Análisis del Dataset Titanic

Proyecto de Programación II

Nombre: Daniel Felipe Sierra Cuellar
Materia: Ciencia de Datos - Programación II
Fecha de entrega: Noviembre 2025
Tiempo de desarrollo: 1 semana

Repositorio GitHub: https://github.com/Danfezinho/taller-pandas-titanic-individual

Tabla de Contenidos

1. Introducción

Este proyecto consiste en un análisis exploratorio del famoso Dataset del Titanic. El objetivo principal fue aplicar lo aprendido en el curso sobre Pandas, visualización de datos y control de versiones con Git/GitHub.

2. Objetivos

Objetivo General

Analizar el dataset del Titanic para identificar los factores que influyeron en la supervivencia de los pasajeros.

Objetivos Específicos

1. Cargar y limpiar los datos del Titanic
2. Hacer estadísticas descriptivas básicas
3. Identificar y analizar datos faltantes
4. Crear variables nuevas que ayuden al análisis
5. Hacer análisis con groupby para comparar grupos
6. Crear visualizaciones que muestren los patrones
7. Determinar qué tipo de personas tuvieron más probabilidad de sobrevivir
8. Practicar el uso de Git y GitHub (ramas, commits, pull requests)

3. Descripción de los Datos

Fuente de Datos

Los datos se descargaron por medio de la pagina web de Kaggle
https://www.kaggle.com/competitions/titanic

Archivos Utilizados

- **train.csv**: 891 pasajeros (con información de si sobrevivieron o no)
- **test.csv**: 418 pasajeros (sin información de supervivencia)

Variables del Dataset

El dataset tiene 12 variables por cada pasajero:

Variable	Qué significa	Ejemplo
PassengerId	Número único del pasajero	1, 2, 3...
Survived	Si sobrevivió (1) o murió (0)	0, 1
Pclass	Clase del ticket (1, 2 o 3)	1=Primera, 3=Tercera
Name	Nombre completo	"Braund, Mr. Owen Harris"
Sex	Sexo (male/female)	male, female
Age	Edad en años	22, 35, 8...
SibSp	Hermanos/cónyuges a bordo	0, 1, 2...
Parch	Padres/hijos a bordo	0, 1, 2...
Ticket	Número del ticket	A/5 21171
Fare	Tarifa pagada en libras (£)	7.25, 71.28...
Cabin	Número de cabina	C85, E46...
Embarked	Puerto donde subieron	S, C, Q

Puertos de embarque:

- S = Southampton (Inglaterra)
- C = Cherbourg (Francia)
- Q = Queenstown (Irlanda)

Estadísticas Básicas

- **Total de pasajeros analizados:** 1,309 (891 train + 418 test)
- **Edad promedio:** 29.88 años
- **Edad más joven:** 2 meses (0.17 años)
- **Edad más vieja:** 80 años
- **Tarifa promedio:** £32.20

4. Análisis Realizado

El proyecto se dividió en 13 preguntas principales que se fueron respondiendo paso a paso. Aquí se resume lo que se hizo en cada una de las preguntas:

Preguntas 1-3: Carga y Exploración Inicial

¿Que se hizo?:

- Cargar los datos utilizando Pandas
- Revisar las columnas que tenía cada dataset
- Calcular estadísticas básicas (promedios, mínimos, máximos)
- Comparar los conjuntos train y test

Resultado importante: Los conjuntos train y test son parecidos, lo que está bien porque significa que la división de los datos fue apropiada.

Pregunta 4: Variable "Familiares"

Lo que hice: Creé una nueva variable llamada "Familiares" que suma SibSp + Parch. Esto me ayudó a saber cuántos familiares tenía cada persona a bordo.

Hallazgos:

- 60% de los pasajeros viajaban **solos** (sin familiares)
- 40% viajaban con al menos 1 familiar
- El máximo de familiares que alguien tenía era 10

Mi opinión sobre esta variable: Me pareció super útil porque es más fácil trabajar con una sola variable en vez de dos. Además, tiene sentido pensar que viajar solo o con familia pudo haber afectado las probabilidades de sobrevivir.

Preguntas 5-6: Unión de Datasets

Lo que se hizo: Se unio los datos de train y test en un solo dataset grande. Antes de unir, agregué una columna llamada "source" para saber de dónde venía cada fila.

Problema encontrado: El dataset test no tenía la columna "Survived", entonces se agregó con valores vacíos (NaN) para poder unir todo.

