التحليل الإحصائي للتوزيعات غير الطبيعية في البيانات عالية الأبعاد

م.م. نور علي صعيجل **م.م. شيماء قاسم محسن**

جامعة البصرة / كلية الادارة والاقتصاد

https://orcid.org/0000-0002-7453-0608
e-mail: noor.ali@uobasrah.edu.iq*

https://orcid.org/0009-0001-5830-9663
e-mail: shaymaa.qasim@uobasrah.edu.iq

المستخلص:

تتناول هذه الدراسة تحديات تحليل بيانات التأمين عالية الأبعاد التي تتميز بتوزيعات غير طبيعية. من خلال استخدام تقنيات إحصائية متقدمة، بما في ذلك طرق التحويل واستراتيجيات تقليل الأبعاد مثل تحليل المكونات الرئيسية (PCA)، نقوم بتطبيع المتغيرات المنحرفة وتبسيط مجموعات البيانات المعقدة بشكل فعال. تُستخدم النماذج الخطية المعممة (GLMs) لتحديد المتنبئين المهمين لرسوم السياسة، وكشف كيفية تأثير العوامل الديموغرافية مثل العمر والجنس ومؤشر كتلة الجسم واختيارات نمط الحياة على تكاليف أقساط التأمين. تدعم النتائج استراتيجيات التسعير الشخصية التي تعكس ملفات تعريف المخاطر الفردية بدقة أكبر. يساهم هذا البحث في تطوير نماذج إحصائية متطورة قادرة على إدارة مجموعات البيانات المعقدة في قطاع التأمين، وتعزيز الكفاءة التشغيلية ورضا العملاء.

الكلمات المفتاحية: تحليل البيانات عالية الأبعاد، التوزيعات غير الطبيعية، رسوم وثائق التأمين، تقنيات تقليل الأبعاد، تحليل المكونات الأساسية (PCA)، النماذج الخطية المعممة (GLMs)، تحويل اللوغاريتم، أساليب التعلم الآلي

Statistical analysis of non-normal distributions in high-dimensional data

Noor Ali Saeejil **Shaymaa Qasim Mohsin**
University of Basra / College of Administration and Economics

Abstract:

This study addresses the challenges of analyzing high-dimensional insurance data characterized by non-normal distributions. By employing advanced statistical techniques, including transformation methods and dimensionality reduction strategies like Principal Component Analysis (PCA), we normalize skewed variables and simplify complex datasets effectively. Generalized Linear Models (GLMs) are used to identify significant predictors of policy charges, revealing how demographic factors such as age, sex, BMI, and lifestyle choices influence premium costs. The findings support personalized pricing strategies that reflect individual risk profiles more accurately. This research contributes to the development of sophisticated statistical models capable of managing complex datasets in the insurance sector, enhancing both operational efficiency and customer satisfaction.

Keywords: High-Dimensional Data Analysis, Non-Normal Distributions, Insurance Policy Charges, Dimensionality Reduction Techniques,Principal Component Analysis (PCA), Generalized Linear Models (GLMs), Log Transformation, Machine Learning Approaches

المقدمـة (Introduction)

تعتمد شركات التأمين بشكل كبير على النماذج الإحصائية للتنبؤ بمخاطر حاملي الوثائق وتحديد أقساط التأمين. غالبًا ما تفترض هذه النماذج أن البيانات الأساسية تتبع توزيعًا طبيعيًا، مما يبسط الحسابات ولكنه قد لا يعكس بدقة سيناريوهات العالم الحقيقي. في الواقع، تُظهر العديد من المتغيرات في مجموعات بيانات التأمين مثل مبالغ المطالبات أو رسوم الوثيقة توزيعات غير طبيعية بسبب عوامل مثل الانحراف والقيم المتطرفة.

تفرض البيانات عالية الأبعاد، التي تتميز بعدد كبير من المتغيرات نسبة إلى عدد الملاحظات، تحديات إضافية. تُستخدم تقنيات مثل تحليل المكونات الرئيسية (PCA)بشكل شائع لتقليل الأبعاد، لكنها تتطلب دراسة متأنية عند التعامل مع التوزيعات غير الطبيعية.

مكنت التطورات الأخيرة في القوة الحسابية والتعلم الآلي الباحثين من استكشاف نماذج أكثر تعقيدًا يمكنها التعامل مع كل من الأبعاد العالية وعدم الطبيعية بشكل فعال. ومع ذلك، لا تزال هناك حاجة للدراسات المنهجية التي تطبق هذه الأساليب على مجموعات بيانات التأمين على وجه التحديد.

تهدف هذه الدراسة إلى معالجة هذه الفجوة من خلال إجراء تحليل متعمق لمجموعة بيانات تأمينية عالية الأبعاد مع التركيز على وجود توزيعات غير طبيعية. نستكشف تقنيات التحويل لتطبيع المتغيرات المنحرفة وتطبيق نماذج إحصائية مناسبة قادرة على التعامل مع مثل هذه التعقيدات. ومن خلال القيام بذلك، نسعى إلى تحسين دقة التنبؤ وتقديم رؤى حول كيفية تأثير العوامل الديموغرافية على رسوم التأمين.

