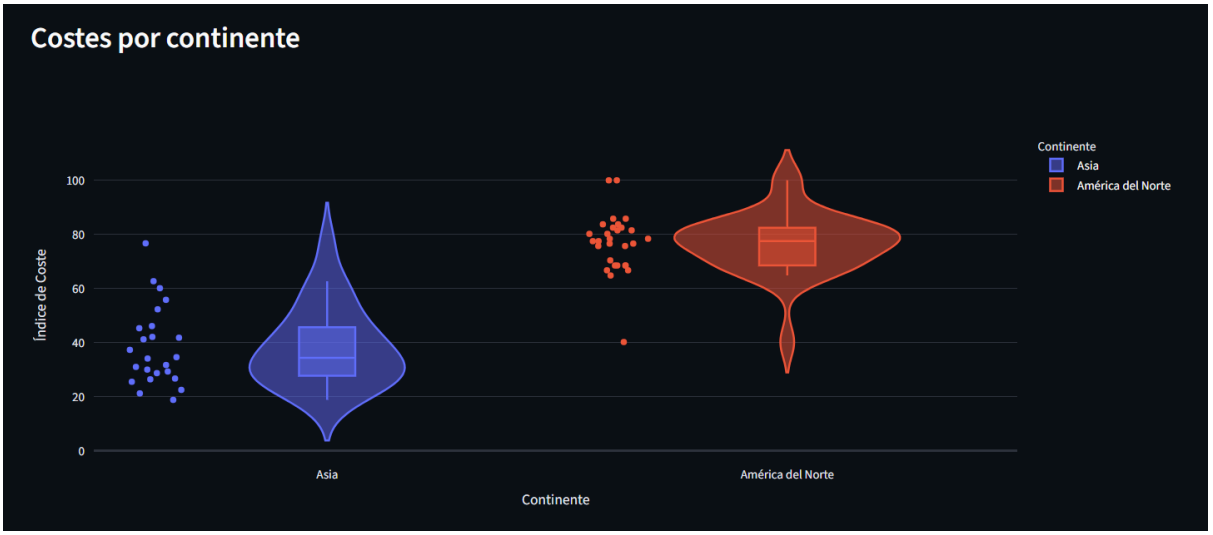

Figura 1: Gráfico de violín que muestra la distribución y densidad del índice de coste de vida segmentado por continentes, facilitando la identificación de la varianza económica regional.

Correlación (Scatter Plot): Se usó un gráfico de dispersión para relacionar Coste de Vida (X) vs Calidad de Vida (Y). Es la forma más efectiva de mostrar la relación entre dos variables cuantitativas.

