¿Cuánto me puede salir estudiar en el exterior?

Spoiler: Depende, pero bastante.

Presentan: Crespi Ramiro, Guardia Axel, Lozano Mateo, Rodriguez Juan Ignacio

Grupo 4

Introducción a la Ciencia de Datos

12/06/2025

Pregunta: ¿Qué factores explican mejor el costo de la matrícula universitaria en el mundo?

Objetivo: Entender el valor de la matrícula en instituciones educativas internacionales, en función de variables como país, nivel educativo, categoría de estudios y costo de vida.

Presentación del dataset

Presenta datos de costos de matrículas universitarias.

- Cada fila del dataset representa **un programa universitario específico**, ofrecido por una universidad en una ciudad y país determinados.
- La unidad del dataset es el Programa por Universidad.
- Dimensión original del dataset: 907 filas y 12 columnas
- Sin NAs.
- Sin datos duplicados.
- En formato Tidy.

Fuente del dataset: Kaggle - Datos reales

Variables del Dataset

Variables Continuas (dbl)	Variables Categóricas (chr)
• **Tuition_USD - Variable Target** • Living_Cost_Index • Rent_USD • Visa_Fee_USD • Insurance_USD • Exchange_rate_USD	• Level • Duration Years • Country • City • Program • University

Limpieza y Adecuacion

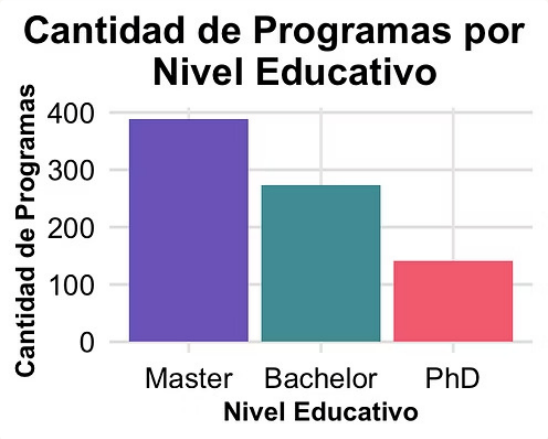
Modificaciones realizadas

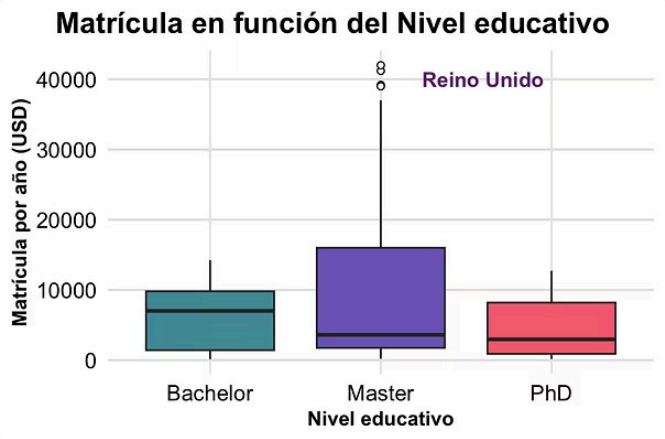
- Creamos la variable Tuition_Per_Year (Tuition_USD / Duration_Years).
- Descartamos la variable Exchange_rate_USD ya que no iba a ser de utilidad para el análisis.
- **Eliminamos** las observaciones en las cuales el **costo de la universidad** era **cero** ya que distorsionarian el analisis del costo.
- Agregamos la columna ***Region*** donde se le asigno a cada pais su continente correspondiente. Dividimos Europa en Occidental y Oriental para analizar mejor comportamientos específicos de esa región (identificados luego de graficar).
- Agregamos la columna ***Category*** donde se agrupan programas en grupos mas grandes.
- Importamos un Dataframe de Índice de Desarrollo Humano (**IDH**) (**Fuente**: ONU) e hicimos un Left Join por paises. Luego la categorizamos en 5 niveles para facilitar el modelado.
- Dataset filtrado: 16 columnas - 804 filas.

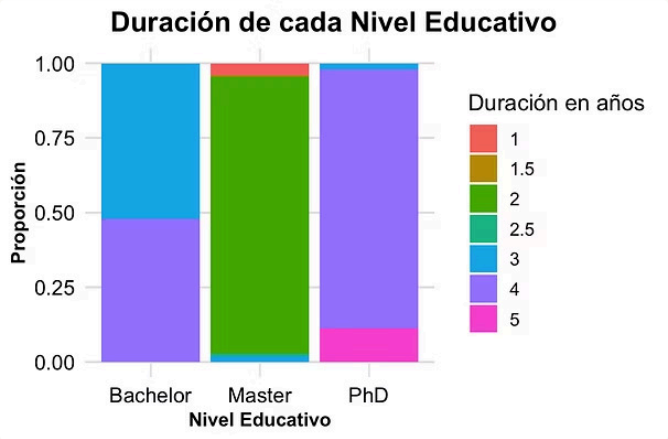
Variables del Dataset

Variables Continuas (dbl)	Variables Categóricas (chr)
• **Tuition_USD** • Tuition_Per_Year • Living_Cost_Index • Rent_USD • Visa_Fee_USD • Insurance_USD • ~~Exchange_rate_USD~~ • *IDH (Importada)*	• Level • Duration Years • Country • City • Program • University • Region (Creada) • Category (Creada) • Nivel IDH (Creada)