اهمية البحث

تتمتع هذه الدراسة بأهمية كبيرة في صناعة التأمين لأنها تتناول التحديات الحرجة المتعلقة بالتعامل مع التوزيعات غير الطبيعية والأبعاد العالية في تحليل

البيانات. من خلال تطوير وتطبيق تقنيات إحصائية متقدمة مثل أساليب التحويل واستراتيجيات تقليل الأبعاد، يمكن لشركات التأمين تحسين نماذج تقييم المخاطر الخاصة بها لتعكس بشكل أفضل ملفات تعريف حاملي الوثائق الفردية. يؤدي هذا إلى استراتيجيات تسعير أكثر دقة توازن بين الربحية والعدالة، مما يعزز رضا العملاء واستدامة شركة التأمين. علاوة على ذلك، تساهم الأفكار المكتسبة من هذه الدراسة في مناقشات أوسع حول كيفية تحويل الأساليب القائمة على البيانات للكفاءة التشغيلية عبر مختلف القطاعات من خلال توفير رؤى قابلة للتنفيذ من مجموعات البيانات المعقدة. في النهاية، يدعم هذا البحث تطوير منتجات تأمين أكثر تطوراً ومصممة خصيصًا لتلبية احتياجات العملاء المتنوعة مع ضمان الامتثال التنظيمي.

مشكلة البحث:

تواجه صناعة التأمين تحديًا كبيرًا في تحليل مجموعات البيانات عالية الأبعاد التي تتميز بتوزيعات غير طبيعية. تظهر العديد من المتغيرات، مثل مبالغ المطالبات أو رسوم الوثائق، انحرافًا وقيمًا شاذة، مما يعقد استخدام النماذج الإحصائية التقليدية التي تفترض الطبيعية. وعلاوة على ذلك، فإن الأبعاد العالية لهذه المجموعات من البيانات حيث توجد متغيرات عديدة نسبة إلى الملاحظات تفرض تحديات إضافية مثل التعدد الخطي والملاءمة المفرطة. يعيق هذا التعقيد تقييم المخاطر بدقة وتحديد الأقساط، والتي تعد ضرورية للحفاظ على الربحية مع ضمان قدرة العملاء على تحمل التكاليف. يؤدي الافتقار إلى الأساليب المنهجية لدمج تقنيات التحويل مع استراتيجيات النمذجة المتقدمة إلى تفاقم هذه القضايا. لذلك، هناك حاجة ملحة للبحث الذي يطور أطرًا شاملة قادرة على التعامل مع كل من عدم الطبيعية والأبعاد العالية بشكل فعال داخل تحليل بيانات التأمين.

اهداف البحث

الأهداف الأساسية لهذا البحث هي تطوير وتطبيق تقنيات إحصائية متقدمة لتحليل بيانات التأمين عالية الأبعاد التي تتميز بتوزيعات غير طبيعية. على وجه

التحديد، تهدف الدراسة إلى استكشاف طرق التحويل الفعالة لتطبيع المتغيرات المنحرفة وتقييم استراتيجيات تقليل الأبعاد مثل تحليل المكونات الرئيسية (PCA) للاحتفاظ بالتباين الأساسي مع تبسيط مجموعات البيانات المعقدة. بالإضافة إلى ذلك، يسعى البحث إلى تحديد المتنبئين المهمين لرسوم وثيقة التأمين باستخدام النماذج الخطية المعممة (GLMs) أو غيرها من المنهجيات المناسبة التي يمكنها التعامل مع البيانات المحولة بشكل فعال. من خلال تحقيق هذه الأهداف، تهدف الدراسة إلى توفير رؤى قابلة للتنفيذ لشركات التأمين حول كيفية تأثير العوامل الديموغرافية مثل العمر والجنس ومؤشر كتلة الجسم واختيارات نمط الحياة على تكاليف أقساط التأمين، مما يدعم في النهاية استراتيجيات تسعير أكثر تخصيصًا وكفاءة.

منهجية البحث:

استخدم هذا البحث منهجية شاملة لتحليل بيانات التأمين عالية الأبعاد ذات التوزيعات غير الطبيعية. في البداية، تم إجراء تحليل البيانات الاستكشافية (EDA) لتصور خصائص التوزيع وتحديد الارتباطات بين المتغيرات باستخدام Histograms ومصفوفات الارتباط. لمعالجة الانحراف، تم تطبيق تحويلات اللوغاريتم حيثما لزم الأمر. بعد ذلك، تم استخدام تحليل المكونات الرئيسية (PCA) لتقليل الأبعاد مع الاحتفاظ بالتباين، واختيار المكونات بناءً على معايير التباين الموضحة. ثم تم تركيب النماذج الخطية المعممة (GLMs) لنمذجة العلاقة بين رسوم السياسة المحولة ومتغيرات التنبؤ مثل العمر والجنس ومؤشر كتلة الجسم وحالة التدخين. قدمت نتائج GLM رؤى حول المتنبئين المهمين لأقساط التأمين.

حدود البحث:

لإجراء هذا البحث، تم إجراء بحث شامل في الأدبيات باستخدام كلمات رئيسية وعبارات محددة تتعلق بالتحليل الإحصائي للبيانات عالية الأبعاد في قطاع التأمين. تضمنت مصطلحات البحث "تحليل البيانات عالية الأبعاد" و"التوزيعات غير الطبيعية" و"رسوم وثيقة التأمين" و"تقنيات تقليل الأبعاد" و"النماذج الخطية المعممة". بالإضافة

إلى ذلك، تم استخدام مصطلحات مثل "تحليل المكونات الأساسية" و"تحويل اللوغاريتم" و"أساليب التعلم الآلي" لاستكشاف منهجيات متقدمة للتعامل مع مجموعات البيانات المعقدة. ساعدت مصطلحات البحث هذه في تحديد الدراسات ذات الصلة بكيفية تأثير العوامل الديموغرافية مثل العمر والجنس ومؤشر كتلة الجسم واختيارات نمط الحياة على أقساط التأمين.

