

Transformando Datos en Conocimiento: Unificación y Análisis Exploratorio de Datasets de Títulos de Educación Superior y de la Secretaría de la Función Pública mediante Python

Autores: Carlos Jein Coronel Rojas, Roberto Carlos Saucedo, Fabricio Villar.

Asesor: Mst. Hector Estigarribia

FCyT UNCA, noviembre 2024

Resumen

En este proyecto se busca obtener información clave mediante la integración de datos provenientes de los registros de Títulos de Educación Superior y la Secretaría de la Función Pública. El enfoque metodológico empleado se fundamenta en la utilización de Google Colab, una plataforma de computación en la nube que optimiza el análisis y procesamiento de datos.

Los objetivos centrales incluyen la integración de diferentes conjuntos de datos con el propósito de identificar patrones y tendencias relevantes en el ámbito de la función pública. La metodología aplicada abarca etapas de limpieza y preprocesamiento de los datos, seguidas de la unificación de los datasets para realizar análisis exploratorios profundos.

Entre los hallazgos más destacados, se identificó que, de los 406.280 documentos registrados en el dataset de títulos académicos, 127.334 también forman parte del cuerpo de 316,761 funcionarios públicos. Asimismo, se observó que las universidades privadas constituyen la principal fuente de formación académica de estos servidores públicos, con predominancia de títulos en áreas como educación y salud. Las instituciones que emplean la mayor cantidad de funcionarios con títulos registrados incluyen el Ministerio de Educación y el Ministerio de Salud, entre otras entidades públicas de relevancia.

Este tipo de estudio pone de relieve la importancia de la ciencia de datos en la gestión pública, demostrando cómo la integración y el análisis de grandes volúmenes de datos pueden generar información valiosa para la toma de decisiones estratégicas en el sector gubernamental.

Palabras clave: Ciencia de Datos, ODS 4 - Educación de calidad, ODS 9 - Industria Innovación e Infraestructura,

Introducción

En Paraguay, la disponibilidad de datos abiertos es un fenómeno relativamente reciente. Estos datos, esenciales para el análisis y la transparencia, son principalmente proporcionados por instituciones públicas gracias a la implementación de la Ley de Acceso a la Información Pública 5282/2014 [1]. Varias instituciones clave han comenzado a liberar sus datos, entre ellas el Ministerio de Educación y Ciencias (MEC) y la Secretaría de la Función Pública (SFP). El MEC, por ejemplo, pone a disposición un conjunto de datos abiertos sobre los registros de títulos de educación superior [2], el cual incluye la lista de todos los egresados de instituciones de educación superior registrados en dicho ministerio. Estos registros comenzaron a publicarse en 2012, excepto los títulos de la Universidad Nacional de Asunción (UNA), que se integraron al portal de datos abiertos en 2018 [3]. Por su parte, la SFP, en cumplimiento de la Ley 5189/2014 [4], publica una nómina de salarios de los funcionarios estatales [5], proporcionando información detallada sobre los salarios percibidos mensualmente por empleados públicos. Es importante mencionar que no todos los entes cumplen con la publicación de los salarios [6] por lo que se buscó un mes en el cual haya una mayor cantidad de entidades en el dataset. En este trabajo propiamente se utilizó el del mes de abril de 2024.

El análisis de datos ha emergido como una herramienta crucial en la optimización de la gestión pública, permitiendo que las decisiones se fundamenten en datos concretos y objetivos. Con el auge de la ciencia de datos, la combinación y procesamiento de grandes volúmenes de información provenientes de diversas fuentes se ha convertido en una estrategia poderosa para obtener análisis profundos y reveladores. Esto es especialmente relevante para la formulación de políticas públicas efectivas que respondan a las necesidades reales de la población.

En este contexto, el presente proyecto tiene como propósito realizar un análisis exploratorio de los datos de registros de títulos de educación superior en combinación con la información sobre los funcionarios públicos de la Secretaría de la Función Pública. Este análisis busca identificar patrones y tendencias en la formación académica de los servidores públicos y cómo estos influyen en la calidad del servicio gubernamental. La investigación de estos patrones tiene el potencial de revelar cómo la educación en distintas disciplinas puede impactar la capacidad de los funcionarios para cumplir con las exigencias y competencias requeridas en el sector público. Este enfoque permite no solo una mayor comprensión de la relación entre la educación y el desempeño en el servicio público, sino también una base de evidencia que puede guiar futuras políticas de formación y desarrollo profesional en el país.

