



Análisis y predicción de tsunamis por año

Auto Arima



ÍNDICE

1. Objetivo
2. Conjunto de datos
3. Limpieza de los datos
4. EDA
5. AUTO ARIMA
6. Conclusión

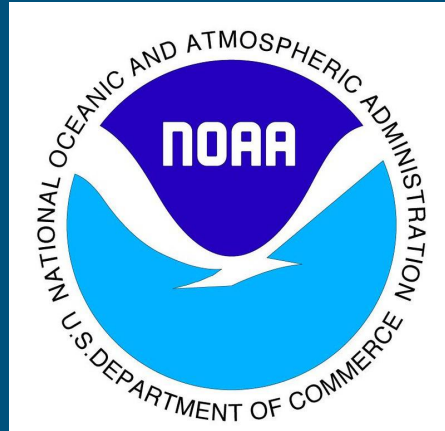


OBJETIVO

Crear un modelo que permita predecir el número de eventos que pueden ser causa de tsunamis por año

CONJUNTO DATOS

Tsunami Causes and Waves



LIMPIEZA DE DATOS

| | SOURCE_ID | YEAR | MONTH | DAY | CAUSE | VALIDITY | REGION_CODE | COUNTRY | LOCATION |
|---|-----------|-------|-------|-----|-------|----------|-------------|---------|---------------------------|
| 0 | 1 | -2000 | NaN | NaN | 1.0 | 1.0 | 50.0 | SYRIA | SYRIAN COASTS |
| 1 | 3 | -1610 | NaN | NaN | 6.0 | 4.0 | 50.0 | GREECE | THERA ISLAND (SANTORINI) |
| 2 | 4 | -1365 | NaN | NaN | 1.0 | 1.0 | 50.0 | SYRIA | SYRIAN COASTS |
| 3 | 5 | -1300 | NaN | NaN | 0.0 | 2.0 | 50.0 | TURKEY | IONIAN COASTS, TROAD |
| 4 | 6 | -760 | NaN | NaN | 0.0 | 2.0 | 50.0 | ISRAEL | ISRAEL AND LEBANON COASTS |

