Análisis de bienes raíces en EU

Autores:

Hugo Moran

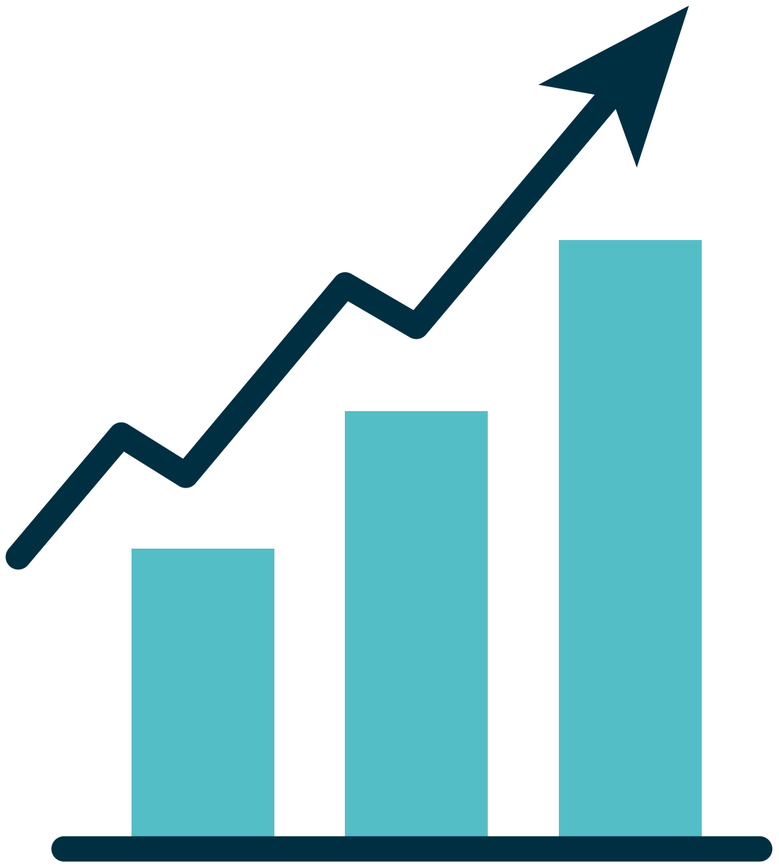
Científico de datos Sr.

Antonio Aguilar

Científico de datos Jr.

Contenido

Introducción

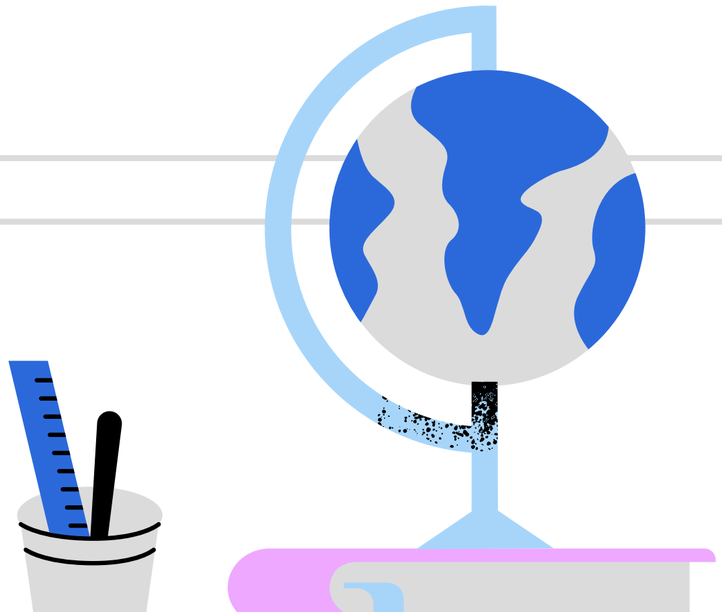


En general el comportamiento que ha tenido este mercado ha sido estable, salvo en algunos períodos excepcionales (Crisis 2007-2008), lo cual convierte a este mercado en uno de los más atractivos para invertir.

No es por nada que la riqueza de una buena parte de los millonarios en Estados Unidos provenga de esta industria. Por esta y otras razones, hacer un buen análisis de este mercado es bastante fructífero en muchos sentidos.

Objetivo

Modelo de valuación de casas



Alcance

El alcance de este proyecto se extiende a trabajar con datos que se obtuvieron de diversas instituciones de forma gratuita.

No contamos con datos actuales, por tanto, nuestra meta se enfoca más en encontrar características sociodemográficas que ayuden a predecir el precio de un inmueble, trazando así un camino que pueda generar una metodología que pueda ser aplicada a datos actualizados y pueda ser puesta en producción.