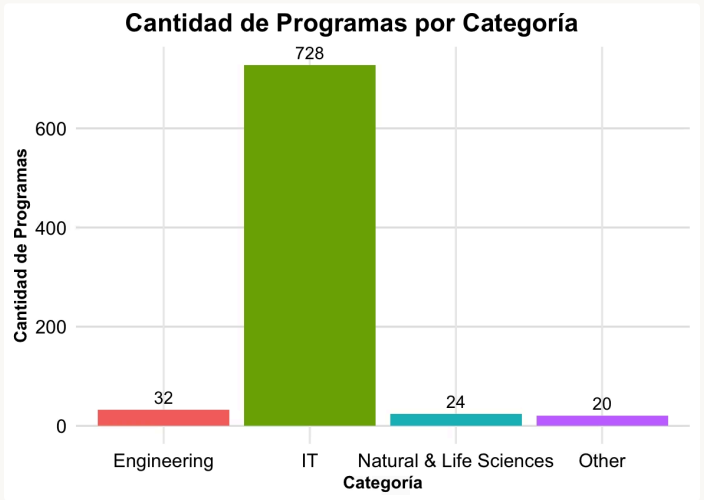
Análisis Exploratorio (EDA):



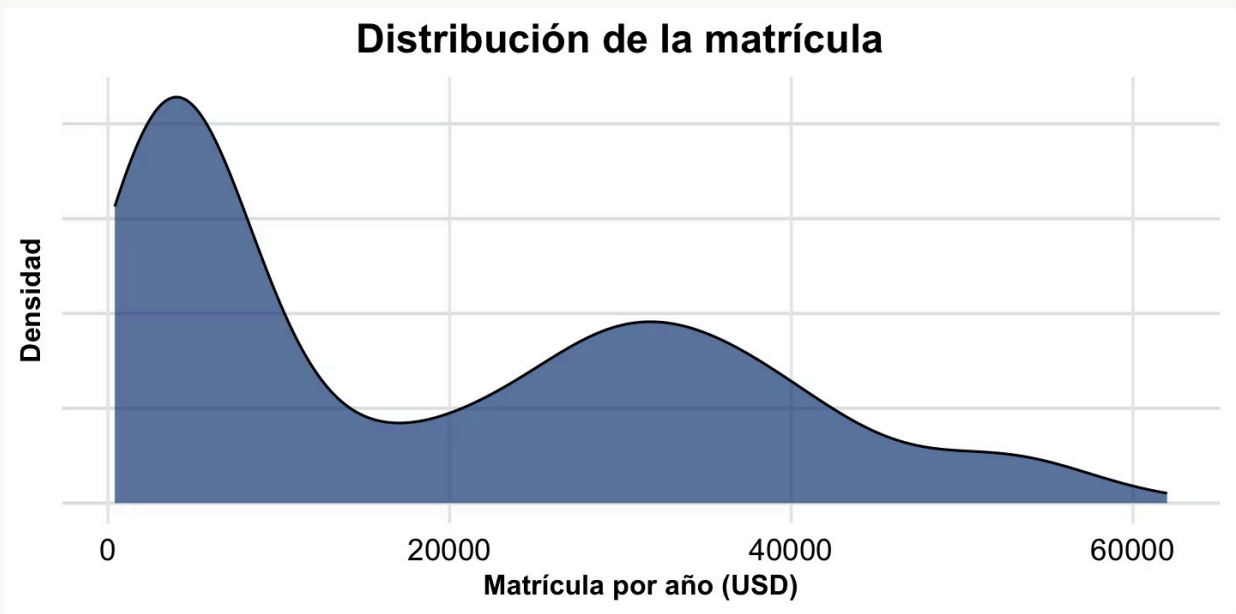




- Los programas Master son los que más aparecen en el dataset, presentan un costo por año más disperso y suelen durar entre 1 y 3 años, aunque la duración más común es 2 años. Se ven algunos outliers del boxplot en matrículas >35.000 por año, pertenecientes a Reino Unido. Tiene el IQR más extenso, lo que implica mayor dispersión.
- Por su parte, los programas Bachelor son los segundos más comunes, tiene la mediana más alta, lo que sugiere un mayor costo. La duración es entre 3 y 4 años.
- Por último, los PhD son los que tienen menos cantidad de programas en el dataset. El IQR es el menor, que muestra distribución del 50% de los casos entre los 2000 USD y 9000 USD aproximadamente. La duración es entre 3 y 5 años.



