# Machine Learning para el Modelamiento y Gestión de Sistemas Complejos

**Análisis preliminar de datos del mercado mayorista de papa de Lima**

40 años
GRADE
Grupo de Análisis para el Desarrollo

# Agenda

- Tratamiento de datos en Python
- Análisis descriptivo

# Tratamiento de datos en Python…

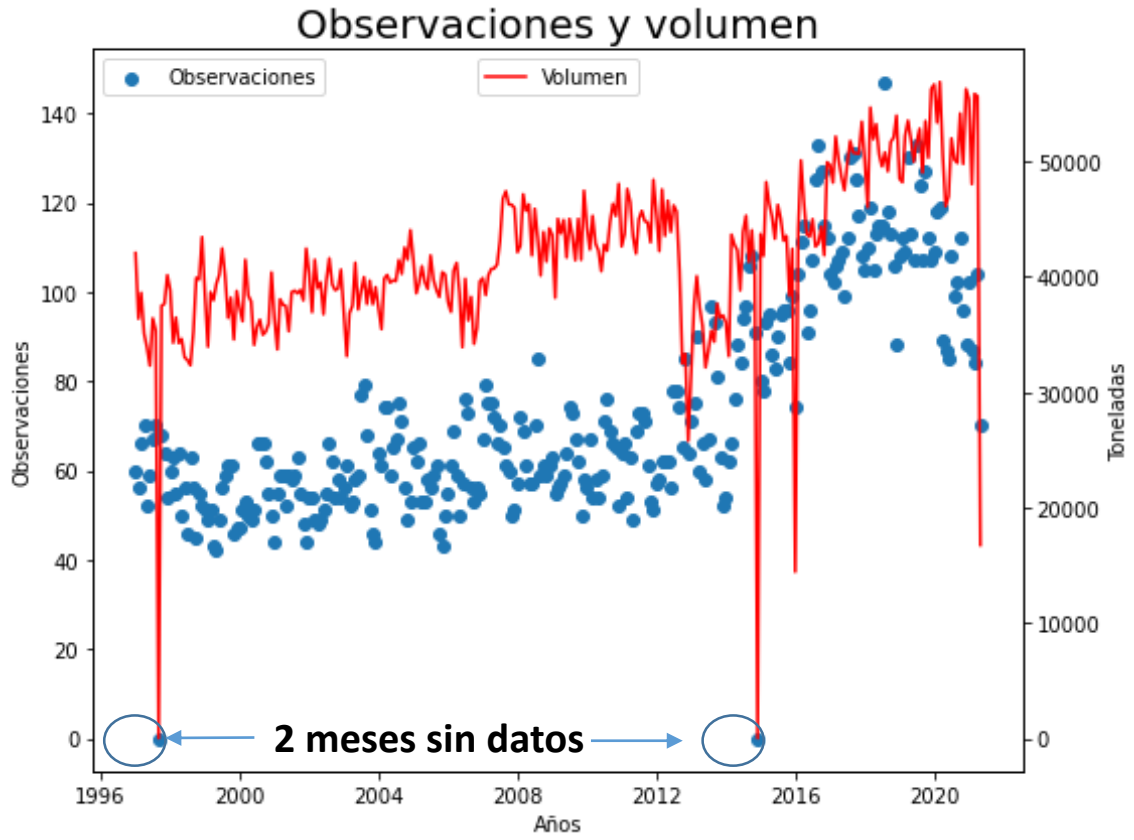
```
# -*- coding: utf-8 -*-
"""
Created on Sat Aug 13 16:30:59 2022

@author: Eduardo Zegarra
"""

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

data=r'C:\Users\Eduardo Zegarra\Documents\__00_papa_new\originales_data/'

df=pd.read_excel(data+'base_completa_1997_2021.xls', sheet_name='volumen')

df.rename(columns={'Variable':'provincia'}, inplace=True)

df_long=pd.melt(df,id_vars=['producto','year','provincia'],
                var_name='mes', value_name='volumen')

dg=pd.read_excel(data+'base_completa_1997_2021.xls', sheet_name='precio')

dg_long=pd.melt(dg,id_vars=['producto','year'],
                var_name='mes', value_name='precio')

df_base=df_long.merge(dg_long, on=['producto','year','mes'], how='outer')
```

- Se importan bases de datos de Excel
- Se pasan de formato ancho a formato largo (*melt*)
- Se juntan (*merge*) en una sola base de datos conjunta
- Notar: base de precios no tiene provincia de origen ya que el precio se forma en el MML