Inicio

1998

Ventana de tiempo

Fin

2018

Marco Estratégico

BASE DE DATOS

Creación de bases de datos con características relevantes

INSPECCIÓN DE DATOS

Limpieza de datos, análisis exploratorio y preprocesamiento

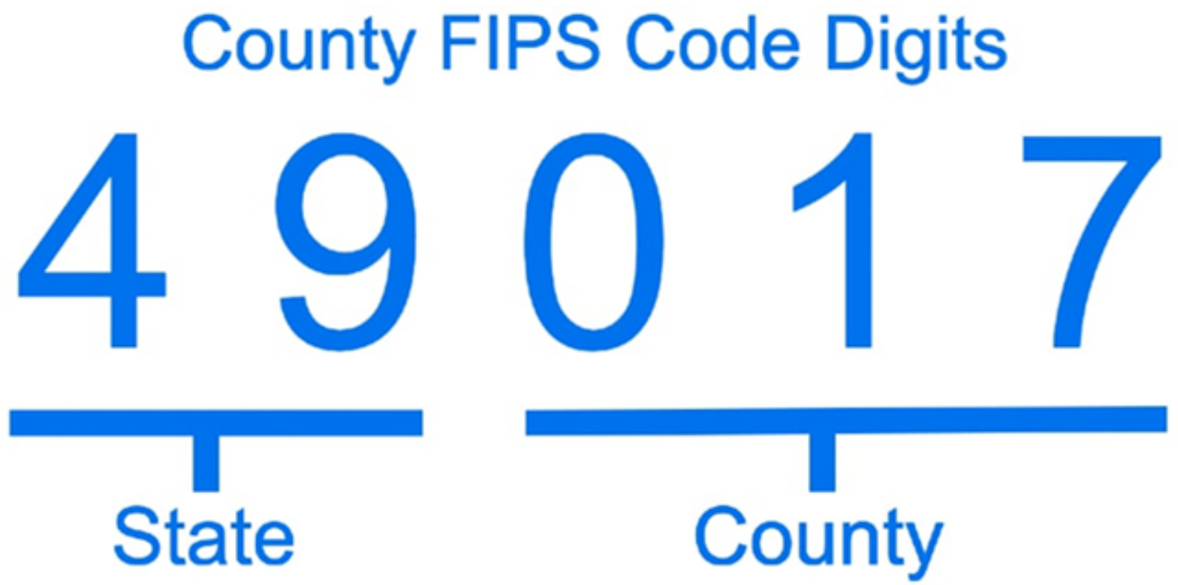
MODELADO

Prueba, evaluación y comparación de modelos

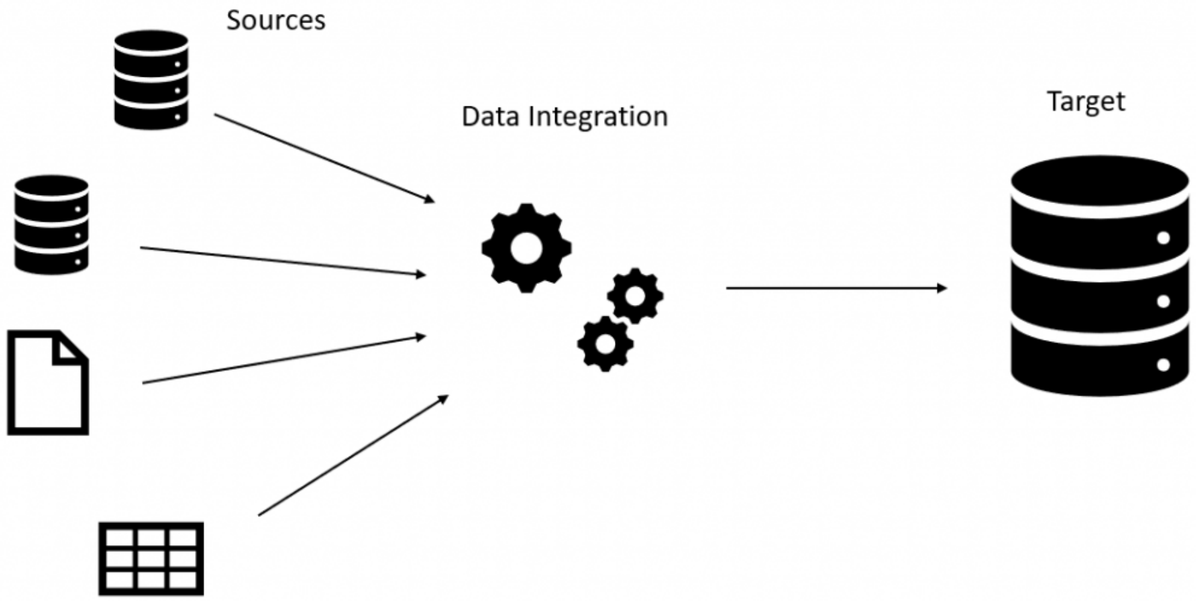
Definiciones

FIPS: El código FIPS tiene como finalidad la de representar numéricamente los condados de Estados Unidos.