Resultado: Dataset completo con 1,309 pasajeros que se uso para el resto del análisis.

Pregunta 7: Datos Faltantes

Lo que se hizo: Se reviso qué variables tenían datos faltantes y se calculo los porcentajes.

Hallazgos:

Variable	% Faltante	Qué significa
Cabin	77.5%	La mayoría no tenía cabina registrada (especialmente 3ra clase)
Age	20.1%	Algunos no reportaron su edad
Embarked	0.2%	Solo 2 personas sin puerto (casi nada)
Fare	0.1%	Solo 1 persona sin tarifa
Survived	31.9%	Normal, es el conjunto test

Conclusión: Los datos faltantes en Cabin son un problema grande, pero en Age es manejable. Se decidió trabajar con los datos disponibles sin hacer imputación.

Pregunta 8: Preguntas Descriptivas

Se respondió 7 preguntas básicas sobre los datos:

8a. Edad promedio: 29.88 años

8b. Supervivencia general:

- Sobrevivieron: 342 personas (38.4%)

- Murieron: 549 personas (61.6%)

8c. Tarifa de primera clase: £84.15 en promedio

- Para comparar: tercera clase pagó £13.68 (casi 6 veces menos)

8d. Pasajeros con familiares: 519 (39.6%) viajaban con familia

8e. Edades extremas:

- Más joven: bebé de 2 meses (Elizabeth Gladys Dean)
- Más viejo: señor de 80 años (Algernon Barkworth)

8f. Pasajeros por puerto:

- Southampton: 914 (69.9%)
- Cherbourg: 270 (20.7%)
- Queenstown: 123 (9.4%)

8g. Solos vs familia: 60% solos, 40% con familia

Pregunta 9: Análisis con GroupBy

Aquí se usó groupby de Pandas para comparar tasas de supervivencia entre diferentes grupos.

9a. Supervivencia por sexo:

- **Mujeres:** 74.2% sobrevivieron
- **Hombres:** 18.9% sobrevivieron
- Las mujeres tuvieron **3.9 veces más probabilidad** de sobrevivir

9b. Niños vs hombres adultos:

- **Niños (<18 años):** 54% sobrevivieron
- **Hombres adultos:** 17.7% sobrevivieron
- Los niños tuvieron **3 veces más probabilidad**

9c. Grupos por edad:

- Menores de 10 años: 61.3%
- Entre 10 y 50 años: 39.1%
- Mayores de 50 años: 34.4%

9d. Supervivencia por puerto:

- Cherbourg: 55.4% (el más alto)
- Queenstown: 39.0%
- Southampton: 33.7%

Nota importante: El puerto no fue un factor causal. Cherbourg tuvo mejor tasa porque de ahí subieron más pasajeros de primera clase.

9e. Supervivencia por clase:

- **Primera clase:** 63.0%
- **Segunda clase:** 47.3%
- **Tercera clase:** 24.2%
- La primera clase tuvo **2.6 veces más probabilidad** que la tercera

Pregunta 10: Grupos Familiares y Cabinas

Lo que se hizo: Se clasifico a los pasajeros en 3 grupos familiares y se reviso cuántos tenían cabina.

Grupos familiares:

- Sin familiares: 790 personas (60%)
- Familia pequeña (1-3): 437 personas (33%)
- Familia grande (4+): 82 personas (6%)

Distribución de cabinas:

- Familias pequeñas: 35% tenía cabina registrada
- Sin familiares: 16.6% tenía cabina
- Familias grandes: 13.4% tenía cabina

Conclusión: La cabina parece más relacionada con la clase social que con el tamaño de la familia. Las familias pequeñas probablemente tenían más dinero en promedio.

Pregunta 11: Análisis Multidimensional

Se cómo interactuaban varios factores al mismo tiempo: sexo, edad, cabina y supervivencia.

Principales hallazgos:

1. El sexo fue lo MÁS importante:

- Mujeres con cabina: 93.3% supervivencia
- Mujeres sin cabina: 66.1% supervivencia
- Hombres con cabina: 43.2% supervivencia
- Hombres sin cabina: 14.5% supervivencia

Dato impactante: Mujeres sin cabina sobrevivieron más que hombres con cabina. El sexo superó incluso a la clase social.