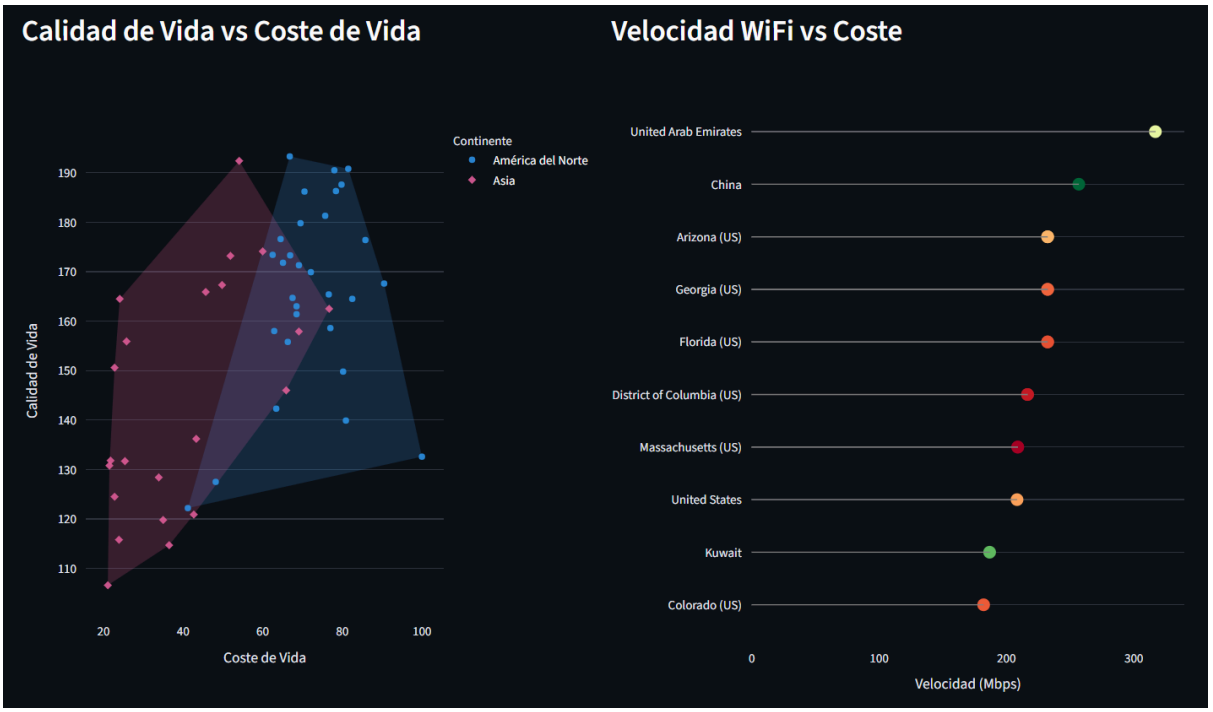

Figura 2: Gráfico de tipo 'Lollipop' que ordena los países según su velocidad media de internet (Mbps), permitiendo una comparación clara de la infraestructura tecnológica sin saturación visual.

Ranking (Lollipop Chart): Para la velocidad de WiFi, se usó un diseño de "Lollipop" en lugar de barras tradicionales.

