Exploratory Data Analysis for Titanic Dataset (Final Project)

Student Names: Mohadeseh Yaghoobi
Mehrdad Sagha Khorasani
Amirhosein Majidi
Ahmadreza Hosseni

Instructor: Rahil Ghane Moosavi

فهرست مطالب

مقدمه

یکی از مهم‌ترین مسائل برای بررسی و تحلیل مواجهات و اقتصاد رفتاری یک جامعه در شرایط بحرانی، بازنگری در وقایع اجتماعی از پیش رخ داده و تحلیل علت‌ها و برآیندهای آن است؛ نظیر حوادث طبیعی یا سیاست‌گذاری‌های اقتصادی و سیاسی که مسائل معیشتی یا فرهنگی یک جامعه را تحت مخاطره قرار می‌دهد.
حادثه تایتانیک، یکی از تراژیک‌ترین و مشهورترین رویدادهای تاریخ دریانوردی است که در ۱۵ آوریل ۱۹۱۲ رخ داد. این کشتی، که به عنوان بزرگ‌ترین و لوکس‌ترین کشتی مسافربری زمان خود شناخته می‌شد، در نخستین سفر خود از بندر ساوت‌همپتون به نیویورک با برخورد به یک کوه یخ غول‌پیکر غرق شد.

این حادثه نه تنها جان بیش از ۱۵۰۰ نفر را گرفت، بلکه سوالات عمیقی درباره ایمنی دریانوردی، طراحی کشتی‌ها و مدیریت بحران را به وجود آورد. بررسی این واقعه نه تنها ما را با جزئیات فنی و انسانی این فاجعه آشنا می‌سازد، بلکه به ما کمک می‌کند تا درس‌هایی از آن بیاموزیم که می‌تواند در بهبود ایمنی و مدیریت خطر در صنعت دریانوردی و سایر حوزه‌ها مؤثر باشد. در این مقاله، به تحلیل علل، پیامدها و تأثیرات فرهنگی و اجتماعی حادثه تایتانیک خواهیم پرداخت.

در این تحلیل داده‌ای (EDA)، به بررسی عوامل مختلفی که منجر به این فاجعه شدند، خواهیم پرداخت. داده‌ها شامل اطلاعات مربوط به مسافران، نظیر اطلاعات شخصی، خانوادگی و اجتماعی آن‌ها می‌باشد. هدف این مطالعه، شناسایی الگوها و روابط میان این عوامل و درک بهتر از چگونگی وقوع حادثه تایتانیک است. با استفاده از تکنیک‌های تحلیل داده، می‌توانیم به بینش‌های ارزشمندی دست یابیم که نه تنها به تاریخ‌نگاری این رویداد کمک می‌کند، بلکه می‌تواند درس‌هایی برای بهبود ایمنی در آینده ارائه دهد.

معرفی دیتاست

در این تحقیق از دیتاستی استفاده شد تا با معیارها و متغیرهای متعدد آن بتوان جنبه‌های مختلف حادثه تایتانیک را مورد ارزیابی قرار دهیم. این متغیرها به داده‌های مربوط به مسافران کشتی تایتانیک اشاره دارند و هر یک از آن‌ها اطلاعات خاصی را درباره مسافران ارائه می‌دهند. در ادامه به توضیح هر یک از این متغیرها پرداخته می‌شود:

۱. **بازماندگان (survived)**: این متغیر نشان می‌دهد که آیا مسافر در حادثه تایتانیک زنده مانده است یا خیر. معمولاً به صورت ۰ (غرق شده) و ۱ (زنده مانده) کدگذاری می‌شود.

۲. **کلاس طبقاتی (pclass)**: این متغیر نمایان‌گر طبقه‌بندی اجتماعی مسافر است. تایتانیک سه طبقه داشت که شامل کلاس اول (لوکس)، کلاس دوم (متوسط) و کلاس سوم (اقتصادی) می‌شود. این متغیر معمولاً به صورت عددی (۱، ۲، ۳) نمایش داده می‌شود.

۳. **جنسیت (sex)** : این متغیر جنسیت مسافر را مشخص می‌کند و معمولاً به صورت "male" (مرد) و "female" (زن) کدگذاری می‌شود.

٤. **سن (age)** : سن مسافر را نشان می‌دهد. این متغیر می‌تواند عددی باشد و به صورت سال بیان شود.

۵. **همشیر و همسر (sibsp)** : این متغیر تعداد خواهر، برادر و همسر مسافر که در کشتی حضور داشتند، نشان می‌دهد.

٦. **والد یا فرزند (Parch)** : این متغیر تعداد والدین و فرزندان را فارغ از جنسیت نشان می‌دهد.

۷. **هزینه (Fare)**: این متغیر نشان‌دهنده میزان مبلغی است که هر مسافر برای سوارد شدن بر تایتانیک پرداخته است.

۸. **محل سوار شدن (embarked)**: این متغیر نشان می‌دهد که مسافر از کدام بندر سوار کشتی شده است. معمولاً شامل مقادیر "C" (شربور)، "Q" (کوینزتاون) و "S" (ساوت‌همپتون) است.

۹. چه کسی (who) : این متغیر معمولاً نشان‌دهنده نوع فرد است، مانند "man" (مرد)، "woman" (زن) یا "child" (کودک).

آشنایی با ابزارها و متدها

۱. کتابخانه‌های موردنیاز: موارد زیر کتابخانه‌های ضروری برای تحلیل داده و تجسم است. برای درک بهتر موضوع هرکدام از این کتابخانه‌ها در ادامه به‌طور خلاصه توضیح داده شده است:

- Numpy: کتابخانه‌ای قدرتمند برای کار با آرایه‌ها و انجام محاسبات عددی با کارایی بالا. ویژگی اصلی آن آرایه‌های چندبعدی است که عملیات ریاضی پیچیده را آسان و سریع می‌کند.
- Pandas: ابزاری برای تحلیل و مدیریت داده‌ها که به‌ویژه برای کار با داده‌های جدولی مانند فایل‌های CSV و دیتابیس‌ها استفاده می‌شود. ساختار اصلی داده‌ها در این کتابخانه، DataFrame و Series است که به سادگی قابل فیلتر، گروه‌بندی و تبدیل هستند.
- Mathplotlib.pyplot: یک کتابخانه جامع برای ایجاد نمودارها و تجسم داده‌ها در پایتون. این کتابخانه امکان طراحی نمودارهای خطی، پراکندگی، هیستوگرام، و بسیاری از انواع دیگر نمودارها را فراهم می‌کند.
- Seaborn: کتابخانه‌ای برای تجسم داده‌ها که بر اساس matplotlib ساخته شده و ابزارهای پیشرفته‌ای برای طراحی نمودارهای آماری و گراف‌های زیبا ارائه می‌دهد. برای مثال، نمودارهای heatmap و pairplot با این ابزار قابل ایجاد هستند.

۲. تنظیم مقادیر پیش‌فرض برای پارامترهای بصری: به جهت یکپارچه‌سازی خروجی‌های پروژه و حفظ نظم و القای یک قالب نظام‌مند بر ماژول‌ها بهتر است که پیش از شروع پروژه اندازه‌های پیش‌فرضی در نظر گرفته شود. به منظور این کار می‌توان از موارد زیر استفاده کرد:

- Plt.rcParams[‘figure.figsize’] که اندازه‌ی پیش‌فرض شکل‌های رسم شده را تنظیم می‌کند.
- Plt.rcParams[‘figure.dpi] که رزولوشن و وضوح اشکال را برحسب نقطه بر اینچ محاسبه و تعیین می‌کند.