El objetivo de este estudio es, por lo tanto, contribuir a la mejora de la administración pública a través de la ciencia de datos, aportando información que permita tomar decisiones informadas para el fortalecimiento de las competencias en el sector público y la optimización de sus recursos humanos.

Breve Marco Teórico

La ciencia de datos se ha establecido como un área multidisciplinaria que combina técnicas estadísticas, algoritmos de aprendizaje automático y habilidades computacionales para el manejo y análisis de grandes cantidades de datos. Por lo tanto, es esencial comenzar con una definición precisa y crear un contexto amplio sobre el fenómeno en cuestión. En este contexto, la ciencia de datos incluye la recolección, depuración, procesamiento y análisis de datos, lo que permite identificar patrones relevantes que apoyan la toma de decisiones estratégicas

En el contexto de este proyecto, la ciencia de datos se convierte en la herramienta clave para integrar y analizar los datos provenientes de los registros de títulos académicos y la función pública. Este enfoque permite no solo identificar tendencias y patrones relevantes, sino también informar decisiones en la gestión pública.

De acuerdo con Ramírez [7], el conocimiento en la investigación científica se origina de la combinación de observaciones empíricas y una base conceptual robusta que permita la interpretación de los datos. En el ámbito de la ciencia de datos, estas observaciones empíricas se generan a partir del análisis de extensos conjuntos de datos, lo que convierte a esta disciplina en una herramienta crucial para sectores como la gestión pública, el marketing, la salud y la educación, entre otros.

Análisis Exploratorio de Datos (EDA)

El análisis exploratorio de datos (EDA) es una técnica fundamental que se utiliza para analizar conjuntos de datos con el fin de resumir sus principales características, a menudo mediante métodos gráficos. John Tukey introdujo el concepto de EDA en su obra de 1977, destacando la importancia de explorar los datos antes de aplicar modelos estadísticos o algoritmos de machine learning. Según Tukey, el EDA proporciona una visión intuitiva de los datos que ayuda a identificar patrones, valores atípicos (outliers) y relaciones que pueden no ser evidentes a través de otros métodos estadísticos.

El EDA es esencial para el proyecto en cuestión, ya que permitirá un análisis profundo de los datasets de títulos de educación superior y de la Secretaría de la Función Pública. A través del uso de herramientas como Python y sus bibliotecas especializadas (Pandas, Matplotlib y Seaborn), se busca obtener una visualización clara de los datos, facilitando la identificación de tendencias y relaciones significativas en el contexto de la función pública.

Uso de Python en la Ciencia de Datos

El lenguaje de programación Python ha ganado notable popularidad en la comunidad científica y tecnológica debido a su facilidad de uso y la amplia gama de bibliotecas especializadas que ofrece para el análisis de datos. Este lenguaje es especialmente eficaz

en el manejo de grandes volúmenes de datos y permite trabajar con ellos de manera versátil en todas las etapas del análisis, desde la recolección y limpieza de datos hasta la modelación predictiva y visualización avanzada. Python proporciona herramientas robustas para la manipulación de datos estructurados y no estructurados, lo que lo convierte en una opción ideal para proyectos de ciencia de datos en sectores variados.

En el estudio de Sagala [8] se demuestra cómo Python es utilizado para llevar a cabo análisis exploratorios y visualización de datos en eventos olímpicos, resaltando la flexibilidad de este lenguaje para gestionar diferentes tipos de datos y facilitar la detección de patrones relevantes. Esta flexibilidad permite realizar análisis exploratorios y visualizaciones que simplifican la interpretación de los resultados, características que son clave en proyectos de integración de datos y análisis en contextos de negocio y administración pública.