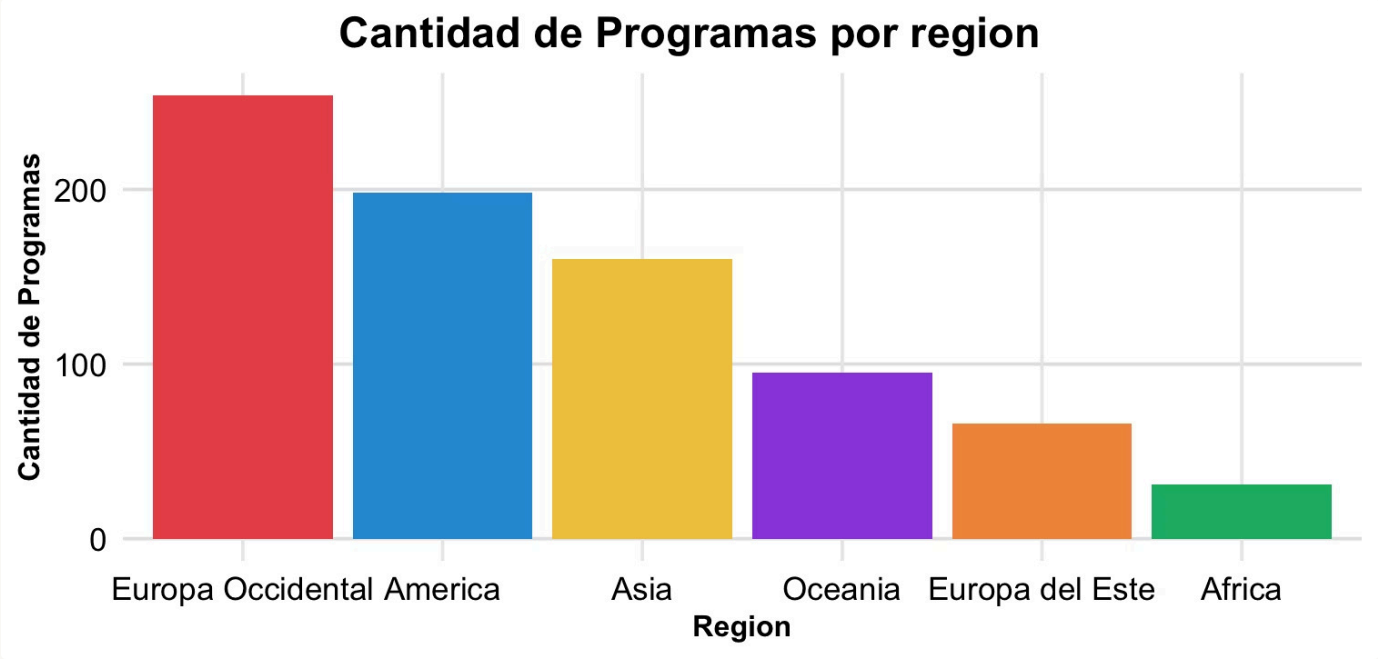
Creamos 4 categorías que agrupaban distintos grupos de programas según su rama de estudios. Al realizar este gráfico notamos que una gran mayoría pertenecían a IT.



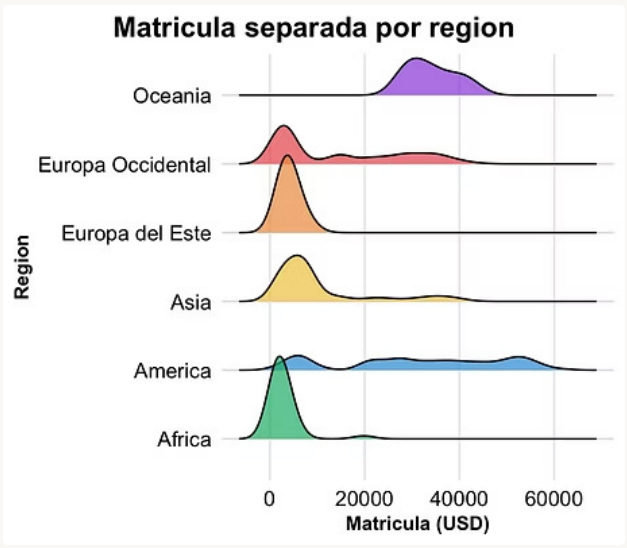
La distribución sugiere una **segmentación** en los costos educativos: Un grupo con un pico al rededor de **10000 USD** y otro al rededor de **30000 USD**

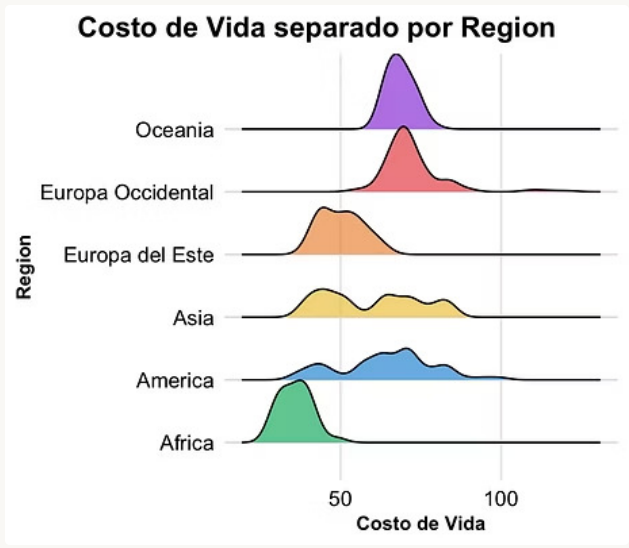
La mayor densidad se encuentra cerca del **primer pico.**

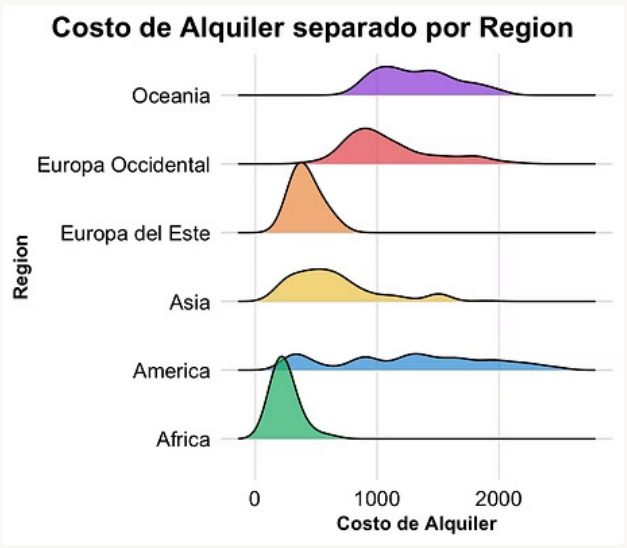
Análisis Exploratorio (EDA):



La región con más programas es Europa Occidental. Inicialmente contabamos con una sola categoría "Europa" lo que concentraba aun más los casos en esta primera barra, sin embargo optamos por dividirla en Europa del Este y Occidental para estudiar las diferencias ya sean culturales, sociales, económicas, etc.