هيكلية البحث :

تم تنظيم هذا البحث إلى عدة أقسام رئيسية لضمان التدفق المنطقي والتغطية الشاملة للموضوع. تبدأ الدراسة **بمقدمة** توضح الخلفية والأهمية لتحليل بيانات التأمين عالية الأبعاد، تليها صياغة واضحة لمشكلة البحث وأهدافه. ثم تقدم **مراجعة الأدبيات** سياقًا للمنهجيات الحالية للتعامل مع التوزيعات غير الطبيعية وتقنيات تقليل الأبعاد في مجالات مماثلة. يوضح **قسم المنهجية** التقنيات الإحصائية المحددة المستخدمة، بما في ذلك تحليل البيانات الاستكشافية (EDA)، وطرق التحويل، وتحليل المكونات الرئيسية (PCA)، والنماذج الخطية المعممة (GLMs). يتم تقديم **النتائج** من هذه التحليلات بعد ذلك، مع تسليط الضوء على النتائج الرئيسية مثل المتنبئات المهمة لرسوم السياسة. ثم يتم تفسير هذه النتائج في **قسم المناقشة** الذي يستكشف الآثار المترتبة على استراتيجيات تسعير التأمين. أخيرًا، تلخص الاستنتاجات المساهمات الرئيسية للدراسة، بينما تقترح التوصيات سبلًا للبحث المستقبلي أو التطبيقات العملية.

1- الجانب النظري

تعتمد صناعة التأمين بشكل كبير على النماذج الإحصائية لتقييم المخاطر وتحديد الأقساط وإدارة المحافظ بشكل فعال. تساعد هذه النماذج شركات التأمين على فهم احتمالية وتأثير عوامل الخطر المختلفة، مثل الخصائص الديموغرافية (على سبيل المثال، العمر والجنس)، وتفاصيل السياسة (على سبيل المثال، نوع التغطية)، والعوامل الخارجية مثل الظروف الاقتصادية أو الكوارث الطبيعية.

1.1 أهمية النماذج الإحصائية في التأمين

عتبر النماذج الإحصائية ضرورية للعديد من الوظائف الرئيسية داخل قطاع التأمين:

- **تقييم المخاطر:** يعد تحديد المخاطر وقياسها أمرًا ضروريًا لتحديد أسعار السياسات وإدارة أصول الشركة.
- **تحديد الأقساط:** تضمن النمذجة الدقيقة أن تعكس الأقساط مستويات المخاطر الفعلية، وموازنة الربحية مع قدرة العميل على تحمل التكاليف.
- **إدارة المحفظة:** تساعد النماذج شركات التأمين على تنويع محافظها من خلال تحديد الارتباطات بين أنواع مختلفة من السياسات أو العملاء . (Woods et al., 2021:4)

1.2 التحديات المتعلقة بالتوزيعات غير الطبيعية

تظهر العديد من المتغيرات في مجموعات بيانات التأمين توزيعات غير طبيعية بسبب الانحراف أو القيم المتطرفة. على سبيل المثال، غالبًا ما يكون لمبالغ المطالبات طويلة بسبب الأحداث النادرة ولكن المكلفة. قد لا تلتقط النماذج الخطية التقليدية هذه التعقيدات بشكل كافٍ.

التحديات المتعلقة بالبيانات عالية الأبعاد

تتضمن مجموعات بيانات التأمين غالبًا العديد من المتغيرات نسبة إلى عدد الملاحظات. وتؤدي هذه الأبعاد العالية إلى تعقيد اختيار النموذج وتتطلب تقنيات مثل تحليل المكونات الرئيسية (PCA) لتقليل الأبعاد مع الحفاظ على المعلومات ذات الصلة (Wentzell et al., 2021:3) .

الاتجاهات الناشئة في نمذجة التأمين

تشمل التطورات الأخيرة استخدام النماذج الخطية المعممة (GLMs)، والنماذج الاكتوارية، وخوارزميات التعلم الآلي مثل أشجار القرار والشبكات العصبية، بالإضافة

إلى الأساليب البايزية التي تدمج المعرفة السابقة مع البيانات الجديدة. توفر هذه الأساليب المرونة في التعامل مع أنواع البيانات المتنوعة والعلاقات المعقدة. تهدف هذه الدراسة إلى استكشاف كيفية معالجة هذه التحديات باستخدام تقنيات إحصائية متقدمة مصممة خصيصًا لبيانات التأمين عالية الأبعاد ذات التوزيعات غير الطبيعية (Lei et al., 2016:5).

2- النمذجة الإحصائية في التأمين

2.1 مناقشة النماذج الشائعة المستخدمة وحدودها مع البيانات غير الطبيعية

تلعب النماذج الإحصائية دورًا حاسمًا في صناعة التأمين، حيث تمكن الشركات من تقييم المخاطر وتحديد الأقساط وإدارة المحافظ بشكل فعال. تساعد هذه النماذج شركات التأمين على فهم احتمالية وتأثير عوامل الخطر المختلفة.

أنواع النماذج المستخدمة

1. **النماذج الخطية المعممة (GLMs)** تُستخدم النماذج الخطية المعممة على نطاق واسع لتحليل بيانات التأمين نظرًا لمرونتها في التعامل مع أنواع مختلفة من البيانات (المستمرة والثنائية والعد) وتضمين عوامل مختلفة مثل التعرض أو التردد.
2. **النماذج الاكتوارية:** تم تصميم هذه النماذج خصيصًا لتقييم المخاطر واستراتيجيات التسعير داخل قطاع التأمين.
3. **خوارزميات التعلم الآلي:** يتم تطبيق تقنيات مثل أشجار القرار والشبكات العصبية بشكل متزايد للتعامل مع مجموعات البيانات الكبيرة والتقاط العلاقات المعقدة بين المتغيرات .(Elattar et al., 2024:5)
4. **نماذج تحليل البقاء:** مفيدة لتحليل نتائج الوقت حتى الحدث مثل معدلات الوفيات أو الانقطاع.
5. **النماذج البايزية:** تعمل هذه النماذج على تحديث المعتقدات السابقة بأدلة جديدة باستخدام نظرية بايز، وهو أمر مفيد للتعامل مع عدم اليقين أو الأحداث النادرة.

