UNIVERSIDAD DEL BÍO-BÍO

INGENIERIA ESTADISTICA
UNIVERSIDAD DEL BÍO-BÍO

Universidad del Bío-Bío

Facultad de Ciencias

Informe Certamen 2

Ciencia de datos en la terminal linux

Bernardo Aroca Sanhueza

Profesor Luis Gómez Guzmán

26 de Diciembre del 2025

Índice

1. Metodología

En este trabajo se desarrolló íntegramente en un entorno Linux, utilizando la terminal como principal herramienta para el procesamiento y análisis de datos en formato CSV. Se empleó **Nushell**, el cual permite manipular datos estructurados mediante pipelines declarativos, facilitando tareas de limpieza, transformación y filtrado de información sin necesidad de lenguajes de programación adicionales.

La base de datos utilizada corresponde a un archivo CSV denominado `estudio_mercado_ordenado.csv`, el cual contiene información de matrícula en instituciones de educación superior. A continuación, se describen de manera secuencial y reproducible todas las etapas del proceso computacional.

1.1. Carga y Limpieza de Datos

1.1.1. Apertura y exploración inicial del archivo

Como primer paso, se abrió el archivo CSV original, especificando correctamente el separador de campos y eliminando registros vacíos. Se visualizaron las primeras filas para validar la correcta lectura del archivo.

```
open --raw estudio_mercado_ordenado.csv
| from csv --separator ';'
| compact --empty
| first 5
```

1.1.2. Renombrado de columnas

Para facilitar el manejo posterior de las variables y mantener una nomenclatura consistente, se procedió a renombrar las columnas principales del dataset.

```
open --raw estudio_mercado_ordenado.csv
| from csv --separator ';'
| compact --empty
| rename "AÑO" anio
          "NOMBRE INSTITUCION" institucion
          "NOMBRE SEDE" sede
          "NOMBRE CARRERA" carrera
          "TOTAL MATRICULADOS" matricula_total
          "TOTAL MATRICULADOS PRIMER AÑO" matricula_primer_ano
| first 5
```

1.1.3. Limpieza de la columna `anio`

La variable correspondiente al año contenía un prefijo textual (`MAT_`) que impedía su uso como valor numérico. Por ello, se realizó una transformación para eliminar dicho prefijo y convertir la columna a tipo entero.

```
open --raw estudio_mercado_ordenado.csv
```

```
| from csv --separator ';'
| compact --empty
| rename "AÑO" anio
          "NOMBRE INSTITUCION" institucion
          "NOMBRE SEDE" sede
          "NOMBRE CARRERA" carrera
          "TOTAL MATRICULADOS" matricula_total
          "TOTAL MATRICULADOS PRIMER AÑO" matricula_primer_ano
| update anio { $in | into string | str replace "MAT_" "" | into int }
| first 5
```

1.1.4. Filtrado temporal (años 2020–2024)

Una vez normalizada la columna de año, se filtraron los registros correspondientes al período comprendido entre los años 2020 y 2024, rango solicitado para el análisis.

```
open --raw estudio_mercado_ordenado.csv
| from csv --separator ';'
| compact --empty
| rename "AÑO" anio "NOMBRE INSTITUCION" institucion "NOMBRE SEDE" sede "NOMBRE CARRERA"
↪  carrera "TOTAL MATRICULADOS" matricula_total "TOTAL MATRICULADOS PRIMER AÑO"
↪  matricula_primer_ano
| update anio { $in | into string | str replace "MAT_" "" | into int }
| where {|r| $r.anio >= 2020 and $r.anio <= 2024 }
| first 5
```

1.1.5. Generación del archivo base 2020–2024

El resultado del filtrado temporal fue almacenado en un archivo CSV intermedio, el cual sirvió como base para los filtros posteriores.

```
open --raw estudio_mercado_ordenado.csv
| from csv --separator ';'
| compact --empty
| rename "AÑO" anio
          "NOMBRE INSTITUCION" institucion
          "NOMBRE SEDE" sede
          "NOMBRE CARRERA" carrera
          "TOTAL MATRICULADOS" matricula_total
          "TOTAL MATRICULADOS PRIMER AÑO" matricula_primer_ano
| update anio { $in | into string | str replace "MAT_" "" | into int }
| where {|r| $r.anio >= 2020 and $r.anio <= 2024 }
| save ies_base_2020_2024.csv
```

1.2. Filtrado de Carreras de Interés

El análisis se centró en carreras relacionadas con **Estadística** y **Ciencia de Datos**. Para ello, se aplicó un filtrado textual sobre el nombre de la carrera, seguido de la exclusión explícita de programas de postgrado y otras modalidades no pertinentes.