2. La combinación de factores importa:

- Mejor escenario: Mujer + niña + cabina = ~90-95% supervivencia
- Peor escenario: Hombre + adulto + sin cabina = ~10-15% supervivencia

3. Por grupo de edad:

- Niños: 61.3% (fueron priorizados)
- Adultos: 38.3%
- Mayores: 36.5%

Pregunta 12: Visualizaciones Avanzadas

Se hizo varios gráficos más avanzados para entender mejor los datos:

Visualizaciones que se hizo:

1. Pirámide poblacional por supervivencia
2. Matriz de correlación entre variables
3. Violin plots de tarifas
4. Análisis por puerto
5. Análisis por títulos (Mr., Mrs., Miss., Master.)
6. Densidad de edad (KDE)
7. Heatmaps multidimensionales
8. Dashboard resumen

Hallazgos de las visualizaciones:

Títulos en los nombres: Se Extrajo los títulos de los nombres (Mr., Mrs., Miss., Master.) y se descubrió que son muy buenos predictores:

- Mrs. (señoras casadas): ~80% supervivencia
- Miss. (señoritas): ~70% supervivencia
- Master. (niños varones): ~58% supervivencia

- Mr. (señores): ~16% supervivencia

Correlaciones:

- Survived vs Pclass: -0.34 (negativa = clase baja murió más)
- Survived vs Fare: 0.26 (positiva = tarifa alta sobrevivió más)
- Pclass vs Fare: -0.55 (primera clase pagó más)

Pregunta 13: ¿Quiénes sobrevivieron más?

Respuesta corta: Las **mujeres y niños de primera clase** tuvieron las mejores probabilidades, mientras que los **hombres adultos de tercera clase** tuvieron las peores.

Jerarquía de factores (de más a menos importante):

1. **SEXO** (diferencia de 55 puntos entre mujeres y hombres)
2. **CLASE SOCIAL** (diferencia de 39 puntos entre primera y tercera)
3. **EDAD** (diferencia de 22 puntos entre niños y adultos)
4. **TAMAÑO FAMILIAR** (familia pequeña fue óptimo)

Perfiles extremos:

Mayor supervivencia:

- Descripción: Mujer o niña de primera clase con 1-3 familiares
- Probabilidad: 90-95%

Menor supervivencia:

- Descripción: Hombre adulto de tercera clase viajando solo
- Probabilidad: 10-15%

5. Principales Hallazgos

Hallazgo #1: "Mujeres y Niños Primero" fue real

Los datos confirman que se siguió el protocolo de evacuación:

- Mujeres: 74.2% vs Hombres: 18.9%
- Niños tuvieron prioridad sobre adultos
- Los hombres adultos fueron el grupo más afectado

Hallazgo #2: La clase social importó mucho

Hubo una desigualdad clara:

- Primera clase: 63% supervivencia
- Tercera clase: 24% supervivencia

¿Por qué?

- Las cabinas de primera clase estaban arriba, cerca de los botes
- El personal probablemente atendió primero esas áreas
- Había barreras físicas que dificultaban que la tercera clase llegara a cubierta

Hallazgo #3: Viajar con familia pequeña fue mejor

- Solos: 30.4% supervivencia
- Familia pequeña (1-3): 50.5% supervivencia
- Familia grande (4+): 16.1% supervivencia

Interpretación: Con familia pequeña había apoyo mutuo pero podían moverse rápido. Familias grandes probablemente tuvieron dificultad para mantenerse juntas en el caos.

Hallazgo #4: El sexo superó a la clase

Descubrimiento sorprendente: una mujer pobre tenía más probabilidades de sobrevivir que un hombre rico.

- Mujeres de 3ra clase: ~46% supervivencia
- Hombres de 1ra clase: ~37% supervivencia

Esto muestra que el protocolo "mujeres y niños primero" se siguió muy estrictamente.

