INFORME PROYECTO FINAL

HENRY/TELECOM - COHORTE 03

GRUPO 7:

Candia, Andres

Fonte, Alejandro

Picca, Pablo

Scorolli, Leandro Damian

1. Contexto y objetivo del Sprint 1

En el Sprint 1 del Proyecto Final de Data Analytics se trabajó sobre el dataset obligatorio de ENACOM, con el objetivo de:

- Construir una **base de datos relacional limpia** a partir de múltiples fuentes CSV.
- Realizar un **Preparación de datos y Análisis Exploratorio de Datos (EDA)** que:
    - Aborde **duplicados, valores faltantes y outliers**.
    - Utilice **gráficos acordes al tipo de variable** (cualitativa vs cuantitativa).
    - Acompañe cada visualización con **análisis y conclusiones**.
- Dejar listo el modelo para:
    - Implementación en SQL Server/Big Query en el Sprint 2.
    - Construcción del dashboard en Looker Studio en Sprints posteriores.

El EDA se documentó con celdas **Markdown** en el notebook, siguiendo un esquema paso a paso y justificando cada decisión (criterios de limpieza, elección de gráficos, manejo de outliers, etc.).

2. Modelo de datos final (tablas y relaciones)

Se consolidaron los siguientes archivos CSV, que forman el modelo relacional:

Tablas de dimensiones (variables cualitativas)

- **Provincias.csv**
    - ID_Provincia
    - 24 provincias + CABA.
    - Variable **cualitativa nominal**.
- **Localidades.csv**
    - ID_Localidad, ID_Provincia, Partido, Localidad
    - Permite bajar el análisis al nivel de localidad.
    - Variables cualitativas: provincia, partido, localidad.

- Variable cuantitativa implícita: cantidad de localidades por provincia.

- **Periodos.csv**
  - ID_Periodo, Año, Trimestre
  - Eje temporal del análisis (2014-T1 a 2025-T1).
  - Año (cuantitativa discreta), Trimestre (ordinal).
- **Tecnologias.csv**
  - ID_Tecnologia
  - ADSL, CABLEMODEM, FIBRA OPTICA, SATELITAL, WIRELESS, WIMAX, DIAL UP, Otros.
  - Cualitativa nominal.
- **Rangos.csv**
  - ID_Rango_Velocidad
  - 0–10 Mbps, 10–30 Mbps, 30–50 Mbps, 50–100 Mbps, 100–300 Mbps, 300–1000 Mbps, >1000 Mbps.
  - Cualitativa ordinal (rangos de velocidad).

Tablas de hechos (variables cuantitativas principales)

- **Conexiones_Periodos.csv**
  - Nivel país por período.
  - Métricas cuantitativas:
    - Banda Ancha Fija, Dial Up, Total(BAF+DU)
    - Accesos por tecnología: ADSL, Cablemodem, Fibra Optica, Wireless, Otros(Tecnologias), Total(Tecnologias)
    - Accesos por rango de velocidad: columnas desde Hasta 512 kbps hasta + 30 Mbps, Total(Velocidades)
    - Ingresos (en miles) y penetración: Accesos cada 100 Hogares, Accesos cada 100 Habitantes
    - ID_Periodo (FK).
- **Conexiones_Provincias.csv**
  - Nivel provincia por período.

- Mismas métricas que la tabla anterior + Mbps (Media de bajada) por provincia y período.
- Claves: ID_Conexion_Prov, ID_Provincia, ID_Periodo.

- **Velocidades_Prov.csv**
  - Accesos por **provincia + velocidad + período + rango**.
  - Claves: ID_Acceso_Prov, ID_Provincia, Velocidad, ID_Periodo, ID_Rango_Velocidad
  - Métrica: Accesos (cuantitativa de conteo).
- **Velocidades_Loc.csv**
  - Accesos por **localidad + velocidad + período**.
  - Claves: ID_Acceso_Loc, ID_Localidad, Velocidad, ID_Periodo
  - Métrica: Accesos.
- **Accesos_Ultimo_Trim.csv**
  - Foto del **último trimestre disponible** (2025-T1) a nivel localidad y tecnología.
  - Claves: ID_Acceso_UT, ID_Localidad, ID_Tecnologia
  - Métrica: Accesos.

Este diseño permite consultas analíticas robustas (por tiempo, provincia, localidad, tecnología y rango de velocidad), cumpliendo el objetivo de preparar la base para BI.

3. Preparación y calidad de datos

3.1. Importación y tipos de dato

Todos los CSV se importaron en Python usando:

- Separador de columnas: ;
- Coma decimal en campos numéricos provenientes de Excel/ENACOM, que se transformó a punto para análisis.
- Conversión de tipos:
  - Variables de conteo (accesos, conexiones) → int64
  - Velocidades medias y métricas derivadas → float64

- Claves, provincias, tecnologías, rangos, períodos → object (categóricas)

3.2. Duplicados

Se verificó:

df.duplicated().sum()

para cada tabla. Resultado:

- **Todas las tablas tienen 0 filas duplicadas.**
- Las claves de negocio (ID_Conexion_Per, ID_Conexion_Prov, ID_Acceso_Prov, ID_Acceso_Loc, ID_Acceso_UT, ID_Periodo, ID_Localidad, ID_Provincia) se comportan como IDs únicos, lo cual es coherente con su uso posterior como claves primarias en SQL.