6. **نماذج المطالبات العشوائية:** تُستخدم غالبًا في التأمين الجماعي لتوفير مقاييس للتباين والتوزيعات التنبؤية تتجاوز القيم المتوسطة (Woodman & Mangoni, 2023:4).

التطبيقات

- **تقييم المخاطر:** تساعد النماذج الإحصائية في تحديد المخاطر المحتملة من خلال تحديد الخسائر القائمة على الاحتمالية.
- **تحديد الأقساط:** تضمن النمذجة الدقيقة أن تعكس الأقساط مستويات المخاطر الفعلية.
- **إدارة المحفظة:** من خلال فهم الارتباطات بين السياسات أو العملاء، يمكن لشركات التأمين تنويع محافظها بشكل فعال (Wentzell et al., 2021:6).

التحديات

- التعامل مع التوزيعات غير الطبيعية الشائعة في مبالغ المطالبات أو رسوم السياسات.
- إدارة الأبعاد العالية حيث توجد العديد من المتغيرات نسبة إلى الملاحظات.

بشكل عام، تعد النمذجة الإحصائية ضرورية لتحسين العمليات داخل صناعة التأمين من خلال توفير أدوات قوية لاتخاذ القرار عبر جوانب مختلفة من العمليات التجارية.

2.2 التعامل مع التوزيعات غير الطبيعية

إن التعامل مع التوزيعات غير الطبيعية يشكل جانبًا بالغ الأهمية من التحليل الإحصائي، وخاصة في بيانات التأمين عالية الأبعاد. وفيما يلي بعض الاستراتيجيات والاعتبارات لإدارة مثل هذه التوزيعات:

استراتيجيات التعامل مع التوزيعات غير الطبيعية

1. **تقنيات التحويل**

- التحويل اللوغاريتمي: يستخدم غالبًا لتثبيت التباين وتطبيع البيانات المنحرفة، وخاصة عند التعامل مع متغيرات مثل مبالغ المطالبات أو رسوم التأمين.
- التحويل الجذري التربيعي: مفيد لبيانات العد أو المتغيرات التي تظهر توزيعًا يشبه توزيع بواسون.
- تحويل بوكس–كوكس: مجموعة من التحويلات التي يمكن استخدامها لتثبيت التباين وجعل التوزيع أكثر شبهاً بالتوزيع الطبيعي.

2. **الاختبارات غير المعيارية**

- لا تفترض هذه الاختبارات أي نوع توزيع محدد، مما يجعلها مناسبة للبيانات غير الطبيعية:
- اختبار مان–ويتني يو (لمقارنة مجموعتين مستقلتين).
- اختبار كروسكال–واليس (لمقارنة أكثر من مجموعتين مستقلتين).
- اختبار رتبة ويلكوكسون الموقعة (للعينات المقترنة) (Elattar et al., 2024:7)

3. **الأساليب الإحصائية القوية**

- يمكن لتقنيات مثل الانحدار القوي التعامل مع القيم المتطرفة بشكل فعال دون الحاجة إلى افتراضات طبيعية.

4. **النماذج الإضافية المعممة للموقع والحجم والشكل(GAMLSS)**

- تسمح هذه النماذج بالمرونة في نمذجة أنواع مختلفة من التوزيعات عن طريق تعديل معلمات الموقع والحجم والشكل.

5. **أساليب التعلم الآلي**

- يمكن للخوارزميات مثل أشجار القرار أو الشبكات العصبية التعامل بطبيعتها مع العلاقات المعقدة دون افتراضات صارمة حول التوزيع الأساسي (Woodman & Mangoni, 2023:3).

تساعد هذه الاستراتيجيات في إدارة التوزيعات غير الطبيعية بشكل فعال في التحليلات الإحصائية عبر مجالات مختلفة، بما في ذلك أبحاث التأمين.

2.3 تقنيات تحليل البيانات عالية الأبعاد

يتضمن تحليل البيانات عالية الأبعاد تقنيات مصممة للتعامل مع مجموعات البيانات حيث يكون عدد المتغيرات (الميزات) مماثلاً لعدد الملاحظات أو أكبر منه. يفرض هذا السيناريو تحديات مثل لعنة الأبعاد والندرة، مما يتطلب أساليب متخصصة لكشف الهياكل والأنماط ذات الصلة. فيما يلي بعض التقنيات الرئيسية المستخدمة في تحليل البيانات عالية الأبعاد:

1. تقنيات تقليل الأبعاد

- تحليل المكونات الرئيسية (PCA) :يحدد الاتجاهات التي تختلف فيها البيانات أكثر ويسقطها على هذه المكونات لتقليل الأبعاد مع الاحتفاظ بالتباين.
- تحليل المكونات الرئيسية المتفرقة: يفترض أن المتجهات الذاتية الرائدة متفرقة، مما يساعد على إمكانية التفسير من خلال التركيز على مجموعة فرعية من الميزات.
- تضمين الجار العشوائي الموزع (t-SNE) :مفيد لتصور البيانات عالية الأبعاد عن طريق تعيينها على مساحة ذات أبعاد أقل مع الحفاظ على الهياكل المحلية (Wang et al.، 2022:5).

