Analisis Exploratorio de Datos (EDA)

Dataset: Titanic - Machine Learning from Disaster

Estudiante: Garcia Manuela
Codigo: 1023774044
Curso: ET0203 - Seminario de la Ciencia de los Datos
Docente: MSc. Luis Esteban Gomez Cadavid
Programa: Ingenieria de Software - Pascual Bravo
Periodo: 2026-1
Fecha: 2026-02-26
Fuente: https://www.kaggle.com/competitions/titanic

Herramientas: pandas, numpy, matplotlib, seaborn, scipy, scikit-learn, statsmodels, missingno

Semilla: random_state=42 | Python 3.10

1. Resumen ejecutivo

Este reporte presenta el Analisis Exploratorio de Datos (EDA) del dataset Titanic de Kaggle, que contiene informacion de 891 pasajeros del RMS Titanic. El objetivo es demostrar dominio de preparacion de datos, estadistica descriptiva e inferencial, y visualizaciones.

- Dataset: Titanic - Machine Learning from Disaster (Kaggle). 891 pasajeros, 12 variables.
- Unidad de observacion: cada fila es un pasajero con sus caracteristicas demograficas y de viaje.
- Variables criticas: Sex (genero), Pclass (clase socioeconomica), Age (edad) y Fare (tarifa).
- Calidad: Cabin tiene 77.1% de faltantes. Age tiene 19.87% faltante. 15 registros con Fare=0. Sin duplicados.
- Hallazgo principal: solo 38.4% sobrevivio. Las mujeres sobrevivieron 74.2% vs 18.9% de hombres. La 1ra clase tuvo 63% de supervivencia vs 24.2% en 3ra clase.
- Relaciones: Pclass-Fare correlacion fuerte (-0.55). Sex-Survived asociacion fuerte (Cramer's V=0.54). Mujeres de 1ra clase: 96.8% de supervivencia.
- Outliers: Fare tiene 116 outliers por IQR, pero son casos validos (suites de lujo). No se eliminan.
- Tratamiento: Winsorizacion y log1p reducen la asimetria de Fare de 4.79 a valores cercanos a 0.
- Inferencia: todas las pruebas significativas (alpha=0.05): t-test (Fare por supervivencia), ANOVA (Age por clase), chi-cuadrado (Sex vs Survived).
- Limitaciones: muestra parcial (891 de 2224), Cabin inutilizable, Age faltante en 20%.

2. Auditoria de calidad

2.1 Valores faltantes

Se identificaron tres variables con valores faltantes: Cabin (687 faltantes, 77.1%), Age (177 faltantes, 19.87%) y Embarked (2 faltantes, 0.22%). Las demas variables estan completas.

Decision: Cabin tiene demasiados faltantes para usarse directamente; se puede crear un indicador binario (tiene_cabina). Embarked se imputa con la moda ('S' - Southampton). Age se analiza con los 714 valores disponibles.

2.2 Duplicados

No se encontraron filas duplicadas (0 de 891).

2.3 Rangos y tipos

Los tipos de datos son correctos. Se verificaron rangos: Age (0.42-80, razonable), Fare (0-512.33, con 15 valores en 0 que podrian ser tripulacion o errores). Survived y Pclass son numericos pero representan categorias (binaria y ordinal respectivamente).

2.4 Cardinalidad

Sex: 2 valores (male=577, female=314). Embarked: 3 valores (S=644, C=168, Q=77). Cabin: 147 valores unicos (alta cardinalidad). Ticket: 681 valores unicos. No se detectaron typos en las categoricas.

3. Estadistica descriptiva y visualizaciones

3.1 Variables numericas

Age: media 29.7, mediana 28, distribucion ligeramente sesgada a la derecha (skew=0.39). Fare: media 32.20, mediana 14.45, fuertemente asimetrica (skew=4.79, kurtosis=33.40), con coeficiente de variacion del 154%. SibSp y Parch: mayoria viajaba solo (mediana=0), distribucion muy sesgada.