3.3. Valores faltantes

Se ejecutó:

df.isna().sum()

En todos los datasets, el conteo de nulos por columna es **0**, es decir:

- No hubo necesidad de imputar valores.

3.4. Conversión numérica y limpieza de formato

En tablas con columnas que llegaban como texto por comas decimales (por ejemplo, Ingresos, Mbps (Media de bajada), Accesos cada 100 Hogares), se aplicó:

df[col] = df[col].astype(str).str.replace(".", "", regex=False)

df[col] = df[col].str.replace(",", ".", regex=False)

df[col] = pd.to_numeric(df[col], errors="coerce")

Esto aseguró que todas las métricas se analizaran como **variables cuantitativas** y se pudieran usar en operaciones estadísticas, correlaciones y gráficos.

4. Outliers (IQR × 3)

Para todas las variables cuantitativas relevantes se aplicó el criterio:

- Q1 y Q3 (cuartiles)
- IQR = Q3 – Q1

- Límite superior = Q3 + 3 × IQR
  (más conservador que el 1.5×IQR estándar)

Se listaron los **10 valores más altos** por columna que superaban ese límite.

4.1. Tablas más relevantes

- **Accesos_Ultimo_Trim.csv**
  - Métrica analizada: Accesos
  - Outliers superiores (IQR×3): **1367 localidades**.
  - Interpretación:
    - Corresponden a grandes localidades o nodos urbanos (ej. AMBA) con muchísimos accesos.
    - No se eliminaron: son valores reales, importantes para entender la concentración de mercado.
- **Velocidades_Loc.csv**
  - Métrica: Accesos
  - Outliers superiores: **2352 registros**.
  - Representan localidades con niveles de penetración de Internet muy superiores al resto para ciertas velocidades.
- **Velocidades_Prov.csv**
  - Métrica: Accesos
  - Outliers superiores: **2038 combinaciones provincia–velocidad–período**.
  - Principalmente en provincias densamente pobladas (Buenos Aires, CABA, Córdoba, Santa Fe) y en rangos de velocidad altos.
- **Conexiones_Provincias.csv**
  - Múltiples columnas con outliers (IQR×3), por ejemplo:
    - Total(BAF+DU): 161 outliers.
    - Banda Ancha Fija: 159 outliers.
    - Cablemodem: 157 outliers.
    - Fibra Optica: 86 outliers.

- De nuevo, corresponden a provincias grandes y períodos recientes donde el nivel de servicio crece fuerte.

- **Conexiones_Periodos.csv**
  - Casi todas las columnas tienen 0 outliers con IQR×3, salvo algunos casos puntuales en ciertos rangos de velocidad (por ejemplo, Hasta 512 kbps, + 10 Mbps - 20 Mbps), lo cual indica cambios de tecnología a lo largo del tiempo.

Decisión

- **No se eliminaron outliers**, porque describen contextos reales del mercado argentino (provincias grandes, crecimiento de fibra, etc.).
- Se decidió mantenerlos y **documentarlos**, mencionando explícitamente en el informe que son parte de la historia que el dashboard deberá contar (brecha entre provincias y concentración en altas velocidades).

5. Variables cualitativas vs cuantitativas y gráficos usados

El EDA se diseñó explícitamente para **mostrar que se comprende la diferencia entre variables cualitativas y cuantitativas**, tal como pidió el mentor:

5.1. Variables cualitativas (nominales / ordinales)

Ejemplos:

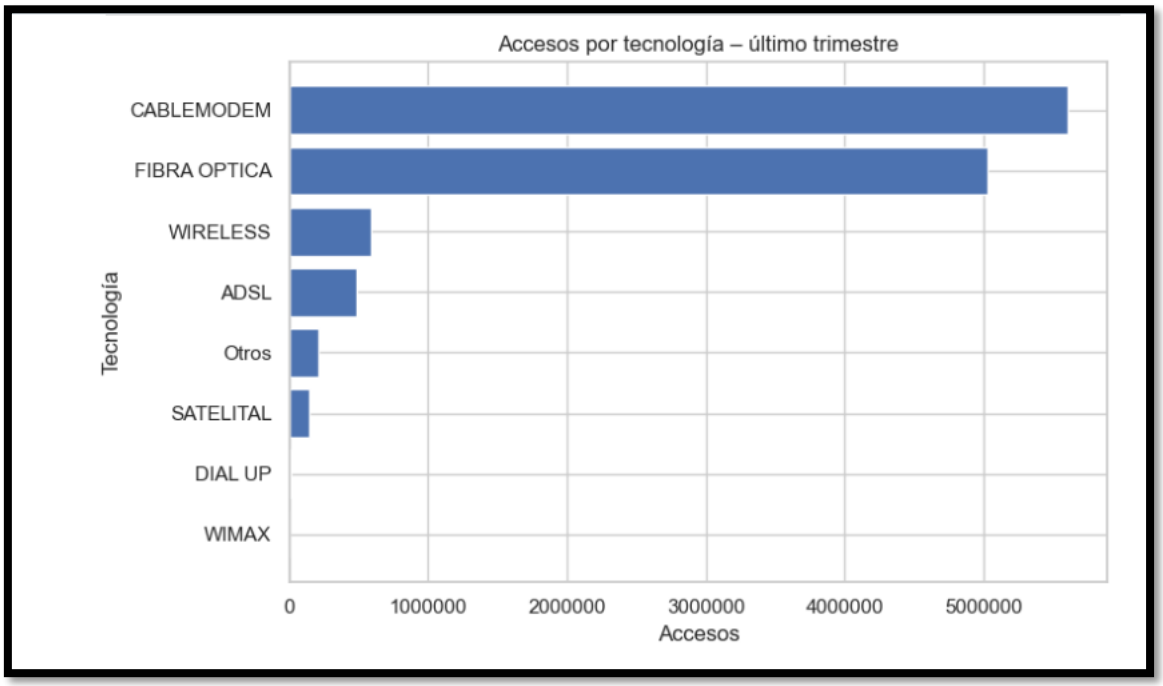
- Provincia (ID_Provincia)
- Localidad (Localidad, Partido)
- Tecnología (ID_Tecnologia)
- Rango de velocidad (ID_Rango_Velocidad)
- Período (ID_Periodo, Año, Trimestre)

