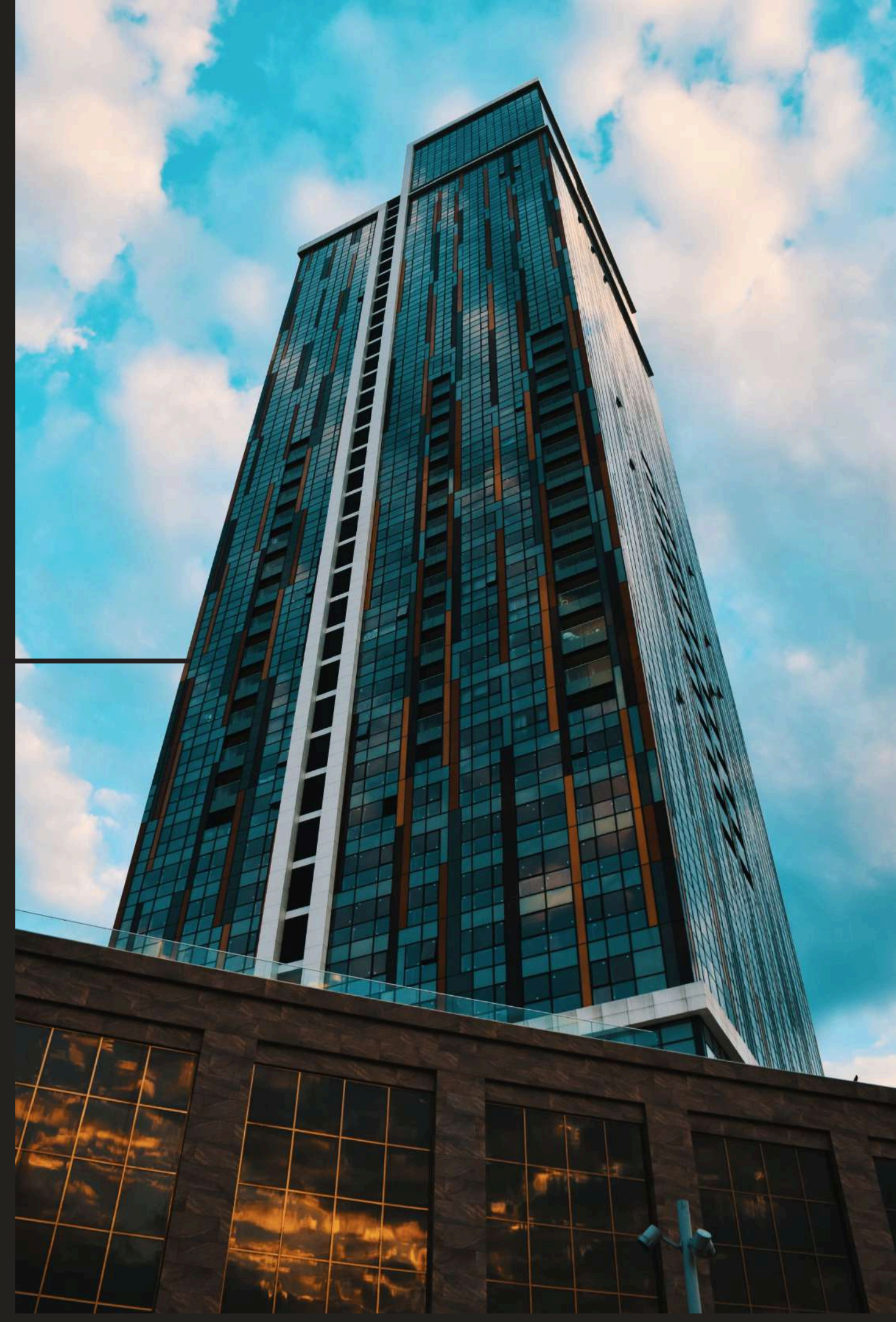




REAL ESTATE PRICE PREDICTION

Datos clave sobre la predicción de
precios de inmuebles en EE.UU





PROPUESTA DE VALOR

Propuesta de valor

Desde nuestro equipo de Data Science **Indrona Data Analytics System**, hemos querido abordar la siguiente casuística.