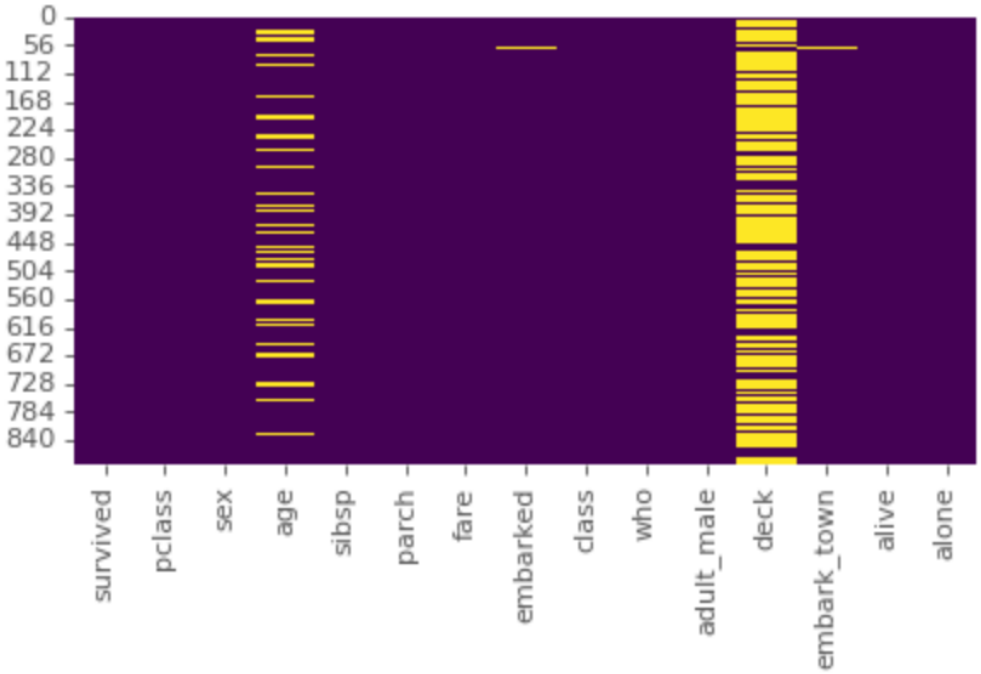
۳. بارگذاری دیتاست تایتانیک: دیتاست تایتانیک یکی از معروف‌ترین دیتاست‌ها در بستر داده‌کاوی و هوش مصنوعی است. این دیتاست آن‌چنان در سطح جهانی مورد استفاده قرار گرفته است که کتابخانه‌ی seaborn نیز آن را بصورت پیش‌فرض در خود دارد و کاربران این امکان را دارند تا این دیتاست را با استفاده از این کتابخانه

فراخوانی و از آن استفاده کنند. این دیتاست اطلاعاتی در مورد مسافران کشتی تایتانیک شامل سن، جنسیت، کلاس سفر و وضعیت نجات دارد. ابعاد کلی این دیتاست شامل ۸۹۱ سطر و ۱۵ ستون می‌باشد که نشان‌دهنده آن است که حامل اطلاعات بیش از هشتصد مسافر تایتانیک است. همین‌طور با استفاده از متد describe می‌توان آمار توصیفی این دیتاست را به‌صورت کامل مشاهده کرد. این متد مقادیر غیرخالی، انحراف معیار، حداقل و حداکثر مقدار هر ستون داده‌ای، چندک‌ها و ... مربوط به دیتاست را در یک جدول منظم نمایش می‌دهد. شکل زیر مربوط به خروجی متد مربوطه می‌باشد:

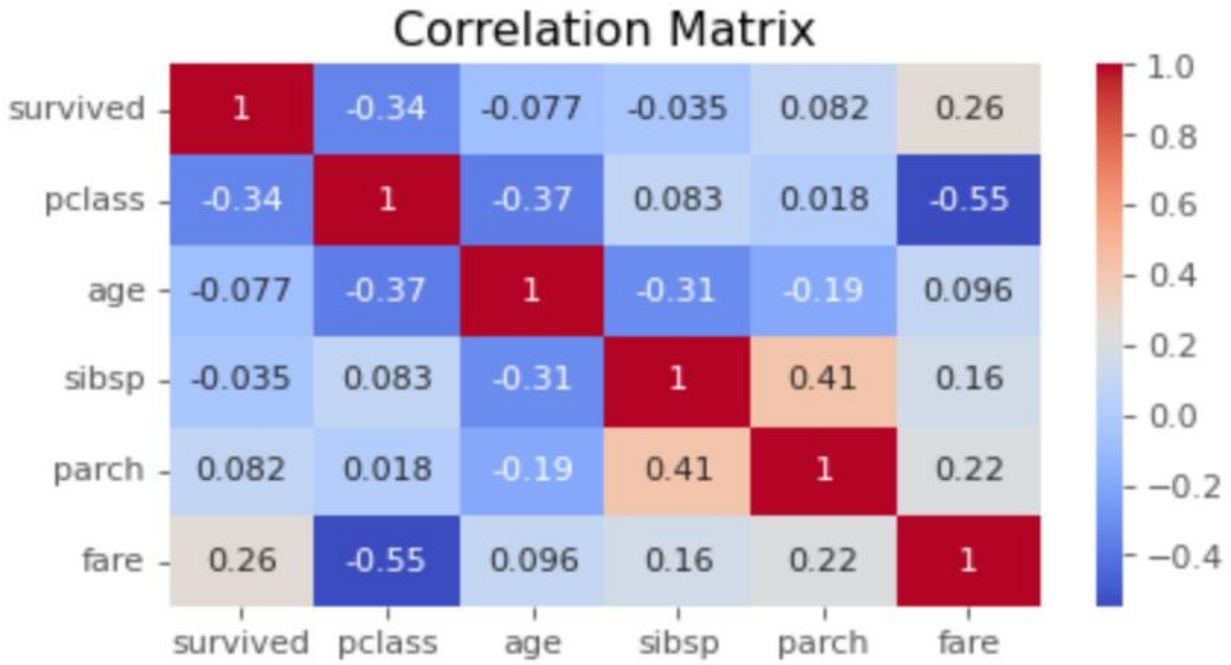
	survived	pclass	age	sibsp	parch	fare
count	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

تحلیل داده‌ها: نقشه حرارتی

نقشه حرارتی یا Heat Map یک نمایش گرافیکی داده است که از یک سیستم کدگذاری رنگ برای نشان دادن مقادیر مختلف استفاده می‌کند. این روش تجسم داده را – که اندازه یک پدیده است – بوسیله‌ی رنگ در دو بعد نشان می‌دهد. تنوع در رنگ ممکن است بر اساس رنگ یا شدت باشد، و این نمودار نشانه‌های بصری آشکاری را در مورد چگونگی خوشه‌بندی یا تغییر پدیده به خواننده نشان می‌دهد. ابتدا ما با استفاده از این نقشه می‌توانیم میزان داده‌هایی که بصورت تهی می‌باشند را شناسایی و از لیست داده‌های خارج کنیم تا در تحلیل نهایی تاثیر نگذارند. در تصویر زیر می‌بینیم که با استفاده از نقشه معلوم می‌شود که به میزان قابل توجهی اطلاعات درمورد شماره اتاق و همچنین سن مسافران در دسترس نیست. به همین دلیل فیلد یا ستون شماره اتاق را به کلی از دیتاست حذف می‌کنیم. اما این امکان برای ستون سن میسر نیست چراکه این ستون از مهم‌ترین داده‌ها در مسیر تحلیل ما هست. پس میانگین سنی مسافران را براساس جنسیت محاسبه می‌کنیم و به هر مسافری که اطلاعات سنی نامشخصی دارد اطلاق می‌کنیم.



بررسی همبستگی میان داده‌ها: در این مرحله ویژگی‌های عددی از دیتاست استخراج شده و سپس میزان همبستگی آن‌ها نیز توسط یک ماتریس همبستگی محاسبه می‌گردد. برای ساخت این ماتریس از کتابخانه seaborn و متد heatmap استفاده می‌کنیم که برای نمایش همبستگی‌ها استفاده شده و هر مقدار در این ماتریس مقداری بین ۱ تا ۱- می‌باشد. در تصویر زیر ماتریس همبستگی مربوط به دیتاست را می‌بینیم که رنگ قرمز در آن نشان‌دهنده‌ی همبستگی مثبت و آبی نشان‌دهنده‌ی همبستگی منفی است.