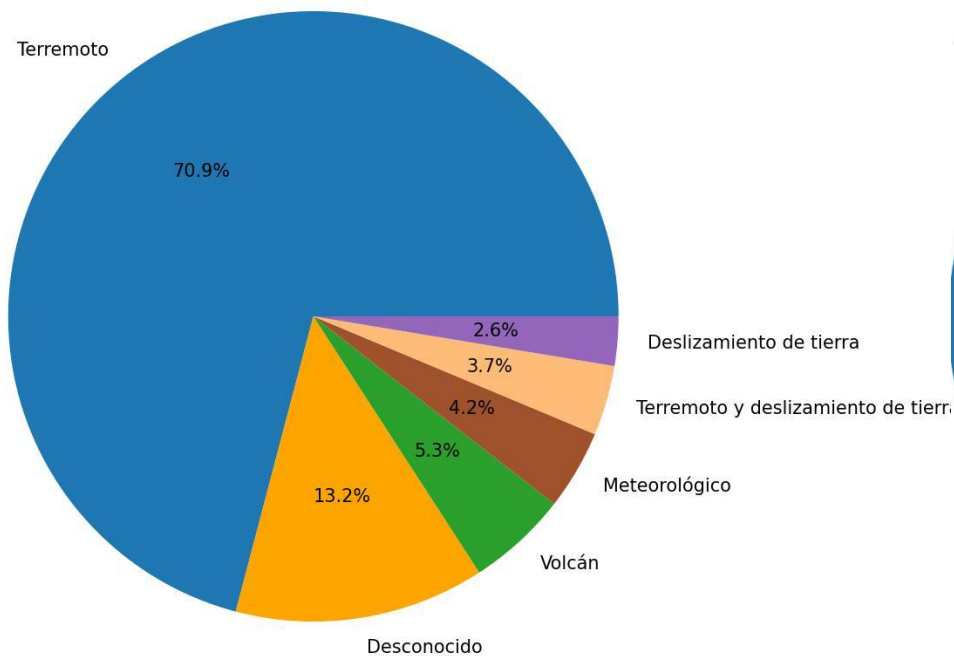
2582 rows x 45 columns

2000 aC

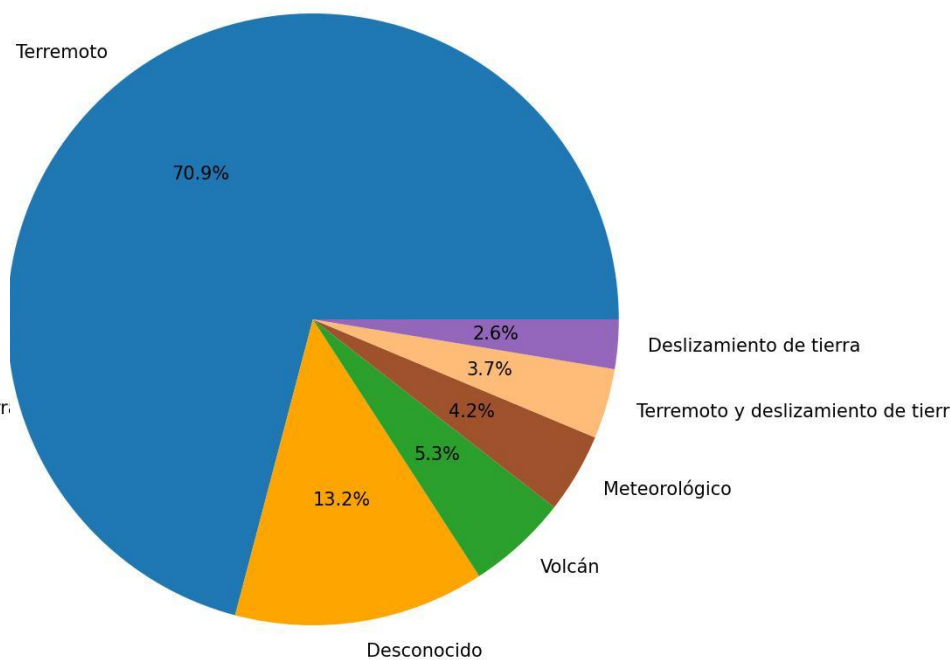
1960 métodos localización computarizados