Los primeros dos dígitos hacen referencia al estado y los últimos tres dígitos hacen referencia al condado en cuestión.



Datos



- Ratio de desempleo
- Hospitales
- Ranting de hospitales

- Escuelas publicas
- Ratio de crímenes
- Empresa Zillow

Descripción de datos

Etiquetas de fila	Cuenta de CountyName
Alabama	9
Alaska	2
Arizona	8
Arkansas	8
California	34
Colorado	12
Connecticut	8
Delaware	2
District of Columbia	1
Florida	36
Georgia	29
Hawaii	3
Idaho	5
Illinois	16
Indiana	16
Iowa	8
Kansas	9
Kentucky	8
Louisiana	9
Maine	1
Maryland	15
Massachusetts	10
Michigan	10
Minnesota	10
Mississippi	5
Missouri	8
Montana	2
Nebraska	2
Nevada	2
New Hampshire	4
New Jersey	20
New Mexico	4
New York	20
North Carolina	28
North Dakota	1
Ohio	18
Oklahoma	9
Oregon	12
Pennsylvania	23
Rhode Island	4
South Carolina	12
South Dakota	1
Tennessee	16
Texas	16
Utah	7
Virginia	27
Washington	12
West Virginia	5
Wisconsin	6
Wyoming	2
Total general	**535**

Número de estados

50

Número de condados

535

2.2 Desarrollo

Limpieza

Talend es una herramienta popular que ayuda a realizar el proceso de extracción, transformación y limpieza. Es "user-friendly"

01

Entradas inválidas

Detección de entradas inválidas en ciertos registros

02

Valores faltantes

Detección de valores faltantes en la columna de rating de hospitales

03

Perfilado numérico

Se dejó a dos decimales los valores flotantes

Integración, Limpieza y Exploración

Python es un lenguaje de programación con el cual es posible manejar grandes cantidades de datos de manera fácil y rápida