در تحلیل این ماتریس می‌توان به نکات زیر اشاره کرد:

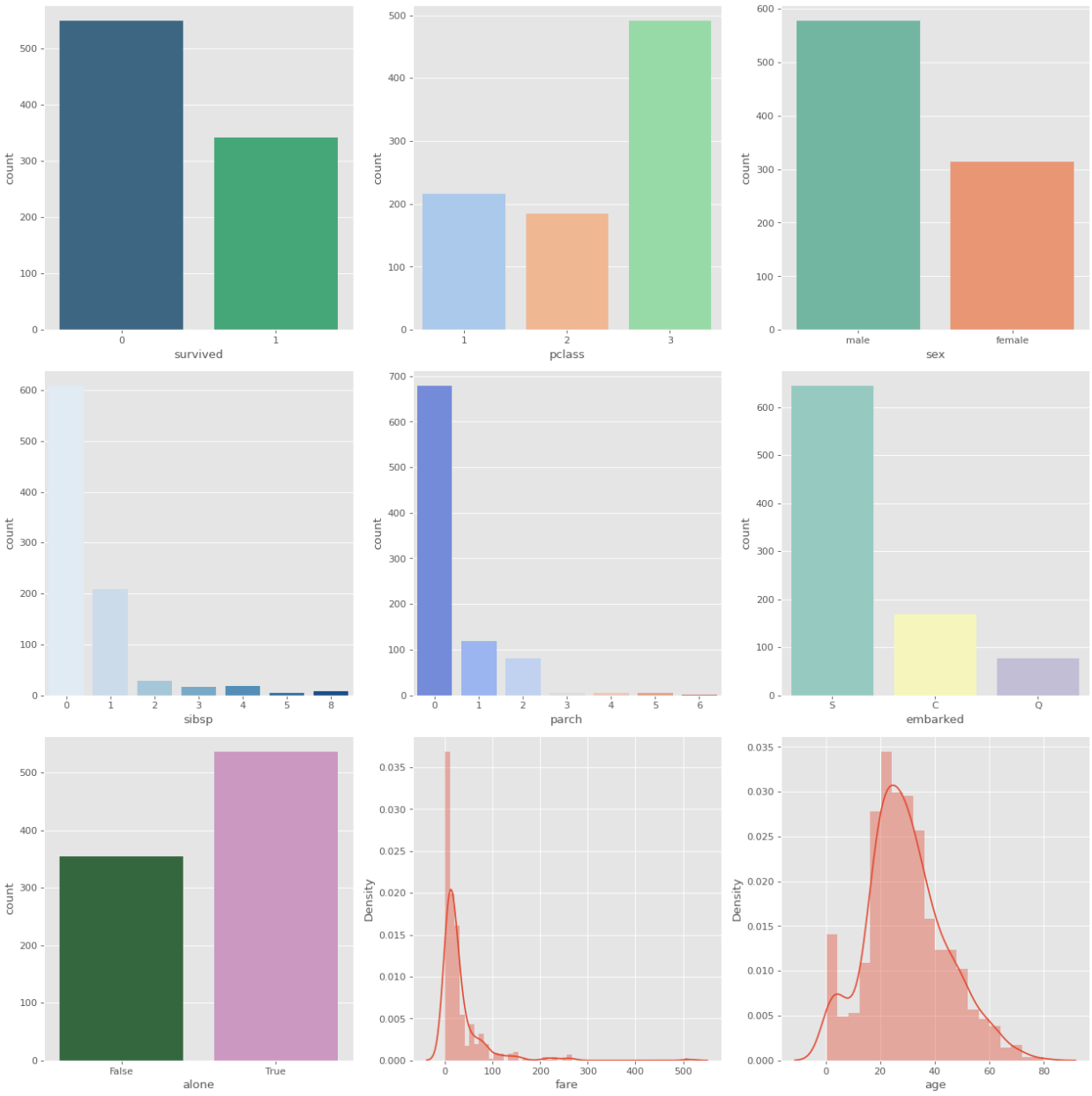
- کلاس‌های fare و pclass دارای همبستگی منفی قوی (۰.۵۵-) هستند که می‌توان نتیجه گرفت هرچه کلاس طبقاتی بالاتر باشد (کلاس یک یا لوکس)، نرخ کرایه بیشتر است.
- کلاس‌های survived و pclass دارای همبستگی منفی متوسطی (۰.۳٤-) هستند که می‌توان گفت افراد در کلاس‌های طبقاتی پایین‌تر به‌طور متوسط شانس کمتری برای نجات داشته‌اند.
- کلاس‌های parch و sibsp دارای همبستگی مثبت (۰.٤۱) هستند که نشان می‌دهد همراه داشتن والدین و یا همسفر می‌تواند همبستگی مثبتی ایجاد کند.

این موارد جزء کوچکی از تحلیل‌هایی است که می‌توان به کمک نقشه‌های حرارتی بدست آورد. اما باید توجه کرد که نقشه‌های حرارتی تنها برای شناخت اولیه از داده‌ها و ارتباط کلی بین آن‌ها استفاده می‌شوند و برای

تحلیل‌های موثرتر باید از روش‌های دیگر که دارای تعمق بیشتر در داده‌ها می‌باشند استفاده کرد. در بخش‌های پیش رو به این گونه از تحلیل‌ها پرداخته می‌شود.

تحلیل داده‌ها: تحلیل‌های تک‌متغیره

تحلیل تک‌متغیره (Univariate Analysis) در EDA به بررسی ویژگی‌های آماری یک متغیر به صورت مجزا می‌پردازد. هدف این تحلیل شناسایی الگوها، توزیع داده‌ها، مقادیر پرت، و خلاصه‌ای از اطلاعات آماری (مثل میانگین، میانه، و انحراف معیار) است. ابزارهای معمول برای این تحلیل شامل جدول فراوانی، نمودارهای هیستوگرام، جعبه‌ای (Box Plot)، و نمودارهای پراکندگی است.



از فواید این روش می‌توان به سادگی و وضوح که به دلیل بررسی تنها یک متغیر، تفسیر نتایج ساده و مستقیم است، نام برد. همچنین این روش در شناسایی داده‌های پرت به کار می‌آید و می‌توان با کمک آن به راحتی مقادیر غیرعادی یا اشتباه در داده‌ها را نشان می‌دهد. همچنین این روش در درک توزیع داده‌ها کمک به‌سزایی می‌کند و به فهم رفتار و توزیع داده‌ها (نرمال یا غیرنرمال بودن) کمک می‌کند.

گرچه این روش دارای کمبودهایی نیز هست. در تحلیل تک‌متغیره فقط یک متغیر را بررسی می‌شود و هیچ اطلاعاتی درباره روابط بین متغیرها ارائه نمی‌شود که همین نیز موجب محدودیت در روابط می‌گردد. همچنین تعاملات پیچیده بین متغیرها نادیده گرفته می‌شود و مفسر را در رسیدن به یک تحلیل جامع ناکام می‌گذارد. برای درک کامل‌تر، باید به تحلیل‌های دو یا چند متغیره نیز روی آورد. در کل، تحلیل تک‌متغیره اولین گام در فرآیند EDA است که دیدگاه اولیه و ارزشمندی از داده‌ها ارائه می‌دهد، اما کافی نیست و باید با تحلیل‌های پیشرفته‌تر تکمیل شود.
در تصویر پیش می‌توان تمام نمودارهای تحلیلی تک متغیره مربوط به دیتاست تایتانیک را مشاهده کرد. با نگرش به این تصویر می‌توان به این تحلیل‌ها رسید:

۱. تحلیل تک متغیره نجات‌یافتگان (survived) : تعداد جان‌باختگان بیشتر از تعداد نجات‌یافتگان است.

۲. تحلیل تک متغیره کلاس طبقاتی (pclass) : بیشترین مسافران در کلاس اقتصادی ۳ قرار دارند که نشان‌دهنده‌ی سطح اقتصادی پایین‌تر است و همچنین کلاس لوکس ۱ کمترین تعداد را دارد.

۳. تحلیل تک متغیره جنسیت (sex) : طبق این نمودار مشخص می‌گردد که تعداد مسافران مرد بیشتر از تعداد مسافران زن است و این تفاوت ممکن است در نرخ نجات تاثیرگذار باشد.

٤. تحلیل تک متغیره تعداد خواهر/ برادر یا همسر (sibsp) : بیشتر مسافران بدون همراهی خواهر، برادر یا همسر خود به سفر با تایتانیک پرداخته‌اند. تعداد کمی با ۱ یا ۲ همراه سفر کرده‌اند و تعداد افراد با ۳ نفر یا بیشتر به ندرت دیده می‌شود.