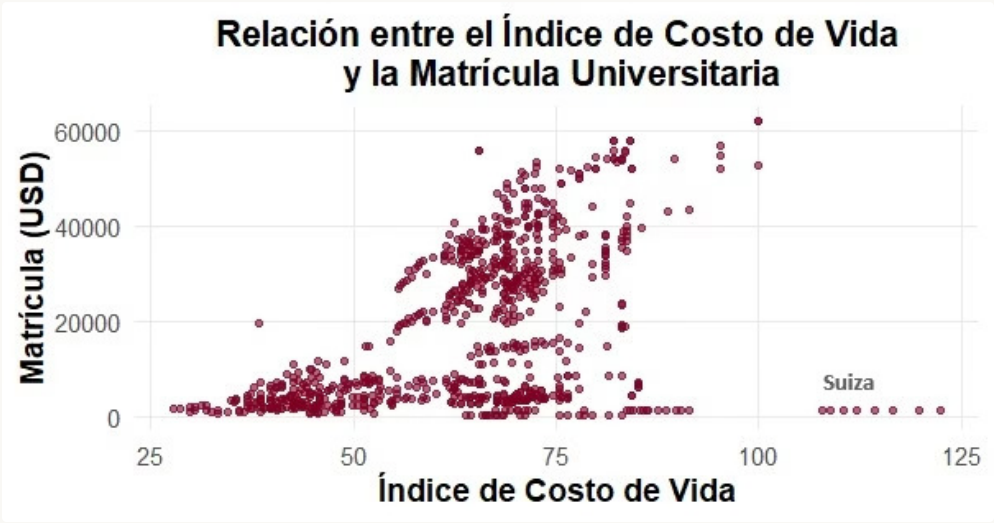


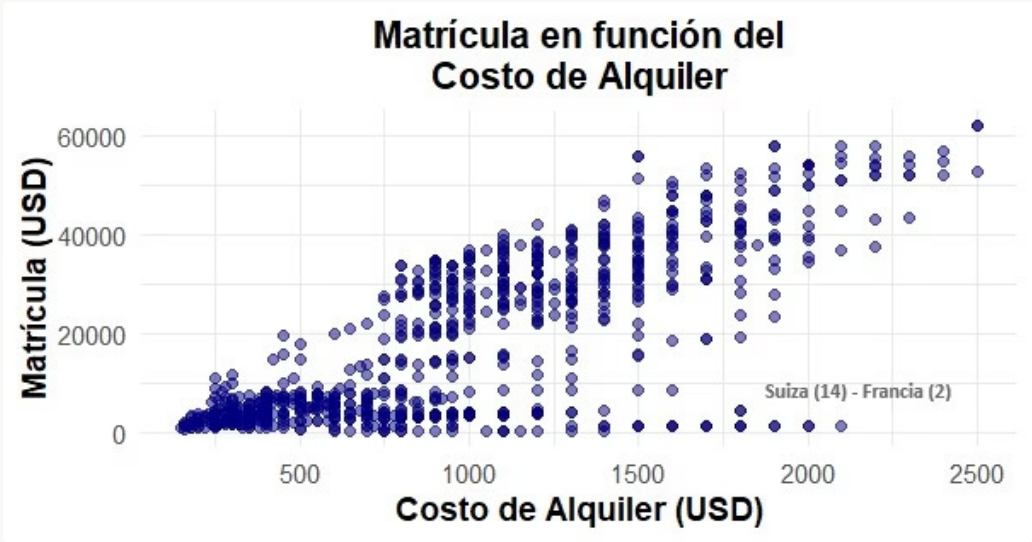




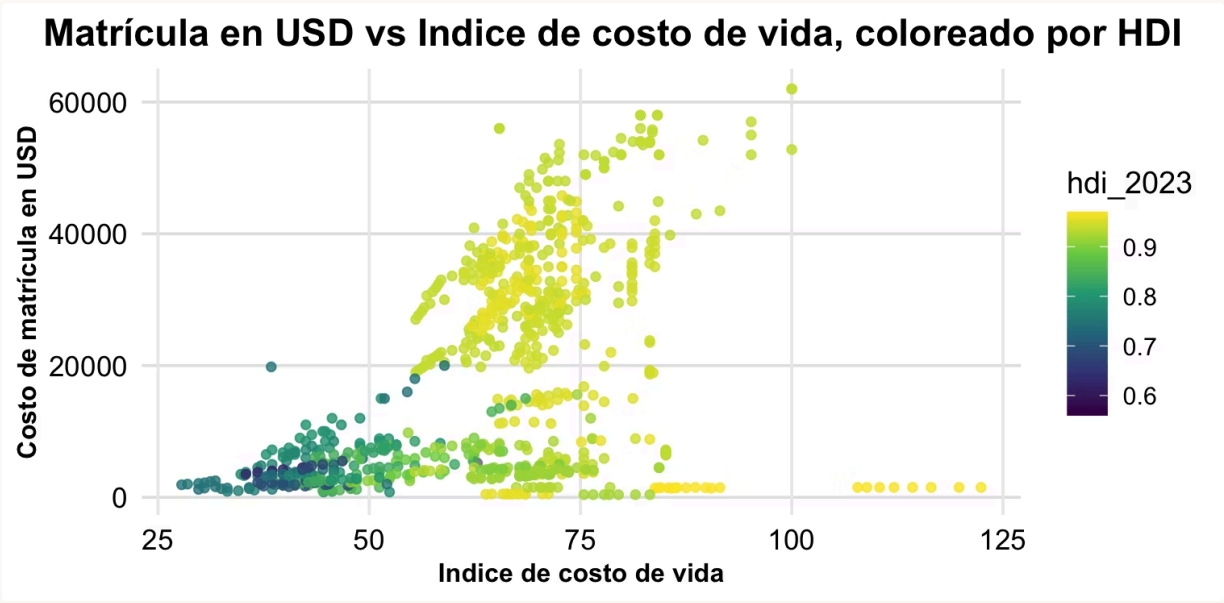
Identificamos un patrón en las regiones: A mayor costo de vida y alquiler mayores costos de matrícula. Las excepciones más importantes se ven en **Europa Occidental**, con costos altos y matrículas más económicas. Un patrón similar se encuentra en Asia, explicado por países como **Japón y Corea del Sur**.