¿Cuál sería el valor adecuado de un inmueble según nuestro salario y el estado en el que queremos vivir?

Seguro que hay muchas inmobiliarias y brokers inmobiliarios preguntándose cómo poder asesorar mejor a sus clientes dependiendo de sus ganancias y su destino soñado.

Aquí es cuando entramos en acción presentando un modelo de machine learning específico para cubrir esta necesidad.





INTELIGENCIA ARTIFICIAL EN REAL ESTATE

Cómo usar la IA en Real Estate

Para tener una mayor eficiencia a la hora de aplicar la inteligencia artificial en el sector inmobiliario, deberemos de tener en cuenta un **enfoque transversal**.

Debe afectar a todos los aspectos involucrados en el proceso de compra-venta de un inmueble.





ATENCIÓN AL CLIENTE

Necesitamos tener un control y seguimiento de nuestros clientes. Es recomendable pilotar todos sus datos y origen del contacto a través de una sola plataforma que nos facilite:

- Att. cliente
- Mensajes / Mails
- CRM
- Producto recomendado
- Customer Journey Map



ANÁLISIS DE DATOS

Necesitamos tener un análisis de los datos que obtenemos a través de nuestro CRM, Web, RR.SS, etc.

Así podemos tener un perfil más detallado del cliente.



AUTOMATIZACIÓN DE TAREAS

Cuando disponemos de diferentes herramientas que nos facilitan y automatizan las acciones de marketing, podemos enfocarnos mucho más en la prospección y compra-venta de inmuebles.

- Mails
- RR.SS
- Web
- Campañas Marketing
- SEO



EXPOSICIÓN PRODUCTO

Con el trabajo en equipo en tiempo real, podrás dividir las tareas y trabajar a la vez para crear una presentación impactante.



BÚSQUEDA DE DATOS

Búsqueda de Datos

Como bien es sabido, EE.UU es la meca del Real Estate, declarándose como el mayor referente mundial en el sector.

Como hemos querido especializarlo en el mercado estadounidense, hemos sacado la información a través de webscraping de las siguientes fuentes:

<https://data.census.gov/> <https://livingwage.mit.edu/states>

<https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset>

<https://www.kaggle.com/datasets/brandonconrady/living-wage-top-100-cities>

<https://www.kaggle.com/code/codename007/dangerous-places-in-usa-2013-2018>

<https://livingwage.mit.edu/>



Búsqueda de Datos

En total hemos encontrado 3 Datasets con la información que necesitábamos para este proyecto.

Hemos juntado los 3 Datasets realizando un *merge* y, del cuál, han salido como resultado 2.000.000 de datos para ser tratados.

Hemos utilizado recursos como Selenium y Cromedriver.exe



An aerial photograph of a city skyline, likely Dubai, featuring numerous skyscrapers and a body of water. The image is overlaid with a teal color filter and white wavy lines in the top-left and bottom-right corners. A white rectangular frame is centered over the text.