۵. تحلیل تک متغیره تعداد والدین/ فرزند (parch) : اکثر مسافران بدون والدین یا فرزند سفر کرده‌اند و همچنین افراد کمی همراه با والد(ین) یا فرزند(ان) خود سفر کرده‌اند.

٦. تحلیل تک متغیره محل سوار شدن (embarked) : بیشتر مسافران از بندر ساوت‌همپتون سوار تایتانیک شده‌اند. تعداد کمتری از چربورگ و کمترین تعداد از کوئینزتاون سوار تایتانیک شده‌اند.

۷. تحلیل تک متغیره تنها یا با همراه (alone) : این نمودار نشان می‌دهد که بیشتر مسافران تنها سفر کرده‌اند و تعداد کمتری همراه داشته‌اند.

۸. تحلیل تک متغیره توزیع کرایه بلیط (fare) : طبق نمودار مربوطه می‌توان استنتاج کرد که کرایه بلیط بیشتر مسافران زیر ۵۰ واحد پولی (بطور مثال پوند) است اما برخی کرایه‌ها (به ندرت) بسیار بالاتر از این میزان مبلغ است. توزیع داده‌ها چولگی مثبت دارد. یعنی بیشتر داده‌ها در سمت چپ قرار دارند و تعداد کمی داده با مقادیر بالا در سمت راست حضور دارند.

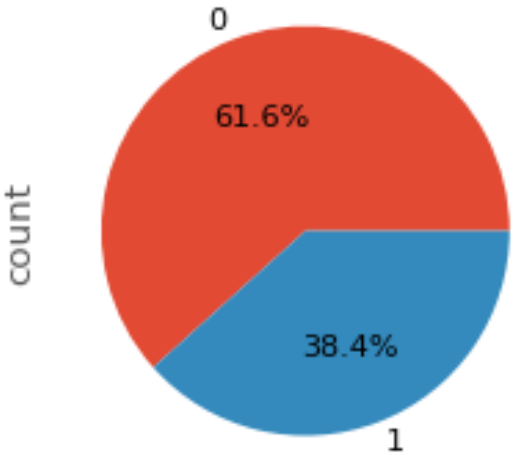
۹. تحلیل تک متغیره سن (age) : توزیع سن دارای شکل زنگوله‌ای است و حول ۳۰ سالگی متمرکز است. تعداد کمی از مسافران در سنین بالاتر از این میانگین و بالای ٦۰ سال حضور دارند.

تحلیل داده‌ها: تحلیل‌های چندمتغیره یا کلاس به کلاس

تحلیل چندمتغیره (Multivariate Analysis) در EDA به بررسی روابط و تعاملات بین دو یا چند متغیر می‌پردازد. این تحلیل به درک بهتر الگوهای پیچیده، همبستگی‌ها، و تأثیر متقابل متغیرها کمک می‌کند. ابزارهای رایج برای این نوع تحلیل شامل ماتریس همبستگی، نمودارهای پراکندگی (Scatter Plot)، جداول محوری (Pivot Tables)، و تحلیل‌های آماری پیشرفته‌تر مثل رگرسیون چندمتغیره یا تحلیل عاملی است. در ادامه به توصیف هر کلاس بصورت مجزا و ارتباطش با متغیرهای دیگر می‌پردازیم:

کلاس **survived** :

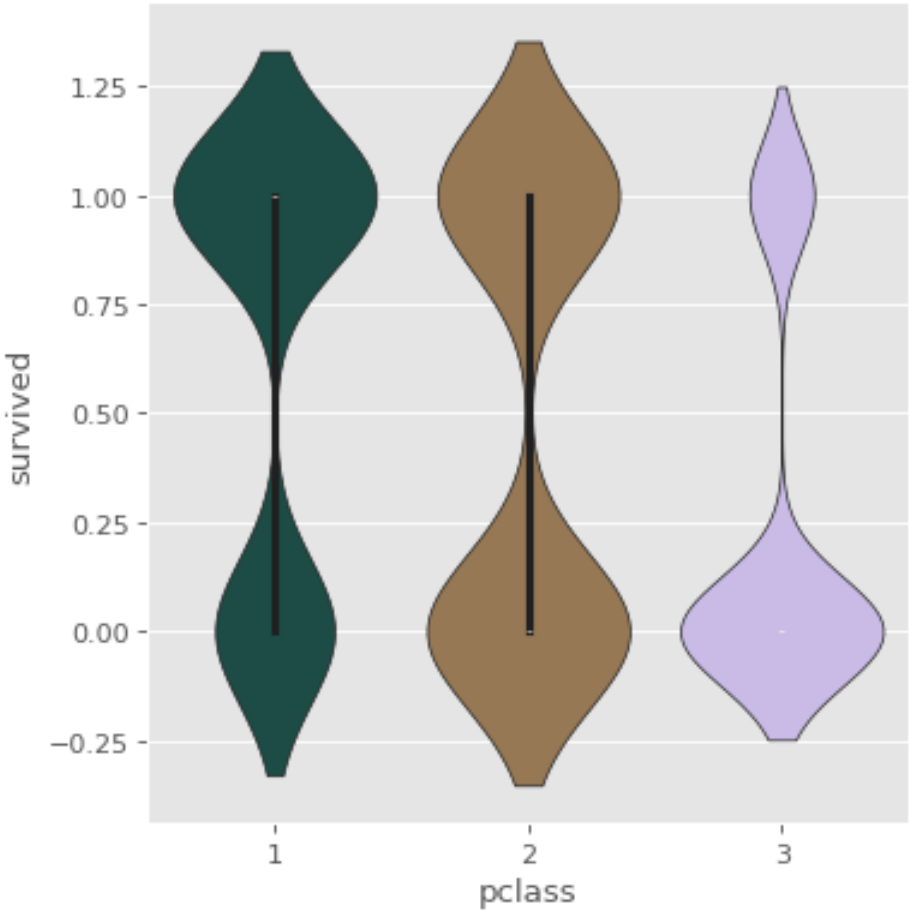
در این کلاس داده‌های مربوط به مسافرانی که از حادثه تایتانیک نجات پیدا کرده‌اند مشخص شده. از تعداد کل ۸۹۱ مسافر، تعداد ۵۴۹ نفر کشته شده‌اند و تنها ۳۴۲ مسافر نجات یافته‌اند. این داده مشخص می‌کند که حدودا ۳۸ درصد جمعیت مسافران نجات یافته و بیشتر از درصد آن‌ها به همراه تایتانیک غرق شده‌اند.



کلاس **pclass** :

این متغیر نمایانگر طبقه‌بندی اجتماعی مسافر است. تایتانیک سه طبقه داشت که شامل کلاس اول (لوکس) کلاس دوم (متوسط) و کلاس سوم (اقتصادی) می‌شود. این متغیر معمولا به صورت عددی (۱و۲و۳) نمایش داده می‌شود.