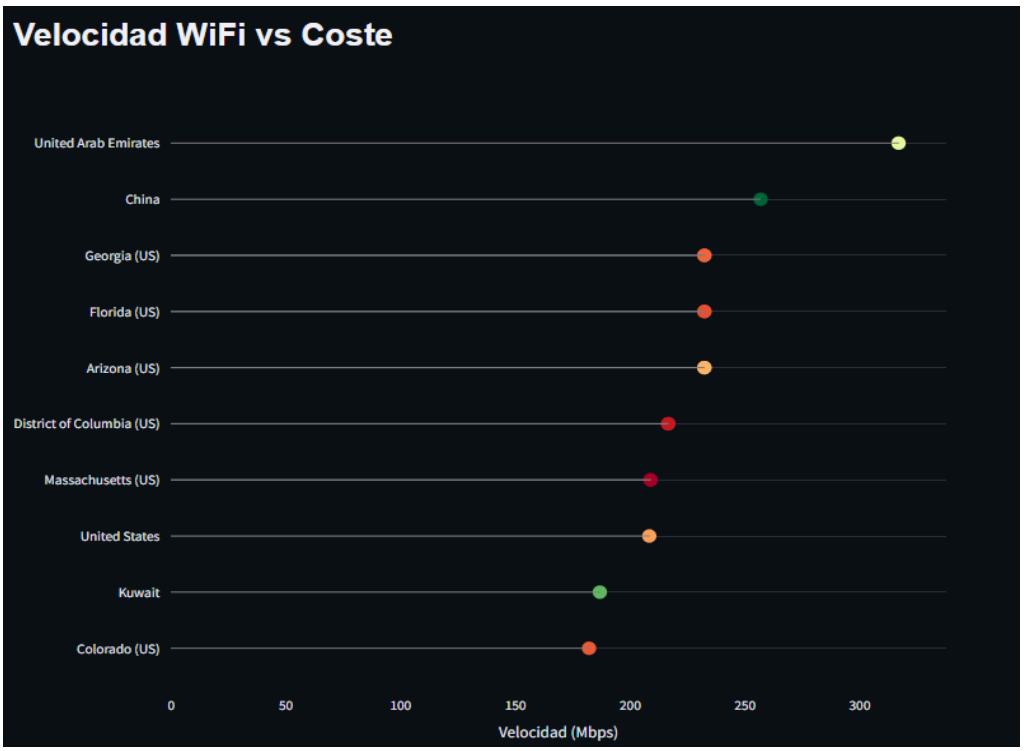

Figura 3: jerarquiza los 15 países con mayor velocidad media de conexión (Mbps). Los marcadores emplean una codificación de color secuencial basada en el 'Índice de Coste', permitiendo al usuario discernir rápidamente la relación costo-eficiencia de la infraestructura digital de cada país.

Scatter Plot: Se empleó un mapa de dispersión para poder observar por continente las variables principales del nómada score.

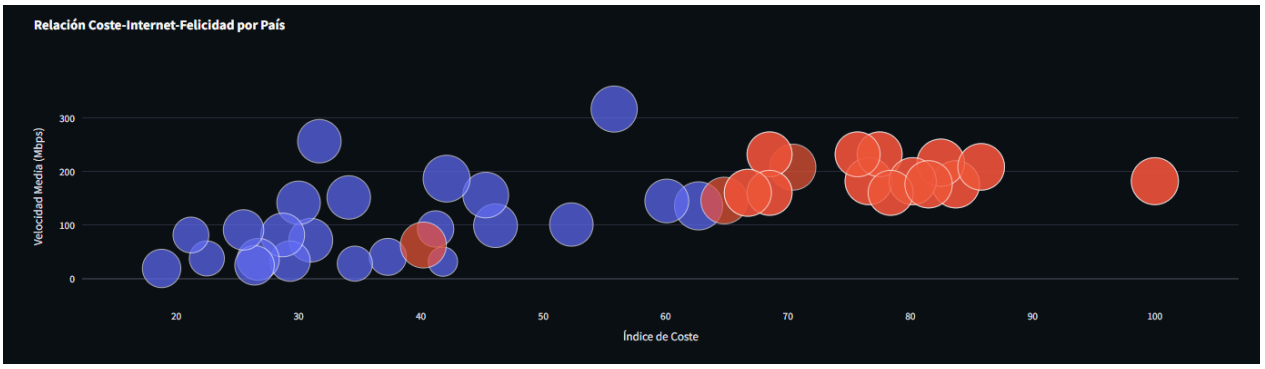

Figura 4: permite identificar qué ciudades ofrecen una alta calidad de vida a bajo precio, mientras que las auras y el tamaño ayudan a comparar qué continentes son más felices en general.

Geo-espacial (Choropleth Map): Se utilizó un mapa coroplético para proyectar el Personal_Score sobre la geografía mundial.