EDA

Causas

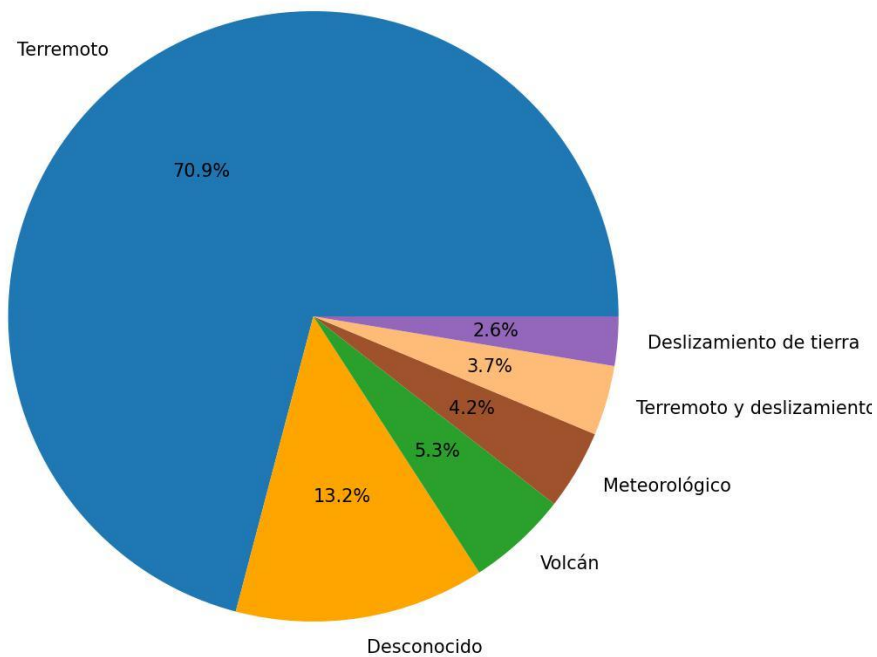


Antes 1960

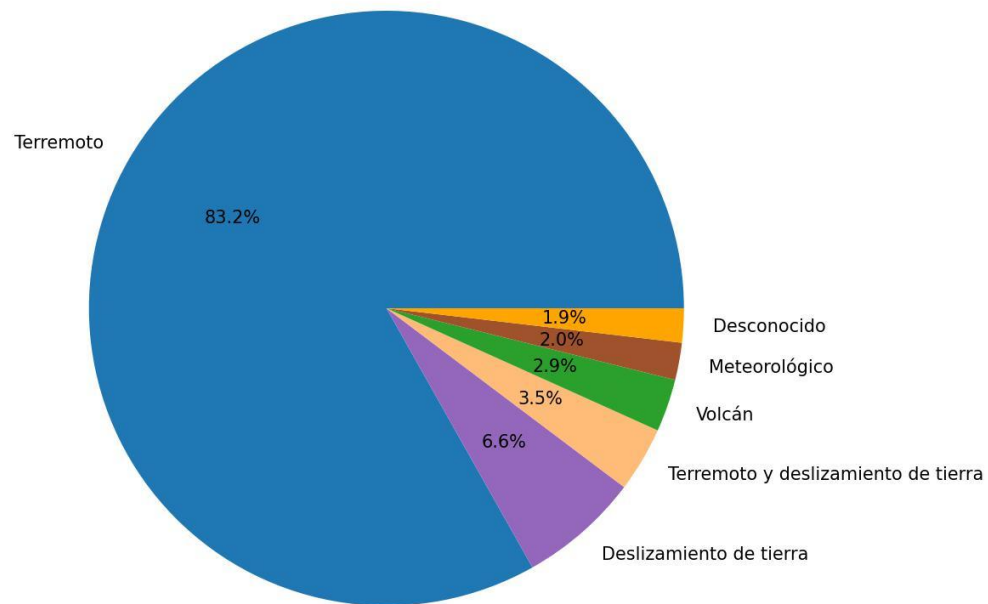


EDA

Causas

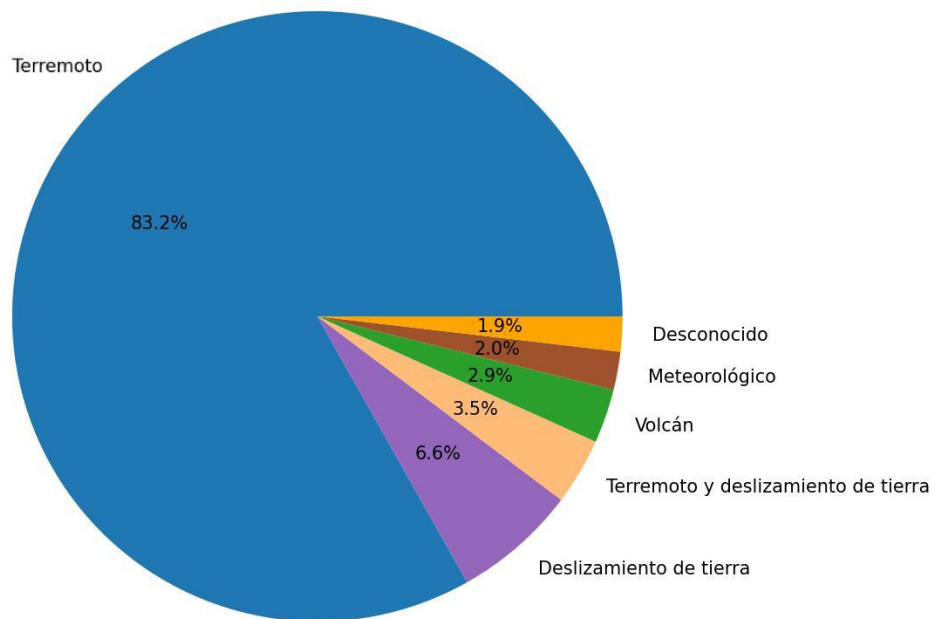


Después 1960

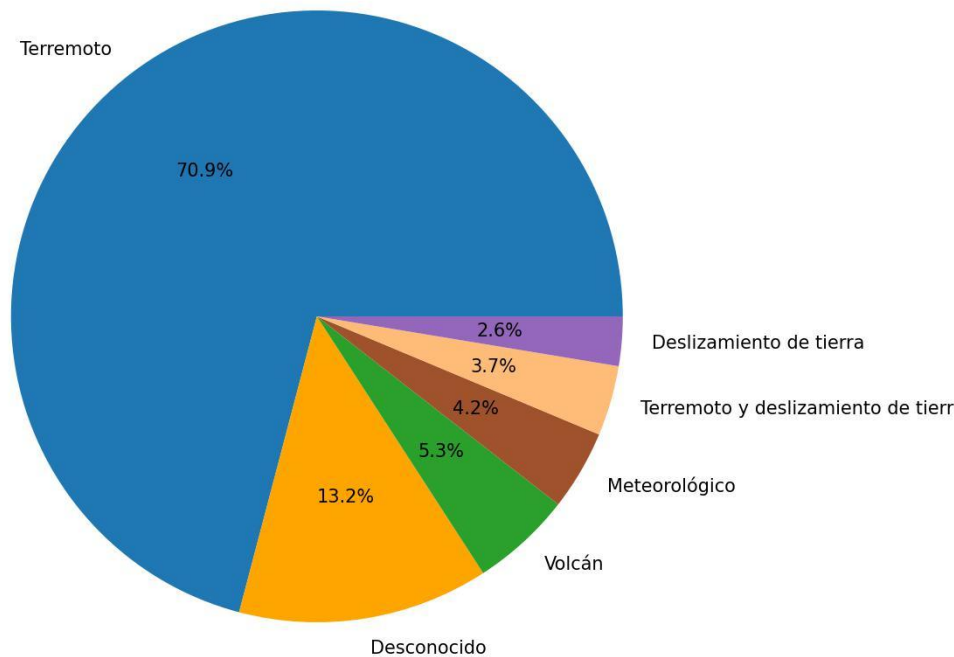


EDA

Después 1960

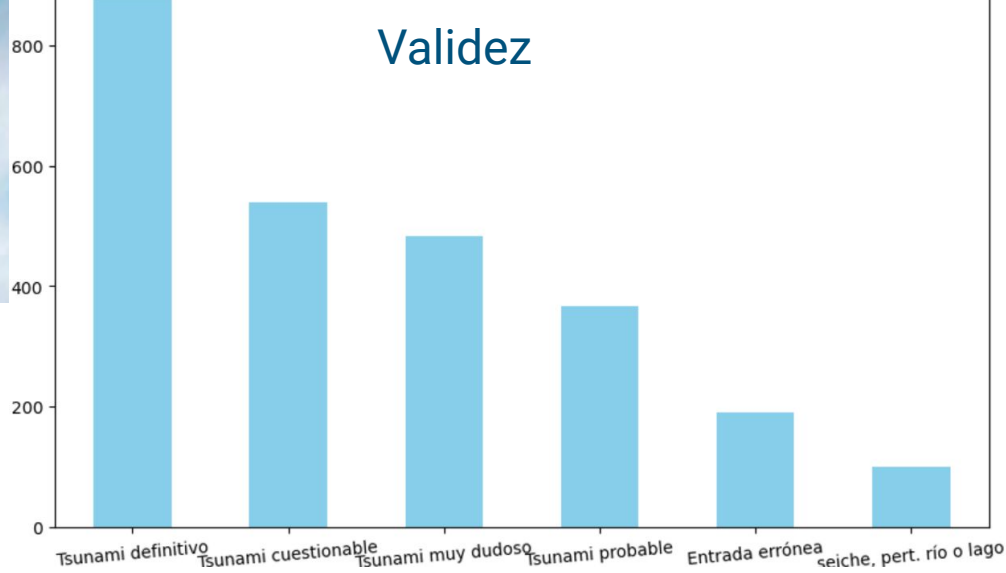


Antes 1960

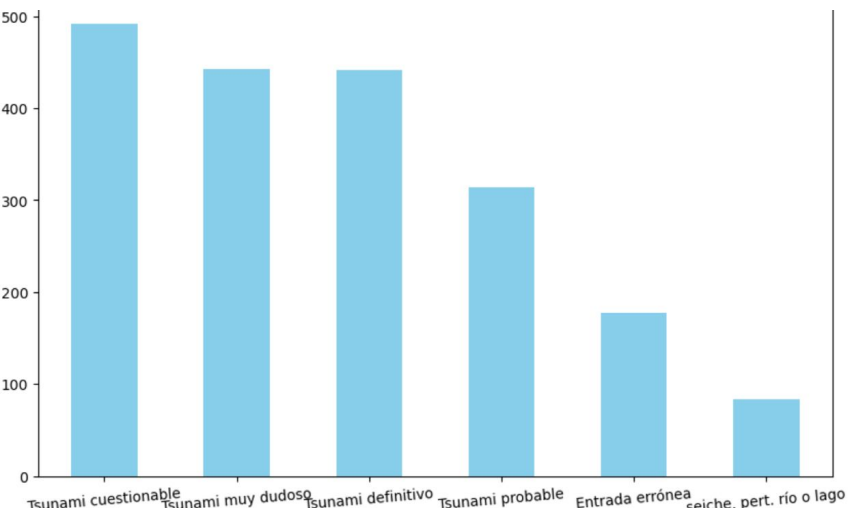


EDA

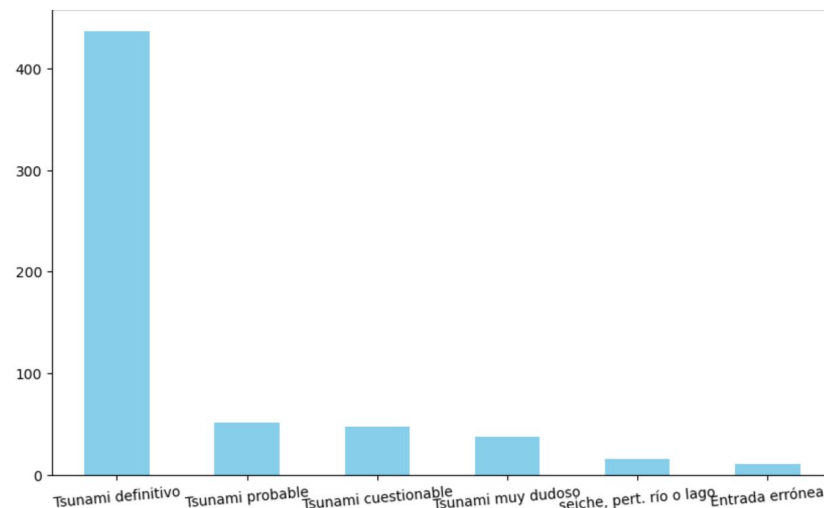
Validez



Antes 1960

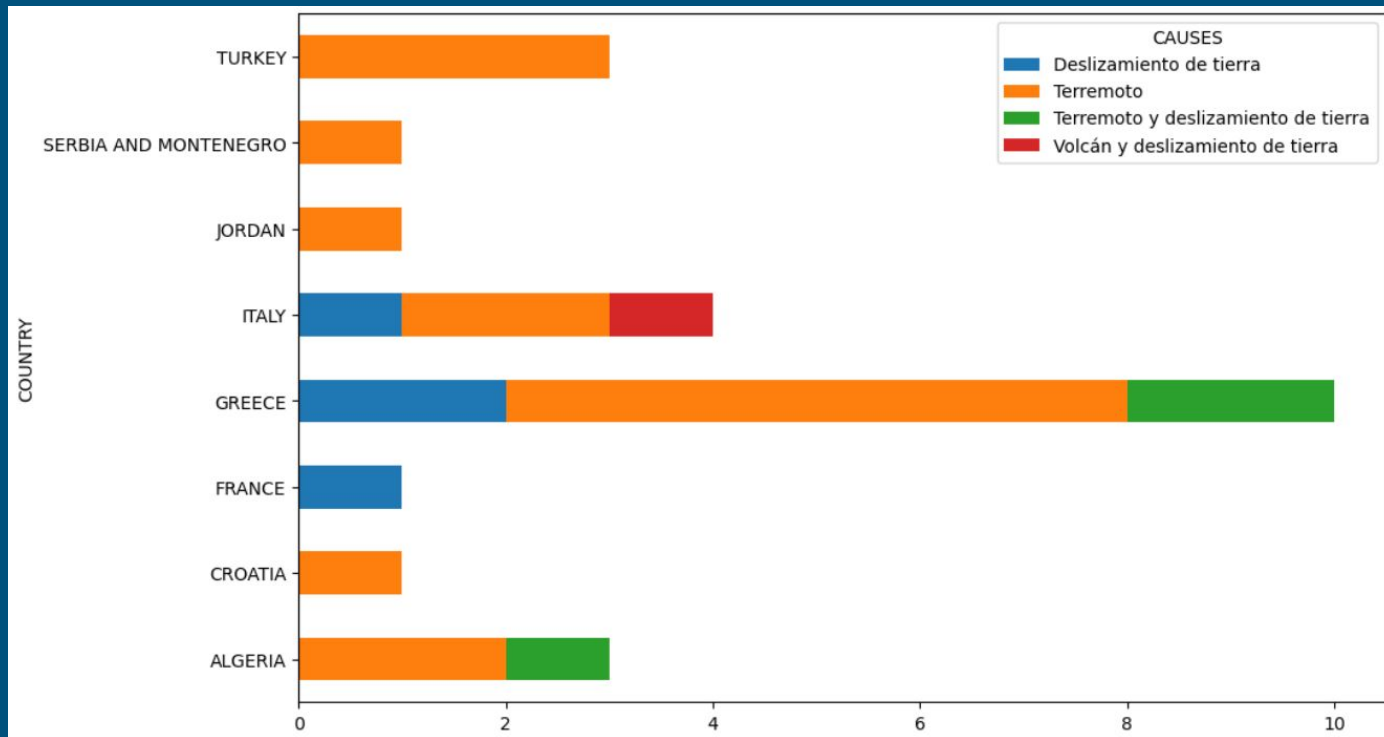


Después 1960



EDA