Hallazgo #5: Los títulos son buenos predictores

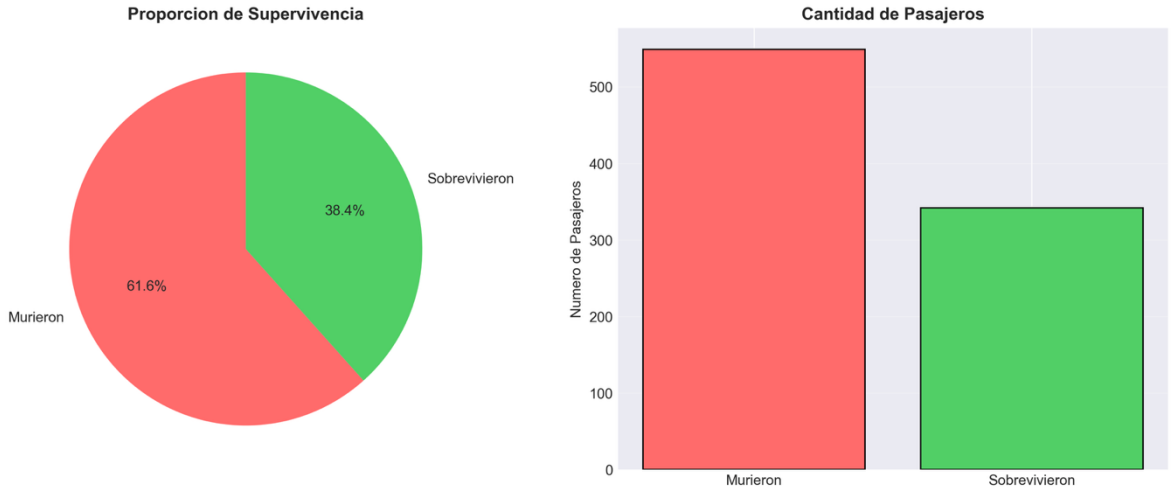
Los títulos (Mr., Mrs., Miss., Master.) encapsulan varios factores:

- Sexo (Mr. = hombre, Mrs./Miss. = mujer)
- Edad aproximada (Master. = niño)
- Estado civil y estatus social

Por eso funcionan tan bien para predecir supervivencia.

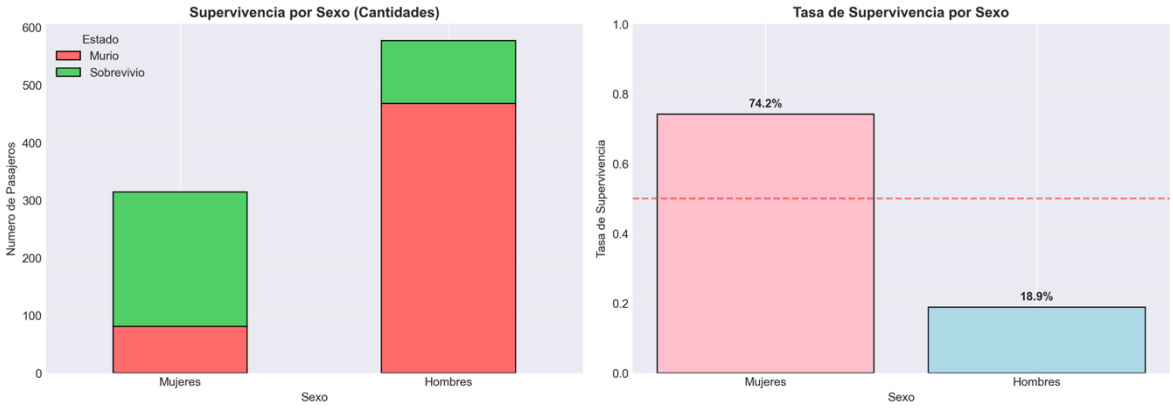
6. Visualizaciones

Gráfico 1: Supervivencia General



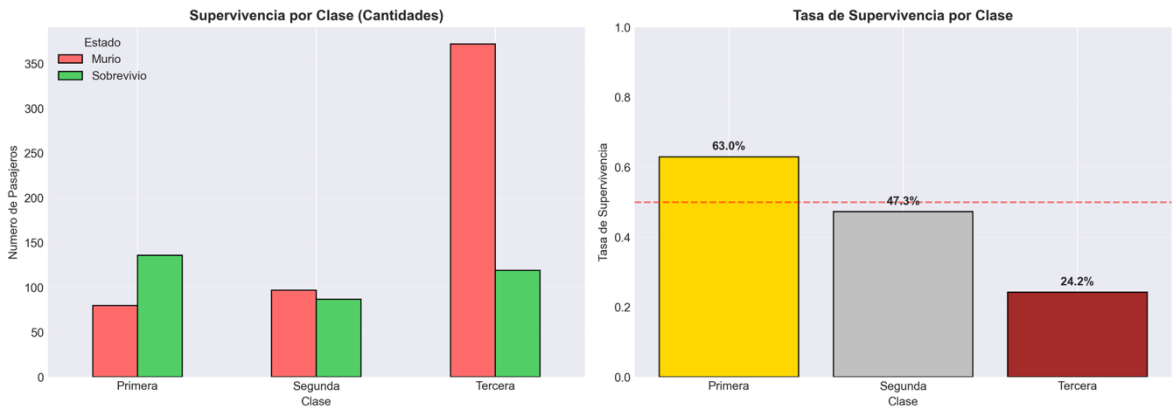
Muestra que solo 38% sobrevivió. La tragedia fue enorme.