01

Eliminación de columnas

Borrado de columnas innecesarias

02

Unión de tablas

Se realizó la unión de tablas mediante la columna FIPS

03

Visualización de gráficos

Se realizó la visualización de estadísticos relevantes

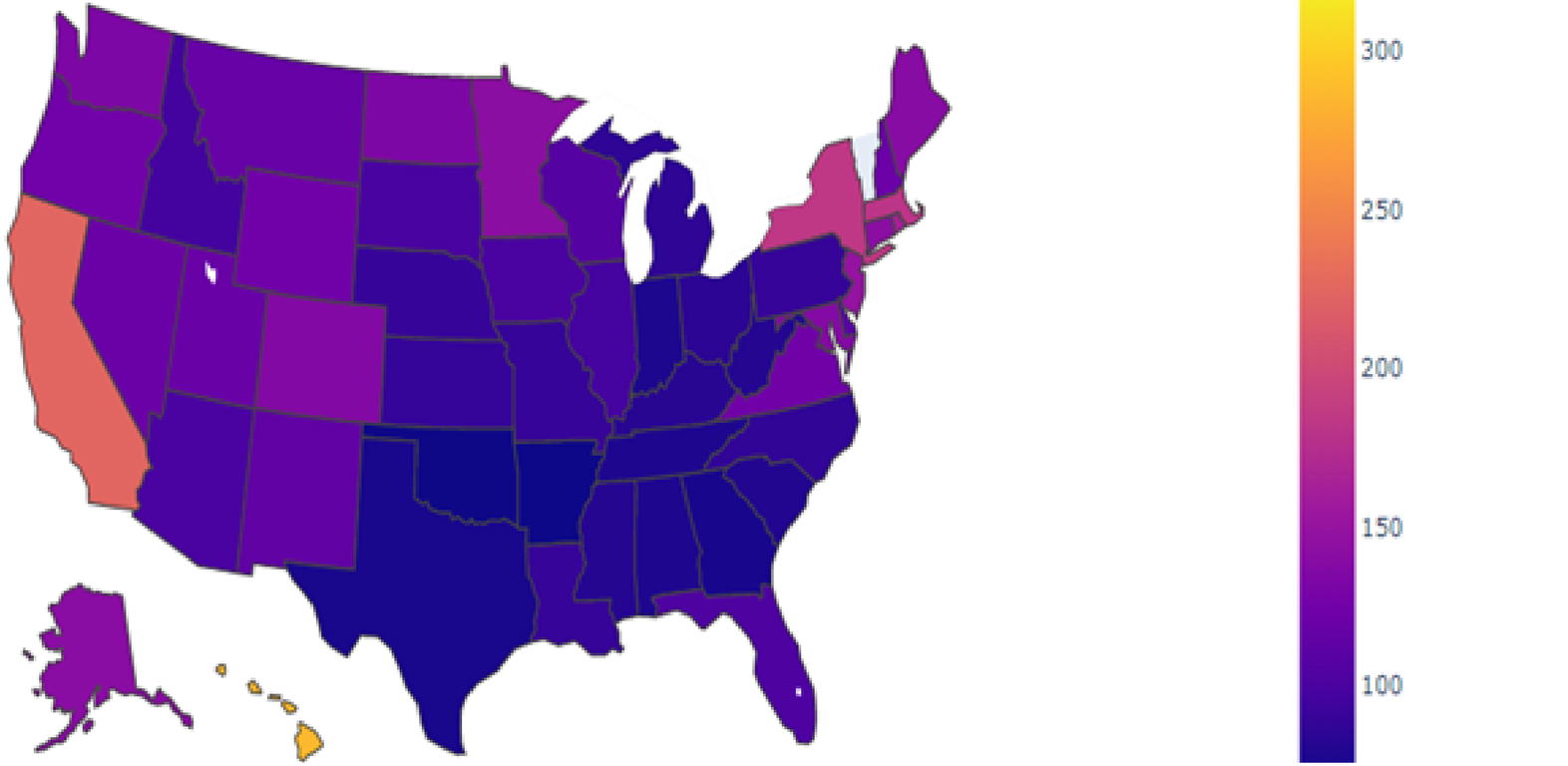
Base de Datos Integrada y Limpia

	FIPS	CountyName	StateName	NumberOfSchools	NumberOfHospitals	AverageHospitalRating	UnemploymentRate	crime_rate_per_100000
0	1001	Autauga	Alabama	15	1	4.00	3.6	251.601926
1	1003	Baldwin	Alabama	47	4	3.00	3.6	228.086325
2	1069	Houston	Alabama	31	5	3.50	4.1	401.281012
3	1073	Jefferson	Alabama	226	14	2.75	3.7	798.357491
4	1081	Lee	Alabama	37	1	4.00	3.6	246.466975

AverageHousePrice	ZHVIPerSqft_AllHomes	MedianRentalPrice_AllHomes
114483.67	72.56	1255.21
164861.69	90.39	1165.94
114941.00	72.94	975.63
99759.77	68.25	855.86
148163.22	74.79	1143.89

Exploración

Valor promedio por pie cuadrado de las casas en EU (Dólares)