3.2 Variables categoricas

Survived: 549 muertos (61.6%) vs 342 sobrevivientes (38.4%). Sex: 577 hombres (64.8%) vs 314 mujeres (35.2%). Pclass: 3ra clase=491 (55.1%), 1ra=216 (24.2%), 2da=184 (20.7%). Embarked: Southampton=644 (72.3%), Cherbourg=168 (18.9%), Queenstown=77 (8.6%).

3.3 Analisis por grupos

Se realizaron agrupaciones por Sex, Pclass, Embarked, Survived, Sex x Pclass (cruzado), y grupo de edad (cuartiles con pd.qcut). Hallazgos principales:

- Mujeres: tasa de supervivencia 74.2% vs 18.9% hombres.
- 1ra clase: 63.0%, 2da: 47.3%, 3ra: 24.2%.
- Cherbourg tuvo mayor tasa (55.4%), por mayor proporcion de 1ra clase.
- Mujeres de 1ra clase: 96.8% de supervivencia. Hombres de 3ra: 13.5%.
- Jovenes (0-20) tuvieron mayor supervivencia (44.4%), por inclusion de ninos.

3.4 Visualizaciones principales

Se generaron 23 graficos en total, incluyendo: histogramas+KDE (3), ECDF (2), boxplots (3), violin plots (3), scatter plots (3), pairplot, heatmap de correlacion, jointplot hexbin, FacetGrid, countplots de supervivencia, heatmap de supervivencia, QQ-plots, y graficos de tratamiento de outliers.