Tsunamis mediterraneo



AUTO ARIMA

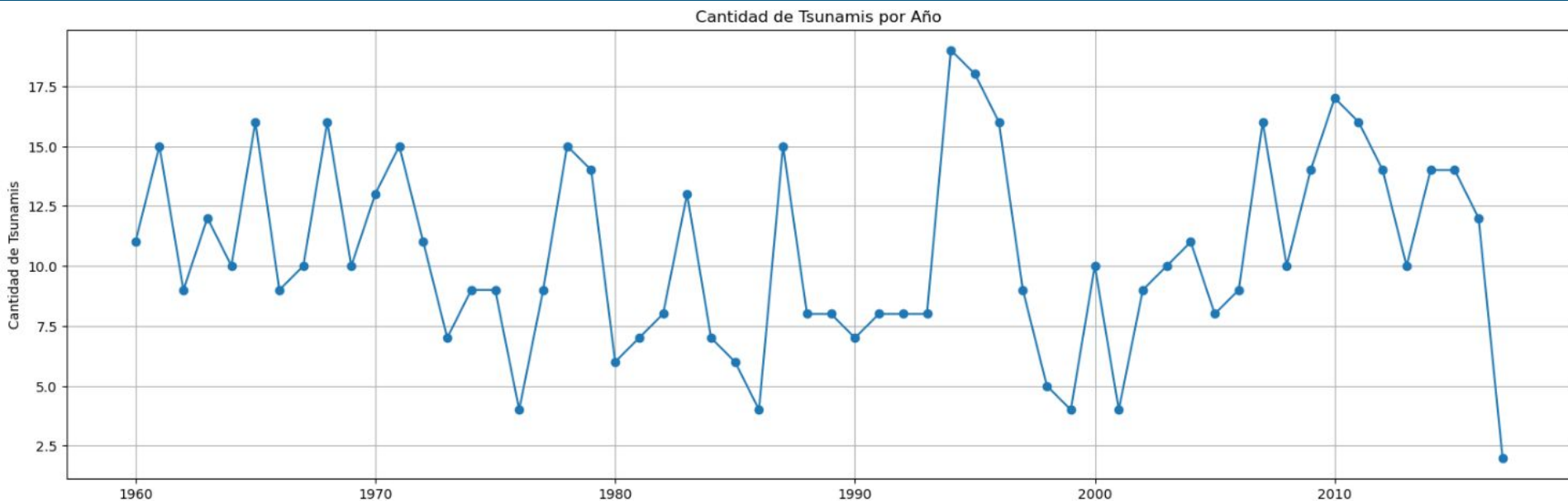
AutoRegressive Integrated Moving Average

```
!pip install pmdarima
```

1. Serie temporal
2. Selección mejores hiper parámetros (auto)
3. Ajuste del modelo
4. Diagnóstico del modelo
5. Predicción
6. Precisión

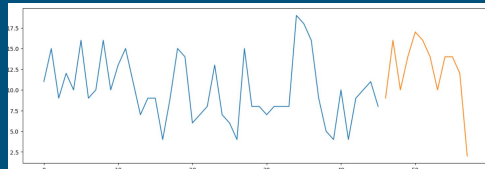
AUTO ARIMA

1. Serie temporal - estacionaria : Dickey-Fuller



AUTO ARIMA

2. Selección mejores hiper parámetros (auto)



```
from pmdarima.arma import auto_arma
```

AutoRegressive Integrated Moving Average

↓ ↓ ↓ ↓

p d q

AUTO ARIMA

Componente Autoregresivo (AR)

Componente de Diferenciación Integrada (I)

Componente de Media Móvil (MA):

"p", indica cuántos períodos anteriores se utilizan para predecir el valor actual

"d", indica cuántas veces se diferencia la serie temporal para hacerla estacionaria

"q", indica cuántos términos de error pasados se utilizan en la predicción actual

"p" es el orden del componente autoregresivo. (correlación)

"d" es el orden de diferenciación integrada.(estacionaridad)

"q" es el orden del componente de media móvil. (Descomposición de series temporales)

AUTO ARIMA

2. Selección mejores hiper parámetros (auto)
3. Ajuste del modelo

```
model = auto_arima(train, start_p=1, start_q=1,  
                  max_p=5, max_q=5,  
                  trace=True,  
                  error_action='ignore',  
                  suppress_warnings=True,  
                  stepwise=True)
```