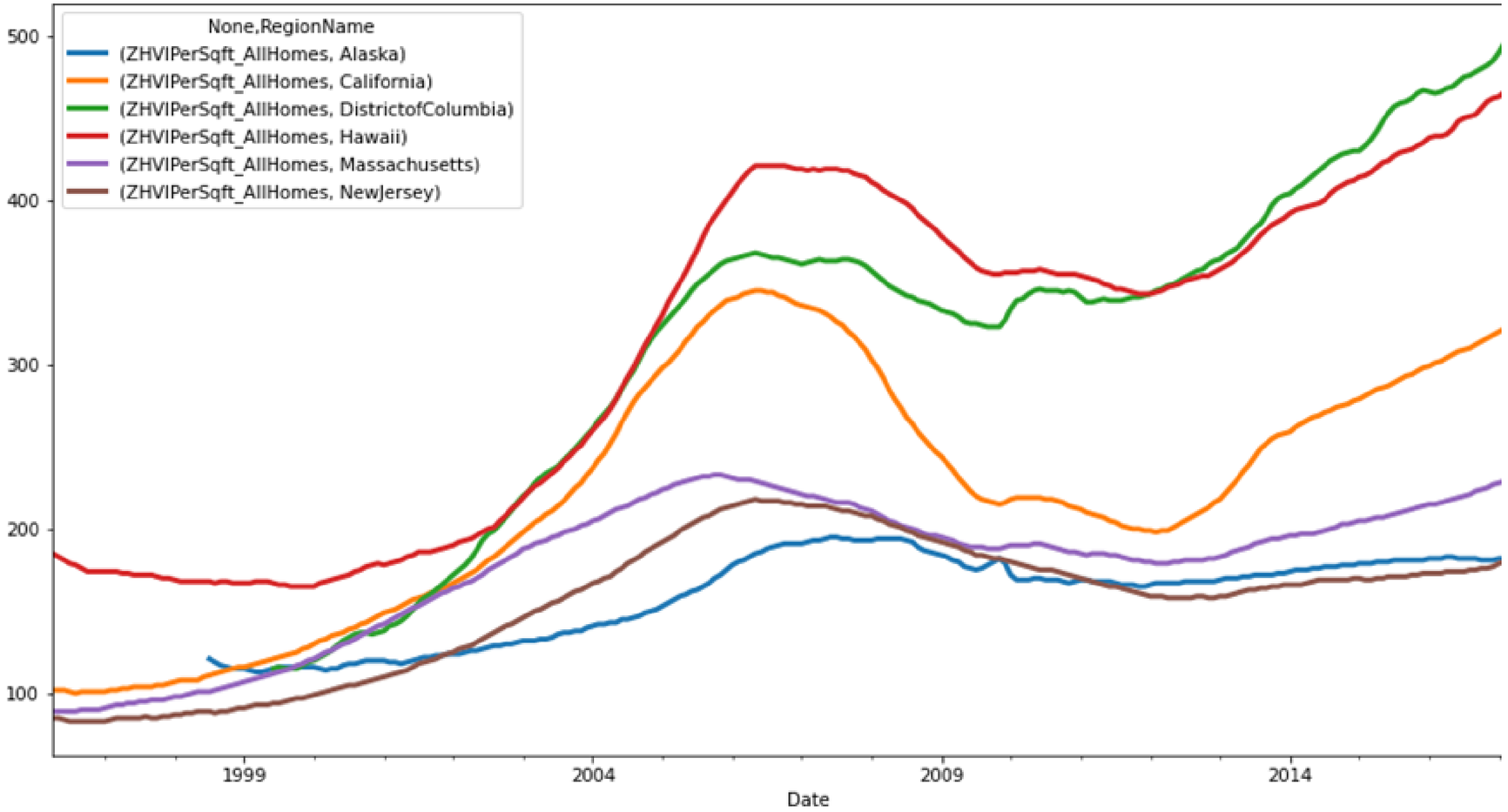


OBSERVACIONES:

- Los estados del sureste del país tienen menor valor comparados con los estados del noreste
- Hawái tiene el valor más grande de todos los estados, seguido de california

Exploración

Serie de tiempo, precio por pie cuadrado para el top de estados más caros (Dólares)