Análisis específico y visualizaciones profundas



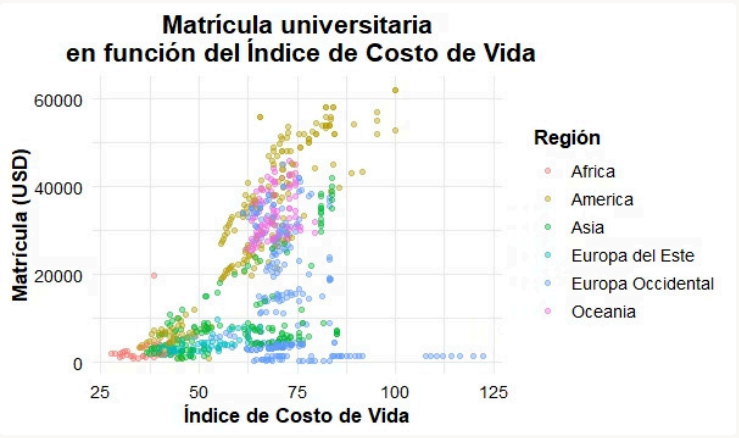


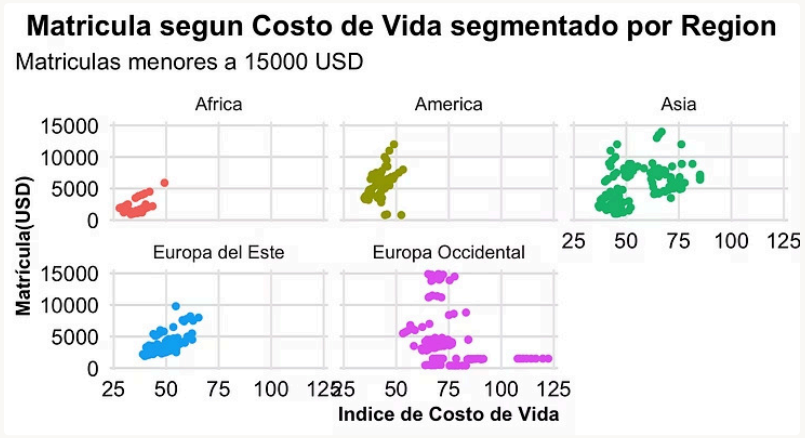
Analizando el comportamiento de las variables continuas identificamos que la tendencia se mantenía a nivel general, con algunas excepciones con altos costos de vida y alquiler y matrículas bajas.



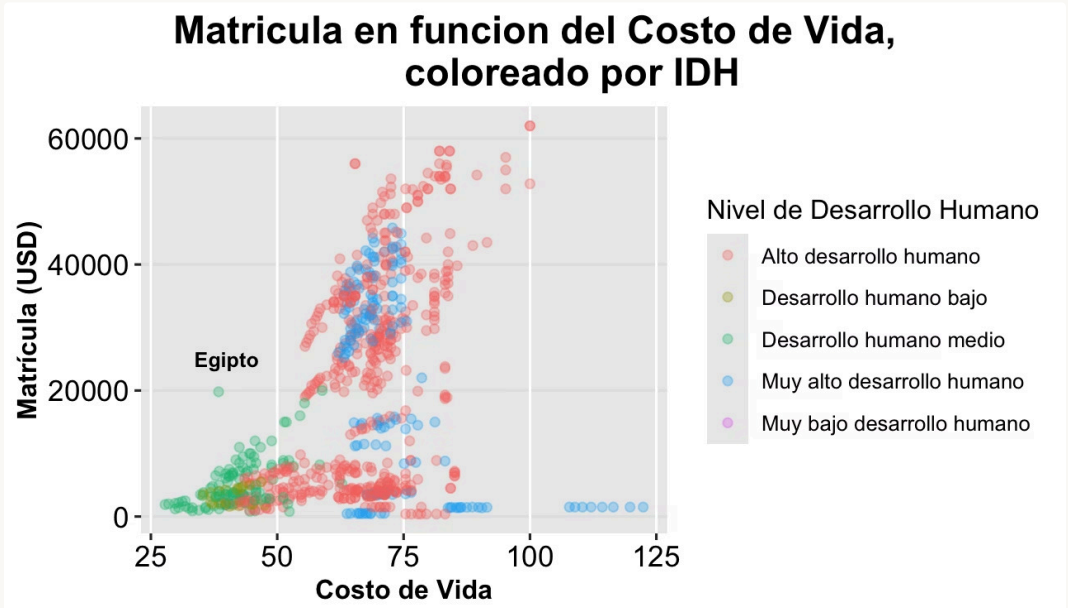
Acá decidimos importar la variable HDI (Indice de Desarrollo Humano) y utilizarla para colorear los puntos. Obtuvimos este gráfico que nos ayudó a identificar que los puntos con este comportamiento (costos altos y matrículas bajas) solían pertencer a países con alto **Indice de Desarrollo Humano.**