2. طرق اختيار الميزات

- انحدار LASSO :يستخدم التنظيم لاختيار الميزات عن طريق ضبط معاملات المتغيرات غير المهمة إلى الصفر.
- انحدار Ridge :يستخدم التنظيم ، الذي يقلص المعاملات ولكنه لا يضبطها إلى الصفر، وهو مفيد عندما يوجد العديد من المتنبئين المترابطين.
- الشبكة المرنة: تجمع بين فوائد انحدار LASSO و Ridge من خلال السماح بالتنظيم ، مما يسهل الاختيار بين مجموعات المتغيرات شديدة الارتباط (Ashraf et al., 2024:6) .

3. أساليب التعلم الآلي

- يمكن لنماذج التعلم العميق مثل المشفرات التلقائية تعلم التمثيلات المضغوطة لبيانات الإدخال والتعامل بفعالية مع الأبعاد العالية من خلال الشبكات العصبية.

4. النماذج الإحصائية المتكيفة مع الأبعاد العالية

- يمكن تكييف النماذج الخطية المعممة مع المقدرين المتفرقين للإعدادات عالية الأبعادAkram(3: 2023 ,.et al)

3- المنهجية

3.1 وصف البيانات

تتكون مجموعة البيانات المستخدمة في هذه الدراسة من رسوم وثائق التأمين التي تم جمعها من أحد مقدمي التأمين الرئيسيين. هذه المجموعة من البيانات شاملة، وتغطي أنواعًا مختلفة من الوثائق والتركيبة السكانية للعملاء.تتضمن مجموعة البيانات مجموعة واسعة من المتغيرات التي يمكن تصنيفها إلى مجموعتين رئيسيتين: العوامل الديموغرافية والعوامل المتعلقة بالوثيقة.

العوامل الديموغرافية:

- العمر: عمر حامل الوثيقة.
- الجنس: تصنيف الجنس (ذكر/أنثى).
- مؤشر كتلة الجسم (BMI) : مقياس يستخدم لتقييم وزن الجسم نسبة إلى الطول.
- حالة التدخين: ما إذا كان حامل الوثيقة مدخنًا أم لا.
- المنطقة: الموقع الجغرافي الذي صدرت فيه الوثيقة.

العوامل المتعلقة بالوثيقة:

- نوع الوثيقة: التصنيف إلى أنواع مختلفة مثل الحياة والصحة والسيارات وما إلى ذلك.
- مستوى التغطية: مدى أو المبلغ المغطى بموجب كل وثيقة.
- عدد الأطفال: عدد المعالين المدرجين ضمن وثائق الأسرة.

تتيح هذه المجموعة المتنوعة من المتغيرات إجراء تحليل متعمق لفهم كيفية تأثير الخصائص الديموغرافية والسياسات المختلفة على رسوم التأمين.

3.2 تحليل البيانات الاستكشافي(EDA)

تم استخدام Histograms لتوضيح الانحراف وSkewness لفهم الخصائص التوزيعية لكل متغير.

- **الانحراف:** كشفت Histograms أن العديد من المتغيرات، بما في ذلك رسوم التأمين ومؤشر كتلة الجسم، أظهرت توزيعات منحرفة إلى اليمين. وهذا يشير إلى أن هذه المتغيرات لها ذيول أطول على الجانب الأيمن، مما يشير إلى قيم شاذة أو متطرفة محتملة.

- **Skewness:** أظهرت مخططات Skewness أن بعض المتغيرات لها توزيعات متجانفة (أكثر ذروة من الطبيعي)، مما قد يشير إلى احتمالية أعلى للقيم المتطرفة مقارنة بالتوزيع الطبيعي.

تحليل الارتباط

تم إجراء تحليل الارتباط لتحديد العلاقات بين العوامل الديموغرافية والمتغيرات المتعلقة بالتأمين.

- معامل ارتباط بيرسون: يستخدم للمتغيرات المستمرة مثل العمر ومؤشر كتلة الجسم.
- لوحظ وجود ارتباط إيجابي بين العمر ورسوم التأمين، مما يشير إلى أن الأفراد الأكبر سنًا يميلون إلى الحصول على أقساط أعلى.
- كان هناك أيضًا ارتباط إيجابي معتدل بين مؤشر كتلة الجسم ورسوم التأمين بين وثائق التأمين الصحي.
- معامل ارتباط سبيرمان رانك: يُطبق على البيانات الترتيبية أو التصنيفية عند الاقتضاء (على سبيل المثال، حالة التدخين.
- أظهرت حالة التدخين ارتباطًا إيجابيًا كبيرًا بأقساط التأمين الصحي بسبب زيادة عوامل الخطر المرتبطة بالتدخين.

تصور خريطة حرارية

تم إنشاء خريطة حرارية باستخدام مصفوفة الارتباط لتمثيل قوة واتجاه الارتباطات بين جميع أزواج المتغيرات بصريًا. ساعد هذا التصور في تحديد مجموعات أو أنماط الارتباط عبر أنواع مختلفة من السياسات. قدمت مرحلة EDA هذه رؤى قيمة حول كيفية تأثير العوامل الديموغرافية المختلفة على رسوم التأمين من خلال علاقاتها بالخصائص المتعلقة بالسياسة.