Gráficos utilizados:

1. **Barras horizontales – Accesos por tecnología (último trimestre)**
   - Dataset: Accesos_Ultimo_Trim.csv
   - Resultado (suma de accesos por tecnología):
     - CABLEMODEM: 5.61 M
     - FIBRA OPTICA: 5.03 M
     - WIRELESS: 0.59 M

- ADSL: 0.48 M
- Otros / SATELITAL / DIAL UP / WIMAX: rezagadas.

- Conclusión:
  - El mercado se concentra en **cablemodem y fibra**, lo que muestra la transición hacia tecnologías de mayor capacidad.
  - ADSL y Dial Up casi han desaparecido.



2. **Barras – Top 10 provincias por accesos totales (último período)**
   - Dataset: Conexiones_Provincias.csv (período 2025-T1).
   - Provincias líderes (por Total(BAF+DU)):
     - Buenos Aires ≈ 5.29 M accesos
     - CABA ≈ 1.48 M
     - Córdoba ≈ 1.13 M
     - Santa Fe ≈ 0.92 M
     - Mendoza ≈ 0.34 M
   - Conclusión:
     - Fuerte concentración en AMBA y provincias grandes.
     - Estas provincias explican una gran parte del parque total de accesos

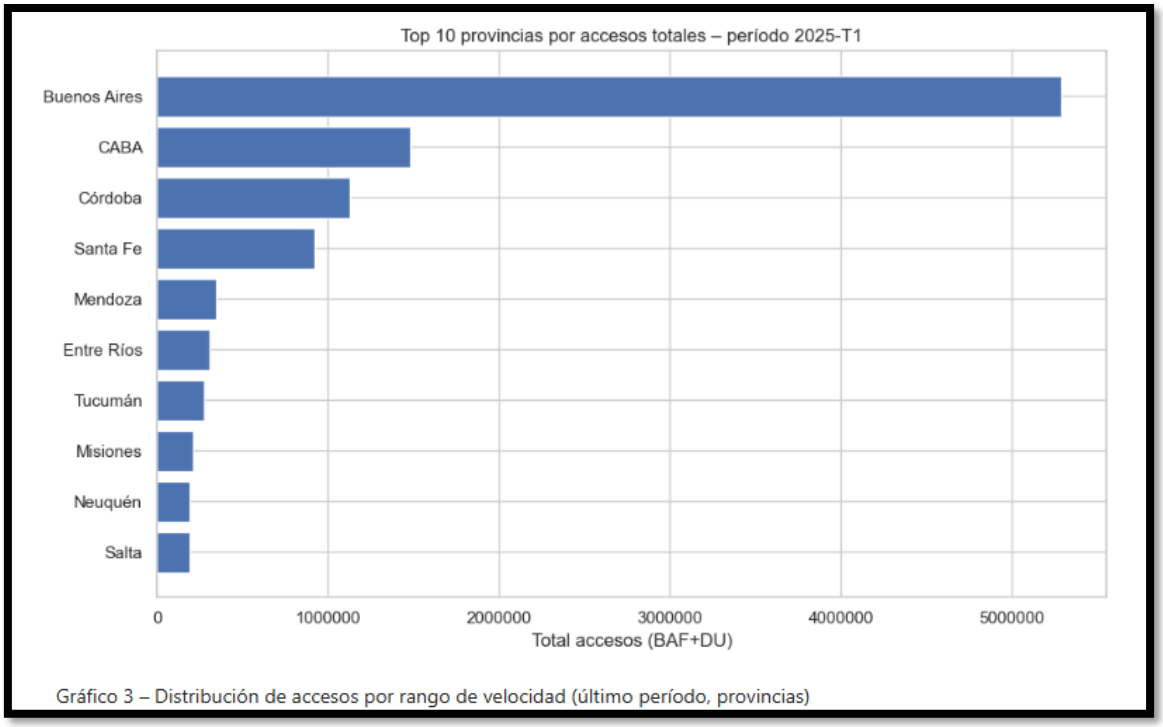
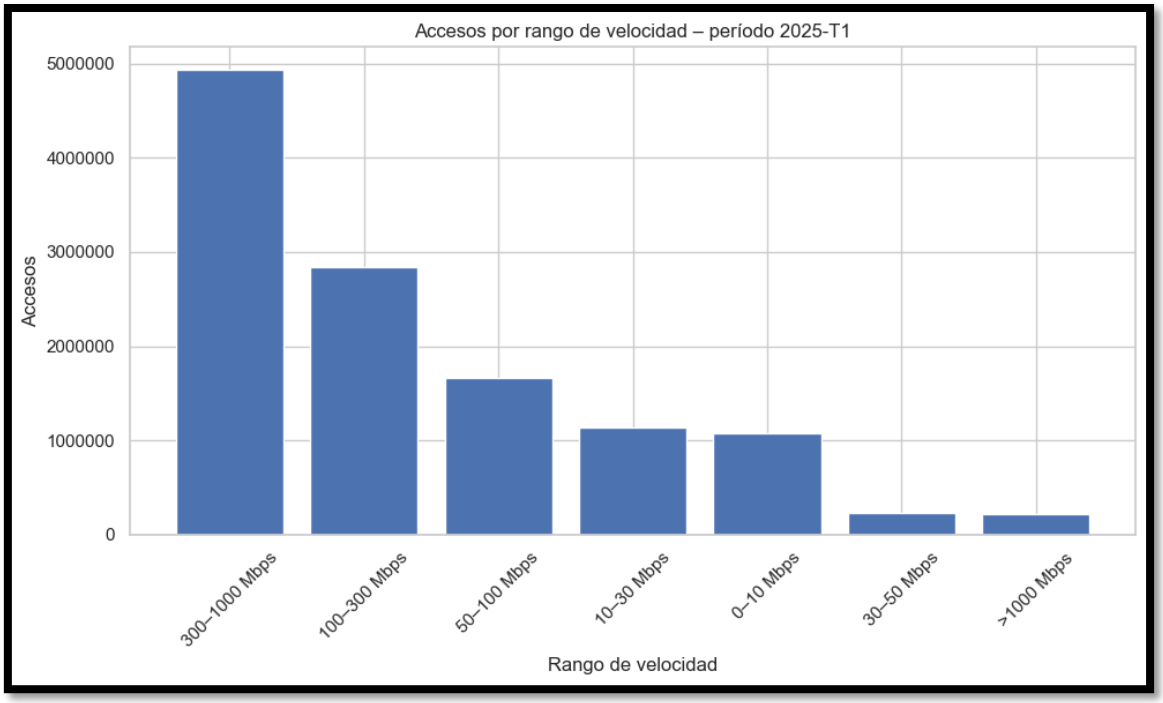