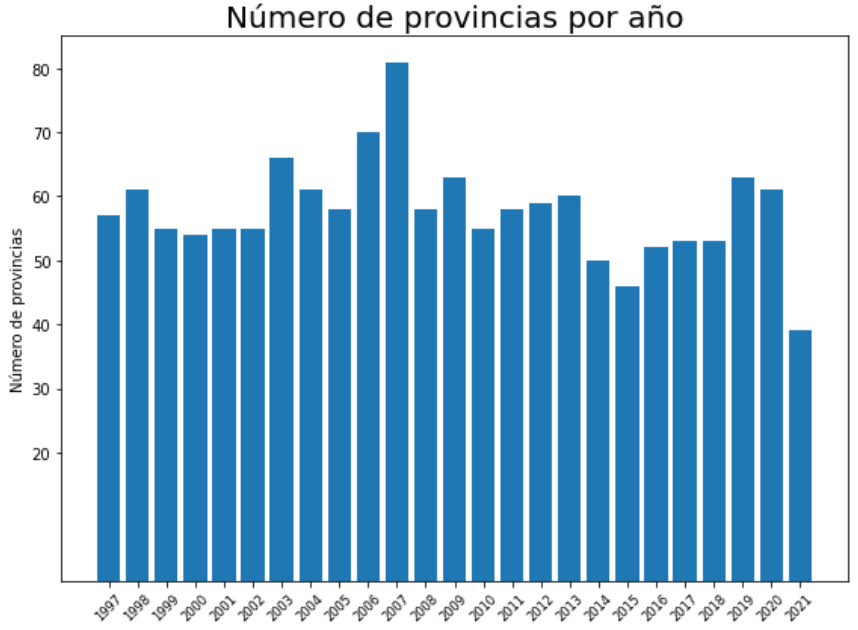
40 años GRADE
Grupo de Análisis para el Desarrollo

```
df_base['mm']=df_base['mes'].str.split('_').str.get(1).astype('int64')

df_base['dia']=1

df_base['mm1']=pd.to_datetime(dict(day=df_base.dia, month=df_base.mm,year=df_base.year),
                              unit='D', format="%d-%m-%Y")

df_base.set_index(df_base['mm1'], inplace=True)
df_base=df_base['Jan-1997':'May-2021']
```

```
## Identificamos meses con ingreso de papa a Lima=0

### numero de observaciones por mes

observ=df_base.groupby(df_base.index)['volumen'].count()
vol=df_base.groupby(df_base.index)['volumen'].sum()

fig,ax=plt.subplots(figsize=(8,6))

ax.scatter(observ.index,observ, label='Observaciones')
ax2=ax.twinx()
ax2.plot(vol.index,vol, color='red', label='Volumen')
plt.title("Observaciones y volumen", fontsize=20)
ax.set_ylabel('Observaciones')
ax2.set_ylabel('Toneladas')
ax.set_xlabel('Años')
plt.xticks(rotation=45, fontsize='x-small')
ax.legend(['Observaciones'])
ax2.legend(['Volumen'], loc='upper center')
plt.tight_layout()
```

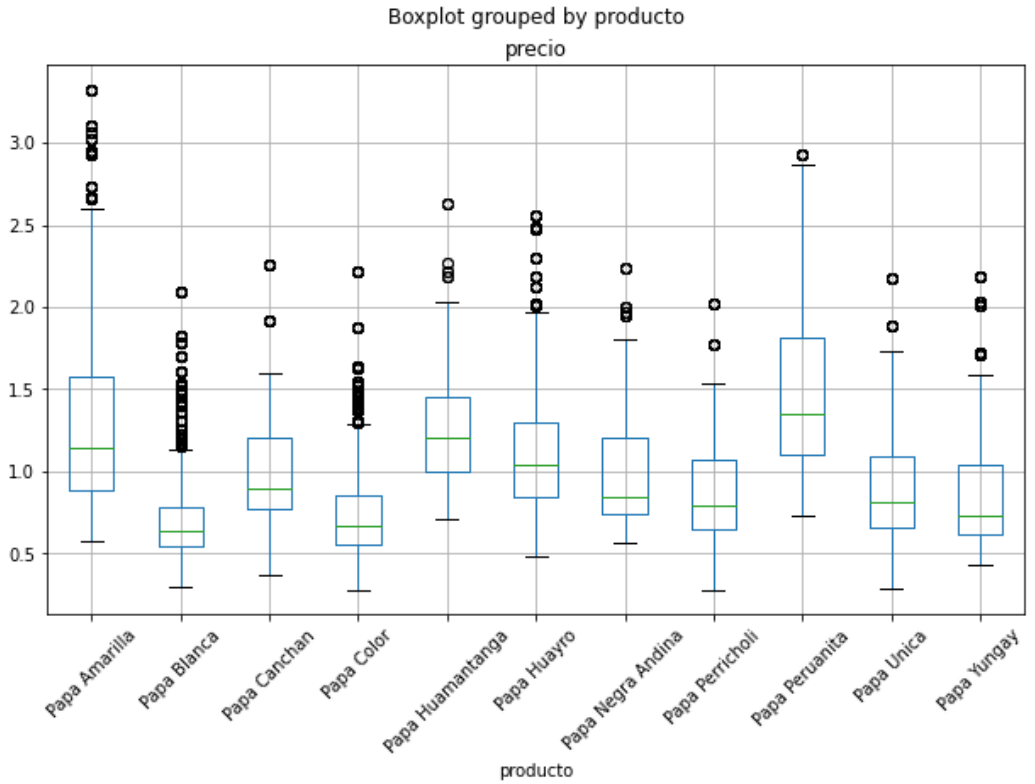



- El número de variedades de papa fue de sólo 4 entre 1997 y 2006, a partir de 2007 se incrementa hasta llegar a 10 en 2010 y luego a 11 a partir de 2013.
- Estos parecen cambios de clasificación (mejoras), no necesariamente que recién ingresan esas nuevas variedades al mercado