EXPLORATORY DATA ANALYSIS

Análisis de variables categóricas

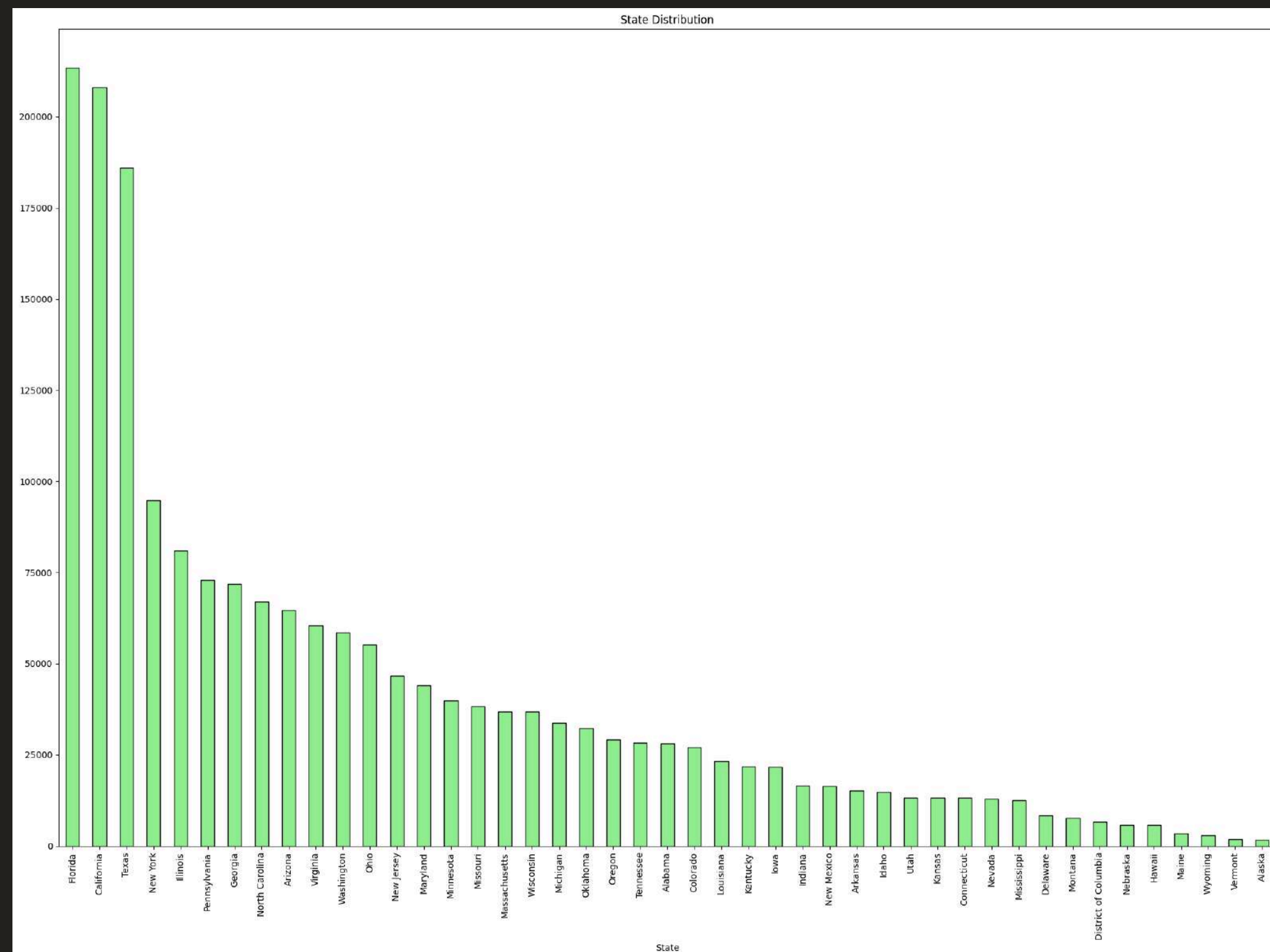
En este gráfico podemos observar dónde se sitúan la mayoría de inmuebles registrados.

principalmente son en los siguientes 3 estados:

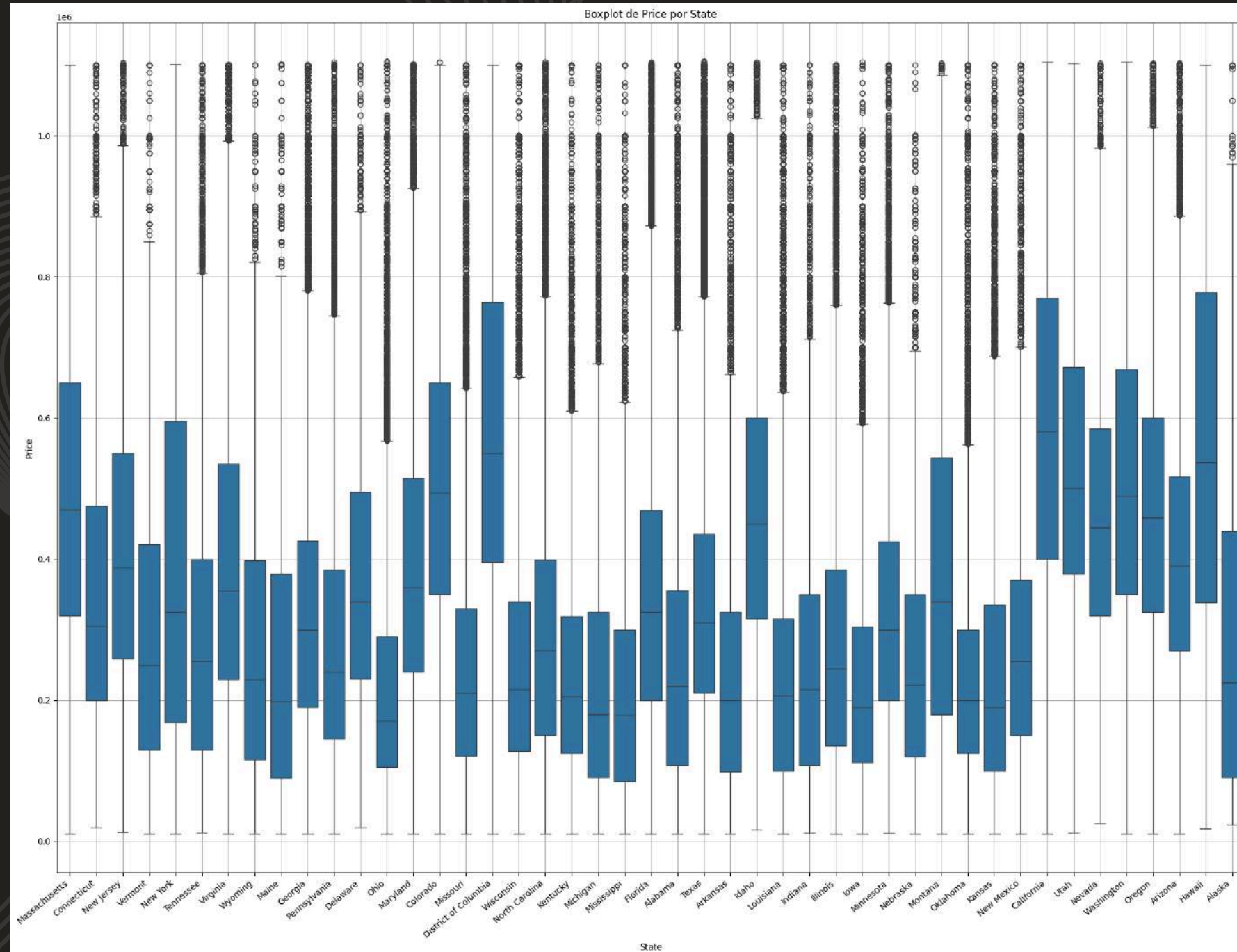
- FLORIDA
- CALIFORNIA
- TEXAS

No es de extrañar que el primero en el ranking sea Florida ya que es considerada la meca del Real Estate en EE.UU.