Análisis específico y visualizaciones profundas

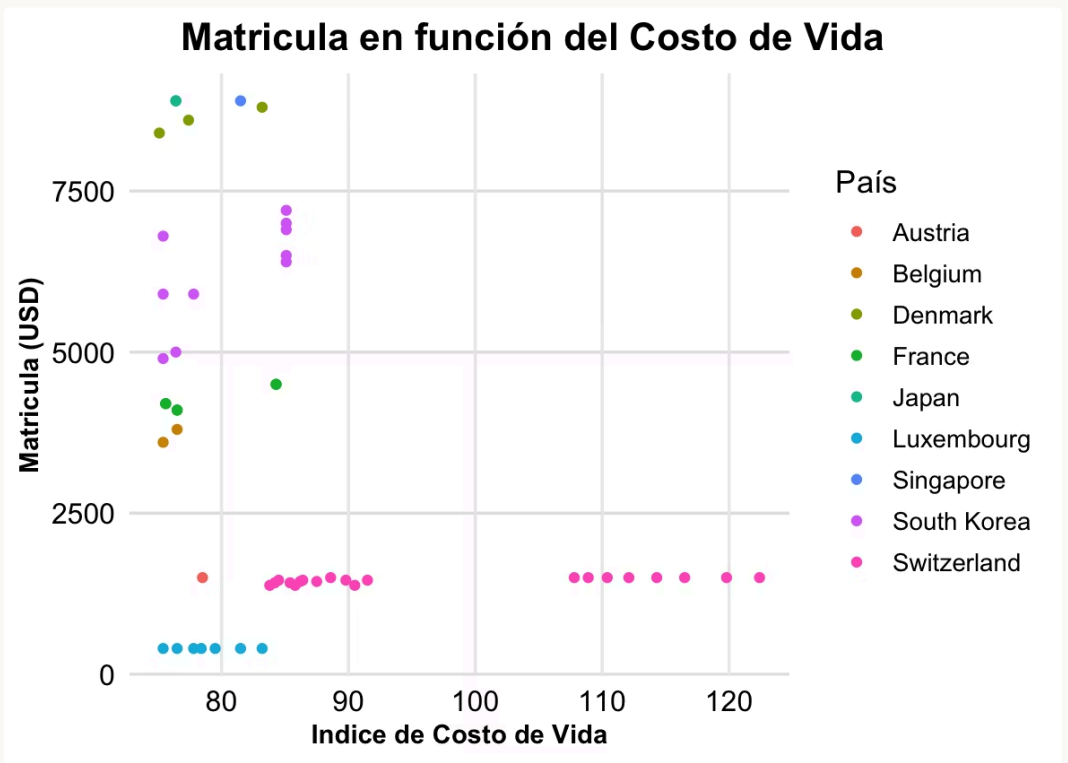




Con el gráfico coloreado por región pudimos corroborar claramente que la tendencia se corresponde con países de la región de Europa Occidental y algunos casos de Asia.



En este gráfico vemos como los países con "Alto desarrollo humano" suelen tener programas más altos en costo de matrícula que los países con "Muy alto desarrollo humano".



"Zoom" en los países con alto Índice de Costo de vida >75 y Costo de Matrícula <9000 USD.

Esto nos permite observar mejor una lista de las observaciones que presentan la excepción de alto costo y baja matrícula

Modelado: Modelo Simple

- Realizamos un modelo al que llamamos "simple" que busca predecir el comportamiento del costo de la matrícula de una manera más general y con un foco en los efectos directos de cada variable.

```
mod_simple = lm(Tuition_USD ~ poly(Living_Cost_Index,2) + region + Nivel_IDH + Rent_USD, data=df)
```

- El modelo ajusta el costo de la matrícula como una función cuadrática del índice de costo de vida, una función lineal de la renta, y considera los efectos principales de la región y el nivel de desarrollo humano (IDH), sin asumir que las relaciones entre estas variables cambian según las combinaciones de las otras.

$$Tuition\_USD = \beta 0 + \beta 1 \cdot LCI + \beta 2 \cdot LCI^2 + \beta 3 \cdot Rent + \beta EfectoRegión + \beta EfectoNivel\_IDH + \epsilon$$

- Métricas:

```
Residual standard error: 7959 on 790 degrees of freedom
Multiple R-squared:  0.7689,    Adjusted R-squared:  0.7654
F-statistic:   219 on 12 and 790 DF,  p-value: < 2.2e-16
```

Modelado: Modelo Complejo

- Realizamos un modelo al que llamamos "complejo" que logró predecir con bastante exactitud el comportamiento de las variables. Sin embargo, resultaba muy difícil de interpretar y explicar.

```
mod_complejo = lm(Tuition_USD ~ poly(Living_Cost_Index,2) * factor(Country) * factor(Level) + Rent_USD, data=df)
```

- El modelo ajusta el costo de la matrícula como una función cuadrática del índice de costo de vida, una función lineal de la renta, y permite que las relaciones varíen completamente según el país y el nivel educativo, incluyendo todas las interacciones posibles entre estas variables.