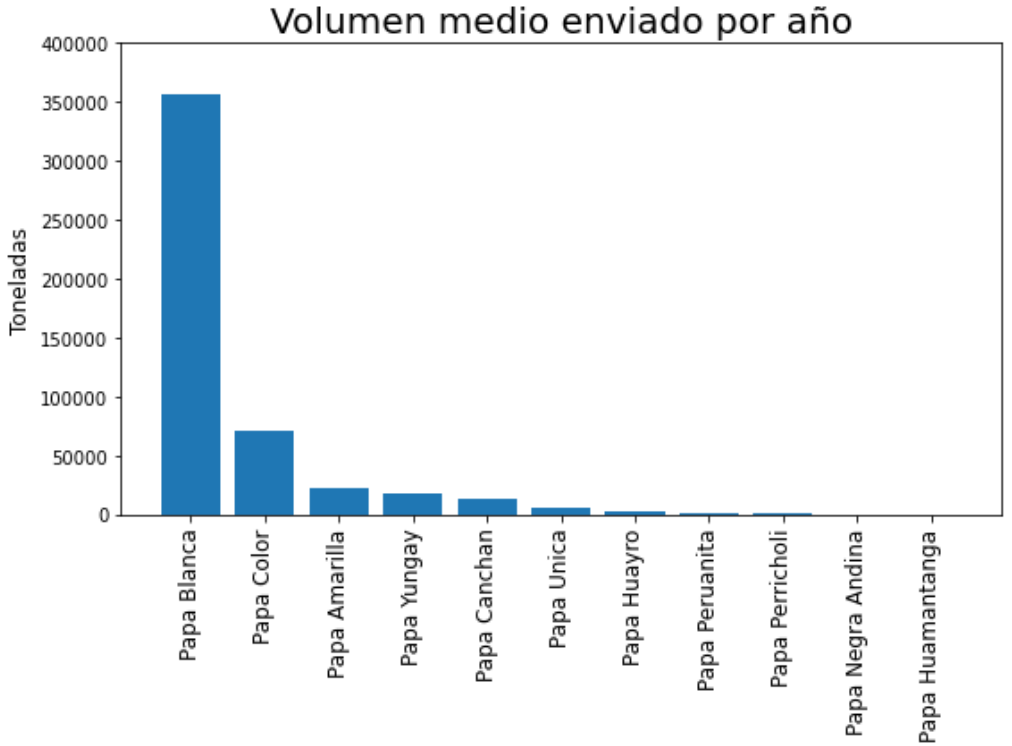

- El número de provincias desde las cuales provienen las variedades de papa tiene fluctuaciones importantes.
- Esto indica cierta dinámica de entrada y salida de provincias al mercado de papa de Lima (2021 sólo hasta mayo)

