

Big Data y Machine Learning (UBA) - 2025

Trabajo Práctico 1: Jugando con APIs y WebScraping

Grupo 1

Francisco Ariel Gonzalez Häberle 903012

David Jimenez Jaldin 909323

Lourdes Sofia Beltramo 888198

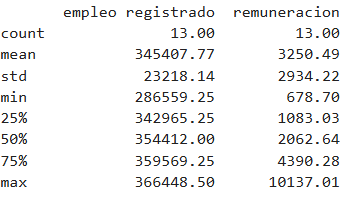
Link repositorio github: https://github.com/davidjjj14082000-png/BigDataUBA-Grupo-1/tree/17ede433b761562dc0b7f4614c41903c9932c528/trabajos\_practicos/tp\_1

**Reporte - Grupo N°1**

**Parte A**

Se seleccionaron de la API solicitada 2 series de periodicidad anual correspondientes a los meses de enero, con 13 observaciones cada, comenzando en el año 2004 y finalizando en el año 2016. La primera de las series contiene información respecto al empleo total registrado en el sector privado agrícola, ganadero y de la pesca. Por su parte, la segunda serie contiene información sobre la remuneración neta promedio del mismo sector.

Se realizo un resumen estadístico de la información obtenida, el cual se puede observar en la siguiente tabla:

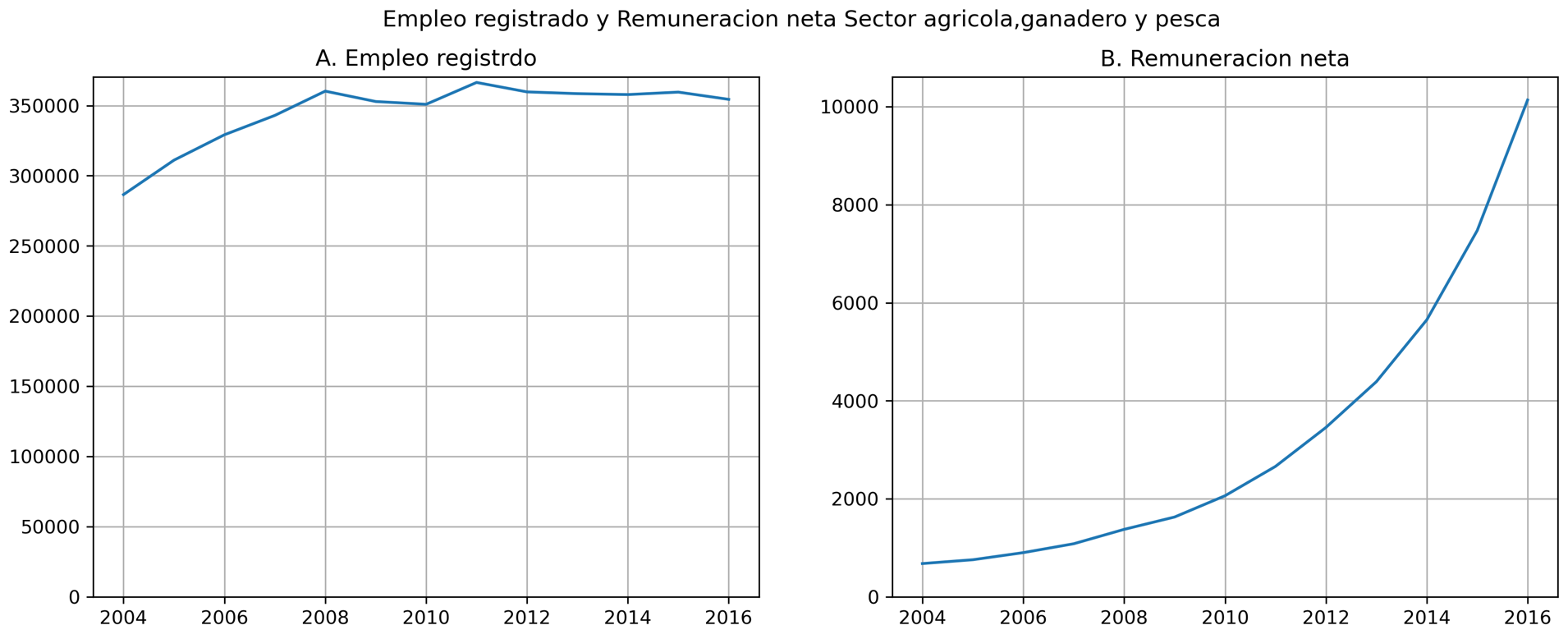


Como se desprende del resumen anterior, se observa a lo largo del período bajo análisis un nivel de empleo promedio de 345.407 asalariados por año. Observamos una variabilidad relativamente baja año a año en base al desvío estándar (con un valor de 23.218,14 que representa aproximadamente un 6,72% del valor promedio. Se puede concluir además en base al resumen estadístico que existe una presencia de valores más extremos por debajo de la mediana (que tiene un valor de 354.412) que, por encima, debido a que el promedio es menor a la misma.