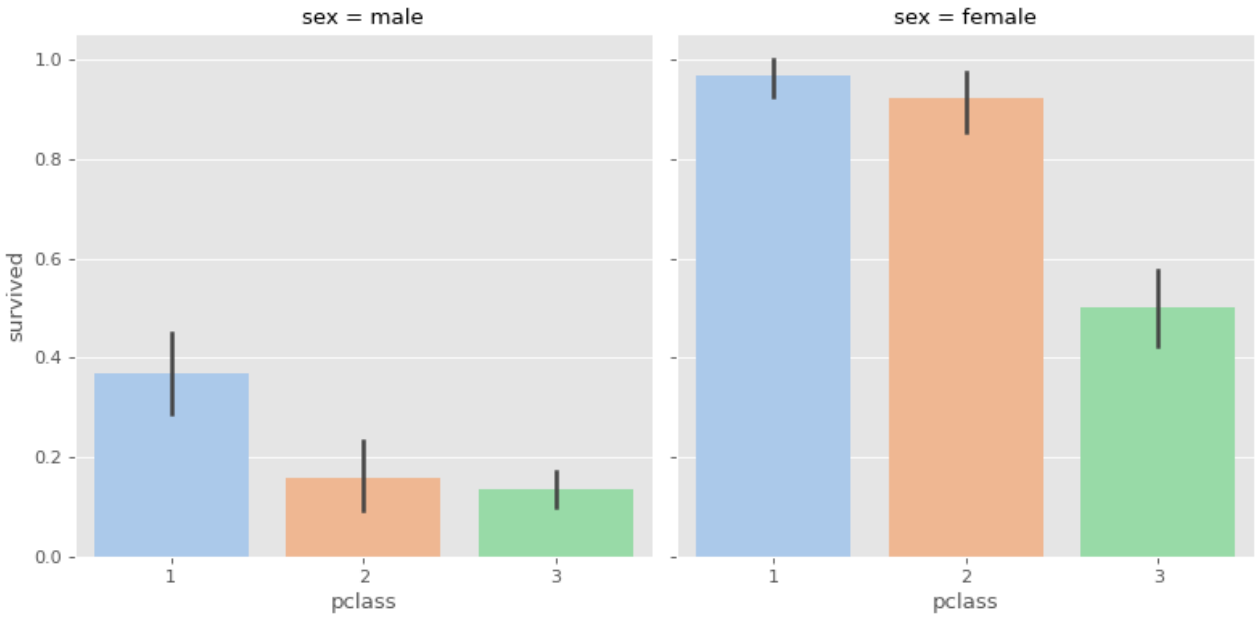
نحوه توزیع و پراکندگی جامعه آماری تایتانیک بدین صورت است که از ۸۹۱ مسافر کشتی، ۲۱٦ نفر در کلاس یک و یا همان کلاس لوکس دسته بندی شده‌اند. و کلاس‌های دو و سه به ترتیب دارای ۱۸٤ و ٤۹۱ مسافر بوده‌اند. به کمک متغیر survived و تلفیق آن با کلاس pclass می‌توانیم به این نتیجه دستیابیم: از۲۱٦ مسافری در کلاس یک بودند ۱۳٦ نفر نجات یافته و ۸۰ نفر غرق شدند. در کلاس دو ۸۷ نفر نجات یافته و ۹۷ نفر غرق شدند. اما در کلاس سه که ٤۹۱ مسافر جز آن بودند تنها ۱۱۹ نفر نجات یافتند و اکثریت جمعیت یعنی ۳۷۲ نفر غرق شدند. می‌توان نتیجه گرفت که اعضای کلاس یک فارغ از جنسیت و سن‌شان تقریبا ٦۰ درصد امکان نجات داشته‌اند. این در حالی است که کلاس متوسط و اقتصادی به ترتیب شانس ٤۷ درصد و ۲۳ درصد داشته‌اند. همچنین با کمک نمودار تراکم یا density می‌توان به سادگی ارزیابی کرد؛ طبقه بندی اجتماعی مسافر بر میزان احتمال زنده ماندنش تاثیر به سزایی دارد و کیفیت خدمات‌رسانی در طبقات با تراکم جمعیتی مسافران در آن کلاس رابطه عکس دارد.



کلاس Sex :

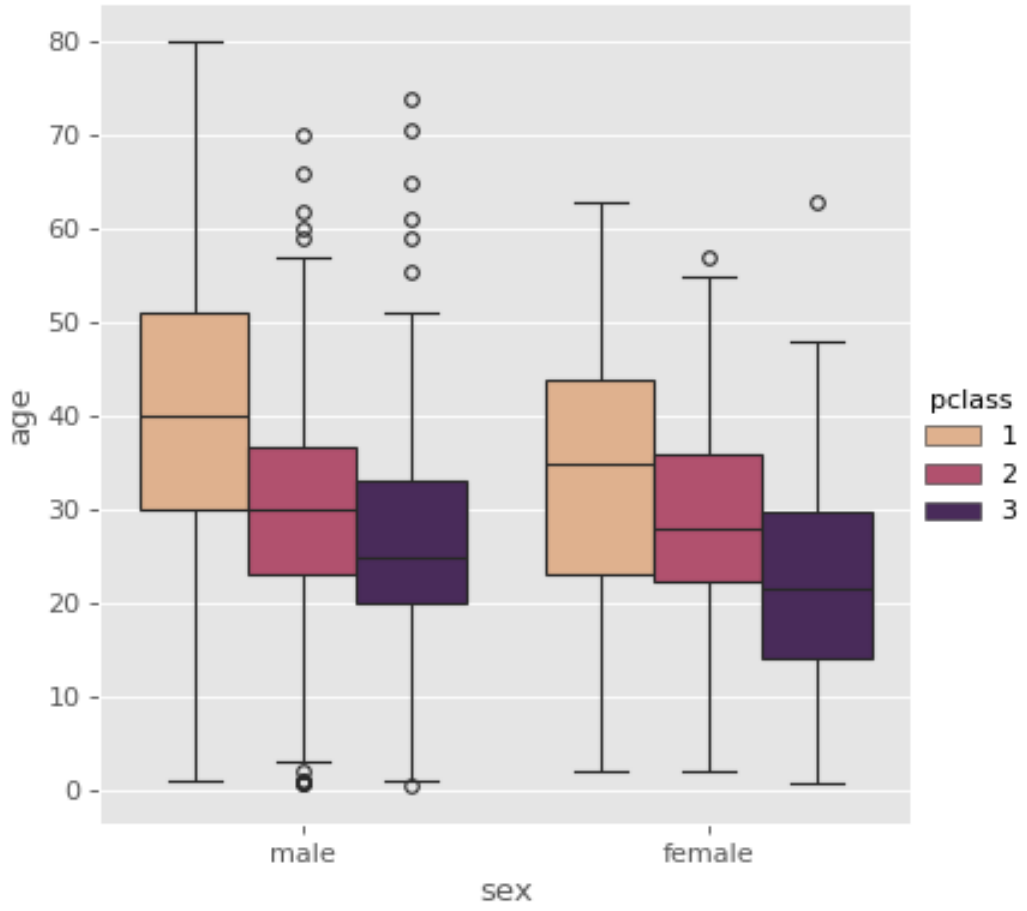
طبق داده‌های موجود، تعداد مسافرین مذکر در تایتانیک ۵۷۷ و تعداد مسافرین مونث ۳۱٤ نفر بوده. به سادگی قابل رویت است که جمعیت مردان تقریبا دوبرابر جمعیت مسافران زن است، و این در حالی است که میزان مرگ و میر آن‌ها نسبت به مسافران زن تقریبا ۵ برابر است. به عبارتی دیگر از ۵۷۷ مسافر مذکر تنها ۱۰۹ نفر نجات پیدا کردند و ٤٦۸ نفر کشته شدند. در صورتی که از ۳۱٤ نفر مسافر زن ۲۳۳ نفر از آن‌ها نجات یافته و ۸۱ نفر از آن‌ها مغروق شدند. هر مرد بر روی کشتی تایتانیک فارغ از سن و کلاس مسافری خود، تقریبا ۱۸ درصد احتمال نجات یافتن داشته و هر زن تقریبا ۷۳ درصد.

با ترکیب سه متغیر survived، sex و pclass به نتایج جالبی می‌رسیم. به عنوان مثال احتمال نجات مسافرین زنی که در کلاس و طبقه لوکس قرار داشتند به نسبت مسافران زنی که در طبقه اقتصادی قرار داشته‌اند میزان قابل توجهی بیشتر است؛ یعنی مسافران زن طبقه لوکس با احتمال ۹۵ درصد امکان زنده ماندن داشته و زنان طبقه اقتصادی تنها ۵۰ درصد احتمال نجات داشته‌اند.