Gráfico 3 – Distribución de accesos por rango de velocidad (último período, provincias)

3. **Barras verticales – Accesos por rango de velocidad**
    - Dataset: Velocidades_Prov.csv (período 2025-T1).
    - Distribución de accesos por ID_Rango_Velocidad:
        - 300–1000 Mbps: 4.93 M (40,8 %)
        - 100–300 Mbps: 2.84 M (23,5 %)
        - 50–100 Mbps: 1.66 M (13,8 %)
        - 10–30 Mbps: 1.14 M
        - 0–10 Mbps: 1.07 M
        - 30–50 Mbps: 0.23 M
        - 1000 Mbps: 0.22 M
    - Conclusión:
        - **≈ 78 % de los accesos tienen velocidades superiores a 50 Mbps**, lo que indica una mejora significativa de la calidad de servicio a nivel país.
        - Aun así, persiste una base no menor en rangos bajos (< 30 Mbps), importante para el análisis de brecha digital.



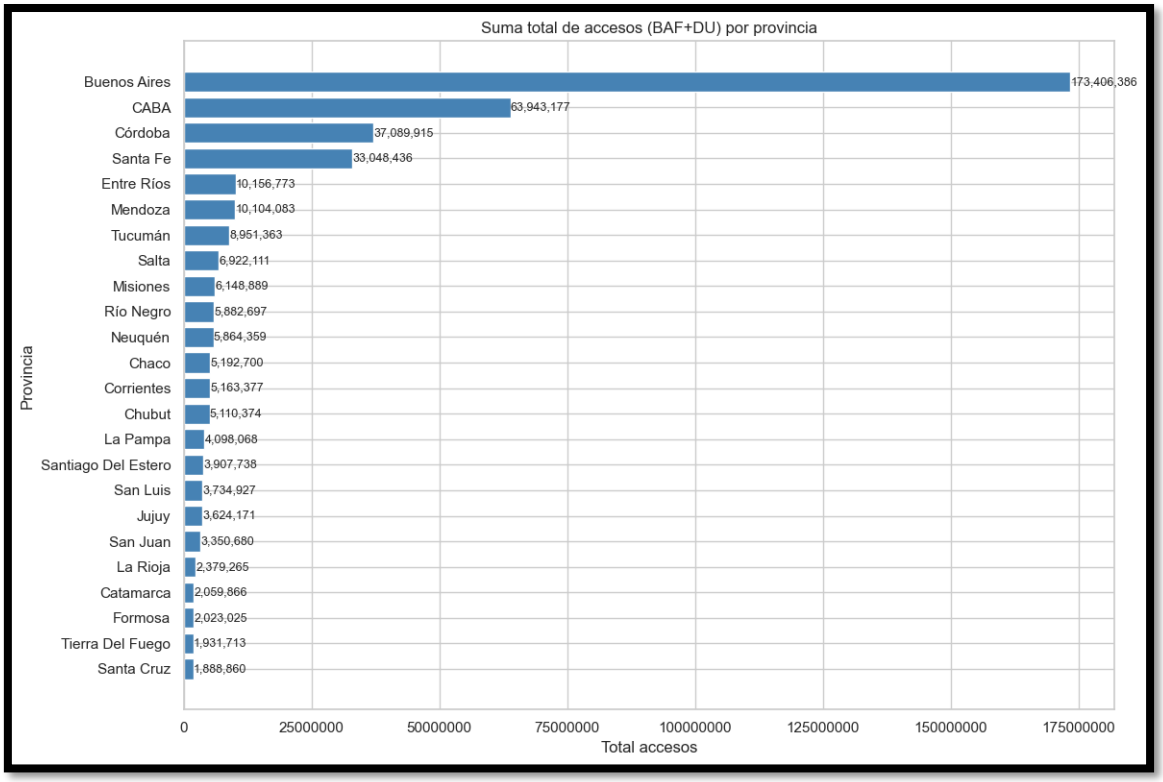
5.2. Variables cuantitativas

Ejemplos:

- Accesos, Total(BAF+DU), Total(Tecnologias), Total(Velocidades)
- Ingresos
- Mbps (Media de bajada)
- Accesos cada 100 Hogares y Accesos cada 100 Habitantes

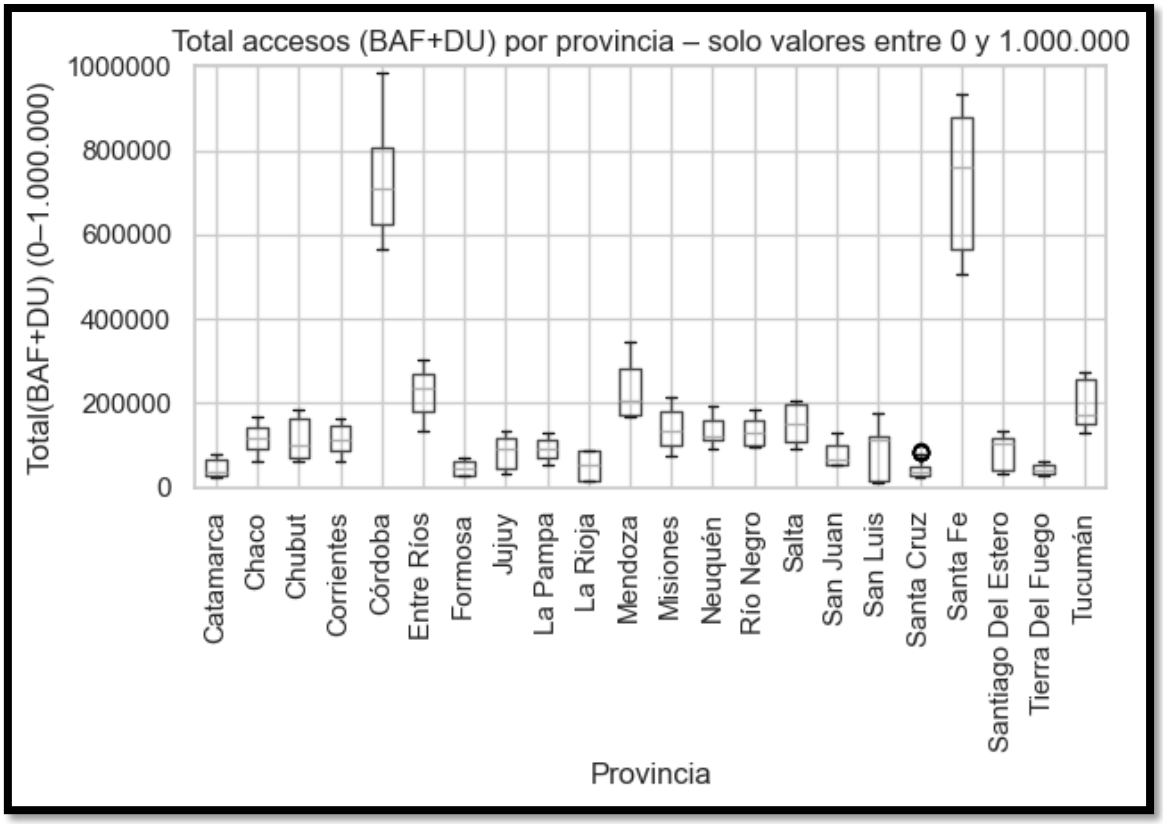
Gráficos utilizados:

1. **Barras horizontales**
    - Para la distribución de accesos por localidad y por provincia.
    - Permiten ver la distribución asimétrica (muchas localidades pequeñas, pocas muy grandes).



2. **Boxplots**

   - Para visualizar la presencia de outliers en Accesos y Total(BAF+DU) por provincia.
   - Complementan el análisis numérico del IQR × 3.



3. **Series de tiempo (line plots)**
    - Dataset: Conexiones_Periodos.csv.
    - Gráfico: Total(BAF+DU) vs ID_Periodo (orden cronológico).
    - Hallazgos:
        - Los accesos totales pasan de ≈ **6.4 M en 2014-T1** a ≈ **12.1 M en 2025-T1**.
        - Esto muestra un **crecimiento casi lineal** con aceleración en años recientes, coincidiendo con el despliegue de fibra óptica y altas velocidades.