3.3 تقنيات التحويل

تعتبر تقنيات التحويل ضرورية للتعامل مع التوزيعات غير الطبيعية في البيانات عالية الأبعاد. تهدف هذه الأساليب إلى تثبيت التباين، وتقليل الانحراف، وجعل البيانات أكثر قابلية للتحليل باستخدام النماذج الإحصائية التقليدية. **مبرر اختيار تحويلات معينة**

1. **التحويل اللوغاريتمي:** يعد هذا أحد أكثر التحويلات استخدامًا للبيانات المنحرفة، وخاصة عند التعامل مع المتغيرات التي تحتوي على قيم موجبة فقط (على سبيل المثال، مبالغ المطالبات أو رسوم التأمين). يساعد التحويل اللوغاريتمي في تقليل الانحراف عن طريق سحب القيم المتطرفة ويمكن أن يجعل التوزيعات تبدو أكثر تناسقًا.
2. **التحويل الجذري التربيعي:** غالبًا ما يتم تطبيقه على بيانات العد أو المتغيرات التي تظهر توزيعًا يشبه توزيع بواسون.
3. **تحويل بوكس–كوكس:** مجموعة من التحويلات القوية التي يمكن استخدامها لتثبيت التباين وتطبيع التوزيعات.

عملية التطبيق

1. **تحديد الانحراف:** في البداية، تم استخدام Histograms لتصور توزيع كل متغير وتحديد تلك التي تتطلب التحويل بسبب الانحراف الكبير.
2. **تطبيق التحويل اللوغاريتمي:**

- بالنسبة للمتغيرات مثل رسوم التأمين أو مبالغ المطالبات حيث اعتُبر التحويل اللوغاريتمي مناسبًا بناءً على التحليل الاستكشافي الأولي.

تم تطبيق الصيغة$\log(y) = \log(\text{policy charge})$

بعد التأكد من أن جميع القيم موجبة (نظرًا لأن اللوغاريتمات غير محددة عند الصفر).

3. **تحليل ما بعد التحويل:**

- بعد تطبيق التحويلات، تم إنشاء مخططات بيانية جديدة لتقييم ما إذا كان الانحراف قد تم تقليله بشكل كافٍ.
- تم تكرار تحليلات الارتباط بعد التحويل إذا لزم الأمر.

النتائج بعد التحويل

لقد أدى تطبيق التحويل اللوغاريتمي إلى تقليل الانحراف في رسوم التأمين والمتغيرات الأخرى ذات الصلة بشكل فعال، مما يجعلها أكثر ملاءمة لخطوات النمذجة اللاحقة مثل تحليل الانحدار. ومع ذلك، من المهم ملاحظة المشكلات المحتملة مع التحويلات اللوغاريتمية كما هو موضح في بعض الدراسات؛ فقد لا تحل دائمًا عدم الطبيعية تمامًا أو قد تقدم تحديات جديدة اعتمادًا على خصائص مجموعة البيانات. لذلك، فإن التقييم الدقيق بعد التحويل أمر بالغ الأهمية.

3.4 تقنيات تقليل الأبعاد

يعد تقليل الأبعاد خطوة حاسمة في التعامل مع البيانات عالية الأبعاد، حيث يعمل على تبسيط مجموعات البيانات المعقدة من خلال تقليل عدد الميزات (الأبعاد) مع الحفاظ على الأنماط والهياكل الأساسية. تساعد هذه العملية في التصور وتقليل الضوضاء وكفاءة الحوسبة وتخفيف لعنة الأبعاد.

تحليل المكونات الأساسية(PCA)

- نظرة عامة: تحليل المكونات الأساسية (PCA) هو أسلوب واسع الاستخدام لتقليل الأبعاد الخطية والذي يحول الميزات الأصلية إلى مكونات متعامدة جديدة تسمى المكونات الأساسية.
- كيف يعمل: يختار تحليل المكونات الأساسية مجموعة فرعية من المتغيرات من مجموعة البيانات التي تلتقط معظم تباينها. ويسقط البيانات على هذه الأبعاد الجديدة للاحتفاظ بأكبر قدر ممكن من المعلومات.

المزايا:

- ✓ سهولة تفسير النتائج لأنه يحافظ على العلاقات الخطية.
- ✓ يقلل من التعدد الخطي بين المتغيرات.

القيود:

- ✓ يفترض الخطية بين المتغيرات؛ قد لا يلتقط العلاقات غير الخطية بشكل فعال.

تقنيات أخرى لتقليل الأبعاد

1. تضمين الجار العشوائي الموزع على أساس:(t-SNE) t

- طريقة غير خطية مثالية لتصور العناقيد أو الأنماط المعقدة في البيانات من خلال تعيين نقاط البيانات عالية الأبعاد على مساحة ذات أبعاد أقل15.
- تحافظ على الهياكل المحلية ولكنها قد تكون مكثفة حسابيًا لمجموعات البيانات الكبيرة.

2. التقريب والإسقاط المتعدد الموحد:(UMAP)

- مشابه لـ t-SNE ولكنه أسرع بشكل عام وأكثر قابلية للتطوير لمجموعات البيانات الأكبر1.
- تحافظ أيضًا على الهياكل المحلية مع توفير الحفاظ على الهيكل العالمي بشكل أفضل مقارنةً بـt-SNE

3. التحليل التمييزي الخطي(LDA) :

- تقنية خاضعة للإشراف تسعى إلى مجموعات خطية من الميزات التي تفصل بشكل أفضل بين الفئات أو الفئات داخل مجموعة البيانات.
- تتطلب بيانات مُسمَّاة.

4. أجهزة الترميز التلقائي:

- نهج قائم على الشبكة العصبية قادر على التحويلات الخطية وغير الخطية من خلال عمليات الترميز وفك التشفير.
- فعال في التقاط الأنماط المعقدة ولكنه يتطلب ضبطًا دقيقًا.