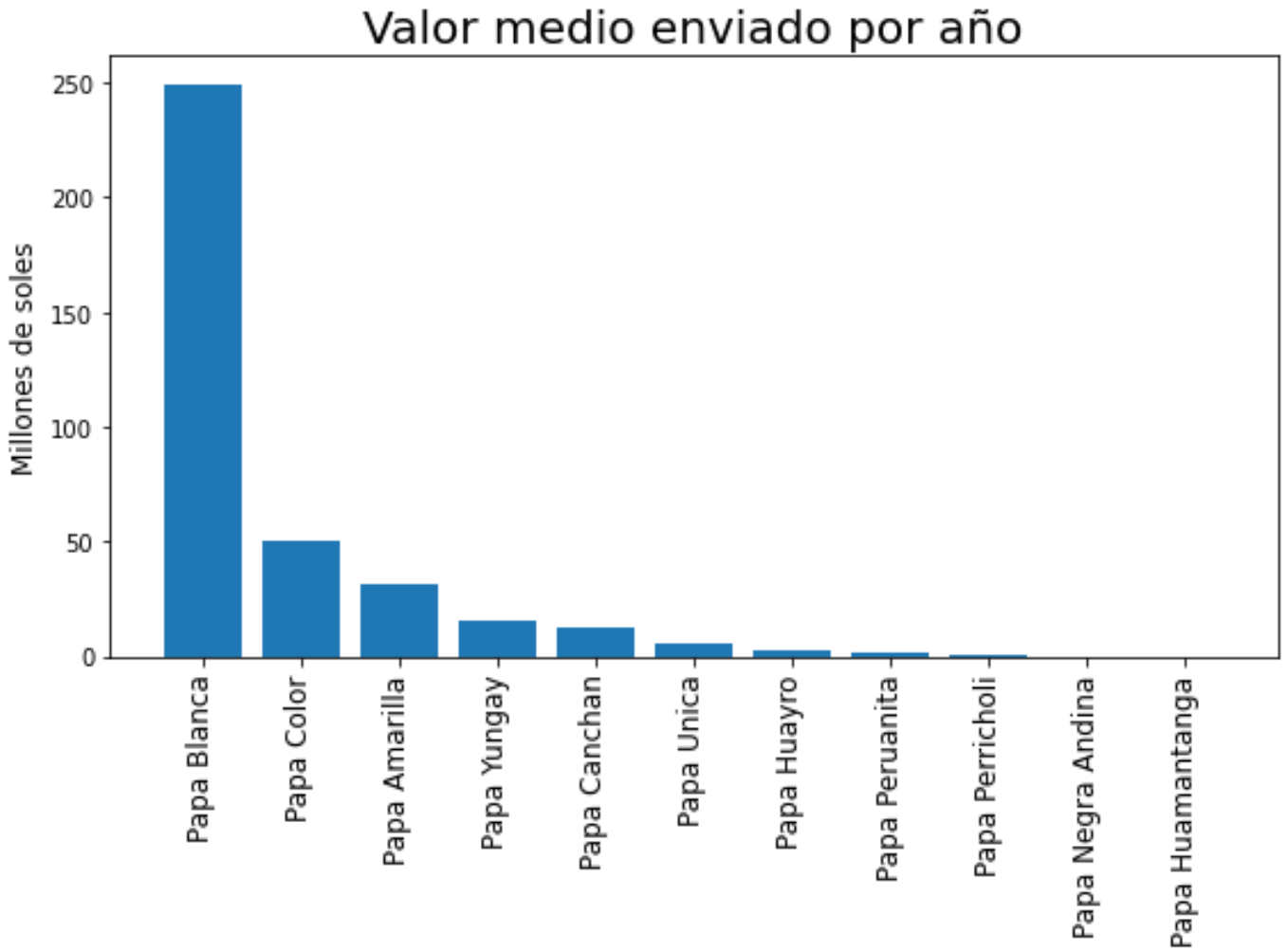
40 años GRADE
Grupo de Análisis para el Desarrollo




```
## ANALISIS DE PRECIOS Y PESO DE CADA VARIEDAD

df_base.boxplot(column='precio', by='producto', rot=45, figsize=(10,6))
```

```
## generamos variable del valor de produccion

df_base['valor']=df_base['precio']*df_base['volumen']/1000

vv_mean=df_base.groupby(
    ['producto', 'year'], as_index=False)['valor','volumen'].sum()
vv_1=vv_mean.groupby(
    ['producto'], as_index=False)['valor','volumen'].mean().sort_values(
        by='valor', ascending=False)

fig,ax=plt.subplots(figsize=(8,6))
ax.bar(vv_1.producto, vv_1.volumen)
ax.set_ylabel('Toneladas', fontsize=12)
plt.title("Volumen medio enviado por año", fontsize=20)
plt.xticks(vv_1.producto,rotation=90, fontsize=12)
plt.yticks(np.arange(0,450000,50000))
plt.tight_layout()
```