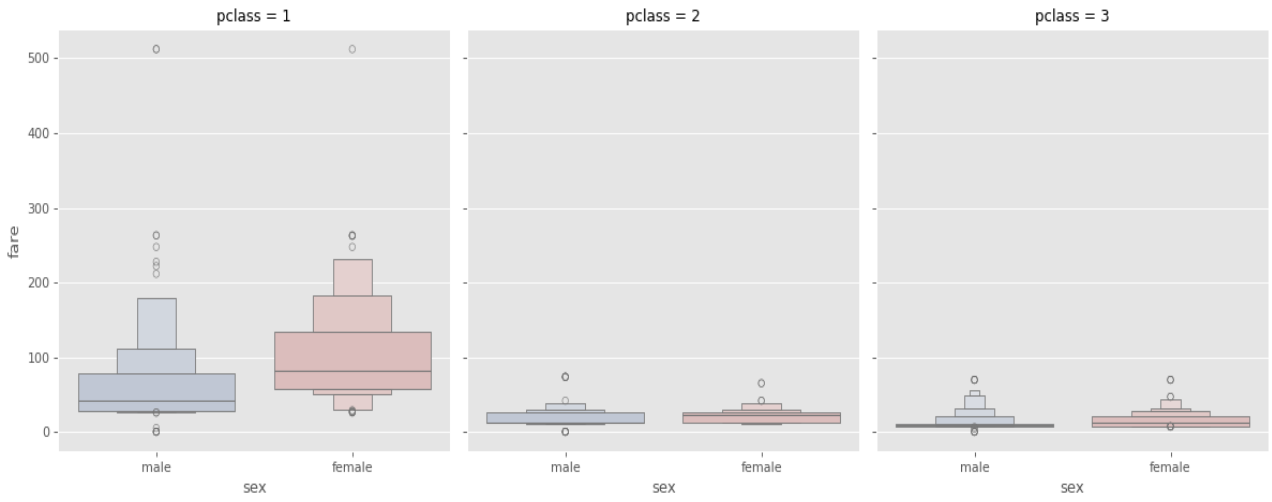
کلاس Age :

اکثریت جمعیت مسافران تایتانیک جوان بوده و در محدوده سنی ۲۰ تا ۳۰ سال قرار داشته‌اند. میانگین سنی زنان بین ۲۰ تا ۳۳ و میانگین سنی مردان بین ۲۲ تا ۴۰ می‌باشد و در نتیجه جمعیت زنان مسافر جوان‌تر از مسافران مرد است. نکته مهم دیگری که از این داده‌ها به دست می‌آید حاصل اتصال دو آرگومان age و pclass می‌باشد. کیفیت کلاس قرارگیری مسافر با سن مسافر رابطه‌ی مستقیمی دارد، یعنی هرچه سن مسافر کمتر باشد احتمال رزرو کردن کلاس لوکس کم‌تر است و این قاعده برای هر دو جنسیت صدق می‌کند. همچنین می‌توان گفت که زنان و افراد نابالغ بطور کلی شانس بیشتری برای زنده ماندن داشتند که این موضوع تاکید بر سیاست‌های نجات‌گری دارد که به حفاظت از این گروه‌ها اولویت می‌دهد. این نکته نشان می‌دهد که در مواقع اضطراری، رفتارهای اجتماعی و فرهنگی می‌تواند بر نتایج تاثیرگذار باشد.



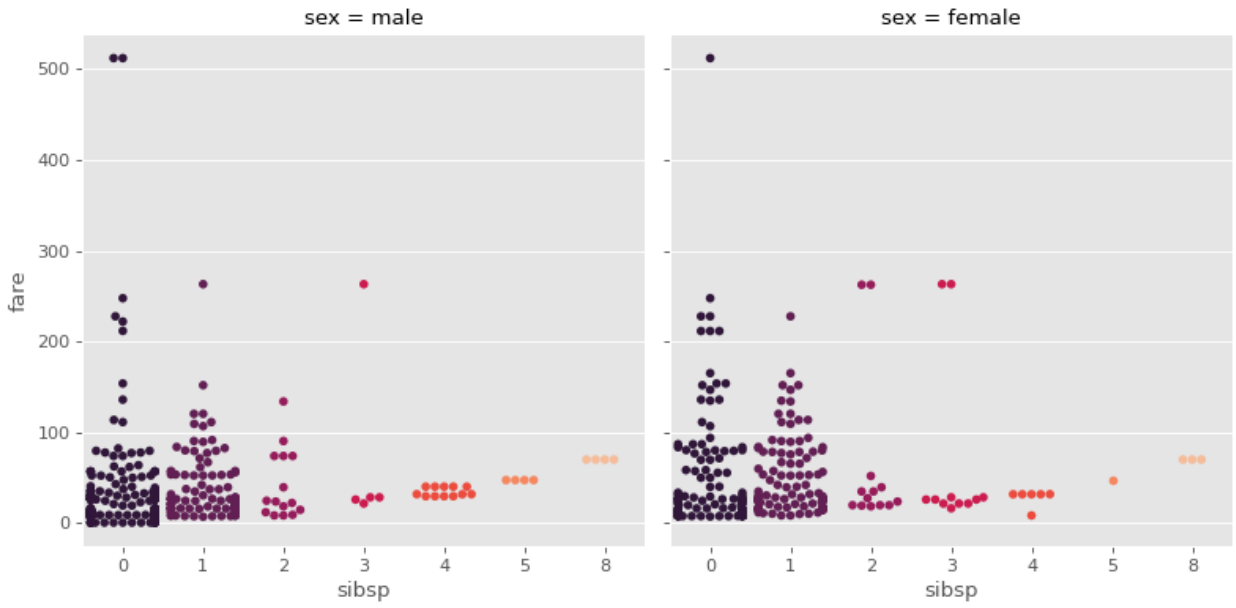
کلاس Fare:

این مولفه مربوط به میزان هزینه‌ای است که هر مسافر برای سوار شدن بر تایتانیک پرداخته است. این داده از تنوع زیادی برخوردار نیست و اکثر مسافران تقریبا هزینه‌ی نزدیک به هم پرداخته‌اند. حدود ٤٠٠ نفر از مسافران مبلغی بین ۱ تا ۱۰ پوند (واحد پول مربوط به دیتاست) پرداخته‌اند، حدود ۳۵۰ نفر تقریبا مبلغی بین ۱۰ تا ۳۰ پوند پرداخته و باقی مسافران چیزی بین ۳۰ تا ۱۵۰ پوند پرداخته‌اند. البته مسافرانی در کشتی حضور داشته‌اند که مبلغی معادل ۵۲۰ پوند پرداخت کرده باشند که می‌توان حدس زد آن‌ها مسافرین ویژه با امتیازات بسیار خاصی بوده‌اند. اگر معیار جنسیت را به تحلیل خود اضافه کنیم به این نتیجه خواهیم رسید که میزان متوسط مبلغی که مردان پرداخت کرده‌اند تقریبا ٤ پوند است و میزان متوسط پرداختی مسافران زن تقریبا ۱۵ پوند که می‌توان نتیجه گرفت زنان مبلغ بیشتری پرداخته‌اند تا از امکان مخصوص جنسیتی برخوردار شوند. این امکانات احتمالا مربوط به وسایل بهداشتی و مراقبتی بوده است. با توجه به آرگومان pclass با نگاهی ساده متوجه می‌شویم که کلاس قرارگیری مسافرین بر میزان مبلغی که پرداخته‌اند تاثیر داشته است. نکته جالب دیگر در این است که تفاوت میزان پرداختی زنان و مردان در کلاس لوکس بسیار بیشتر از تفاوت میزان پرداخت آن‌ها در کلاس‌های معمولی و اقتصادی است. پس احتمال استفاده از امکانات مخصوص مسافران بیشتر مربوط به زنانی بوده که در کلاس لوکس بوده‌اند. نکته جالب دیگر این است که میزان پرداخت هزینه در تایتانیک هیچ ارتباطی با میزان احتمال زنده ماندن مسافران نداشته است.



کلاس **sibsp (Siblings and Spouses)** :

این کلاس به طور مشخص آمار مسافرینی را در اختیار ما می‌گذارد که به همراه همسر و یا خواهر و برادر خود بر تایتانیک بوده‌اند. از ۸۹۱ مسافری که بر تایتانیک بوده‌اند ٦٠٨ نفر از آن‌ها نه خواهر و برادری داشته‌اند که در مسافرت با آن‌ها همراه بوده باشد و نه همسری. ۲۰۹ مسافر یک همراه داشته‌اند که احتمالا این همراه همسر آن‌ها بوده است (گرچه که این تنها یک حدس است و داده‌ای موثق برای مشخص گردن این مدعا وجود ندارد). ۲۸ نفر دو همراه داشته‌اند (که احتمالا موضوع بحث خواهران و برادران هستند زیرا چند همسری در بریتانیا از لحاظ اجتماعی و اخلاقی از دیرباز مورد پسند نبوده). ۱۸ مسافر ٤ همراه، ١٦ مسافر ۳ همراه، ۷ مسافر ۸ همراه و پنج مسافر پنج همراه داشته‌اند. با دقت در این آمار می‌توان متوجه شد که در تایتانیک دو خانواده پرجمعیت (به صورت تقریبی) هشت نفره و پنج نفره وجود داشته است.