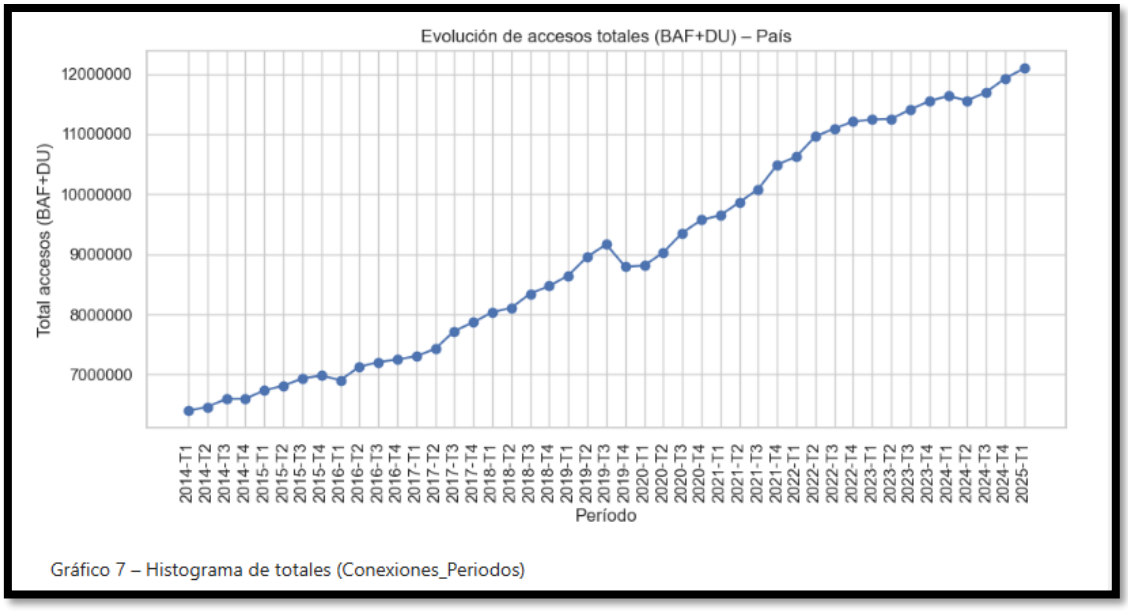


Gráfico 7 – Histograma de totales (Conexiones_Periodos)

4. **Heatmap de correlaciones**
    - Sobre variables numéricas de Conexiones_Periodos y Conexiones_Provincias.
    - Conclusiones típicas:
        - Fuerte correlación entre Banda Ancha Fija y Total(BAF+DU).
        - Correlación positiva entre Fibra Optica y Mbps (Media de bajada) a nivel provincial.
        - Correlación esperable entre Total(Velocidades) y Total(Tecnologias).

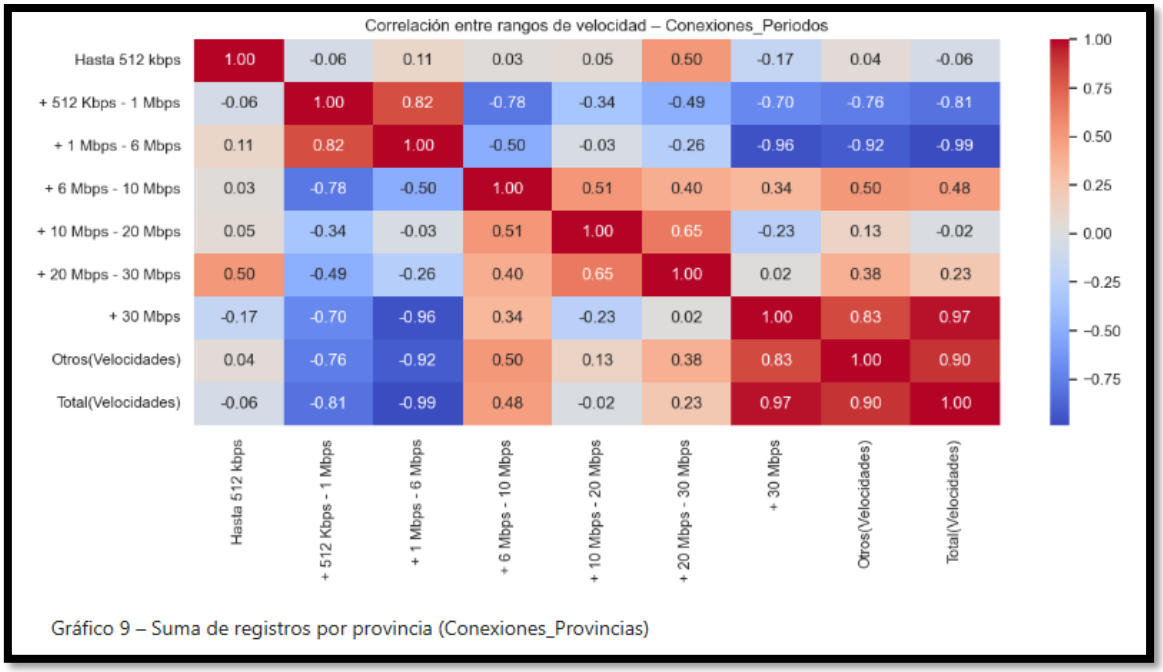


Gráfico 9 – Suma de registros por provincia (Conexiones_Provincias)

6. Ajuste y edición del dataset en Excel a partir de los archivos limpios

Una vez finalizada la etapa de limpieza y EDA en Python, el siguiente paso del Sprint 1 consistió en **ajustar y editar el dataset en Excel** para llegar al modelo relacional final utilizado en el proyecto. Esta instancia fue clave para transformar el *Dataset Obligatorio original* (Dataset Obligatorio.xlsx) en el *Dataset Obligatorio FINAL* (Dataset Obligatorio FINAL.xlsx), pensado ya como base de datos analítica.