Gráfico 2: Supervivencia por Sexo



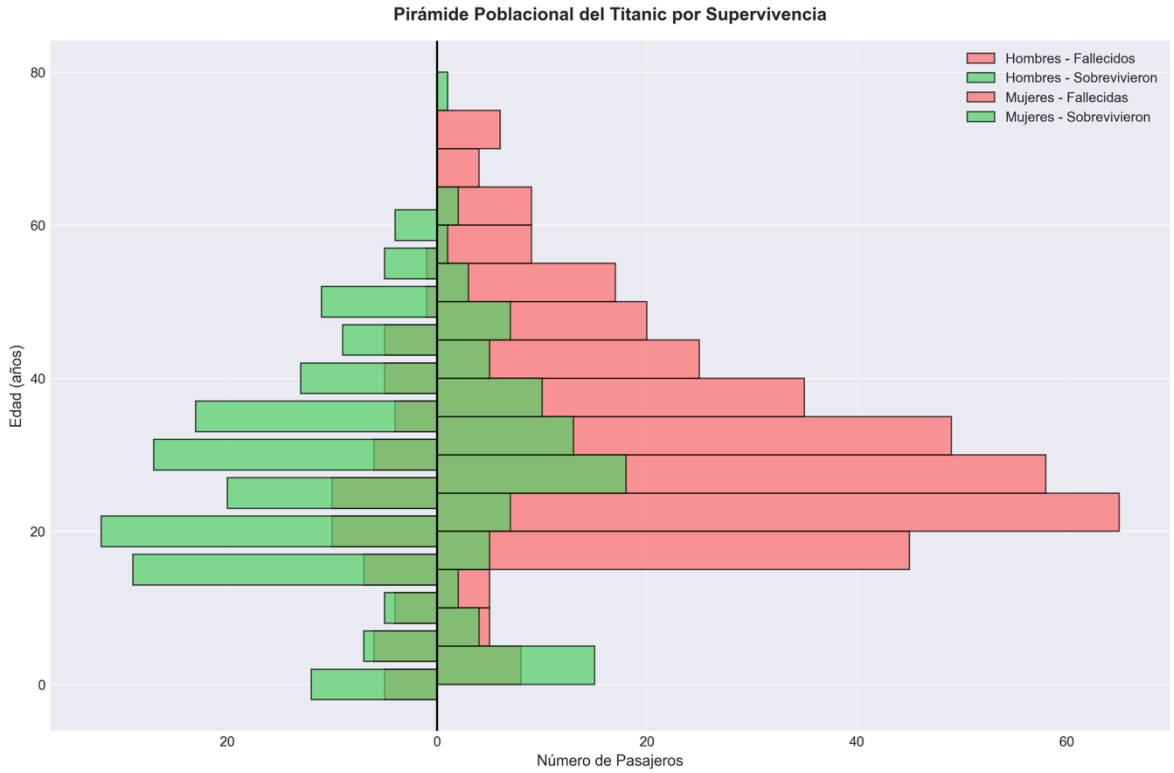
Se ve claramente la diferencia entre mujeres (74%) y hombres (19%). El protocolo de evacuación priorizó a las mujeres.

Gráfico 3: Supervivencia por Clase



Muestra la desigualdad: primera clase tuvo el doble de probabilidad que tercera clase.