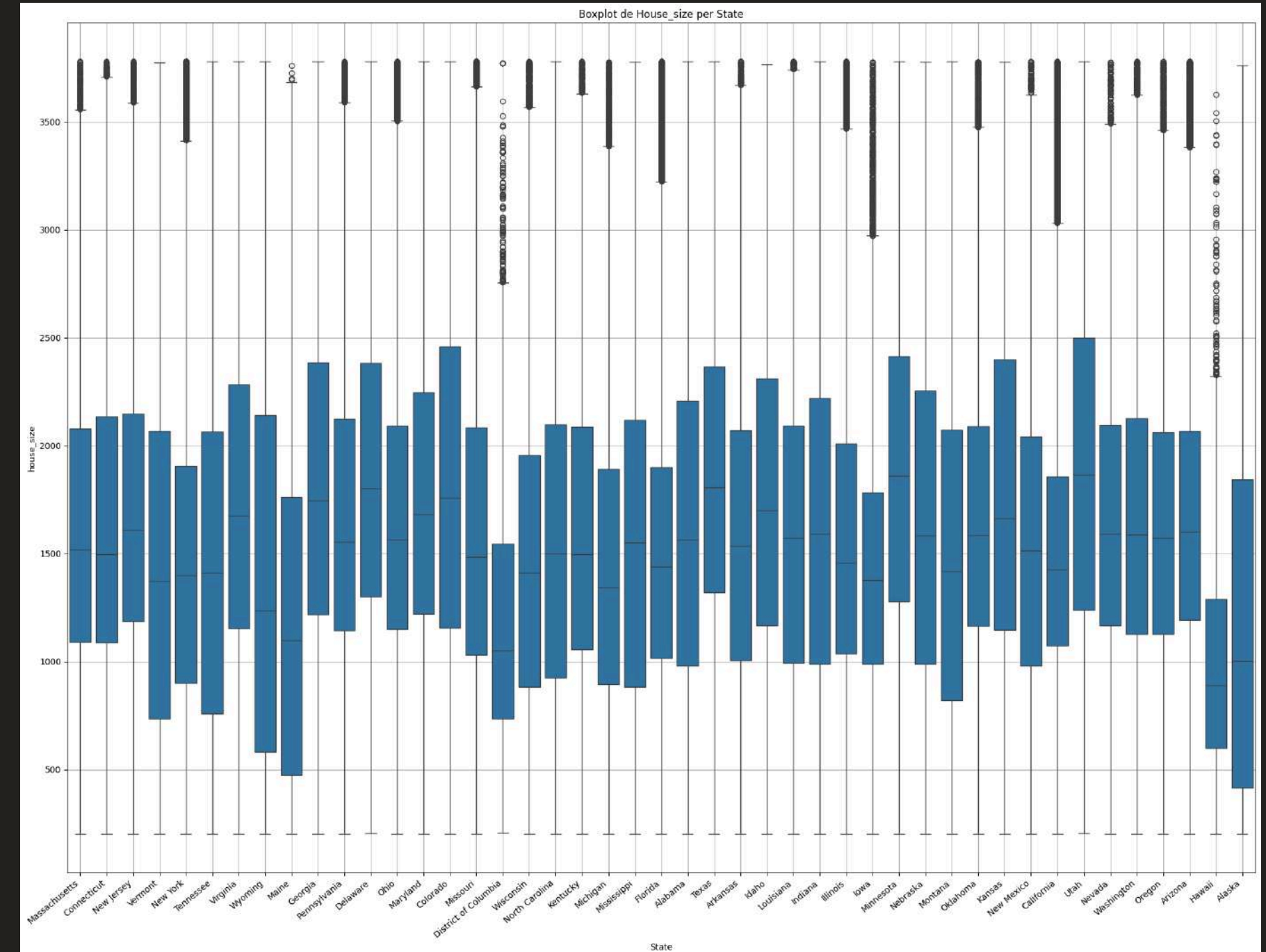
Por otro lado California y Texas son de los estados más grandes a nivel territorial y, donde se concentra gran parte de la población que habita en el país.



Análisis de variables numéricas



1. PRICE DISTRIBUTION



2. HOUSE_SIZE DISTRIBUTION

Análisis de variables numéricas

1. Price Distribution

En general, la mayoría de los estados tienen precios que se agrupan en valores relativamente bajos, pero hay algunos estados que muestran una variabilidad significativa con algunos precios extremadamente altos, que se representan como outliers.

La escala del eje Y es bastante alta, alcanzando hasta valores de $1e8$ (100 millones), lo que indica que algunos inmuebles pueden estar valorados en decenas o cientos de millones de dólares.

2. House_size Distribution

El gráfico indica una gran desigualdad en los precios por pie cuadrado dentro de la mayoría de los estados. Aunque la mayoría de las propiedades están agrupadas dentro de un rango bajo, hay propiedades que son considerablemente más caras, lo que se refleja en la presencia de numerosos outliers.

Los outliers podrían estar representando áreas de lujo o zonas urbanas con precios extremadamente altos, como Manhattan en Nueva York o Beverly Hills en California.