6.1. Exportación desde Python e importación controlada en Excel

A partir de los scripts desarrollados en Python:

- Se generaron archivos **CSV limpios**, con:
    - números en formato estándar (punto decimal interno),
    - conversión de texto a tipo numérico,
    - separación consistente de columnas.
- Antes de importarlos a Excel se definió:
    - **separador de columnas** ;,
    - **separador decimal** , según configuración regional,
    - evitando que Excel interpretara decimales como enteros o aplicara formatos automáticos incorrectos.

Este paso aseguró que los valores de accesos, velocidades e ingresos se mantuvieran coherentes con lo observado en el EDA de Python.

6.2. Normalización de nombres de tablas y columnas

Sobre el archivo ya importado en Excel se realizó un trabajo específico de **ajuste semántico y de nomenclatura**:

- Se renombraron hojas y tablas con nombres **claros y descriptivos**, por ejemplo:
    - Conexiones_Periodos, Conexiones_Provincias, Velocidades_Prov, Velocidades_Loc, Accesos_Ultimo_Trim, etc.
- Se ajustaron nombres de columnas para:
    - eliminar caracteres especiales (paréntesis, signos “+”, etc.),
    - mantener un criterio uniforme (mismas etiquetas para el mismo concepto en todas las tablas),
    - hacer explícito el significado de cada campo (ej.: Total(BAF+DU), Total(Tecnologias), Total(Velocidades), Mbps (Media de bajada), Accesos cada 100 Hogares, etc.).

Este trabajo permitió que cualquier integrante del equipo pudiera interpretar el dataset sin necesidad de revisar continuamente la documentación original de ENACOM.

6.3. Construcción y unificación de tablas (modelo relacional)

A partir de las hojas originales se definió un **modelo relacional más limpio**, separando:

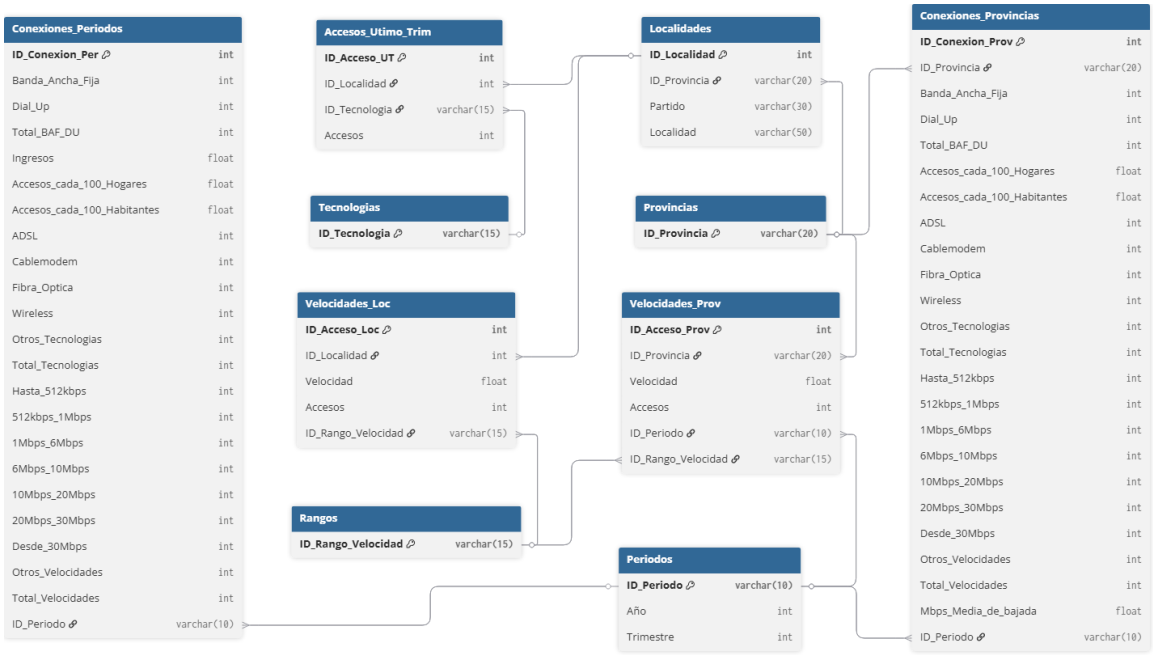
- **Tablas de dimensiones**:
    - Provincias: lista única de provincias y CABA.
    - Localidades: combinación única de ID_Localidad, provincia, partido y localidad.
    - Periodos: ID_Periodo, Año y Trimestre.
    - Tecnologias: catálogo de tipos de acceso (ADSL, Cablemodem, Fibra Óptica, Wireless, etc.).
    - Rangos: rangos de velocidad de descarga.
- **Tablas de hechos**:
    - Conexiones_Periodos: métricas agregadas a nivel país por período.

- Conexiones_Provincias: métricas a nivel provincia–período.
- Velocidades_Prov: accesos por provincia–periodo–rango de velocidad.
- Velocidades_Loc: accesos por localidad–periodo–velocidad.
- Accesos_Ultimo_Trim: detalle del último trimestre disponible.