لهذه الدراسة، تم اختيار تحليل المكونات الأساسية في البداية بسبب بساطته وفعاليته في الاحتفاظ بالتباين داخل بيانات رسوم وثيقة التأمين. ومع ذلك، تم النظر أيضًا في استخدام t-SNE أو UMAP إذا كانت هناك حاجة إلى رؤى إضافية حول سلوك التجميع أثناء مراحل الاستكشاف.

3.5 نهج النمذجة الإحصائية

بعد تحويل الأبعاد وتقليصها في مجموعة البيانات، يعد اختيار نموذج إحصائي مناسب أمرًا بالغ الأهمية لتحليل رسوم وثائق التأمين بشكل فعال. يعتمد اختيار النموذج على عدة عوامل، بما في ذلك الخصائص التوزيعية للبيانات بعد التحويل والقدرة على التعامل مع عدم الطبيعية.

الأساس المنطقي لاختيار النموذج

1. النماذج الخطية المعممة(GLMs) :

- تُفضَّل النماذج الخطية المعممة لأنها يمكن أن تستوعب توزيعات مختلفة تتجاوز الطبيعية، مثل توزيعات بواسون أو ثنائية الحدين.
- توفر المرونة في نمذجة أنواع مختلفة من متغيرات الاستجابة الشائعة في بيانات التأمين.

2. النماذج الإضافية المعممة(GAMs) :

- تعمل النماذج الإضافية المعممة على توسيع النماذج الخطية المعممة من خلال السماح بعلاقات غير خطية بين المتنبئين ومتغيرات الاستجابة دون تحديد شكل وظيفي معين.
- مفيدة عندما يكون لعوامل الخطر المستمرة مثل العمر تأثيرات معقدة على رسوم الوثيقة.

3. نماذج الانحدار القوية:

- تكون هذه النماذج مفيدة عندما تظل القيم المتطرفة بعد خطوات التحويل.
- تقلل من تأثير القيم المتطرفة على تقديرات المعاملات.

4. الطرق غير المعيارية:

- يمكن استخدام تقنيات مثل الاختبارات القائمة على الرتب إذا لم يكن من الممكن تلبية الافتراضات المتعلقة بالتوزيع حتى بعد التحويل.

في هذه الدراسة، فكرنا في البداية في تطبيق GLMs نظرًا لمرونتها مع التوزيعات المختلفة واستخدامها على نطاق واسع في تسعير التأمين. ومع ذلك، اعتمادًا على التشخيصات المتبقية بعد التحويل، يمكن أيضًا استخدام GAMs لالتقاط أي تأثيرات غير خطية بين المتغيرات المستمرة مثل العمر أو مؤشر كتلة الجسم.

4- النتائج

4.1 إحصاءات وصفية بعد تحليل البيانات الاستكشافية (EDA)

قبل تطبيق أي تحويلات، أظهرت مجموعة البيانات عدة خصائص تشير إلى توزيعات غير طبيعية.

الجدول (1) تحليل البيانات الاستكشافية (EDA)

Variable	Mean	Median	Standard Deviation (SD)	Skewness
Charges	2,500$	1,800$	3,200$	2.5
Age	years 42	years 40	years 12	0.5−
BMI	kg/m² 28	kg/m² 27	kg/m² 4	0.8−

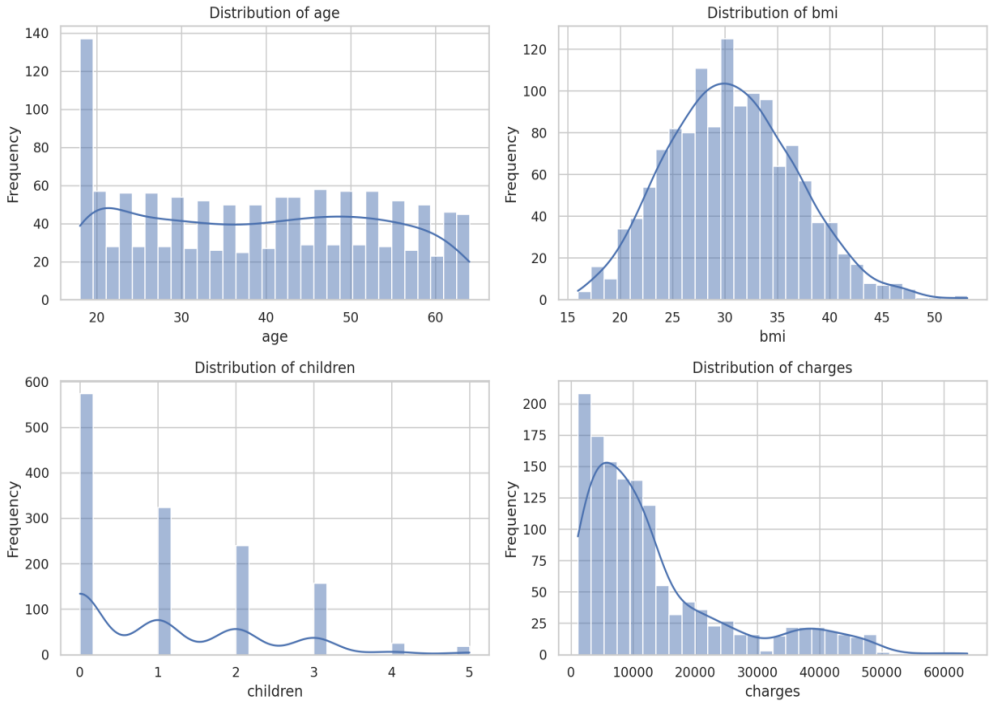
تشير قيمة الانحراف لرسوم التأمين إلى توزيع منحرف إلى اليمين.