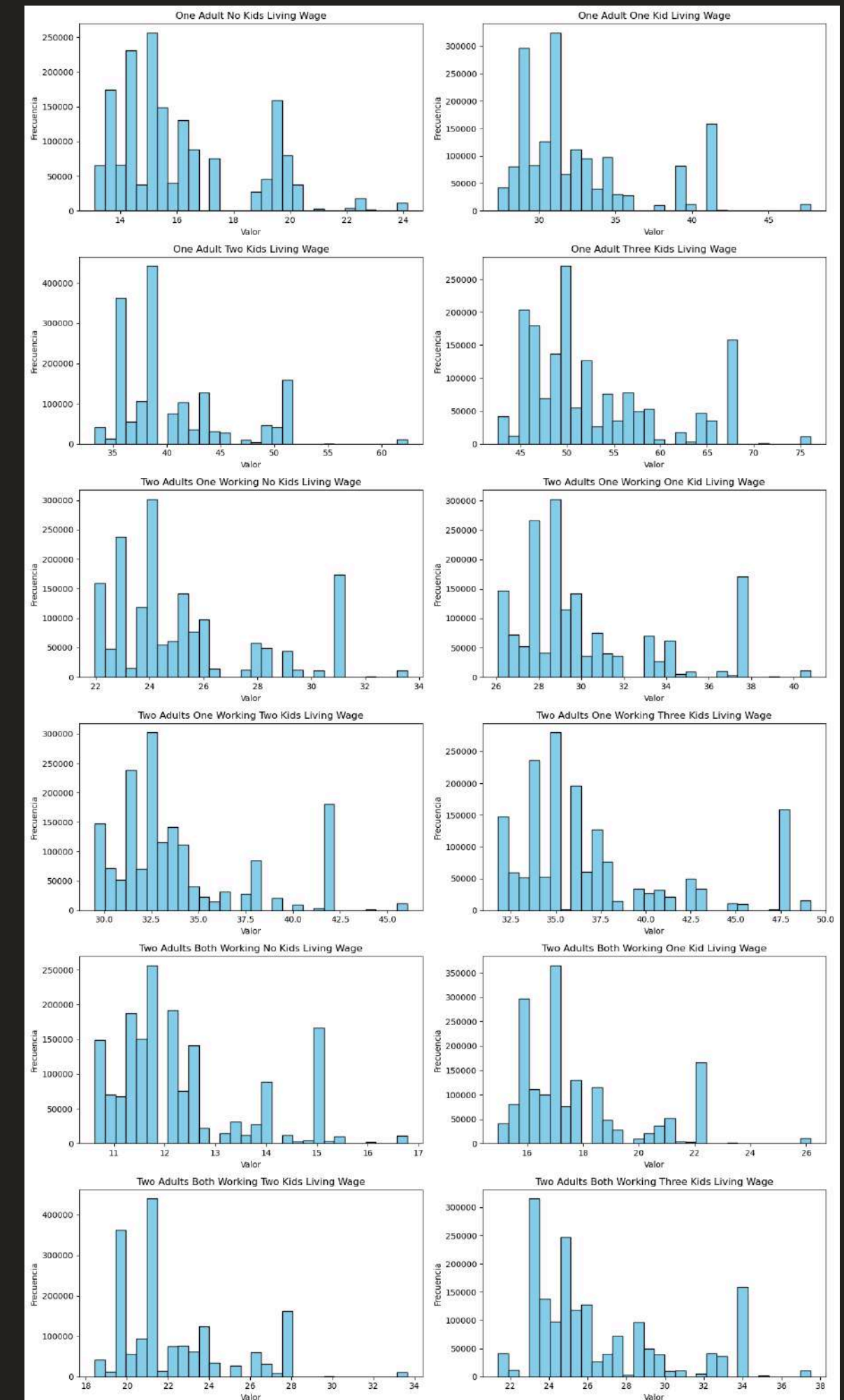
Análisis de variables numéricas

Living wage

La necesidad de un salario digno es menor cuando ambos adultos están empleados.

Los niveles de salario digno son más altos cuando solo uno de los adultos trabaja, reflejando la carga financiera adicional.