En el contexto de este trabajo, Python se empleará para realizar la unificación de conjuntos de datos y el análisis exploratorio de los registros de títulos académicos y de funcionarios públicos. La capacidad de Python para automatizar procesos de limpieza y transformación de datos permite integrar información proveniente de diversas fuentes, facilitando el análisis de relaciones entre la formación académica y el desempeño en el servicio público. Las bibliotecas de Python, como Pandas, NumPy y Matplotlib, juegan un papel central en el procesamiento y la visualización de datos, mientras que herramientas como Seaborn permiten realizar análisis más complejos y visualizaciones detalladas. En conjunto, estas herramientas ofrecen una plataforma robusta que respalda el análisis de patrones y tendencias significativas, subrayando la relevancia de Python en el desarrollo del presente proyecto.

Metodología:

El enfoque metodológico de este proyecto es cuantitativo, empleando técnicas de análisis exploratorio de datos (EDA) para extraer patrones y tendencias significativas a partir de los datasets combinados. El proceso comienza con la recolección de datos de varias fuentes fiables, seguido por una etapa de limpieza para asegurar la calidad y consistencia de los datos utilizados. Los datos fueron procesados utilizando scripts de Python, que incluyen pasos de preprocesamiento como normalización y eliminación de duplicados.

Posteriormente, se unificaron los datasets para permitir un análisis comparativo e interrelacional entre las variables de interés. Se utilizaron técnicas avanzadas de visualización de datos para facilitar la identificación de patrones de interés. Finalmente, se realizó un análisis detallado de los datos, enfocándose en la detección de coincidencias y relaciones relevantes.

Los materiales principales utilizados incluyen el acceso a Google Colab, que facilitó la programación en Python y el manejo de datasets robustos en un entorno de computación en la nube. Se emplearon librerías de Python como Pandas para la gestión y el procesamiento de los datos, y Matplotlib junto Seaborn para la visualización.

Los métodos de recolección implicaron la descarga de datasets abiertos de los registros oficiales del Ministerio de Educación y la Secretaría de la Función Pública. Los datos fueron almacenados en forma de archivos CSV y luego importados a Google Colab para su procesamiento. El análisis incluyó la utilización de técnicas de limpieza de datos para asegurar su calidad, como la imputación de datos faltantes y la detección de outliers.

Resultados

Se procedió a la descarga de los dataset dentro del cuaderno convirtiéndolos en dataframe [9], luego se procedió a la limpieza del campo documento para evitar duplicaciones, después se procedió a la unión de ambos dataframe en el campo documento para obtener los documentos que figuren en ambos datasets. Estos son los primeros resultados:

*Tabla 1: Principales características de cada dataset y coincidencias en documentos.
Elab. propia*

Descriptor	Df1 (títulos MEC)	Df2 (Salarios SFP)
Cantidad de filas y columnas	553.407 filas, 16 columnas	979.807 filas, 35 columnas
Cantidad de documentos únicos	406.280	316.761
Cantidad de coincidencias en documentos encontradas	127.334	

Luego se procedió a buscar los datos en cada dataset correspondientes a los documentos coincidentes, obteniéndose los siguiente principales resultados:

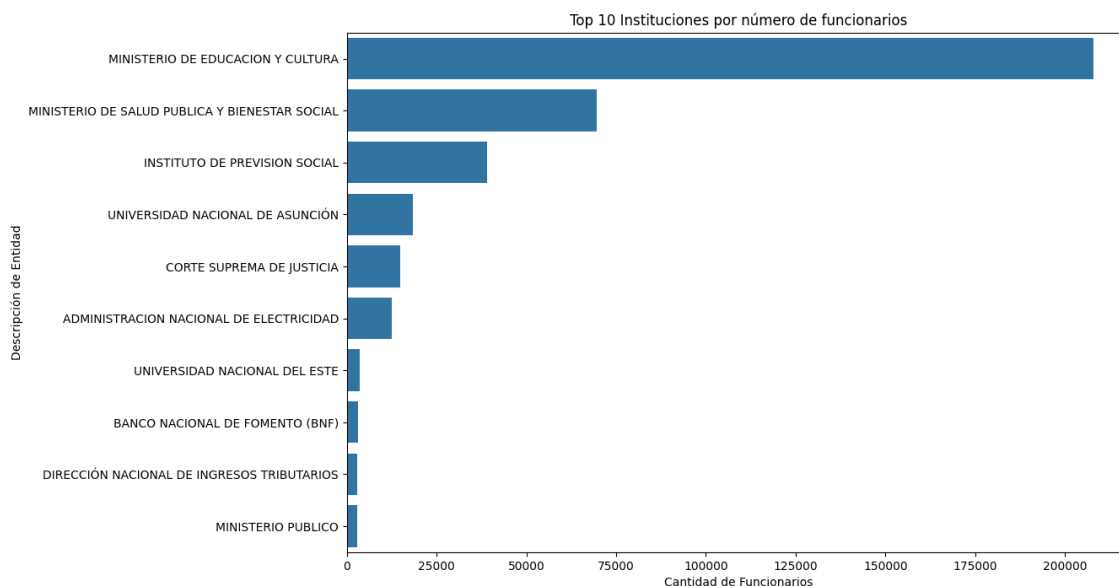


Ilustración 1: Principales instituciones donde figuran egresados de la E.S. por cantidad. Elaboración propia.

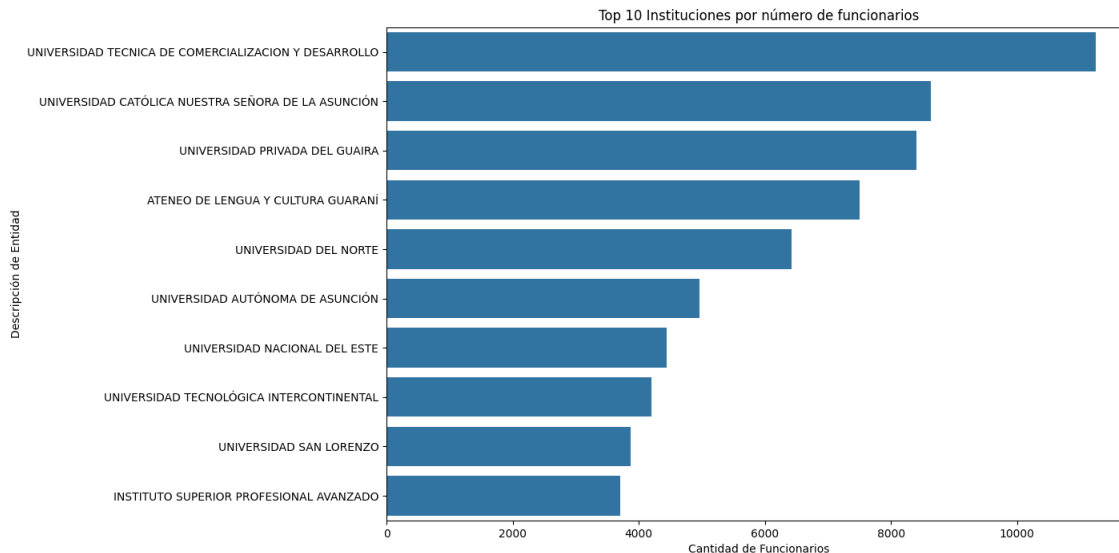
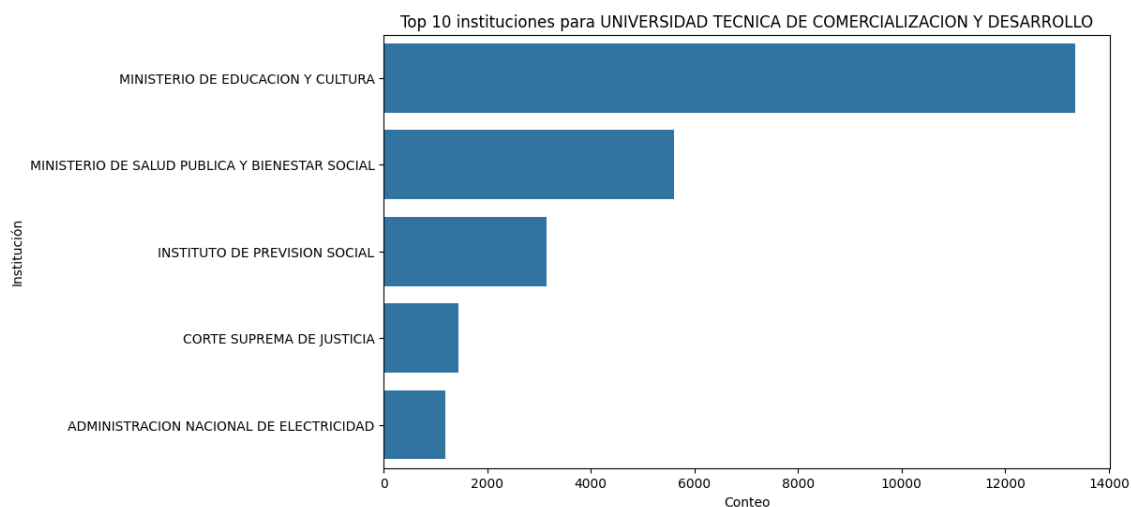