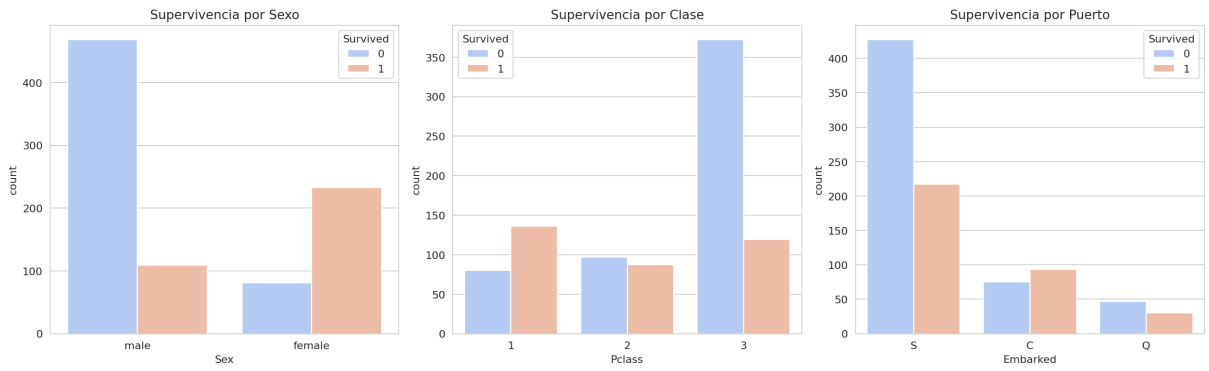

Figura 1: Supervivencia por Sexo, Clase y Puerto de embarque

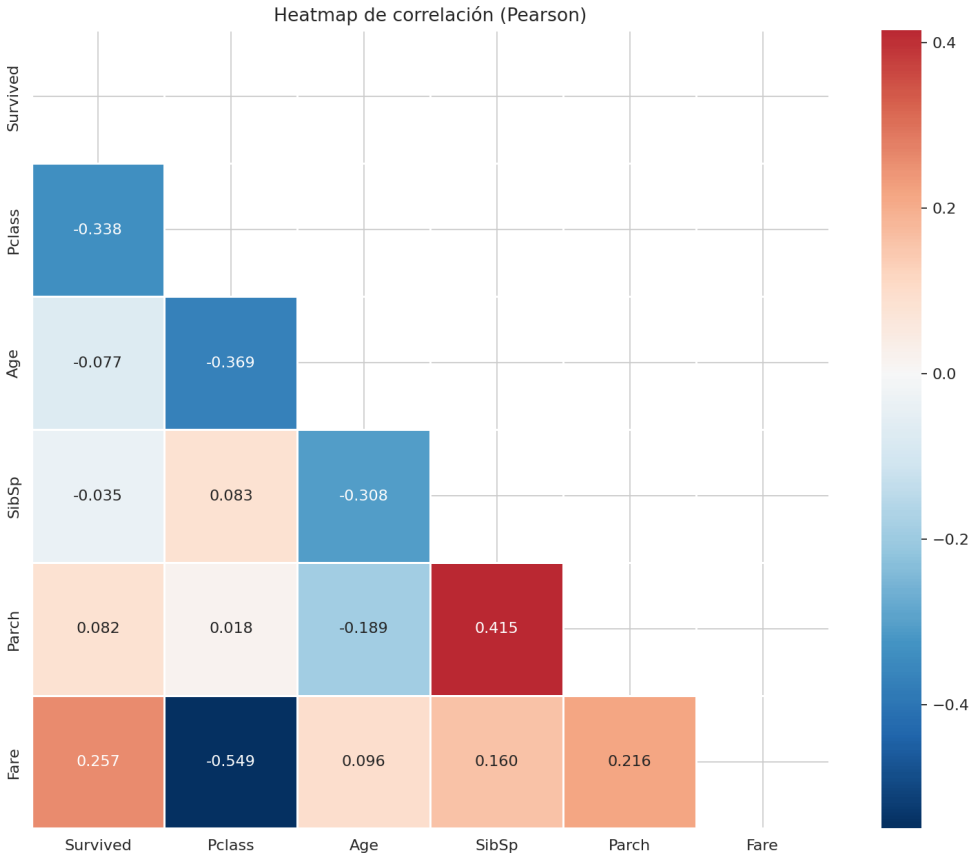

Figura 2: Heatmap de correlacion de Pearson entre variables numericas

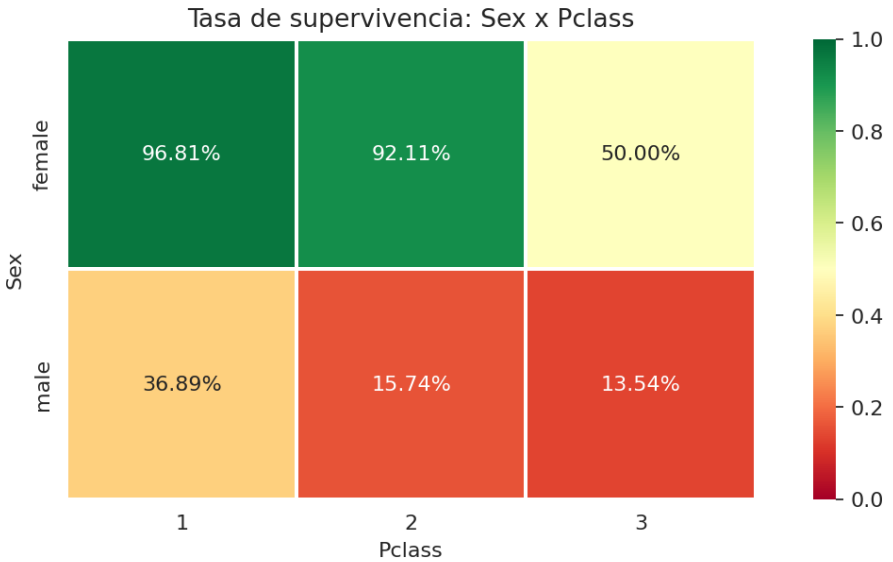

Figura 3: Tasa de supervivencia cruzada Sex x Pclass

4. Deteccion y tratamiento de outliers

4.1 Deteccion univariada

Se aplicaron 4 metodos univariados a las variables Age, Fare, SibSp y Parch:

- IQR 1.5x: Fare=116 outliers, SibSp=46, Parch=213, Age=11.
- IQR 3.0x: Fare=53, SibSp=12, Parch=30, Age=1.
- Z-score >3: Fare=20, SibSp=6, Parch=4, Age=1.
- Modified Z-score (MAD) >3.5: Fare=116, SibSp=46, Age=6, Parch=0 (MAD=0).