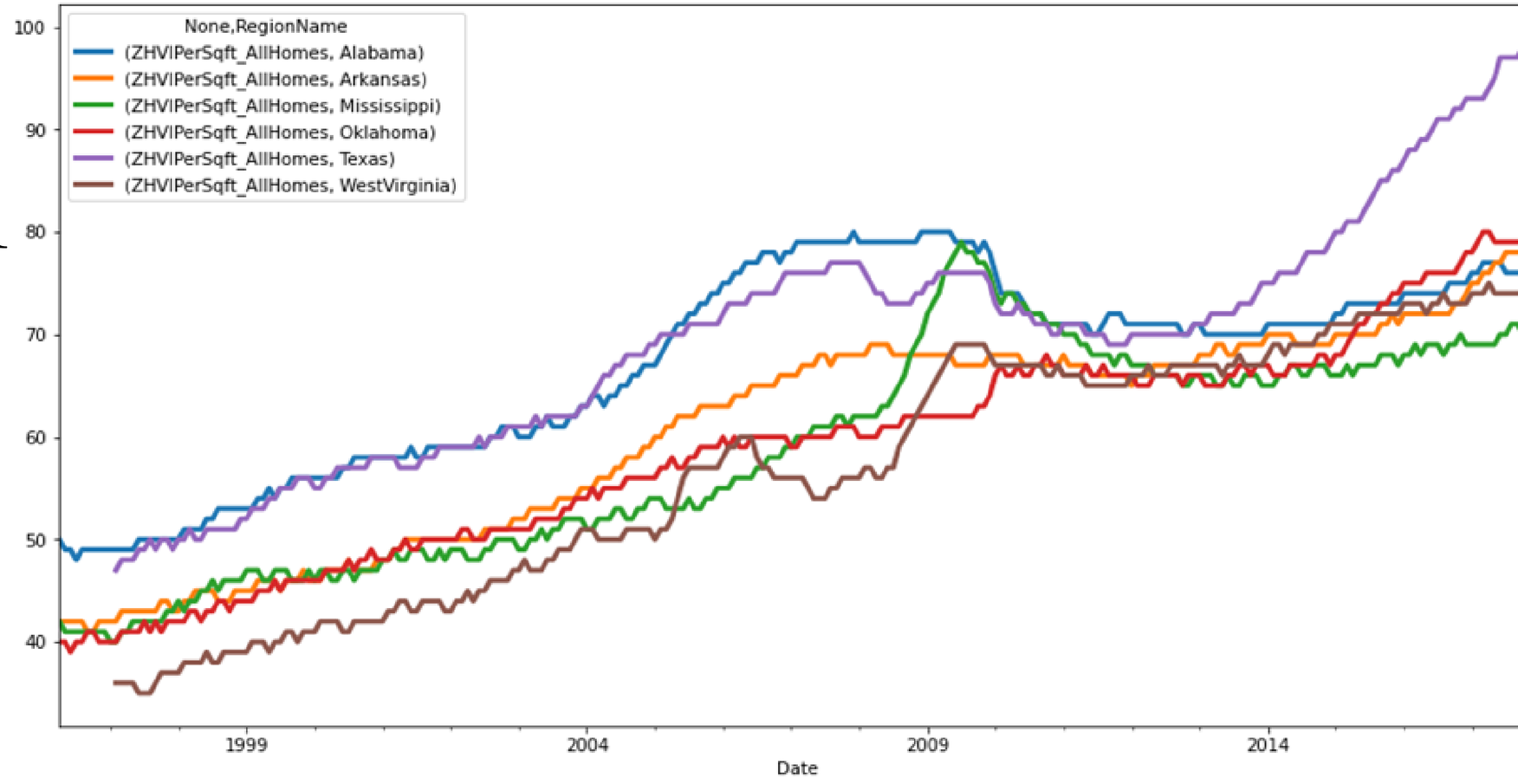


OBSERVACIONES:

- Se nota la crisis inmobiliaria (2007) y como después del 2012 se empieza a recuperar los precios
- Al estado de New Jersey le ha costado más recuperarse de esa crisis

Exploración

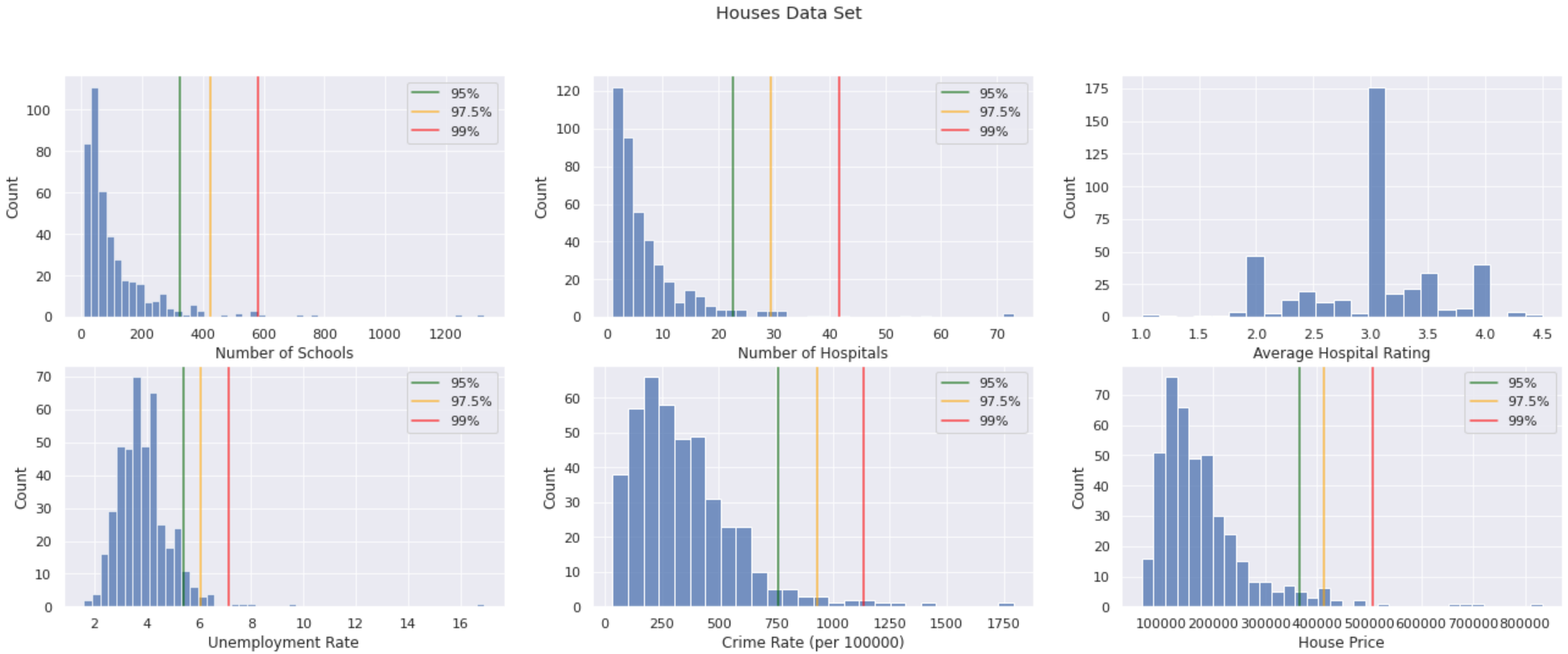
Serie de tiempo, precio por pie cuadrado para el top de estados más baratos (Dólares)



OBSERVACIONES:

- Ocurre lo contrario al caso anterior, durante la crisis tienen un aumento significativo en el precio
- La caída de precio tiene un desfase en comparación con la gráfica anterior