Ilustración 2: Principales instituciones cuyos egresados aparecen en el dataset de salarios. Elaboración propia.

Finalmente, se procedió a buscar dentro de las 3 primeras entidades públicas el origen principal de los funcionarios de las mismas, así como dentro de las 3 principales instituciones de educación superior, en donde se encuentran trabajando principalmente sus egresados. A continuación, los resultados:



*Ilustración 3: Principales entidades en las que trabajan los egresados de la UTCD.
Elab. propia.*

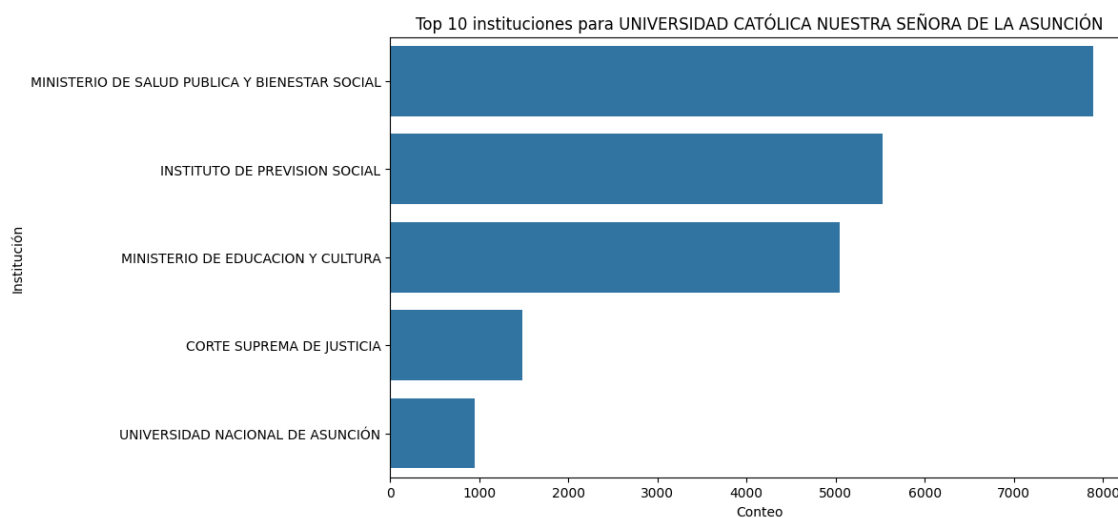


Ilustración 4: Principales entidades en las que trabajan los egresados de la UC. Elab. propia.