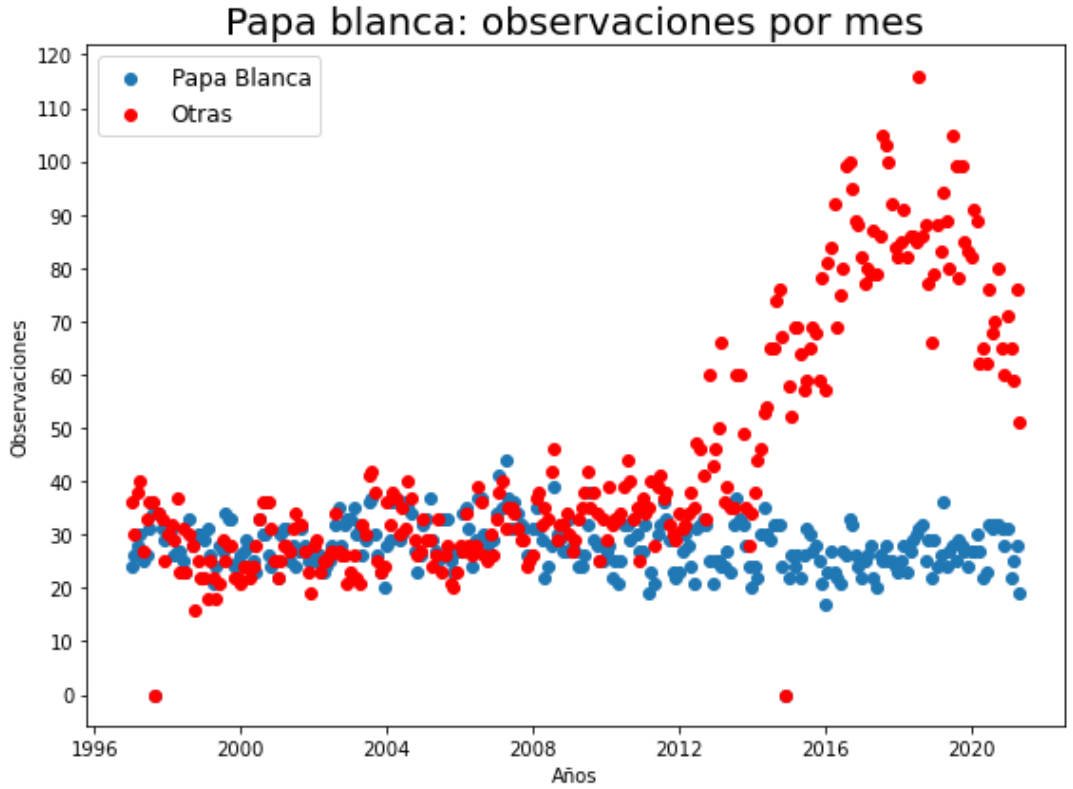



```
In [7]: peso=vv_1.iloc[0].get(1)/vv_1.valor.sum()
   ...: print("Peso de papa blanca en total:",peso)
Peso de papa blanca en total: 0.6733455744199748
```

- El valor total de papa blanca representa el 67.3% del valor total del mercado de papa transada en el MML entre 1997 y mayo 2021

40 años GRADE
Grupo de Análisis para el Desarrollo

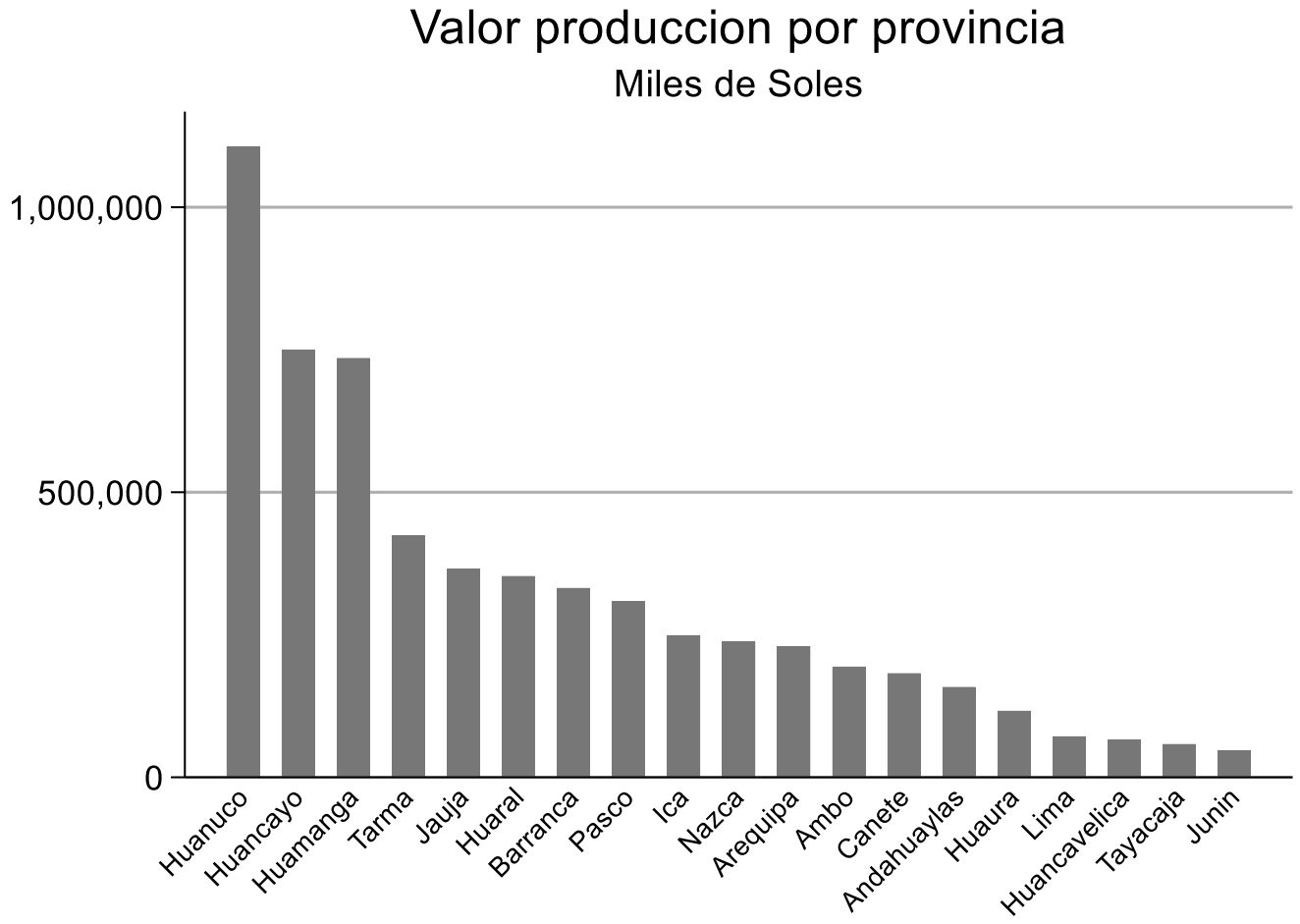
```
df_blanca=df_base[df_base['producto'].isin(["Papa Blanca"])]
df_otras=df_base[~df_base['producto'].isin(["Papa Blanca"])]

observ = df_blanca.groupby(df_blanca.index)['volumen'].count()
vol = df_blanca.groupby(df_blanca.index)['volumen'].sum()

observ1 = df_otras.groupby(df_otras.index)['volumen'].count()
vol1 = df_otras.groupby(df_otras.index)['volumen'].sum()

fig,ax=plt.subplots(figsize=(8,6))
ax.scatter(observ.index,observ, label='Papa Blanca')
ax.scatter(observ1.index,observ1, label='Otras',color='red')
plt.title("Papa blanca: observaciones por mes", fontsize=20)
ax.set_ylabel('Observaciones')
ax.set_xlabel('Años')
ax.legend(fontsize=12)
plt.yticks(np.arange(0,130,10))
plt.tight_layout()
```