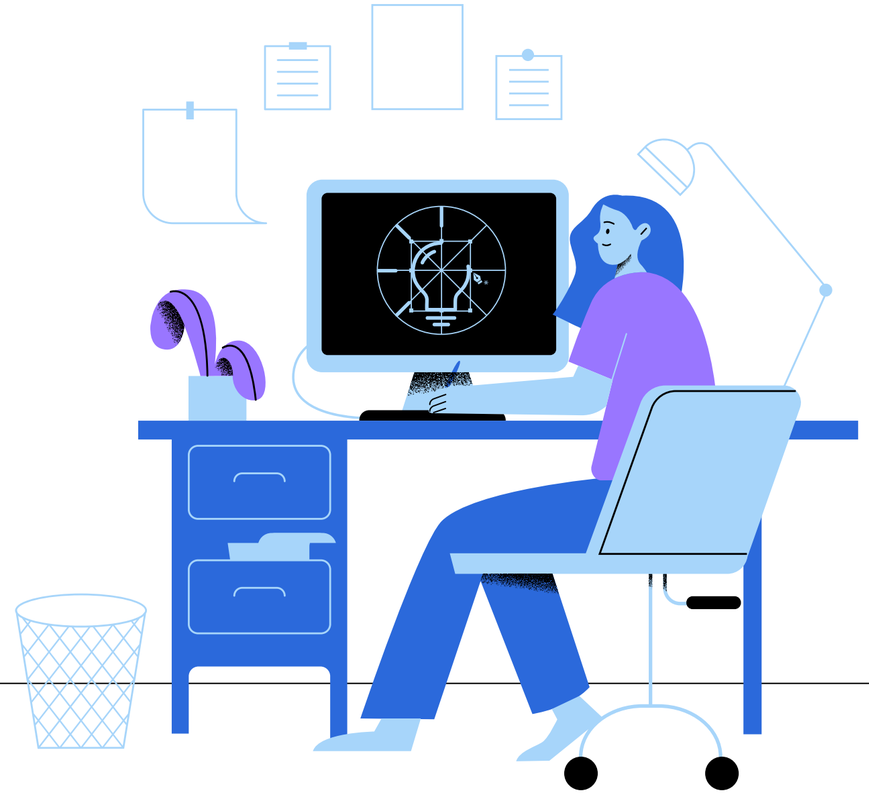
3.1 Modelos



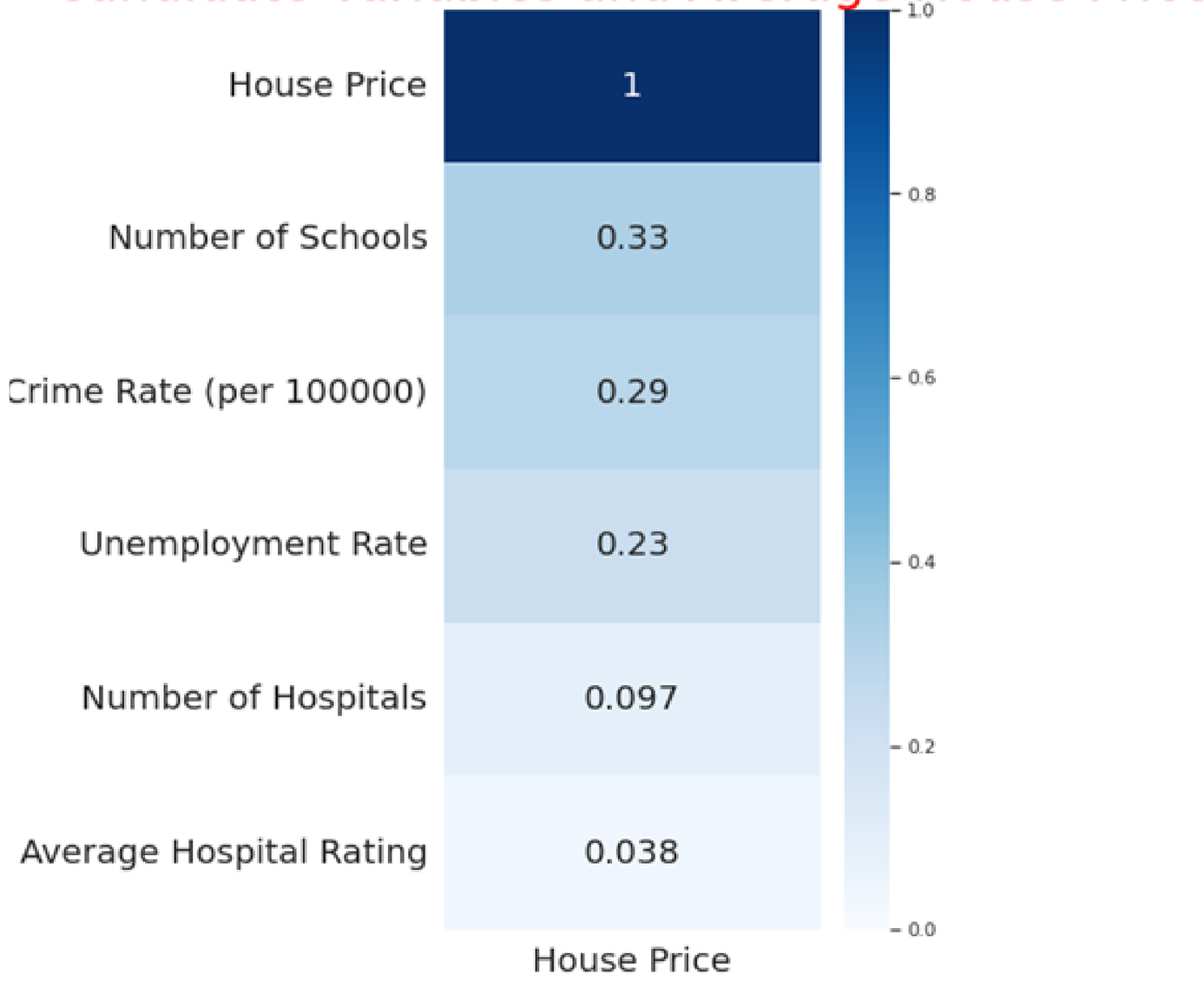
Inspección de Valores atípicos y Distribuciones