Los datos refuerzan la importancia de considerar la estructura familiar y el número de personas que contribuyen al ingreso del hogar al evaluar las necesidades salariales.





MODELADO: ENTRENAMIENTO Y DECISIÓN

Modelado: Entrenamiento y Decisión

Decisión Final sobre el Modelo: XGBoost

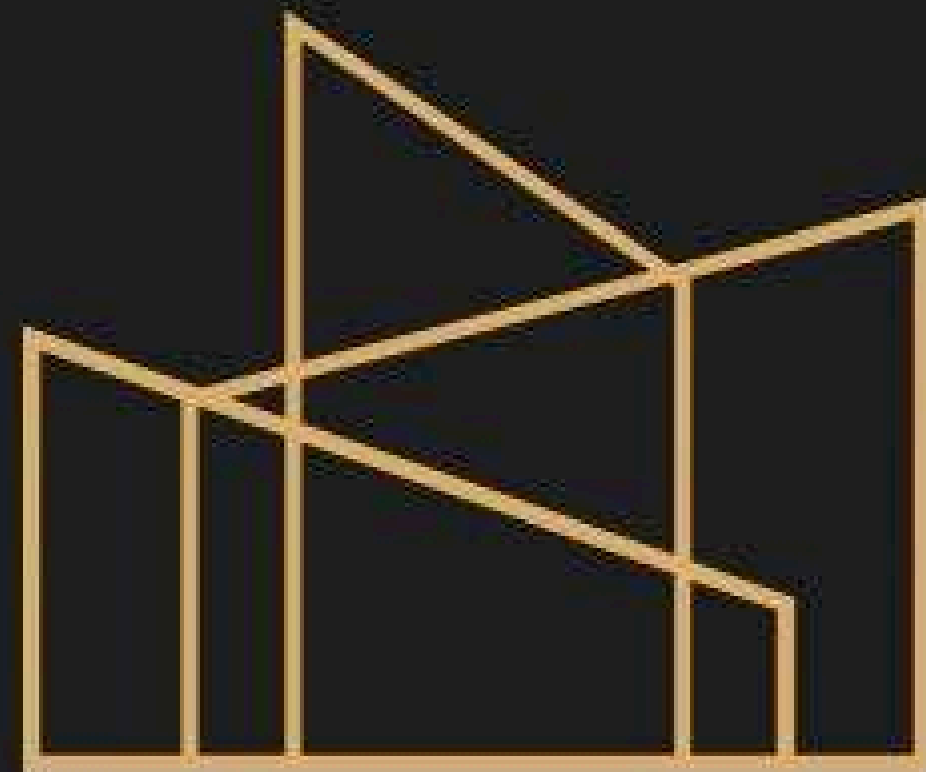
Columna1	Evaluating Linear Regression	Evaluating Lasso	Evaluating Ridge	Evaluating Random Forest	Evaluating XGBoost
RMSE Train	0,4869	0,5525	0,4869	0,1270	0,3774
RMSE Test	0,4864	0,5524	0,4864	0,3177	0,3794
MAE Test	0,3547	0,4186	0,3548	0,2052	0,2714
R ² Score Train	0,6249	0,5170	0,6249	0,9745	0,7747
R ² Score Test	0,6252	0,5167	0,6252	0,8401	0,7720

Durante la evaluación de los modelos, se observó que aunque **Random Forest** presentó un rendimiento superior en términos de **RMSE** y **R²**, también mostró ciertas desventajas significativas que afectan su aplicabilidad en el problema actual:

1. Inestabilidad
2. Sobrefit (sobreajuste): Con un **R² Train** de 0.974 y un **R² Test** de 0.840
3. Tiempo de Modelado

Ventajas de XGBoost:

- **Mejor estabilidad**: Menos propenso al sobreajuste.
- **Menor tiempo de entrenamiento**: Lo que facilita la optimización del modelo.
- **Rendimiento robusto**: Aunque el **R² Test** (0.772)



INDRONA

DATA ANALYTICS
SYSTEM

AGRADECIMIENTOS

DANIELA BLANCO

GONZALO C.B.

Gerente General