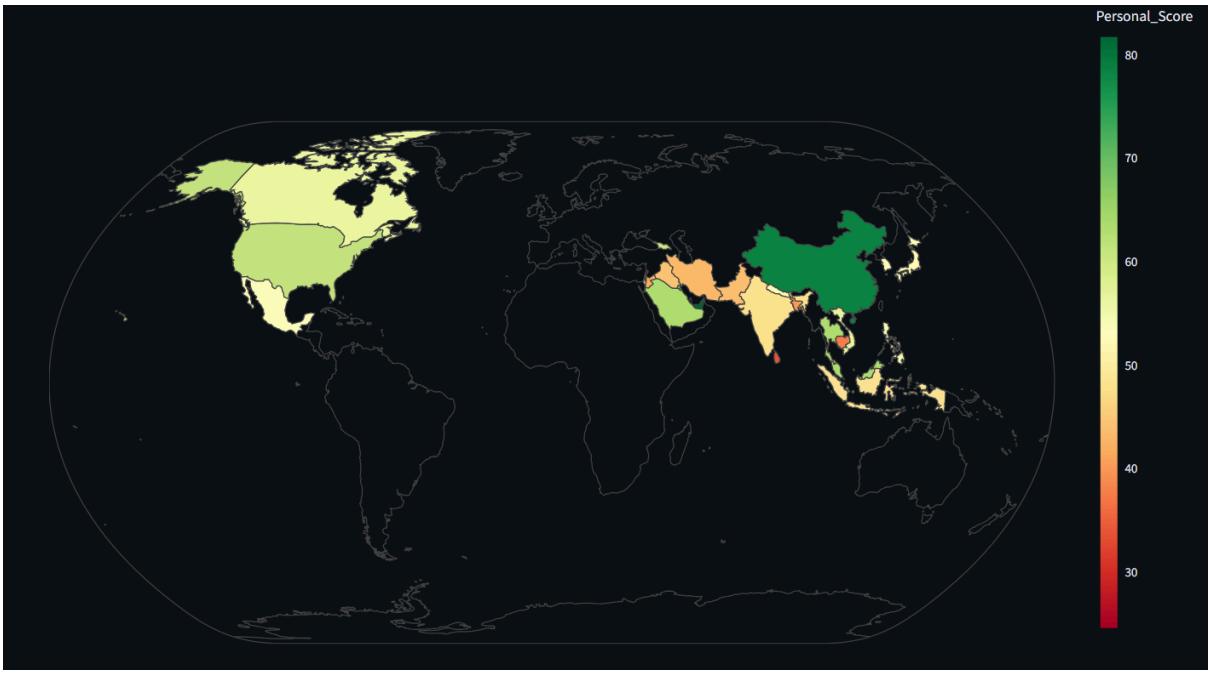

Figura 5: Mapa coroplético global que representa la 'Puntuación Nómada' por país, utilizando una escala de color divergente para identificar visualmente los destinos con mayor equilibrio entre coste e ingresos.

- **Violin Plot:** Justificado porque permite ver si la distribución es modal (ej. un continente con países muy baratos y muy caros, sin término medio).
- **Lollipop:** Maximiza el *Data-Ink Ratio* (Tufte). Las barras gruesas ocuparían demasiado espacio visual innecesario; el punto final es lo que importa para comparar magnitudes de velocidad.
- **Map:** Esencial para este dominio, ya que la ubicación geográfica es intrínseca al concepto de "Nómada Digital".

6. Refinado (Refine)

Se aplicaron técnicas de diseño para mejorar la legibilidad y el atractivo.

- **Gestión del Color:** Se usaron paletas semánticas divergentes (Rojo a Verde) donde el verde indica "positivo" (bajo coste o alta felicidad) y el rojo "negativo".
- **Opacidad (Alpha):** En el gráfico de dispersión, se aplicó opacity=0.2 a las "auras" (Convex Hulls) para que las regiones de fondo no ocultaran los puntos de datos individuales (Layering).

El refinado asegura que la atención del usuario se dirija a los datos y no a los elementos estructurales. El uso de Dark Mode no es solo estético, sino funcional para el público objetivo. La opacidad gestiona la oclusión en gráficos densos, permitiendo ver la intersección entre continentes (ej. dónde se solapan los costes de Europa y América del Sur).

7. Interacción (Interact)

La visualización estática se transforma en una herramienta de exploración mediante:

- **Tooltips (Hover):** Al pasar el ratón por los puntos del Violin Plot o el Mapa, se despliega información detallada (hovertemplate) que no está visible a primera vista.
- **Sliders de Ponderación Dinámica:** La característica más avanzada es el slider que distribuye el 100% de importancia entre Coste, WiFi y Felicidad.