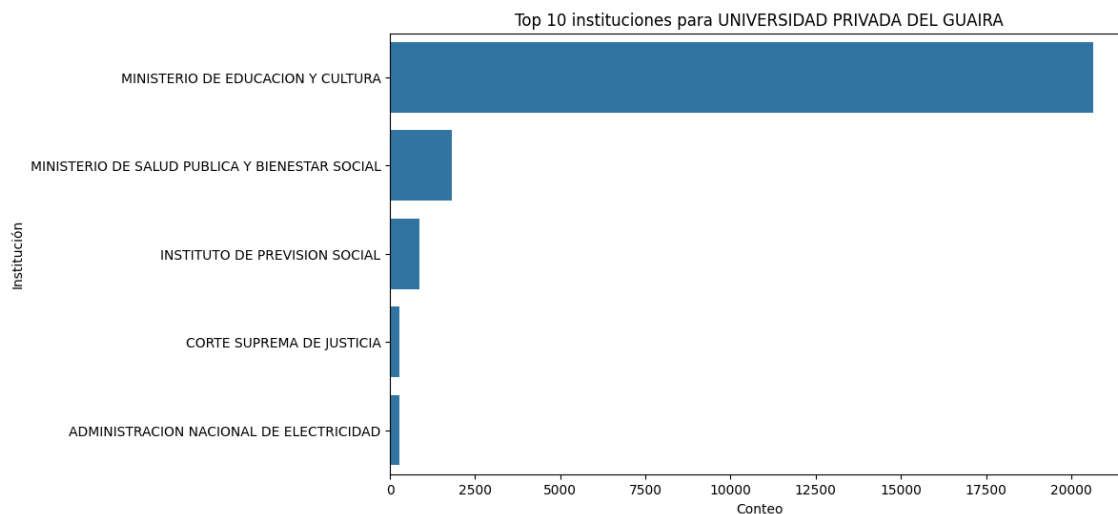


Ilustración 5: Principales entidades en las que trabajan los egresados de la UPG. Elab. propia.

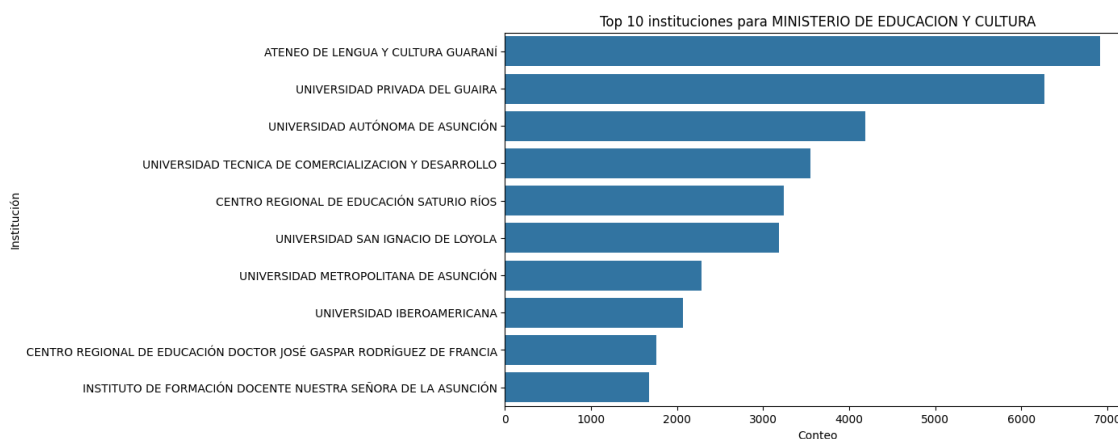


Ilustración 6: Principales instituciones de donde provienen los empleados del MEC. Elab. propia.