Métricas:

```
Residual standard error: 3510 on 437 degrees of freedom
Multiple R-squared:  0.9752,      Adjusted R-squared:  0.9543
F-statistic: 46.85 on 366 and 437 DF,  p-value: < 2.2e-16
```

Comparación entre modelos

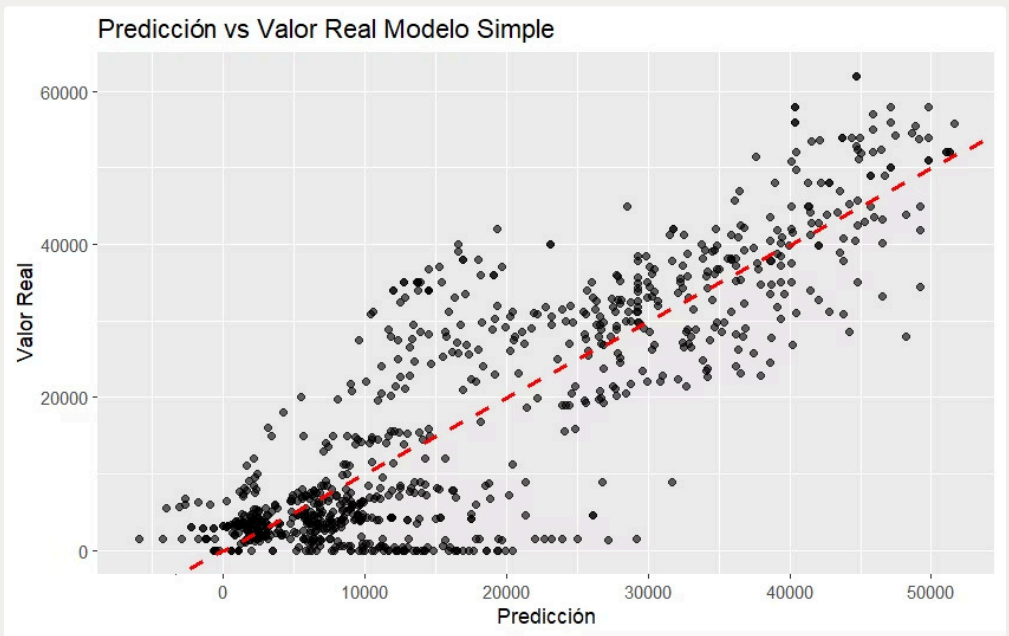
Analysis of Variance Table

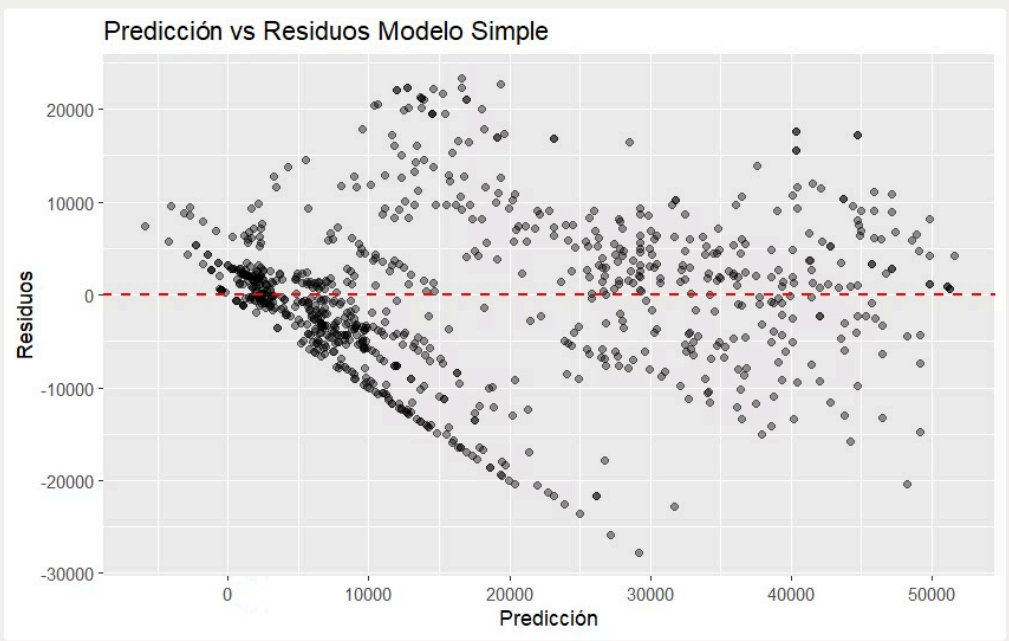
Model 1: Tuition_USD ~ poly(Living_Cost_Index, 2) + region + Nivel_IDH + Rent_USD
Model 2: Tuition_USD ~ poly(Living_Cost_Index, 2) * factor(Country) * factor(Level) + Rent_USD
  Res.Df   RSS Df   Sum of Sq   F   Pr(>F)
1   790 50801667752
2   437 5384979440 354 45416688311 10.411 < 2.2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