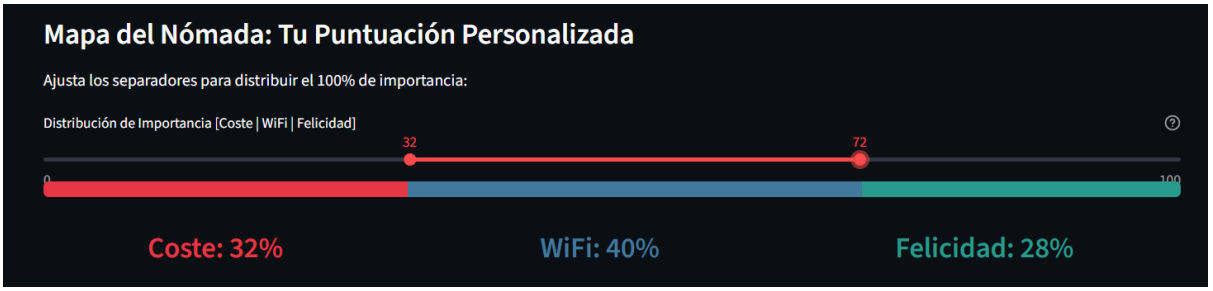

Figura 5: Interfaz de usuario interactiva que permite el ajuste personalizado de pesos (Coste vs. WiFi vs. Felicidad) para recalcular la idoneidad de los destinos en tiempo real.

- **Feedback Visual:** Al mover los sliders, se actualizan instantáneamente las barras de porcentaje de colores y el mapa se repinta.

La interacción convierte al usuario pasivo en activo. El slider de ponderación permite realizar preguntas del tipo "¿Qué pasa si solo me importa el internet y no el coste?" (What if analysis). Esto empodera al usuario para encontrar su "óptimo personal" en lugar de imponerle un ranking estático predefinido por el autor.

Vídeo descriptivo

El vídeo sigue una estructura narrativa en tres actos: presentación del dilema del nómada, exploración de datos regionales e identificación de oportunidades mediante el dashboard interactivo.

Conclusiones

El desarrollo del Dashboard de Geo-Arbitrage representa la culminación práctica de los fundamentos teóricos adquiridos, demostrando que una visualización efectiva no es meramente estética, sino el resultado de una planificación y estructuración veraz de la información.

En primer lugar, la arquitectura del proyecto se ha cimentado sobre las 7 fases de Ben Fry (Adquisición, Formateado, Filtrado, Minado, Representación, Refinado e Interacción). Este flujo ha garantizado que no solo "mostremos datos", sino que transformemos información cruda en conocimiento procesable. Siguiendo el principio de "Conoce a tu audiencia", el diseño se ha centrado en el usuario "Nómada Digital", adaptando la complejidad y el lenguaje visual a sus necesidades de decisión económica y vital.

Desde la perspectiva de la percepción visual y la creación de "buenos gráficos", se han seleccionado cuidadosamente los canales de comunicación visual. Se ha evitado la sobrecarga cognitiva simplificando la interfaz y utilizando atributos preatentivos (color y posición) para destacar patrones clave, como la relación entre coste y calidad de vida.

En cuanto a la elección de modelos gráficos, hemos cumplido estrictamente la máxima de "Di no a las tartas", sustituyéndolas por alternativas más eficientes para la comparación de magnitudes, como los diagramas de violín y los gráficos de dispersión. Estas elecciones evitan los errores comunes de distorsión y facilitan una comparación honesta y precisa de los datos.

Finalmente, el dashboard se alinea con la clasificación estudiada, funcionando como un sistema híbrido. Comienza como una visualización de Presentación (mostrando hechos claros sobre los costes globales) y evoluciona hacia una herramienta exploratoria, permitiendo al usuario interactuar con los datos para descubrir sus propias respuestas. De este modo, el proyecto cumple con el objetivo último de la asignatura: crear sistemas de soporte a la decisión que sean, ante todo, herramientas de veracidad y utilidad.