Para lograrlo se aplicaron en Excel, apoyados por los resultados obtenidos en Python:

- **Unificación de registros repetidos** mediante:
  - eliminación de duplicados,
  - obtención de listas únicas (por ejemplo, lista única de provincias y de localidades).
- **Generación de nuevas tablas** a partir de desgloses del dataset original, separando un único Excel "plano" en varias tablas especializadas:
  - una tabla para conexiones por período,
  - otra para conexiones por provincia,
  - otras específicas para velocidades, rangos y tecnologías.

El resultado fue un **modelo en estrella** preparado para análisis en SQL, Power BI o Looker Studio.



6.4. Codificación y tratamiento de ceros en claves y categorías

Durante esta etapa también se hizo un trabajo fino de **ajuste de códigos**:

- Se revisaron campos donde aparecían valores 0 en columnas que debían actuar como **claves o códigos** (por ejemplo, identificadores de períodos, provincias, tecnologías o rangos).
- En los casos donde el 0 representaba un valor **inválido o incompleto**, se lo sustituyó por:
    - un **código único consistente** con el resto de la tabla.
- En columnas de **conteo**, como Accesos, los ceros que representaban "no hay conexiones" se mantuvieron como valor válido, diferenciándolos de los casos de "dato no informado".

De este modo se evitó que claves "0" contaminaran relaciones entre tablas o generaran problemas posteriores al construir el modelo en SQL.

6.5. Validación de outliers mediante filtros y revisiones manuales

Si bien el EDA en Python ya había detectado y documentado outliers utilizando el criterio IQR×3, en Excel se realizó una **segunda instancia de validación**:

- Se utilizaron **filtros por columnas** y ordenamientos descendentes para:
    - revisar manualmente los valores más altos de accesos, ingresos y velocidades,
    - confirmar que correspondían a provincias y períodos esperables (ej. Buenos Aires, CABA y años recientes).
- Se aplicaron filtros y resúmenes (por ejemplo, "Top 10" en Excel) para:
    - verificar que los outliers respondieran a realidades de mercado (provincias grandes, despliegue de fibra, aumento de velocidades),
    - detectar posibles errores de carga (por ejemplo, un número mal separado por miles o decimales).

Solo tras esta doble verificación (Python + filtros en Excel) se validó definitivamente que los outliers **debían mantenerse** en el dataset por ser representativos del fenómeno analizado.

6.6. Resultado de la instancia de ajuste y edición

Como resultado de este proceso:

- El **Dataset Obligatorio FINAL.xlsx** quedó estructurado como un conjunto de tablas:
    - limpias,
    - consistentes,
    - con nombres descriptivos,
    - listas para ser **importadas a SQL Server/Big Query** y utilizadas en el Sprint 2.
- Se redujo notablemente la complejidad del archivo original, desarmando estructuras anchas y repetitivas en un **modelo relacional claro**, donde:
    - cada tabla tiene una función bien definida,
    - las claves están listas para actuar como PK y FK,
    - las métricas numéricas ya fueron depuradas y validadas.

7. Uso de Markdown y documentación del proceso

En el notebook se siguió una estructura de Markdown como la siguiente:

- Secciones por tabla y por tema:
    - *Importación de datos*
    - *Chequeo de tipos*
    - *Nulos y duplicados*
    - *Outliers (IQR×3)*
    - *Gráficos por variable cualitativa*
    - *Gráficos por variable cuantitativa*
- Debajo de cada gráfico:
    - **Descripción del gráfico** (qué muestra, qué ejes).
    - **Conclusión interpretando el negocio** (no solo repetir lo que se ve).

Ejemplo de bloque Markdown debajo de un gráfico:

Título de gráfico – Distribución de accesos por rango de velocidad (2025-T1)
Se observa que el 78 % de los accesos se concentra en velocidades superiores a 50 Mbps, lo que indica una fuerte mejora en la infraestructura de banda ancha fija. Sin embargo, persiste un volumen significativo en rangos inferiores a 30 Mbps,

especialmente en provincias menos pobladas, lo que será clave para analizar la brecha digital en los próximos sprints.

---

8. Conclusiones generales del Sprint 1

1. Se construyó una **base relacional limpia, completa y coherente**, con 10 tablas bien definidas y claves listas para ser implementadas en SQL Server.
2. No se detectaron **nulos ni duplicados** en las tablas finales, lo que simplifica el modelado y mejora la confiabilidad del análisis.
3. Se identificaron **outliers** significativos en variables de accesos y conexiones (especialmente en provincias grandes y localidades con alta penetración), que se conservaron por ser parte importante de la historia de negocio.
4. Se distinguieron claramente **variables cualitativas y cuantitativas**, utilizando gráficos apropiados para cada tipo:
    - Barras y apilados para provincias, tecnologías, rangos.
    - Histogramas, boxplots y series de tiempo para métricas numéricas.
5. Se obtuvieron insights relevantes:
    - El total de accesos a Internet fija en Argentina crece de ~6,4 M (2014-T1) a ~12,1 M (2025-T1).
    - La mayoría de los accesos actuales se concentra en rangos de velocidad altos (≥50 Mbps).
    - El mercado está dominado por **cablemodem y fibra óptica**.
    - Buenos Aires, CABA, Córdoba y Santa Fe explican gran parte del parque de accesos.

Con esto, el Sprint 1 cumple las consignas del proyecto final:
preparación de datos, calidad, EDA sólido, comprensión de tipos de variables, outliers, duplicados y valores faltantes, con un análisis justificado y documentado en Markdown.