informe_extracción

November 4, 2025

1 Informe de Extracción de datos:

2 Fondecyt 11241304, Re-conociendo las des-igualdades de la Academia Chilena: Un análisis interseccional de género y clase en las trayectorias de personas con grado de doctorado

2.1 Sexo

La variable sexo fue construida utilizando un enfoque combinado de herramientas automatizadas y supervisión humana. Inicialmente, se empleó el paquete `genderizeR` de R (Realizado en Rstudio), una herramienta que permite inferir el género a partir de nombres propios. Este paquete utiliza bases de datos externas y modelos estadísticos para asignar probabilidades de género (masculino, femenino u otros) a cada nombre analizado.

Sin embargo, dado que los resultados de herramientas automatizadas pueden contener errores o sesgos, se incorporó un proceso de supervisión humana individual.

Variables añadidas: sexo, es_mujer

2.1.1 Recursos:

Detalle del trabajo en: - https://github.com/RodrigoMolinaAvila/Fondecyt-11241304/tree/main/Genderizaci%C3%B3n - https://github.com/RodrigoMolinaAvila/Fondecyt-11241304/blob/main/Genderizaci%C3%B3n/Namesplit%20-%20Gender.R

2.2 Pregrado

Los datos utilizados en este análisis provienen de diversas fuentes, incluyendo el Portal del Investigador, una plataforma oficial que centraliza información sobre investigadores y proyectos en Chile. Además, se realizó una exploración manual llevada a cabo por el personal técnico del proyecto Fondecyt, utilizando recursos como LinkedIn, Scopus y sitios web oficiales de universidades.

Para la calidad de los establecimientos de pregrado se realizó un procesamiento manual para establecer criterios categóricos y de localidad, las fuentes utilizadas fueron la pagina oficial de la CNA (Comisión Nacional de Investigación).

Variables añadidas: acreditacion_pregrado, años_acreditacion_regrado, elite, nombre_pregrado, region_pregrado

Como consideraciones, de los 3145 registros totales de becarios, solo 3075 valores contienen información verificable sobre el paso por la institución de pregrado. Este número fue aceptado como

tope de la busqueda posible

2.3 Secundaria

Los datos sobre la educación secundaria se basa en datos extraídos del Ministerio de Educación mediante técnicas de webscraping, utilizando Python y Selenium. Este proceso permitió recopilar 2619 contienen información válidos asociados a becarios que estudiaron en Chile.

El análisis cruza el valor RBD del establecimiento educacional secuidario de los becarios con el extraido de los datos del Ministerio de educación. Este valor fue otorgado por el Portal del Investigador.

Variables añadidas: nombre_secundaria, region_secundaria, comuna_secundaria, gse_secundaria

Como consideracion, parte de la muestra incluye individuos que cursaron su enseñanza secundaria en el extranjero, lo que introduce una variable adicional en el análisis.

2.3.1 Recursos:

Detalle del trabajo en: - https://github.com/RodrigoMolinaAvila/Fondecyt-11241304/tree/main/webscraping%20RBD - https://github.com/RodrigoMolinaAvila/Fondecyt-11241304/blob/main/webscraping%20RBD/webscraping_rbd.py

2.4 Shanghái Rank

Los datos fueron obtenidos directamente desde la página oficial del Ranking Shanghái, utilizando técnicas de webscraping. Este proceso se llevó a cabo mediante la ejecución de scripts en JavaScript a través de la consola del explorador, lo que permitió extraer información estructurada directamente desde el sitio web.

El ranking corresponde a su edición del año 2024.

De los 3143 becarios, se identificaron 2112 registros asociados a universidades clasificadas en el Ranking Shanghái.

Asumimos que no todas las universidades de postgrado a las cuales accedieron están incluidas en este ranking.

Se construyó una variable que identifica si una universidad pertenece al top 100 del Ranking Shanghái.

Variables añadidas: university_shanghai, shanghai_rank, top_100_shanghai + variables asociadas a las métricas del ranking.

2.4.1 Recursos:

Detalle del trabajo en: - https://github.com/RodrigoMolinaAvila/Fondecyt-11241304/tree/main/university_ranking - https://github.com/RodrigoMolinaAvila/Fondecyt-11241304/blob/main/university_ranking/shanghai_extract.js

2.5 QS Rank

Los datos fueron descargados directamente desde la página oficial del Ranking QS. Este proceso se llevó a cabo mediante la descarga de la base de datos disponible en el sitio web, lo que permitió recopilar información estructurada sobre las universidades clasificadas en este ranking.

El ranking corresponde a su edición del año 2025.

De los 3143 becarios, se identificaron 3077 registros asociados a universidades clasificadas en el Ranking QS.

Se entiende que no todas las universidades de postgrado a las cuales accedieron están incluidas en este ranking.

Se construyó una variable que identifica si una universidad pertenece al top 100 del Ranking QS.

Variables añadidas: university_qs, qs_rank, top_100_qs + variables asociadas a las métricas del ranking.

2.5.1 Recursos:

Detalle del trabajo en: - https://github.com/RodrigoMolinaAvila/Fondecyt-11241304/tree/main/university_ranking - https://github.com/RodrigoMolinaAvila/Fondecyt-11241304/blob/main/university_ranking/universities_qs.py

2.6 Scopus - Author API

La extracción de datos se realizó utilizando la API de programación de Elsevier, atraves del endpoint `author_retrieval`. Este proceso fue habilitado mediante el uso de credenciales de acceso, incluyendo una `API_KEY` y un `INST_TOKEN`, que permiten realizar solicitudes a los servidores de Elsevier bajo una cuota de extracción definida.

La identificación y validación del Author ID os becarios se llevó a cabo por el personal técnico del proyecto, combinando herramientas automatizadas y supervisión humana. De los 3143 registros totales de becarios, se identificaron 2704 valores.

Se asume que que no todos los becarios optan por carreras académicas ni publican exclusivamente en revistas indexadas.

Variables añadidas: author_id, indexed_name, first_publication_year, last_publication_year, document_count, cited_by_count, coauthor_count, h-index, subject_areas, affiliation_display_name, affiliation_parent_id, affiliation_parent_name, affiliation_id, affiliation_name, affiliation_city, affiliation_country, asjc, affiliation_history, scopus_publications

2.6.1 Recursos:

Detalle del trabajo en: - https://github.com/RodrigoMolinaAvila/Fondecyt-11241304/tree/main/scopus - https://github.com/RodrigoMolinaAvila/Fondecyt-11241304/blob/main/scopus/scripts/api_extract_authors.py

2.7 Scopus - Abstact API

La extracción de datos mediante la API de programación de Elsevier a través del endpoint `abstract_retrieval`. Este proceso fue habilitado mediante el uso de credenciales de acceso,

incluyendo una `API_KEY` y un `INST_TOKEN`, que permiten realizar solicitudes estructuradas a los servidores de Elsevier bajo una cuota de extracción definida.

De un total de 29549 publicaciones identificadas en la suma de la varibale document_count de los datos de Author API, se logró extraer información completa de 28438 publicaciones.

Variables añadidas: scopus_id, title, abstract, keywords, year, month, day, grant_year, diff_year, period, publicationname, aggtype, subtype, citedbycount, doi, url

2.7.1 Recursos:

Detalle del trabajo en: - https://github.com/RodrigoMolinaAvila/Fondecyt-11241304/tree/main/scopus - https://github.com/RodrigoMolinaAvila/Fondecyt-11241304/blob/main/scopus/scripts/api_extract_abstract.py