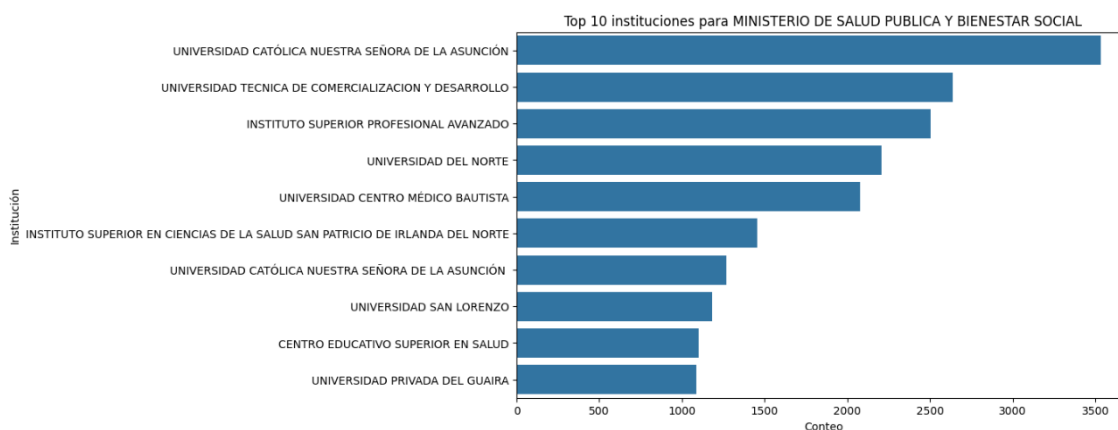
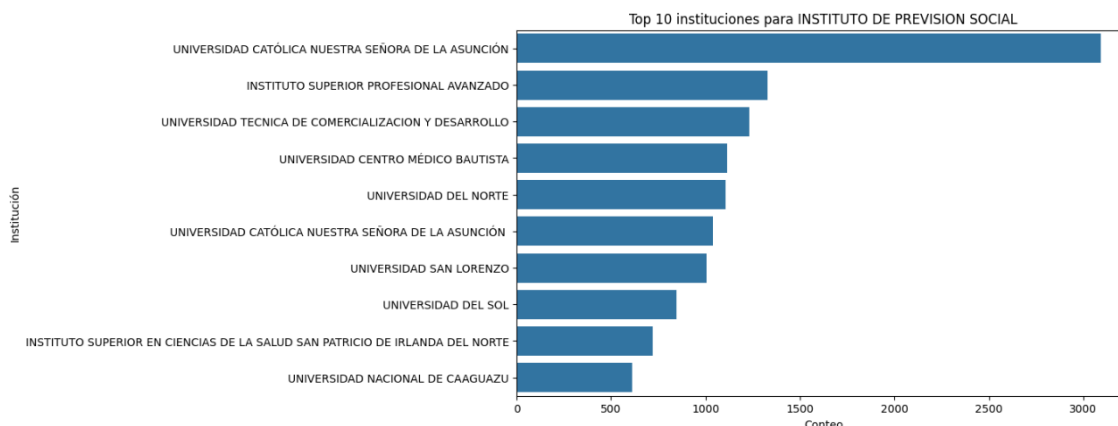


Ilustración 7: Principales instituciones de donde provienen los empleados del MSP y BS. Elab. propia.



*Ilustración 8: Principales instituciones de donde provienen los funcionarios del IPS.
Elab. propia.*

Discusión

No se ha encontrado en la bibliografía consultada ningún trabajo similar al presente por lo que no se puede hacer una comparación. Se espera que este sirva de base para futuros estudios similares. Sobre el dataset de títulos si hay algunos trabajos que brindan información y conocimiento a partir de los datos publicados tales como [3].

Se intentó buscar una correlación entre el año de egreso en la institución y el ingreso a la entidad como funcionario, pero no se pudo establecer un patrón o tendencia al encontrarse que muchos ya eran funcionarios al momento de egresar, en varios casos con bastantes años de antigüedad, por lo que se puede especular que ingresaron a trabajar antes de estudiar o que egresaron antes de que se empiecen a publicar los datos por el MEC.

Además, las coincidencias entre ambos dataset solo representan un 40% de los documentos únicos del total de funcionarios, por lo que en principio no se pueden sacar conclusiones que afecten al total del plantel.

Conclusiones

Gracias al código utilizado el cual está disponible como fuente [10] se tienen las siguientes informaciones:

* De un total de 406.280 egresados de la E.S. a partir del 2012, 127.334 se encuentran trabajando en la función pública, un 31% del total.

* Estos 127.334 representan un 40% del total de 316.761 funcionarios públicos cuyos datos aparecen en el dataset utilizado.

* Las principales entidades en las que trabajan los egresados del conjunto hallado son el MEC (con mucha diferencia el primero), el MSP y BS, el IPS, la UNA y la CSJ. A continuación el análisis de las 3 primeras:

* Las principales instituciones donde egresaron los funcionarios del MEC son el Ateneo, la UPG, la UAA, la UTCD, el Centro Regional Saturio Ríos, la USIL, la UMA, la UNIBE, casi todas ellas privadas.

* Las principales instituciones donde egresaron los funcionarios del MSP son la UCA, la UTCD, el ISPA, la UNINORTE, la UCMB, todas privadas.

* Las principales instituciones donde egresaron los funcionarios del IPS son la UCA, el ISPA, la UTCD, la UCMB, UNINORTE, UNISAL, todas privadas.

* Las principales instituciones que más “aportan” funcionarios públicos son la UTCD, la UCA, la UPG, el Ateneo Guaraní y la UNINORTE, todas privadas. A continuación, el análisis de las 3 primeras:

* Las principales entidades donde trabajan los egresados de la UTCD son el MEC, el MSPyBS, el IPS y la CSJ.