Performing stepwise search to minimize aic

```
ARIMA(1,0,1)(0,0,0)[0] intercept : AIC=257.820, Time=0.12 sec  
ARIMA(0,0,0)(0,0,0)[0] intercept : AIC=257.484, Time=0.03 sec  
ARIMA(1,0,0)(0,0,0)[0] intercept : AIC=255.963, Time=0.09 sec  
ARIMA(0,0,1)(0,0,0)[0] intercept : AIC=255.972, Time=0.05 sec  
ARIMA(0,0,0)(0,0,0)[0]           : AIC=350.600, Time=0.02 sec  
ARIMA(2,0,0)(0,0,0)[0] intercept : AIC=257.665, Time=0.11 sec  
ARIMA(2,0,1)(0,0,0)[0] intercept : AIC=258.897, Time=0.42 sec  
ARIMA(1,0,0)(0,0,0)[0]           : AIC=274.796, Time=0.03 sec
```

Best model: ARIMA(1,0,0)(0,0,0)[0] intercept

Total fit time: 0.875 seconds

```
model.summary()
```

SARIMAX Results

| | | | |
|-------------------------|------------------|--------------------------|----------|
| Dep. Variable: | y | No. Observations: | 46 |
| Model: | SARIMAX(1, 0, 0) | Log Likelihood | -124.981 |
| Date: | Sun, 18 Feb 2024 | AIC | 255.963 |
| Time: | 17:20:42 | BIC | 261.449 |
| Sample: | 0 | HQIC | 258.018 |
| | - 46 | | |
| Covariance Type: | opg | | |

| | coef | std err | z | P> z | [0.025 | 0.975] |
|-----------|---------|---------|-------|-------|--------|--------|
| intercept | 7.2994 | 1.576 | 4.632 | 0.000 | 4.210 | 10.388 |
| ar.L1 | 0.2695 | 0.144 | 1.873 | 0.061 | -0.013 | 0.552 |
| sigma2 | 13.3892 | 3.453 | 3.877 | 0.000 | 6.621 | 20.157 |

| | | | |
|--------------------------------|------|--------------------------|------|
| Ljung-Box (L1) (Q): | 0.04 | Jarque-Bera (JB): | 2.35 |
| Prob(Q): | 0.85 | Prob(JB): | 0.31 |
| Heteroskedasticity (H): | 1.56 | Skew: | 0.51 |
| Prob(H) (two-sided): | 0.40 | Kurtosis: | 2.58 |