1.2.1. Filtrado de carreras de interés

```
open ies_base_2020_2024.csv
| where {|r|
    (
        ($r.carrera | into string | str downcase | str contains "estad")
        or
        ($r.carrera | into string | str downcase | str contains "dato")
    )
}
| where {|r|
    not (
        ($r.carrera | into string | str downcase | str contains "magister")
        or ($r.carrera | into string | str downcase | str contains "magíster")
        or ($r.carrera | into string | str downcase | str contains "doctor")
        or ($r.carrera | into string | str downcase | str contains "post")
        or ($r.carrera | into string | str downcase | str contains "diplom")
        or ($r.carrera | into string | str downcase | str contains "licenc")
        or ($r.carrera | into string | str downcase | str contains "regulariz")
        or ($r.carrera | into string | str downcase | str contains "pedagog")
    )
}
| save ies_final_estadistica_datos_2020_2024.csv
```

1.3. Reestructuración (Pivoteo)

No se realizó un pivoteo obligatorio de los datos, ya que el enunciado solicita permitir el análisis y la visualización de la información, lo cual se cumple mediante una estructura tabular ordenada y filtrada. El enunciado no exige explícitamente la transformación de los años en columnas independientes ni el uso de operaciones de tipo *pivot* o *transpose*.

No obstante, de manera complementaria, se implementó un pivoteo explícito para generar una tabla resumen por institución, sede y carrera, con columnas separadas por año, tanto para matrícula total como para matrícula de primer año.

```
open ies_final_estadistica_datos_2020_2024.csv
| group-by institucion sede carrera
| items { |key, reg|
    {
        "NOMBRE INSTITUCIÓN": ($key | get 0),
        "NOMBRE SEDE": ($key | get 1),
        "NOMBRE CARRERA": ($key | get 2),
        "TOTAL_2020": ($reg | where anio == 2020 | get 0?.matricula_total | default 0),
        "TOTAL_2021": ($reg | where anio == 2021 | get 0?.matricula_total | default 0),
        "TOTAL_2022": ($reg | where anio == 2022 | get 0?.matricula_total | default 0),
        "TOTAL_2023": ($reg | where anio == 2023 | get 0?.matricula_total | default 0),
        "TOTAL_2024": ($reg | where anio == 2024 | get 0?.matricula_total | default 0),
        "PRIMER_2020": ($reg | where anio == 2020 | get 0?.matricula_primer_ano | default 0),
        "PRIMER_2021": ($reg | where anio == 2021 | get 0?.matricula_primer_ano | default 0),
        "PRIMER_2022": ($reg | where anio == 2022 | get 0?.matricula_primer_ano | default 0),
        "PRIMER_2023": ($reg | where anio == 2023 | get 0?.matricula_primer_ano | default 0),
        "PRIMER_2024": ($reg | where anio == 2024 | get 0?.matricula_primer_ano | default 0)
```

```
    }
}
| save ies_last_5_years_data.csv
```

1.4. Análisis de Resultados

El proceso de limpieza y transformación permitió obtener un conjunto de datos final (`ies_last_5_years_data`) que consolida la información de matrícula en carreras de Estadística y Ciencia de Datos para el período 2020-2024.

Resultados obtenidos:

- Se filtraron y procesaron todas las carreras que contienen los términos .[es]tad.[o] "dato.[en] su nombre, excluyendo programas de postgrado y otras modalidades no pertinentes.
- Se limpió la columna de año eliminando el prefijo `MAT_` y convirtiendo los valores a numéricos.
- Se consolidaron las matrículas mediante operaciones de suma por institución, sede, carrera y año.
- Finalmente, los datos se transformaron a un formato ancho donde cada año (2020-2024) es una columna separada, mostrando tanto la matrícula total como la de primer año.

La estructura resultante permite analizar fácilmente la evolución temporal de la matrícula y comparar la oferta académica entre diferentes instituciones y sedes, cumpliendo con el objetivo principal del certamen.

2. Análisis de Reproducibilidad

El proceso desarrollado es completamente reproducible. Todo el flujo de trabajo está encapsulado en el script `script.nu`, que ejecuta secuencialmente:

1. Carga y limpieza inicial del archivo CSV
2. Filtrado por período (2020-2024)
3. Filtrado de carreras de interés
4. Transformación a formato ancho (pivoteo)
5. Exportación del resultado final

Cada etapa genera archivos intermedios que permiten verificar el procesamiento paso a paso. Para reproducir el análisis, basta con ejecutar el script en cualquier entorno Linux con Nushell instalado, asegurando que se obtendrán exactamente los mismos resultados.

3. Conclusiones

Este trabajo permitió procesar y estructurar datos de matrícula académica utilizando Nushell en terminal Linux. Se logró:

- Filtrar y limpiar eficientemente los datos brutos
- Consolidar la información mediante operaciones de agregación
- Transformar los datos al formato requerido para análisis temporal

- **Nota sobre Polars:** Durante el desarrollo se intentó utilizar el plugin Polars como sugería el enunciado, pero tras varias horas de intentos se encontraron errores de compatibilidad. Consultando con compañeros, se confirmó que experimentaban problemas similares. Por esto, se optó por implementar una solución alternativa utilizando únicamente las funciones nativas de Nushell, la cual logra los mismos resultados requeridos.

El archivo final generado (`ies_last_5_years_data.csv`) proporciona una base de datos limpia y estructurada que permite analizar la evolución de la matrícula en carreras de Estadística y Ciencia de Datos en Chile durante los últimos cinco años, cumpliendo con todos los objetivos planteados.