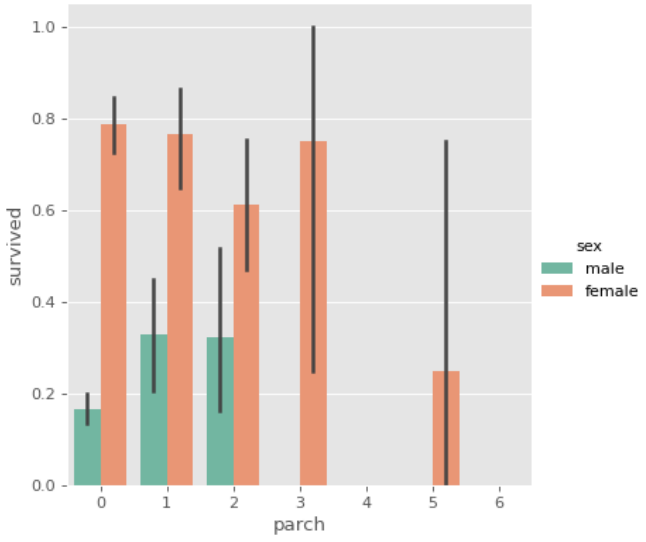
با توجه به آرگومان survived به نکات بسیار جالبی می‌رسیم: مسافرینی که بدون هیچ همراهی به کشتی تایتانیک آمده بودند، احتمال مرگ نسبتا بالایی داشته‌اند اما مسافرینی که روی کشتی همراهی داشتند این احتمال را کاهش داده‌اند. گرچه هرچه تعداد همراهان بیشتر می‌شود میزان مرگ و میر مجددا بیشتر می‌شود تا جایی که افزایش میزان همراه به مرگ حتمی دچار می‌شود. تمامی اعضای هر دو خانواده‌ی پرجمعیتی که پیشتر از آن‌ها

یاد شد در کشتی تایتانیک کشته شدند. پس همان‌طور که می‌توان نتیجه گرفت بودن همراهان (یک تا دونفر) بواسطه دلایل فطری و خواست خیر ارجع اجتماعی از احتمال مرگ در شرایط بحرانی می‌کاهد؛ نیز می‌توان نتیجه گرفت که میزان بالای داشتن همراه نتیجه عکس می‌دهد و این موضوع احتمال عاملی روان‌شناختی داشته باشد که اگر فردی خود و تمامی اعضای خانواده خود را در وضعیتی خطرناک و تهدیدگر بیابد احتمال تعقل و تفکر منطقی از او سلب می‌شود و همچنین مدیریت بحران کاری بسیار دشوارتر می‌گردد.

نکته دیگر در این بخش رابطه جالب نسبت به طبقه اجتماعی و میزان جمعیت است. خانواده‌های پرجمعیت عموما در طبقه سوم و اقتصادی هستند و اکثر مسافرینی که به تنهایی سفر کرده‌اند یا دارای خانواده کم جمعیت هستند در طبقه لوکس بوده‌اند.

کلاس **parch(Parents and Children)** :

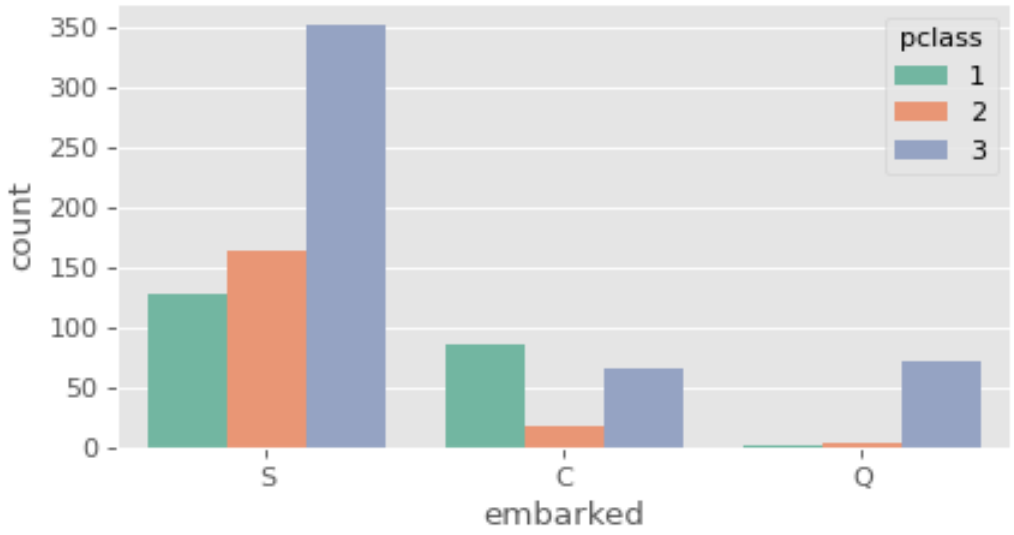
این کلاس نشان‌دهنده‌ی آمار مسافرانی است که روی کشتی فرزند یا والدی را به همراه خود داشته‌اند. در نگاه اول این داده‌ها به دست می‌آیند: ٦٧٨ مسافر از ٨٩١ مسافر بدون هیچ والد یا فرزندی بر روی کشتی سوار بوده‌اند. ١١٨ نفر از مسافران یا با یکی از والدین خود بر کشتی بوده و یا یک فرزند خود را به همراه خود آورده بوده‌اند. ٨٠ مسافر یا با والدین خود در سفر بوده و یا دو فرزند خود را به تایتانیک آورده بودند. در ادامه آمار بدین‌گونه می‌باشد که ٥ مسافر دارای ٥ واحد از این کلاس، ٥ مسافر دارای ٣ واحد از این کلاس، ٤ مسافر دارای ٤ واحد از این کلاس و ١ مسافر دارای ٦ واحد از این کلاس هستند.



پس بصورت کلی می‌توان گفت اکثریت قریب به یقین جمعیت مسافران بدون حضور والدین یا فرزندان خود به این سفر آمده بودند. مانند کلاس قبلی یعنی sibsp می‌توان گفت که مسافرینی که تنها به سفر آمده بودند احتمال زنده ماندن کمتری نسبت به افرادی داشتند که با والدین یا فرزندان خود به سفر آمده بودند. اما مجددا هرچه تعداد اعضای خانواده افزایش یافته احتمال مرگ و میر بیشتر شده است. نکته جالب دیگر در این کلاس این است که فارغ از تعداد اعضای خانواده، مسافرین مونث شانس بیشتری برای نجات یافتن داشته‌اند، حتی در خانواده‌هایی با جمعیت بالا.