4. Diagnóstico del modelo

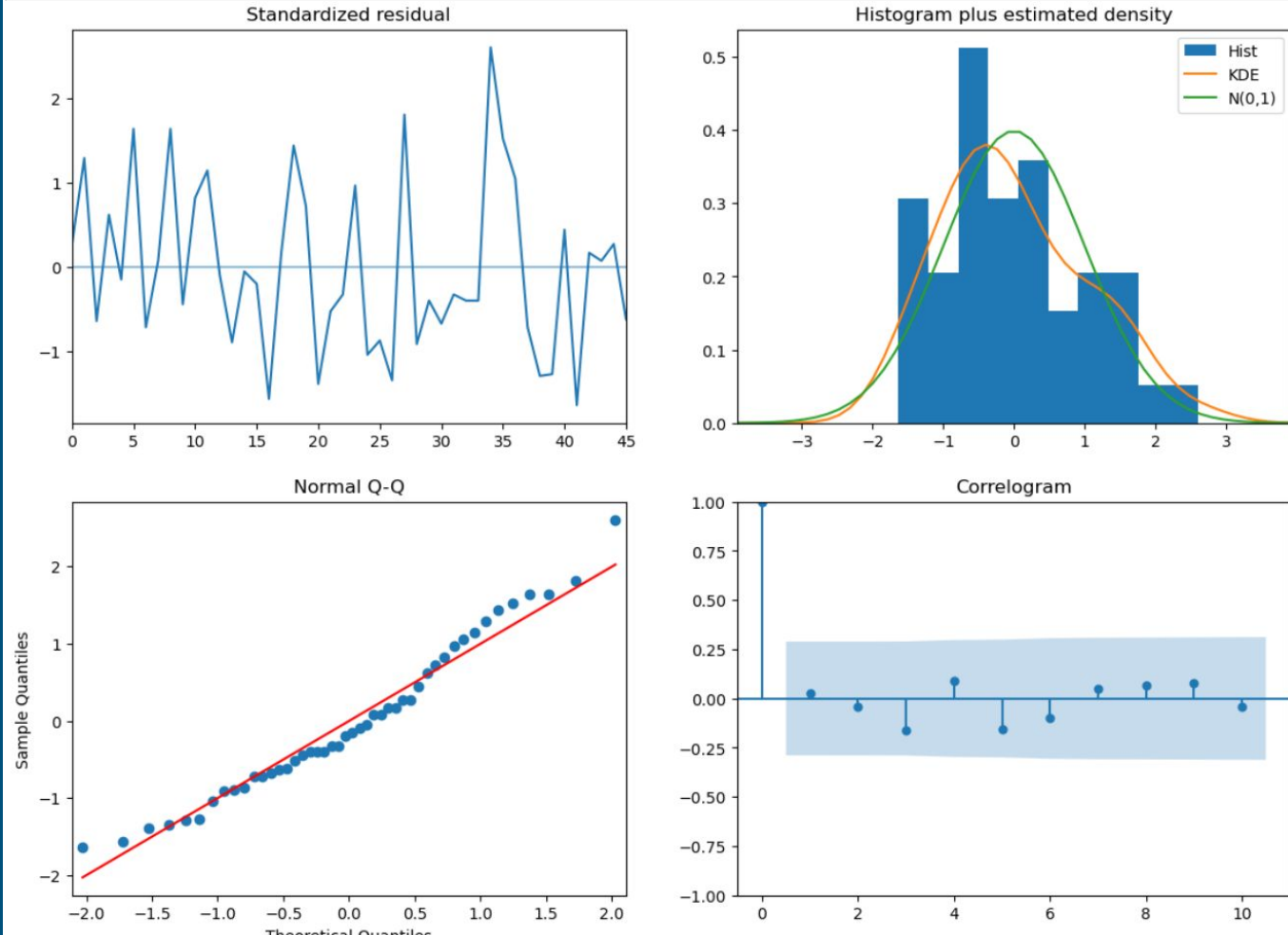
Estadísticas de Ajuste del Modelo

Coeficientes

Adecuación del modelo a los datos.

- Ljung-Box (L1) (Q) autocorrelación
- Jarque-Bera (JB) normalidad
- Heteroskedasticity (H) presencia de heterocedasticidad en los residuos.
- Skew y Kurtosis: asimetría y forma


```
model.plot_diagnostics(figsize=(14,10))  
plt.show()
```



AUTO ARIMA

5. Predicción

```
prediction, confint = model.predict(n_periods=len(test), return_conf_int=True)
```



AUTO ARIMA

6. Precisión

```
mse = mean_squared_error(test, prediction)
r2 = r2_score(test, prediction)

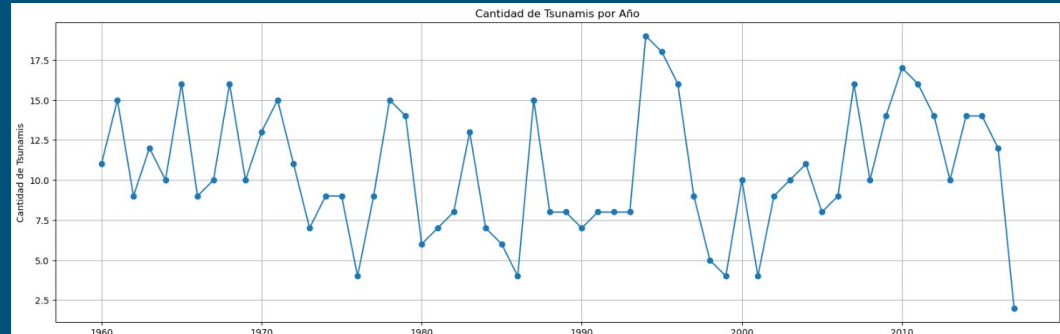
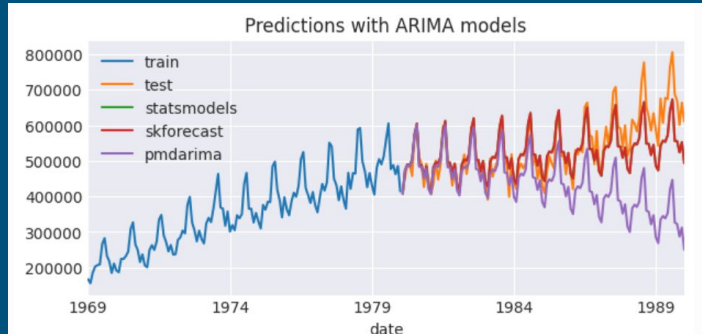
print("Error Cuadrático Medio (MSE):", mse)
print("Coeficiente de determinación (R^2):", r2)
```

```
Error Cuadrático Medio (MSE): 21.296681684902037
Coeficiente de determinación (R^2): -0.354559259110377
```

el modelo no es útil

CONCLUSIÓN

- Falta cantidad de datos: de 2582 filas a 46
- Serie temporal
- Más variables a tener en cuenta





¡MUCHAS GRACIAS!

Github: \Nuriarique

LinkedIn: \nuriarique