Fare es la variable con mas outliers. Los valores extremos (>300 libras) corresponden a pasajeros de 1ra clase en suites de lujo: son casos raros validos, no errores. SibSp/Parch tienen outliers por familias grandes (eventos reales).

4.2 Deteccion multivariada

Se aplicaron 3 algoritmos multivariados sobre las variables transformadas con Yeo-Johnson: DBSCAN (ruido), Isolation Forest (300 arboles), y LOF (25 vecinos). Se uso un sistema de votacion: los outliers consensuados (detectados por 2+ metodos) corresponden principalmente a pasajeros con familias muy grandes y tarifas extremas.

Decision: no se eliminan outliers porque representan eventos reales del naufragio. Para modelado se recomienda winsorizar Fare y transformar con log.

4.3 Tratamiento (antes vs. despues)

Se aplico tratamiento a Fare (variable con mayor asimetria, skew=4.79):

- Winsorizacion al 1%: reduce skewness moderadamente, preserva escala original.
- log1p: reduce drasticamente el skewness, acercando a normalidad. Mejor balance normalizacion-interpretabilidad.
- Yeo-Johnson: mejor simetria (skew cercano a 0), pero pierde interpretabilidad.

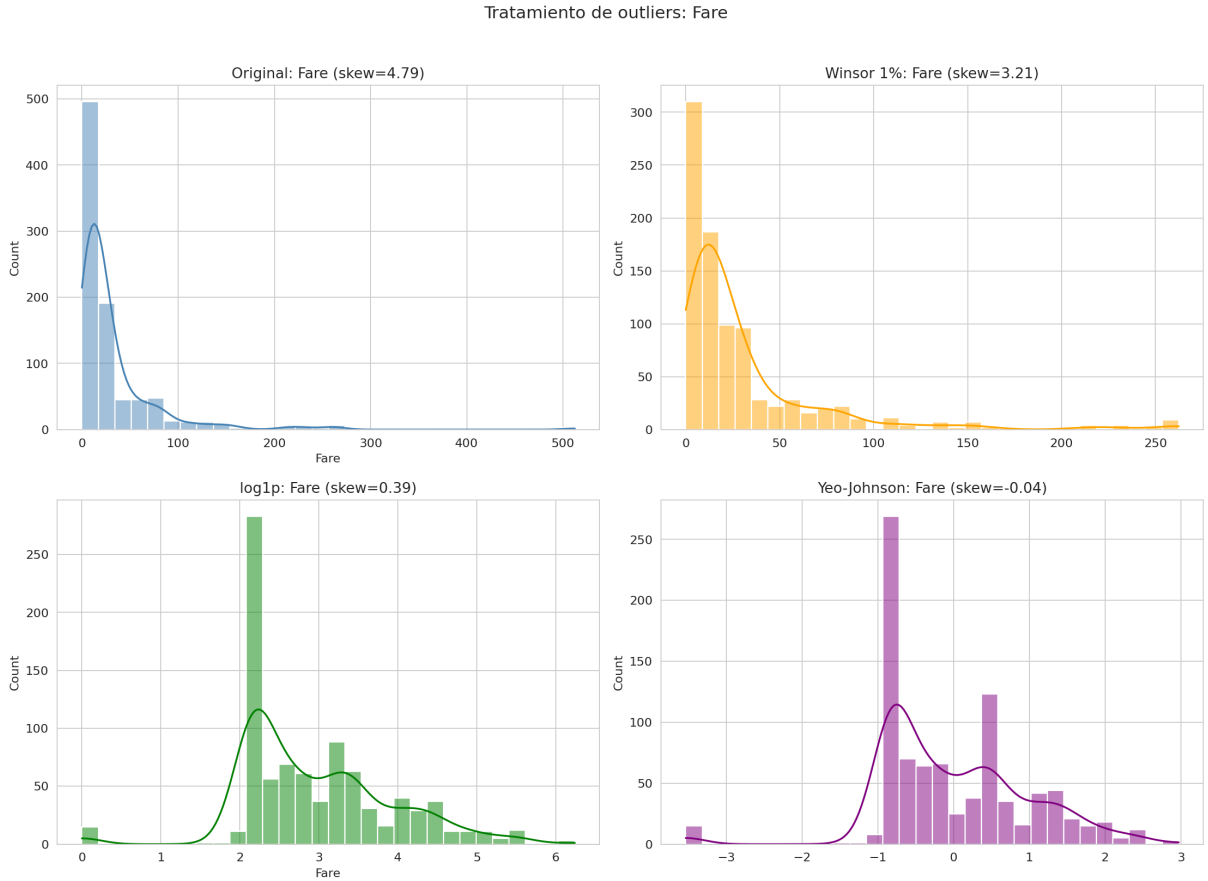

Figura 4: Comparacion de tratamientos aplicados a Fare

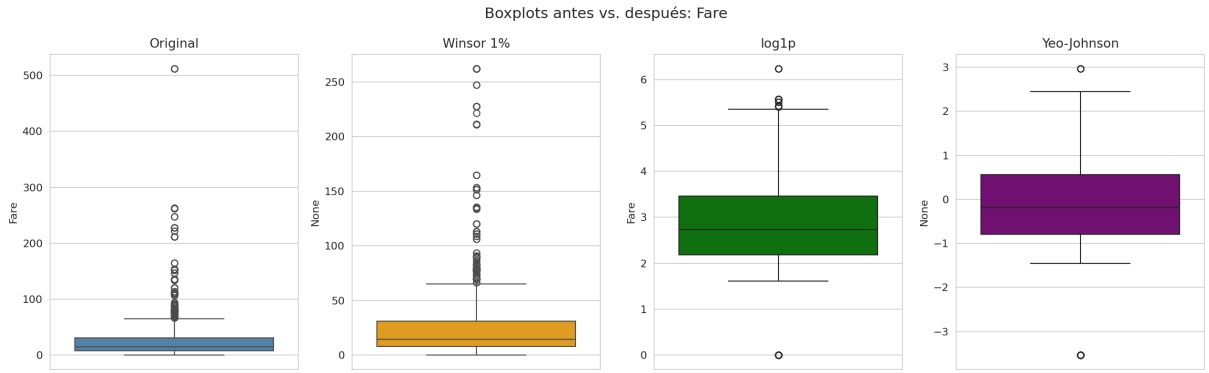

Figura 5: Boxplots antes vs. despues del tratamiento de Fare

5. Inferencia estadistica

Todas las pruebas se realizaron con nivel de significancia alpha = 0.05.

5.1 Normalidad - Shapiro-Wilk

H0: la variable sigue una distribucion normal. H1: no es normal.
Age: W=0.9815, p=7.34e-08. Conclusion: rechazamos H0, Age no es normal.
Fare: W=0.5219, p=1.08e-43. Conclusion: rechazamos H0, Fare no es normal.
Esto justifica usar pruebas robustas como Welch t-test y pruebas no parametricas.

5.2 t-test de Welch - Fare por Survived

H0: no hay diferencia en Fare entre sobrevivientes y no sobrevivientes.
H1: hay diferencia significativa.
Sobrevivientes: media=48.40. No sobrevivientes: media=22.12.
Resultado: t=6.84, p=2.70e-11. Conclusion: rechazamos H0.
Los sobrevivientes pagaron tarifas significativamente mas altas, porque la clase alta tenia prioridad en los botes salvavidas.