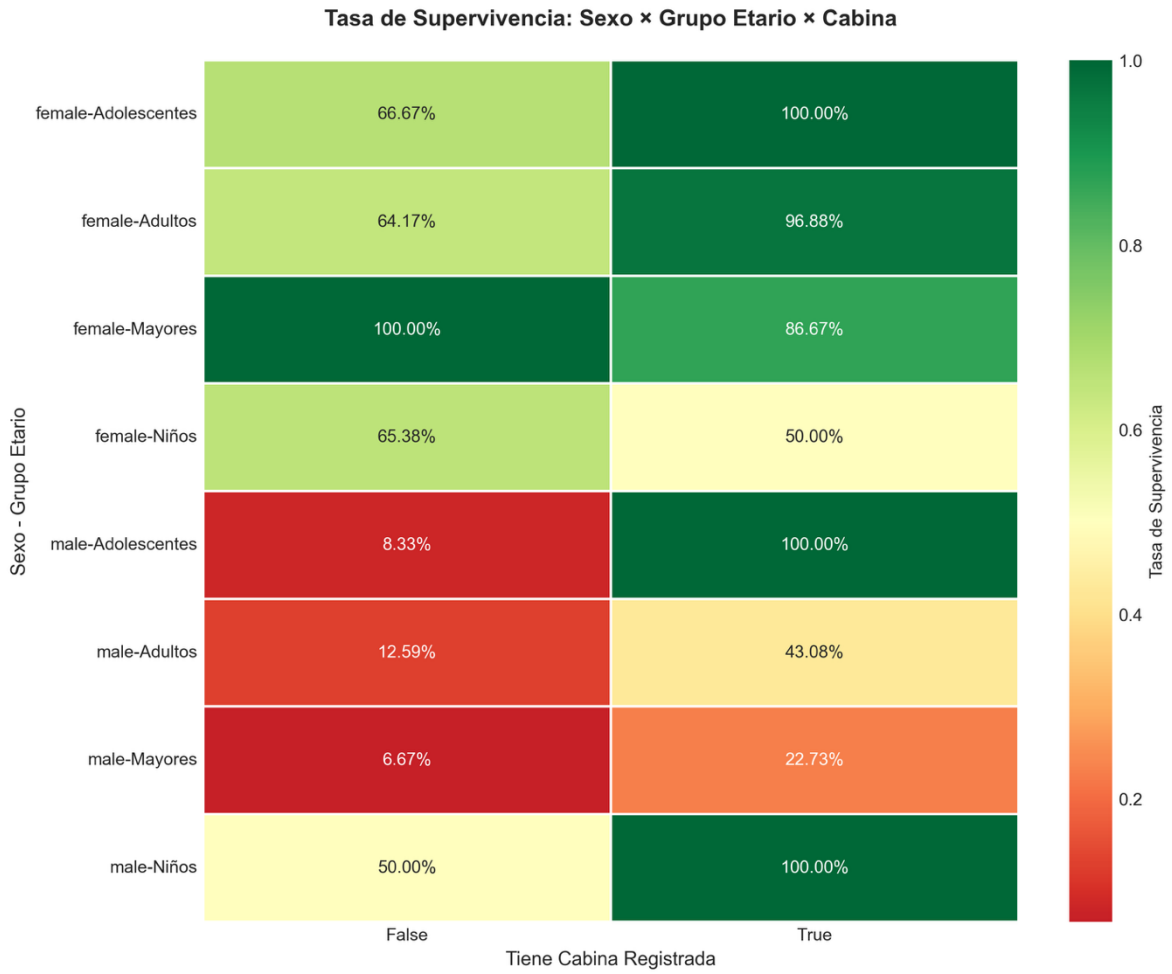
Gráfico 4: Pirámide Poblacional



Visualización impactante que muestra por edad y sexo quién sobrevivió (verde) y quién murió (rojo). Se ve claramente que:

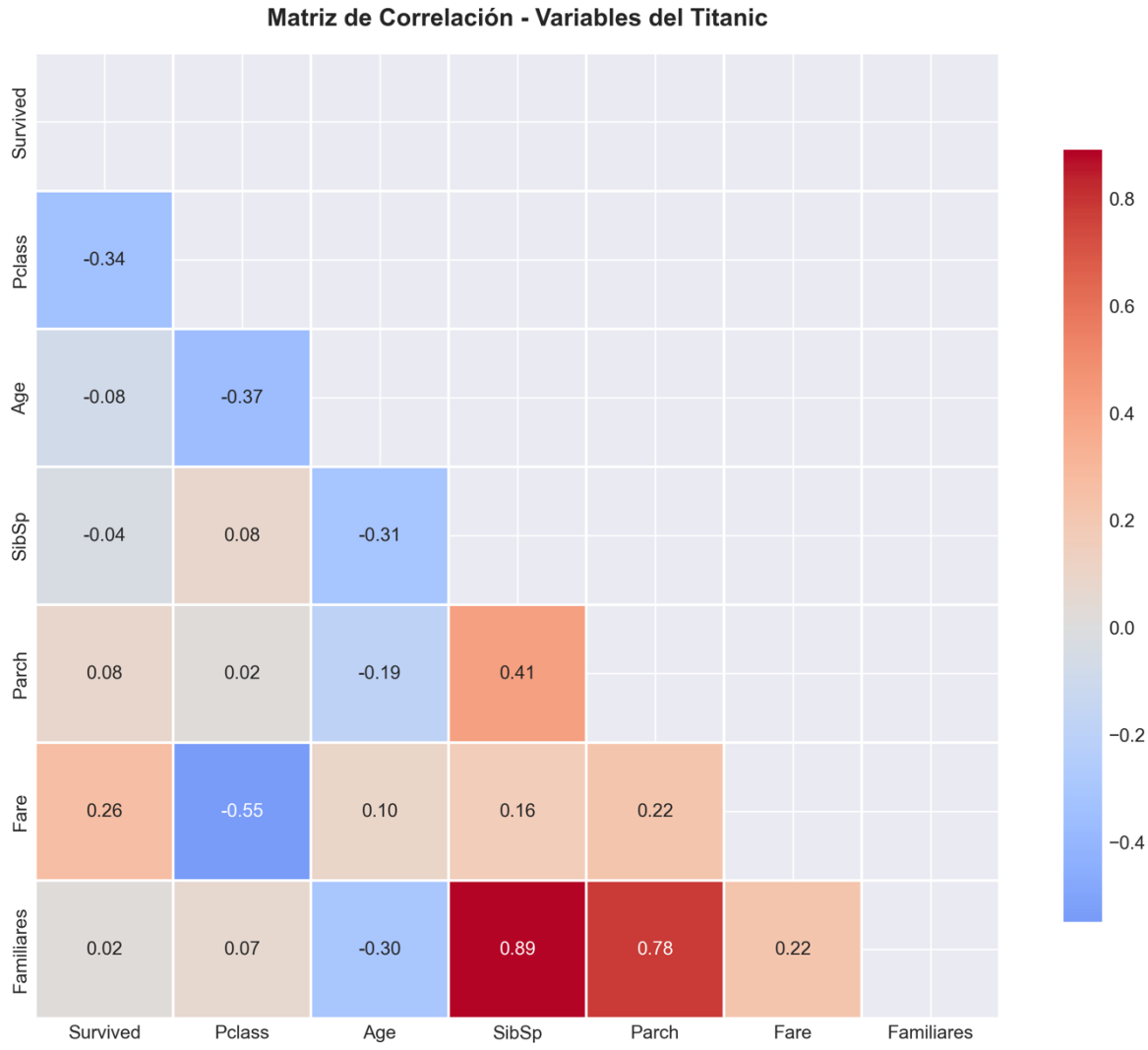
- Las mujeres (lado izquierdo) tienen más verde
- Los hombres (lado derecho) tienen más rojo
- Los niños pequeños tienen bastante verde en ambos lados

Gráfico 5: Heatmap Multidimensional



Este gráfico muestra las tasas de supervivencia considerando sexo, edad y si tenían cabina al mismo tiempo. Los colores van de rojo (baja supervivencia) a verde (alta supervivencia).

Gráfico 6: Matriz de Correlación



Muestra cómo se relacionan las variables numéricas. Lo más importante:

- Pclass y Survived tienen correlación negativa (-0.34)
- Fare y Survived tienen correlación positiva (0.26)

7. Trabajo en GitHub

Una parte importante del proyecto fue practicar el uso de Git y GitHub.

Estructura del Repositorio

Organicé el proyecto así:

```
taller-pandas-titanic-individual/
├── README.md
```

```
├── .gitignore
├── requirements.txt
├── data/
│   └── README.md
├── notebooks/
│   └── analisis_titanic.ipynb
├── outputs/
│   ├── graficos/  (14 gráficos PNG)
│   └── tablas/    (5 tablas CSV)
└── docs/
    ├── informe_final.pdf
    └── capturas_conflicto/
```

Commits Realizados

Se hicieron 23 **commits** en total durante el proyecto:

- Configuración inicial
- Carga de datos
- Análisis descriptivo
- Visualizaciones
- Documentación final

Ejemplo de mensajes de commit:

- "Configuración inicial del proyecto"
- "Agregar carga de datos y exploración inicial"
- "Completar preguntas 8-9: análisis descriptivos y groupby"
- "Resolver conflicto en README.md"

Resolución de Conflicto

Como parte del proyecto, simulé y resolví un conflicto de Git:

¿Qué pasó? Dos ramas modificaron la misma línea del README.md (la sección de "Estado del Proyecto").

¿Cómo lo resolví?