Las líneas verticales de color verde, amarillo y rojo representan los cuantiles al 95%, 97.5% y 99% respectivamente

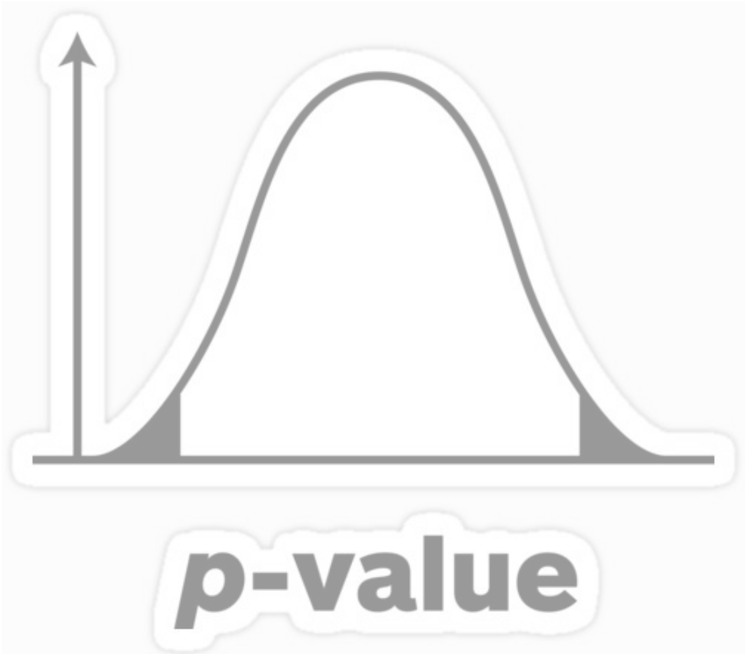
Análisis de correlación



Absolute Correlation Between Candidate Variables and Average House Price



Análisis de la variable Estado



H0: La variable del **estado NO** contiene información relevante respecto al precio de las casas

172.25

Estadístico

8.8 e-17

P-valor

Por lo tanto, hay evidencia estadística que indica que la variable de estado contiene información que nos intersa

¿Qué modelos consideraremos?

Veamos los modelos que usamos...

Regresión Lineal

La regresión lineal es un modelo clásico en este tipo de problemas sin mencionar la simplicidad que ofrece durante la implementación e interpretación

Árbol de decisión

Los árboles de decisión generalmente destacan en problemas de clasificación, sin embargo, tambien suelen utilizarse en la regresión

XGBoost

Arboles de decisión "boosteados" por gradientes

Resultados

Modelo	R^2	MSE
Regrsión Lineal	0.34	0.158
Árbol de decisión	0.18	0.18
XGBoost	0.554	0.102

- El modelo de árbol de decisión tuvo el peor desempeño con respecto al r^2-score

- Aunque el XGBoost se basa en áboles de decisión, gracias a sus parámetros y de su función objetivo, obtiene un mejor resultado

- Por último la regresión lineal se lleva el segundo lugar, encontrándose justo a la mitad de los otros dos modelos

Conclusiones

Se exploró una metodología para tratar de predecir los precios de inmuebles en EU, si bien no se obtuvieron grandes resultados, se trazó el camino para seguir explorando más características, que puedan aportar información a los modelos

Para trabajo a futuro, pensamos en conseguir la información del ingreso promedio por condado, esto nos ayudará a comprender que tanto "presupuesto" tienen las personas de ese lugar para invertir en una casa.