- La evolución del número de observaciones, por mes, de papa blanca es más estable
- En otras variedades pueden haber ocurrido cambios en la definición o haber aparecido nuevas variedades entrando al mercado (ambos procesos hacen aumentar el número de observaciones mensuales). Este comportamiento se observa desde inicios del 2012 en adelante. Es decir, para el análisis de las otras variedades de papa se debe dilucidar este quiebre y su origen (si es un tema de clasificación, o de cambios reales en entrada de nuevas variedades a Lima).

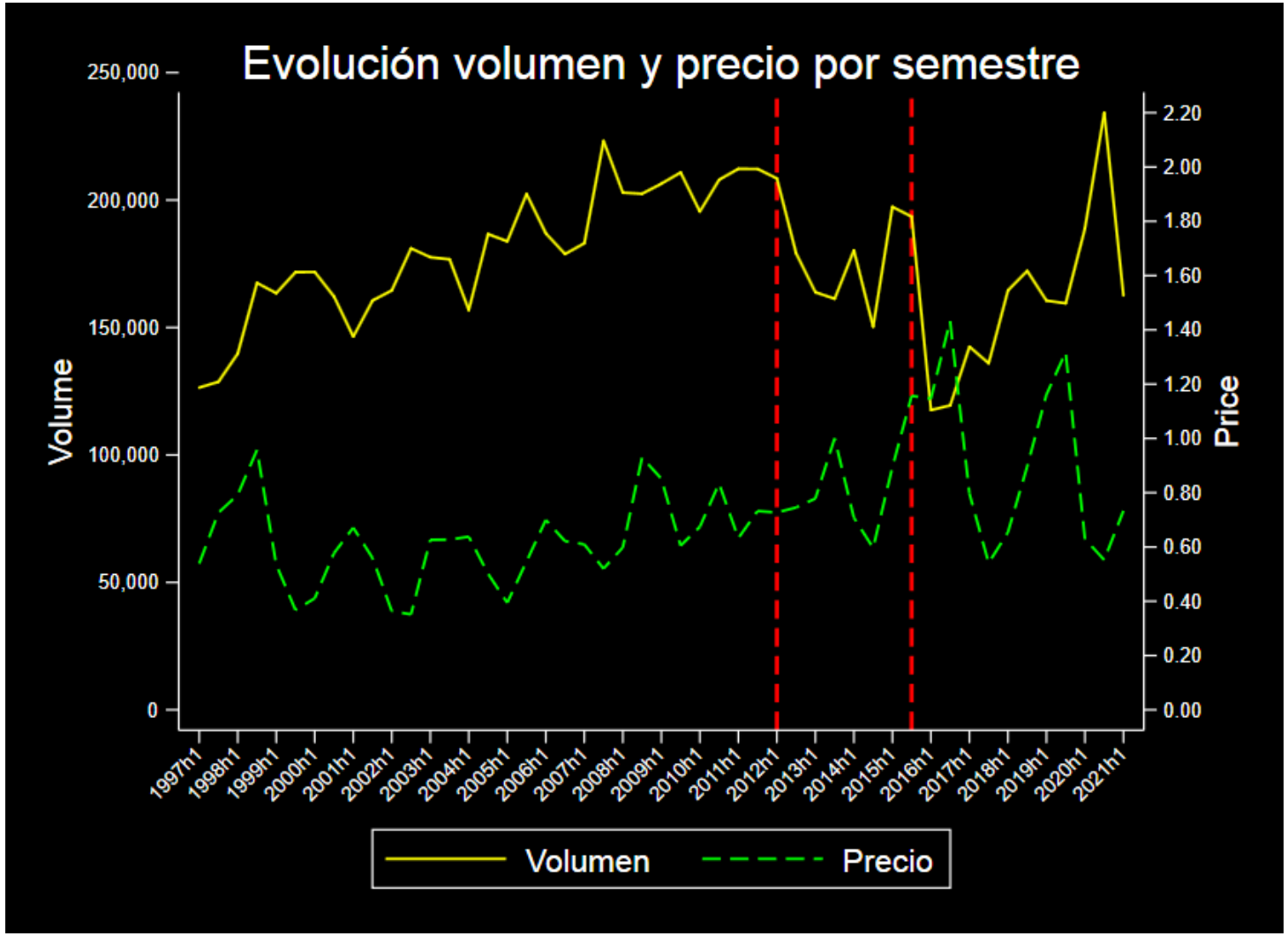
40 años GRADE
Grupo de Análisis para el Desarrollo