1. Git me mostró el conflicto con las marcas `<<<<<<<`, `=======`, `>>>>>>>`
2. Abrí el archivo en VS Code
3. Decidí qué versión quería mantener
4. Eliminé las marcas de conflicto
5. Hice commit del merge
6. Documenté todo con capturas de pantalla

Capturas incluidas en docs/capturas_conflicto/:

- Conflicto en GitHub.com
- Conflicto en GitHub Desktop
- Marcas de conflicto en VS Code
- Archivo después de resolver
- Commit exitoso

8. Conclusiones

1. El protocolo "mujeres y niños primero" fue real

Los datos lo confirman claramente. Las mujeres tuvieron casi 4 veces más probabilidad de sobrevivir que los hombres. Los niños también fueron priorizados.

2. La clase social fue el segundo factor más importante

Ser de primera clase vs tercera clase marcó una diferencia de 39 puntos porcentuales. Esto refleja la desigualdad de la época (1912) donde tu posición social literalmente determinaba si vivías o morías.

3. La combinación de factores fue clave

No fue solo un factor, sino la combinación:

- Una mujer rica tenía ~90% de probabilidad
- Un hombre pobre tenía ~10% de probabilidad

4. Los datos faltantes fueron un desafío

Especialmente Cabin (77% faltante) limitó algunos análisis. Aprendí que los datos del mundo real casi siempre tienen problemas y hay que trabajar con lo que hay.

5. Las visualizaciones ayudan mucho

Al principio solo veía números en tablas, pero cuando hice los gráficos todo se entendió mucho mejor. Por ejemplo, la pirámide poblacional muestra de forma super clara el patrón de supervivencia.

9. Reflexión Personal

¿Qué aprendí?

Técnicamente:

- Usar Pandas para análisis de datos (groupby, agg, pivot_table)
- Crear visualizaciones con Matplotlib y Seaborn
- Manejar datos faltantes
- Trabajar con Git y GitHub (ramas, commits, PRs, conflictos)
- Organizar un proyecto de datos

Sobre el Titanic:

- La historia detrás de los datos es importante. No son solo números, fueron personas reales
- El desastre mostró lo mejor (sacrificio) y lo peor (desigualdad) de la humanidad
- El protocolo de evacuación fue seguido, pero no de forma igual para todas las clases

10. Referencias

Fuente de Datos

- Kaggle Titanic Competition
  https://www.kaggle.com/competitions/titanic

Documentación Técnica

- Pandas Documentation
  https://pandas.pydata.org/docs/
- Matplotlib Documentation
  https://matplotlib.org/
- Seaborn Documentation
  https://seaborn.pydata.org/
- Git and GitHub Guides
  https://docs.github.com/

Recursos Utilizados

- Pandas Cheat Sheet
  https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf
- Data-to-Viz (guía de visualizaciones)
  https://www.data-to-viz.com/
- Stack Overflow (para resolver dudas específicas)

Anexos

A. Resumen de Archivos Generados

Tablas CSV (5):

- datos_completos.csv (1,309 registros)
- resumen_datos_faltantes.csv
- resultados_pregunta_8.csv
- resultados_pregunta_9.csv
- analisis_multidimensional.csv

Gráficos PNG (14):

1. distribucion_familiares.png
2. comparacion_train_test.png
3. datos_faltantes.png
4. supervivencia_general.png
5. tarifas_por_clase.png
6. pasajeros_por_puerto.png
7. supervivencia_por_sexo.png
8. supervivencia_edad_extrema.png
9. supervivencia_por_clase.png
10. cabinas_por_grupo_familiar.png
11. heatmap_supervivencia_multidimensional.png
12. analisis_multidimensional_detallado.png
13. piramide_poblacional_supervivencia.png
14. matriz_correlacion.png

B. Estadísticas del Repositorio

- **Commits totales:** 24
- **Pull Requests:** 9
- **Ramas creadas:** 8
- **Conflictos resueltos:** 1
- **Líneas de código en notebook:** ~1,500
- **Tiempo de desarrollo:** 1 semana

C. Tabla Resumen de Supervivencia

Grupo	Total	Sobrevivieron	Tasa
Por Sexo			
Mujeres	314	233	74.2%
Hombres	577	109	18.9%
Por Clase			
Primera	216	136	63.0%
Segunda	184	87	47.3%
Tercera	491	119	24.2%
Por Edad			
Niños (<18)	113	67	59.3%
Adultos (18-49)	644	239	37.1%
Mayores (≥50)	64	26	40.6%

Fecha de finalización: Noviembre 2025
Proyecto desarrollado en: 1 semanas
Total de páginas: 20

FIN DEL INFORME