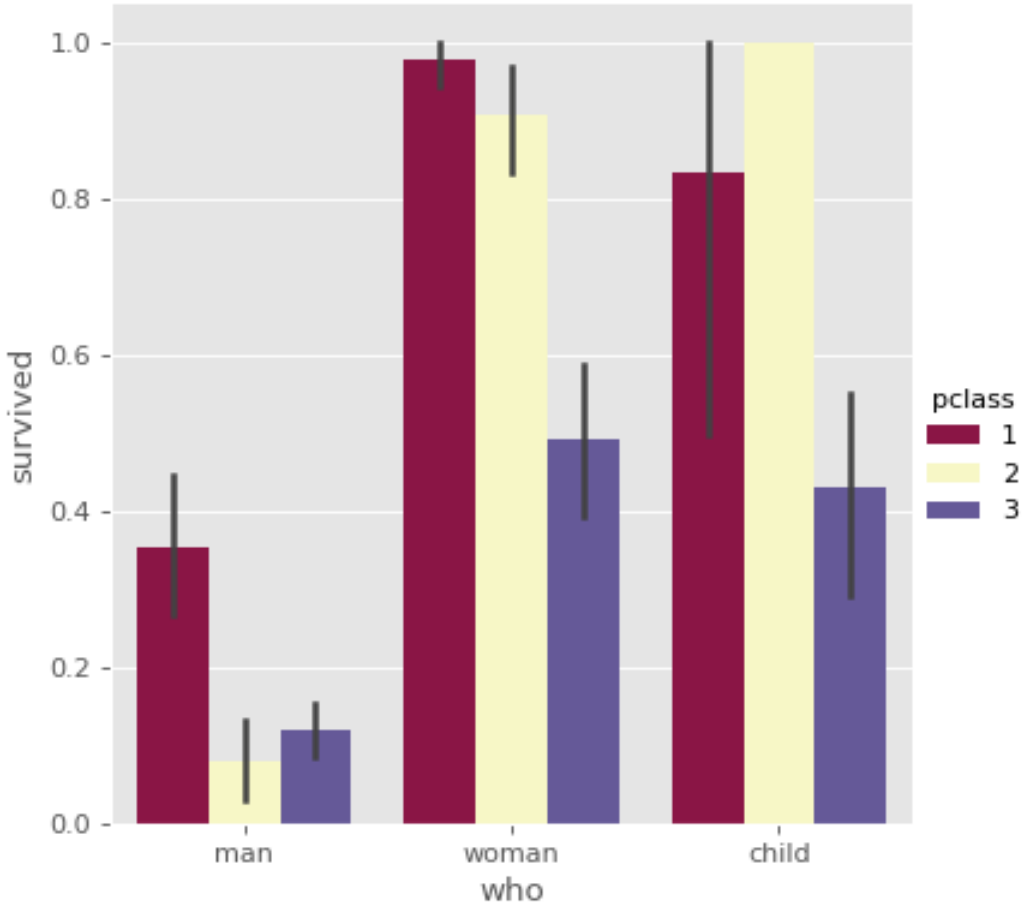
کلاس **Embarked**:

داده‌های این کلاس نمایانگر سه توقف‌گاهی بوده که کشتی در آن مکان‌ها توقف کرده تا مسافرین سوار شوند. سه داده‌ی مشخص این کلاس چربورگ، کوییزتاون و ساوت‌همپتون می‌باشد. ٦٤٤ مسافر در توقف‌گاه ساوت‌همپتون سوار کشتی شده‌اند. ١٦٨ مسافر در توقف‌گاه چربورگ و ٧٧ مسافر در ایستگاه کوییزتاون. اطلاعات دو مسافر در دسترس نیست که از لیست داده‌ها حذف می‌شوند و احتمالا به واسطه اینکه ساوت‌همپتون اولین توقف‌گاه بوده و شروع حرکت قطار در این مکان بوده، پس کارکنان قطار نیز جز همین ٦٤٤ نفر می‌باشند. مهم‌ترین نکته در مورد این کلاس می‌توان این باشد که میزان آمار نجات یافتگان بیشتر مربوط به گروهی از مسافران هست که در توقف‌گاه چربورگ سوار شده‌اند. نکته جالب‌تر این‌که این گروه به نسبت دیگر گروه‌ها دارای بیشترین میزان مسافران در کلاس لوکس بوده و همچنین اکثریت قاطع جمعیت مونث را دارا بوده است.



کلاس Who :

این کلاس مسافران را به سه دسته‌ی مرد، زن و کودک تقسیم کرده است تا بتوان در مرحله تحلیل گروه‌های مسافرین را با دسته‌بندی دقیق‌تری مورد ارزیابی قرار داد و بین مسافران بالغ و نابالغ تفاوت مشخصی قائل شد. طبق داده‌های ما در کشتی تایتانیک ۵۳۷ مسافر مرد، ۲۷۱ مسافر زن و ۸۳ کودک وجود داشته است. میزان آمار بازماندگان به نسبت جان‌باختگان در گروه مردها تقریبا ۱ به ۵ است در گروه زن‌ها ۳ به ۱ است و در کودکان تقریبا ۸ به ۷ است. نکته قابل تامل در این این است که با این‌که نسبت کودکان بازمانده از کودکان جان‌باخته بیشتر است اما این تفاوت بسیار ناچیز است و تنها زمانی که متوجه می‌شویم اکثر کودکان تایتانیک در گروه اقتصادی بوده‌اند می‌توان نتیجه‌گیری کرد که حتی تفاوت طبقاتی گریبان کودکان را نیز گرفته و کودکانی که از بضاعت خوبی برخوردارنبوده به احتمال بسیار بالاتری جان خود را از دست داده‌اند.



مزایای استفاده از تحلیل چندمتغیره:

- شناسایی روابط: ارتباطات و همبستگی‌های خطی یا غیرخطی بین متغیرها را آشکار می‌کند.
- تشخیص تعاملات: تأثیر متغیرها بر یکدیگر و بر متغیر هدف (در صورت وجود) را نشان می‌دهد.
- تحلیل جامع‌تر: برای مسائل پیچیده‌تر و درک بهتر رفتار سیستم بسیار کاربردی است.

کمبودهای استفاده از تحلیل چندمتغیره:

- پیچیدگی بالا: تحلیل چندمتغیره می‌تواند پیچیده باشد و نیازمند دانش آماری و محاسباتی بیشتری است.
- مشکلات در تفسیر: درک نتایج تحلیل‌های چندمتغیره، به ویژه در مدل‌های پیچیده، می‌تواند چالش‌برانگیز باشد.
- حساسیت به داده‌ها: داده‌های پرت یا ناقص می‌توانند به شدت بر نتایج تأثیر بگذارند.

مقایسه با تحلیل تک‌متغیره:

در حالی که تحلیل تک‌متغیره برای خلاصه‌سازی اولیه داده‌ها مناسب است، تحلیل چندمتغیره ارزش بیشتری برای پیش‌بینی‌ها و درک عمیق‌تر از ساختار داده‌ها ارائه می‌دهد و معمولاً در مراحل پیشرفته‌تر EDA انجام می‌شود.

نتیجه‌گیری

نتیجه‌گیری از بررسی داده‌های حادثه تایتانیک با توجه به متغیرهای مختلف مانند بقا (survived)، کلاس (pclass)، جنسیت (sex)، سن (age)، تعداد خواهر و برادر یا همسر (sibsp)، تعداد والدین و فرزندان (parch)، کرایه (fare) و بندر سوار شدن (embarked) به ما بینش‌های ارزشمندی درباره عوامل مؤثر بر بقا در این فاجعه تاریخی ارائه می‌دهد:

۱. تأثیر کلاس اجتماعی: مسافران کلاس اول با شانس بیشتری برای بقا نسبت به مسافران کلاس‌های پایین‌تر مواجه بودند. این نشان‌دهنده اهمیت دسترسی به منابع و امکانات در شرایط بحرانی است.

۲. جنسیت و سن: زنان و کودکان به طور کلی شانس بیشتری برای زنده ماندن داشتند، که تأکید بر سیاست‌های نجات‌گری دارد که به حفاظت از این گروه‌ها اولویت می‌دهد. این نکته نشان می‌دهد که در مواقع اضطراری، رفتارهای اجتماعی و فرهنگی می‌توانند بر نتایج تأثیرگذار باشند.

۳. خانواده و روابط اجتماعی: وجود اعضای خانواده (sibsp و parch) می‌تواند به عنوان یک عامل حمایتی در بقا عمل کند. مسافران با خانواده ممکن است بهتر بتوانند از یکدیگر حمایت کنند و شانس بقا را افزایش دهند.

٤. تأثیر کرایه و محل سوار شدن: متغیر fare نیز نشان‌دهنده وضعیت اقتصادی مسافران است و می‌تواند به نوعی با شانس بقا مرتبط باشد. همچنین، بندر سوار شدن (embarked) می‌تواند به تفاوت‌های فرهنگی و اجتماعی میان مسافران اشاره کند. در نهایت، این تحلیل نشان می‌دهد که بقا در حادثه تایتانیک تحت تأثیر ترکیبی از عوامل اجتماعی، اقتصادی و فرهنگی قرار داشته است. این یافته‌ها نه تنها به ما کمک می‌کند تا درک بهتری از این فاجعه داشته باشیم، بلکه می‌توانند به عنوان درس‌هایی برای مدیریت بحران‌ها و بهبود روش‌های نجات در آینده مورد استفاده قرار گیرند.