# Analizaremos solamente el mercado de papa blanca dado su peso económico y homogeneidad

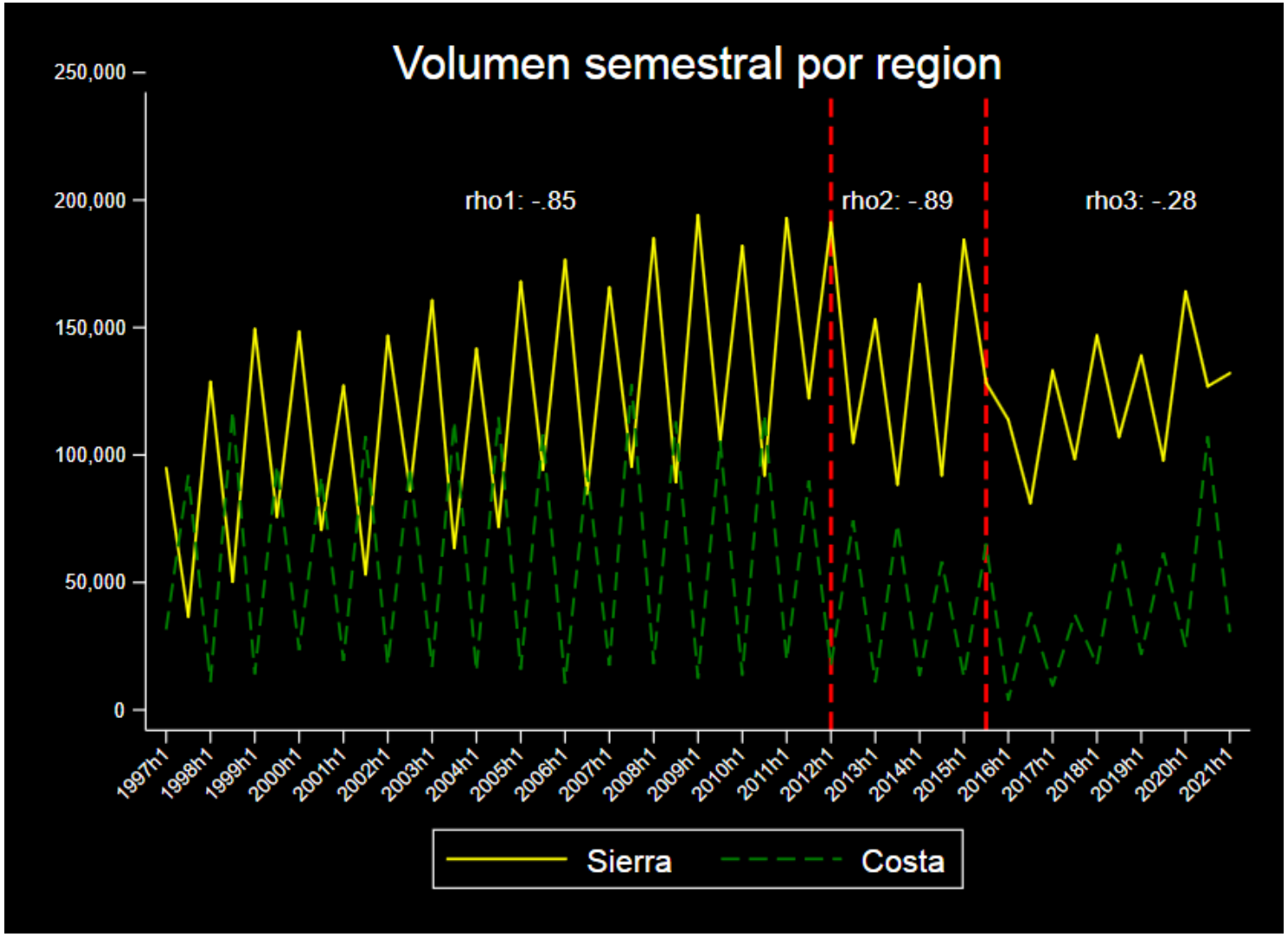



- 19 provincias concentran el 96% del valor total de las transacciones de papa blanca (de un promedio de unas 50 provincias) para 1997-2021
- Seleccionamos a esas 19 provincias para el análisis