* Las principales entidades donde trabajan los egresados de la UCA son el MSPyBS, el IPS, el MEC y la CSJ.

* Las principales entidades donde trabajan los egresados de la UPG son el MEC, el MSPyBS, el IPS y la CSJ.

De esta manera se puede establecer como principal conclusión que una gran cantidad de egresados de instituciones privadas terminan trabajando como empleados públicos, y a la vez la mayoría de las instituciones públicas tienen su plantel conformado por egresados de instituciones privadas, siempre teniendo en cuenta las observaciones mencionadas en la sección de discusión.

Bibliografía

- [1] Congreso Nacional del Paraguay, «Ley N° 5282 / LIBRE ACCESO CIUDADANO A LA INFORMACIÓN PÚBLICA Y TRANSPARENCIA GUBERNAMENTAL,» Leyes Paraguayas, setiembre 2014. [En línea]. Available: <https://www.bacn.gov.py/leyes-paraguayas/3013/libre-acceso-ciudadano-a-la-informacion-publica-y-transparencia-gubernamental>.
- [2] MEC, «Registro de Títulos,» [En línea]. Available: https://datos.mec.gov.py/data/registros_titulos. [Último acceso: 2024].
- [3] H. Estigarribia Barreto, «Análisis de la oferta académica y egreso de carretas TIC en la educación superior en Paraguay, basado en el estudio de datos abiertos,» *ACADEMO Revista de investigación en ciencias sociales y humanidades*, vol. 11, n° 2, pp. 201-208, 2024.
- [4] Congreso Nacional del Paraguay, «Ley N° 5189 / ESTABLECE LA OBLIGATORIEDAD DE LA PROVISIÓN DE INFORMACIONES EN EL USO DE LOS RECURSOS PÚBLICOS SOBRE REMUNERACIONES Y OTRAS RETRIBUCIONES ASIGNADAS AL SERVIDOR PÚBLICO DE LA REPÚBLICA DEL PARAGUAY,» Leyes Paraguayas, mayo 2014. [En línea]. Available: <https://www.bacn.gov.py/leyes-paraguayas/2962/ley-n-5189-establece-la-obligatoriedad-de-la-provision-de-informaciones-en-el-uso-de-los-recursos-publicos-sobre-remuneraciones-y-otras-retribuciones-asignadas-al-servidor-publico-de-la-republica-del-paraguay>.
- [5] SFP, «Nómina de Funcionarios,» 2024. [En línea]. Available: <https://datos.sfp.gov.py/data/funcionarios>. [Último acceso: octubre 2024].
- [6] Diario UH, «El 78% de entes cumplen a medias publicación de salarios,» Diario última hora, 23 agosto 2024. [En línea]. Available: <https://www.ultimahora.com/el-78-de-entes-cumplen-a-medias-publicacion-de-salarios>.
- [7] A. V. Ramírez, «La teoría del conocimiento en investigación científica: una visión actual,» *An. Fac. med.*, 2009. [En línea]. Available: <http://www.scielo.org.pe/pdf/afm/v70n3/a11v70n3.pdf>.
- [8] N. & A. F. Sagala, «Exploratory Data Analysis (EDA): A Study of Olympic Medallist,» 2022. [En línea]. Available: [https://www.semanticscholar.org/paper/Exploratory-Data-Analysis-\(EDA\)%3A-A-Study-of-Olympic-Sagala-Aryatama](https://www.semanticscholar.org/paper/Exploratory-Data-Analysis-(EDA)%3A-A-Study-of-Olympic-Sagala-Aryatama).

- [9] W. McKinney, «pandas.DataFrame,» Python documentation Library, abril 2024.
[En línea]. Available: <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html>.
- [10 C. J. Coronel Rojas, R. Saucedo, F. Villar y H. R. Estigarribia Barreto, «EDA
] títulos MEC y SFP,» Notebook de Google Colab, octubre 2024. [En línea].
Available:
https://colab.research.google.com/drive/1CFNqwQ_vhKbqtpzImjOq2sysVGnpY3Y3?usp=sharing.
- [11 E. Sansegundo Hernando, Guía práctica de introducción al Análisis Exploratorio
] de Datos, 2021.