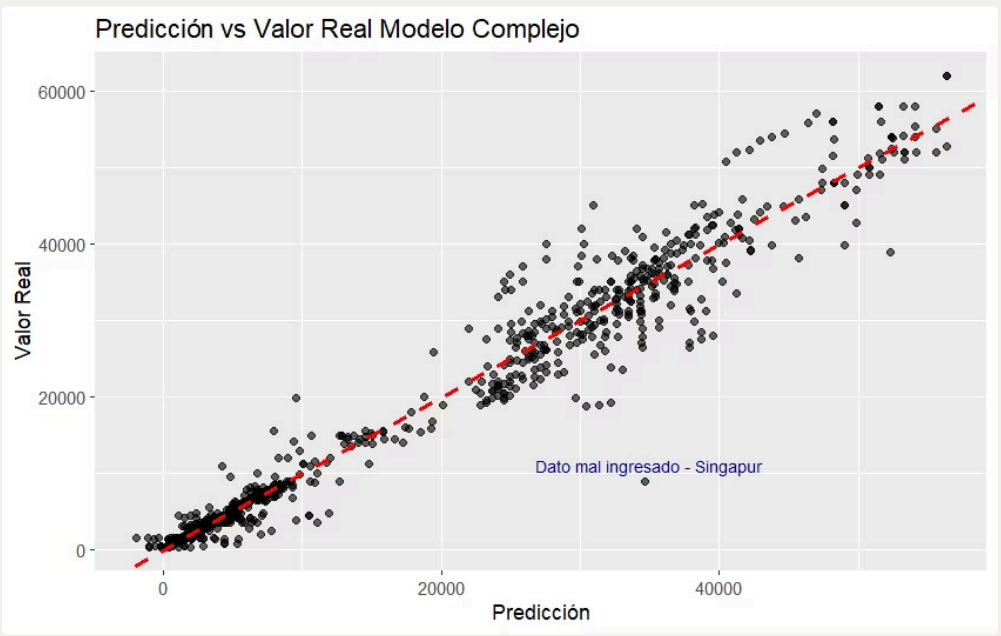
Comparativa de gráficos

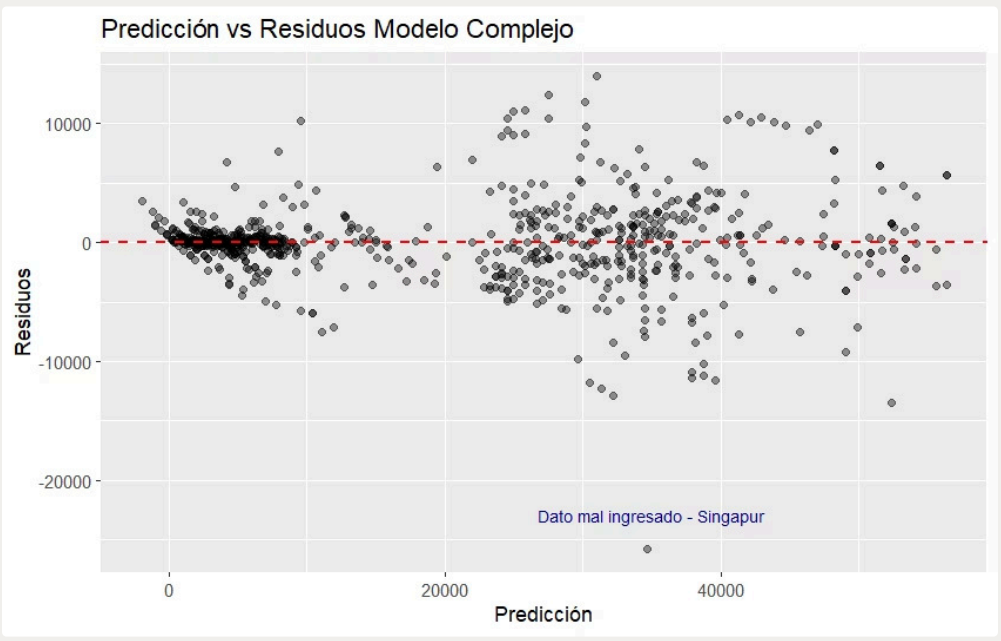
Modelo Simple





Modelo Complejo





Comparación de modelos

Característica	mod_simple	mod_complejo
Variable Categórica 1	region (Región geográfica)	factor(Country) (País específico)
Variable Categórica 2	Nivel_IDH (Nivel de Índice de Desarrollo Humano)	factor(Level) (Nivel educativo)
Términos de interacción	Sin interacción entre variables.	Todas las interacciones entre poly(LCI,2), Country y Level.
Parámetros	Menos parámetros: 14	Más parámetros: 367
Interpretación	Más fácil de interpretar; los efectos de cada variable se asumen aditivos e independientes.	Más difícil de interpretar directamente debido a las interacciones.
Grados de libertad	Más grados de libertad para el error.	Menos grados de libertad para el error ().
Factor explicativo principal	Se focaliza en los efectos promedio de costo de vida, renta, región y nivel de IDH sobre la matrícula, sin asumir que sus efectos varían según las combinaciones de otras variables categóricas.	Busca entender si las relaciones entre el costo de vida y la matrícula varían significativamente por país y nivel educativo, y cómo estos factores interactúan.

Conclusiones

Respondiendo a la pregunta inicial:

- Distintos factores explican el valor de la matrícula universitaria, los más **significativos** son:

1. **Costo de vida de la ciudad**
2. **Costo del alquiler de la vivienda**
3. **Nivel educativo**
4. **Nivel de IDH del país**
5. **Datos geográficos como Región / País**

- En líneas generales podemos expresar que hay una correlación positiva entre el costo de la matrícula universitaria con el costo de vida, alquiler y nivel de desarrollo humano del país, con la excepción de los países con IDH muy elevado (e.g. Suiza, Luxemburgo, etc.), que suelen ofrecer matrículas a costos muy bajos.
- Contamos con dos modelos que nos facilitan la predicción del costo de la matrícula universitaria, uno con mayor complejidad y otro más simple. Descartamos algunas variables como Category, Visa_Fee_USD e Insurance_USD.

Próximos pasos:

- Incluir datos macroeconómicos como porcentaje de reinversión en educación universitaria de los Estados o Porcentaje del financiamiento universitario que proviene del Estado , para entender si los bajos valores en matrículas tienen relación con estas variables.
- Segmentar y analizar con mayor detalle las regiones de los continentes relevantes (Asia y América) para evaluar posibles relaciones.
- Explorar otras técnicas de modelado que capten relaciones no necesariamente lineales.

FIN

¡ Gracias por su atención!