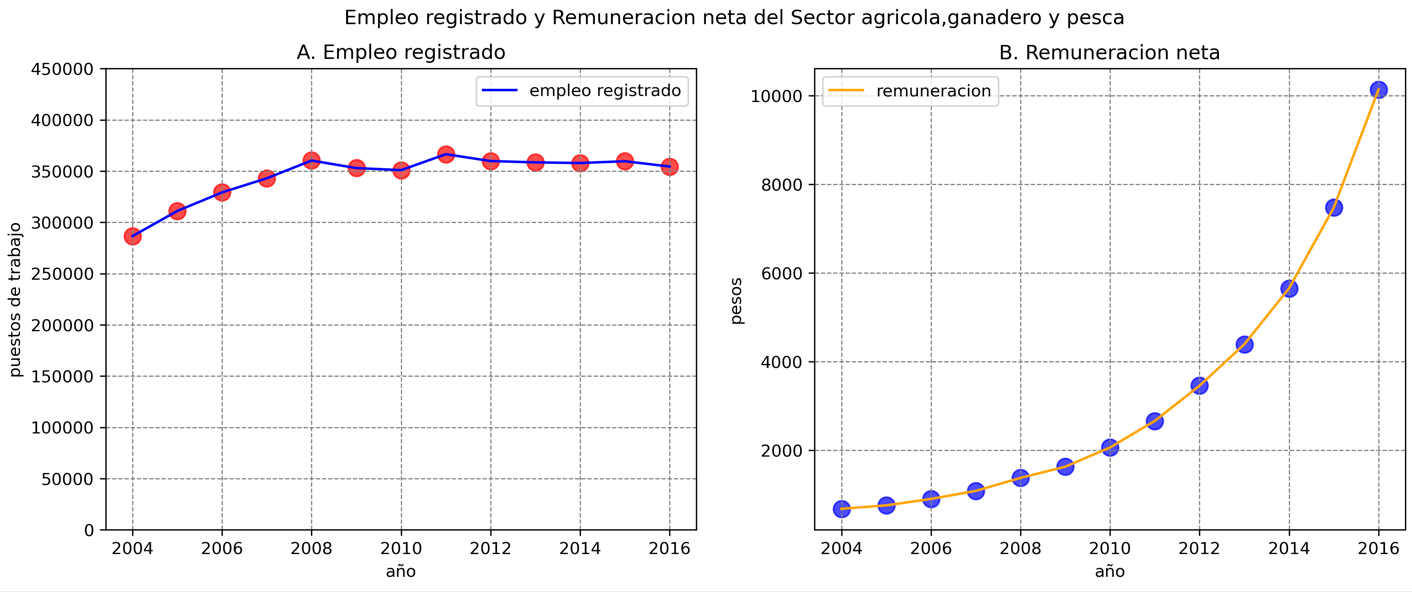
Respecto a la información de la segunda serie, correspondiente a la remuneración promedio de los sectores bajo análisis, se concluye que no es correcto ni relevante un análisis de los valores tal cual se presentan por tratarse de montos nominales en pesos, que reflejan principalmente la evolución de la inflación a lo largo del período, como se desprende de los gráficos correspondientes en la siguiente sección del reporte. Esto puede observarse en que los valores mínimo y máximo de la serie se corresponden con la primera y última observación de la misma. Para realizar un análisis más correcto en términos económicos seria necesario deflactar la serie con un indicador de precios.

A continuación, se presentan las gráficas creadas a partir de las mismas:

Estilo pyplot:



Estilo orientado a objetos:



**Parte B**

*Ejercicio 4 - Documentando noticias locales y en el extranjero*

1)





2) Diferencias/similitudes al extraer titulares en diarios de Argentina versus diario extranjero

Al analizar titulares de Clarín (Argentina) y CNN (extranjero), se observaron similitudes y diferencias en la estructura HTML y las clases CSS empleadas. Los encabezados de ambos sitios están dentro de etiquetas <a> que contienen el enlace. Sin embargo, las etiquetas específicas para el texto del titular (<h2>, <h3>, <span>) y las clases CSS asociadas a los contenedores de titulares (container\_\_link en el caso de CNN) difieren entre los sitios. Esto implica una meticulosa revisión del código fuente de cada página web para ajustar los selectores de BeautifulSoup y extraer la información de manera adecuada demostrando que, aunque el método general de web scraping es parecido los detalles de implementación dependen en gran medida de la estructura particular de cada sitio web.

3) Dificultades y cómo se resolvieron dichas dificultades.

Durante el proceso de extracción de titulares, las principales dificultades surgieron al identificar las etiquetas HTML y las clases CSS específicas que contenían los titulares y sus enlaces en cada sitio web. La estructura de Clarín y CNN no es la misma, por lo que fue necesario examinar el código fuente manualmente en cada página para identificar los patrones adecuados (por ejemplo, las etiquetas <h2> con enlaces en Clarín y las etiquetas <a> con clases específicas y elementos anidados como <span> o <h3> en CNN). Además, se encontró un error de NameError porque una celda que definía una variable necesaria no se ejecutó a tiempo; esto se resolvió moviendo la definición de la variable a la celda donde se utilizaba para asegurar su disponibilidad durante la ejecución.