5.3 ANOVA - Age por Pclass (3 grupos)

H0: la edad media es igual en las 3 clases (mu1 = mu2 = mu3).
H1: al menos una clase tiene edad media diferente.
Clase 1: media=38.23. Clase 2: media=29.88. Clase 3: media=25.14.
Resultado: p=7.49e-24. Conclusion: rechazamos H0.
Los pasajeros de 1ra clase eran mayores en promedio, consistente con mayor poder adquisitivo.

5.4 Correlacion Pearson/Spearman - Age vs Fare

H0: no hay correlacion (rho = 0). H1: existe correlacion (rho != 0).
Pearson: r=0.096, p=0.010. Spearman: rho=0.135, p=0.0003.
Conclusion: correlacion debil pero significativa. La tarifa depende mas de la clase que de la edad.

5.5 Chi-cuadrado - Sex vs Survived (adicional)

H0: Sex y Survived son independientes. H1: existe asociacion.
chi2=260.72, p=1.20e-58, Cramer's V=0.54 (asociacion fuerte).
Conclusion: rechazamos H0. El sexo esta fuertemente asociado con la supervivencia, confirmando la politica de 'mujeres y ninos primero'.

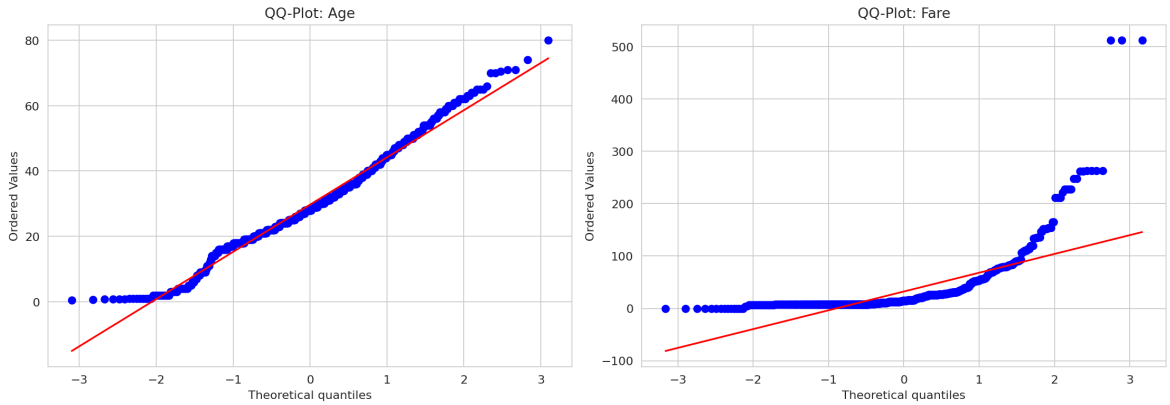

Figura 6: QQ-Plots de Age y Fare vs. distribucion normal teorica

6. Limitaciones y reproducibilidad

6.1 Limitaciones

- Muestra parcial: solo 891 de 2224 pasajeros (no incluye tripulacion completa).
- Cabin tiene 77.1% de faltantes, lo que impide analizar ubicacion en el barco.
- Age faltante en 19.87% puede sesgar analisis relacionados con edad.
- No se dispone de informacion sobre ubicacion al momento del impacto ni orden de evacuacion.
- Fare=0 en 15 registros podria ser error de registro o tripulacion sin pago.

6.2 Reproducibilidad

- Python >= 3.10 con pandas, numpy, matplotlib, seaborn, scipy, scikit-learn, statsmodels, missingno.
- Semilla fija: random_state=42 en todos los muestreos y algoritmos.
- Ruta de datos: data/train.csv (891 filas x 12 columnas).
- Exportables: exports/tabla_resumen.csv, exports/tests.json, exports/figuras/ (23 PNG).
- Notebook reproducible: Kernel -> Restart & Run All ejecuta sin errores.