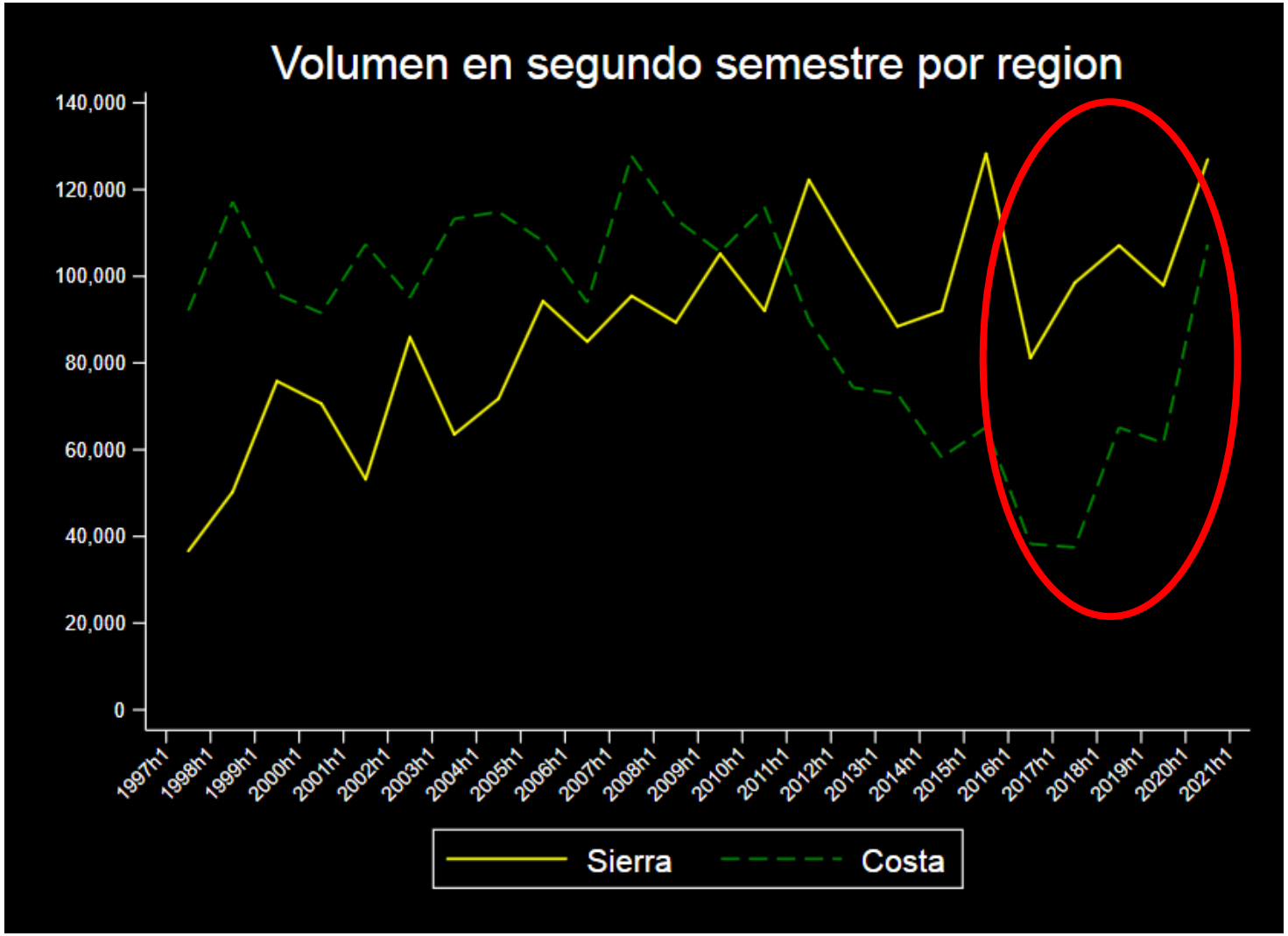
40 años GRADE
Grupo de Análisis para el Desarrollo




40 años
GRADE
Grupo de Análisis para el Desarrollo




40 años
GRADE
Grupo de Análisis para el Desarrollo




40 años
GRADE
Grupo de Análisis para el Desarrollo

¡Muchas gracias!

40 años
GRADE
Grupo de Análisis para el Desarrollo