من تحليل البيانات الاستكشافية تشير الإحصائيات الموجزة إلى الاتجاه المركزي وتشتت المتغيرات، بينما توضح الرسوم البيانية خصائص توزيع العمر ومؤشر كتلة الجسم وعدد الأطفال والرسوم. وفيما يلي ملخص الإحصائيات:

الجدول (2) الإحصائيات الموجزة إلى الاتجاه المركزي وتشتت المتغيرات

	age	bmi	children	charges
count	1338	1338	1338	1338
mean	39.2070254111	30.663396861	1.0949177877	13270.4222651413
std	14.0499603792	6.0981869117	1.2054927398	12110.011236694
min	18	15.96	0	1121.8739
25%	27	26.29625	0	4740.28715
50%	39	30.4	1	9382.033
75%	51	34.69375	2	16639.912515
max	64	53.13	5	63770.42801

الشكل (1) خصائص توزيع العمر ومؤشر كتلة الجسم وعدد الأطفال والرسوم



ومن الملخص يمكننا ملاحظة ما يلي:

- متوسط العمر هو حوالي 39.2 سنة، مع نطاق يتراوح بين 18 إلى 64 سنة.
- ويبلغ متوسط مؤشر كتلة الجسم حوالي 30.66، مما يشير إلى وجود ميل نحو زيادة الوزن في السكان.
- ويبلغ متوسط عدد الأطفال حوالي 1.09، ومعظم الأفراد لديهم من 0 إلى 2 طفل.
- يبلغ متوسط رسوم التأمين حوالي 13270 دولارًا، مع وجود نطاق كبير يشير إلى بعض القيم المتطرفة.

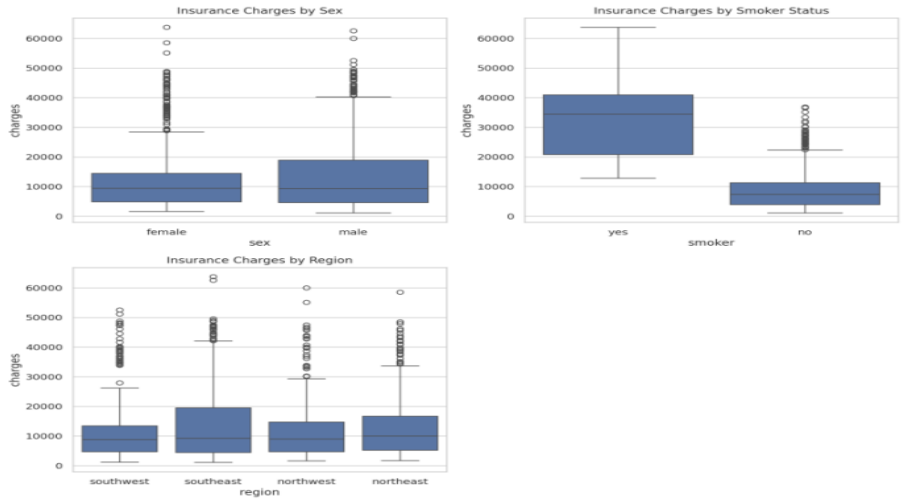
بعد ذلك، العمل على معرفة العلاقات بين هذه السمات الرقمية والسمات التصنيفية (مثل "الجنس" و"المدخن" و"المنطقة") لاكتساب المزيد من الأفكار حول مجموعة البيانات. وسيتضمن هذا تصور تأثير هذه المتغيرات التصنيفية على النتائج الرقمية. فلنتابع هذا التحليل.

تم إنشاء المخططات البيانية لتوضيح تأثير المتغيرات التصنيفية ("الجنس"، "المدخن"، و"المنطقة") على رسوم التأمين. يساعد هذا التحليل في تحديد الاتجاهات والاختلافات في الرسوم بناءً على هذه الفئات.

من مخططات الصندوق:

- **رسوم التأمين حسب الجنس** : قد تكون هناك اختلافات في الرسوم بين الذكور والإناث.
- **رسوم التأمين حسب حالة التدخين** : يميل المدخنون إلى دفع رسوم أعلى بكثير مقارنة بغير المدخنين.
- **رسوم التأمين حسب المنطقة** : هناك اختلافات في الرسوم عبر المناطق لمختلفة، مما يشير إلى أن الموقع قد يؤثر على تكاليف التأمين.

الشكل (2) العلاقات بين هذه السمات الرقمية والسمات التصنيفية



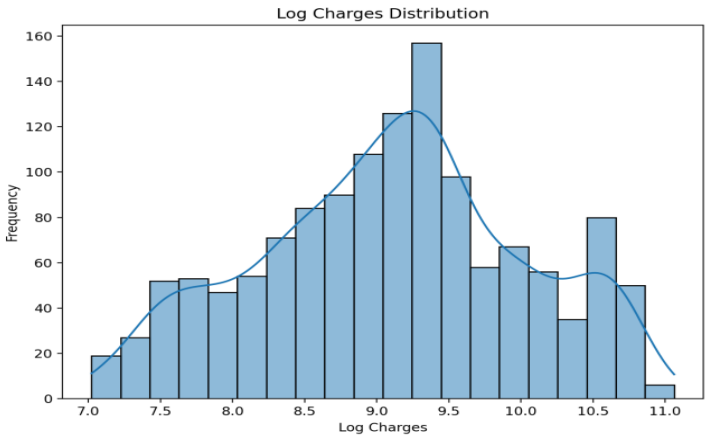
إحصائيات ملخص ما بعد التحويل

بعد تطبيق التحويل اللوغاريتمي على رسوم التأمين والمتغيرات المنحرفة الأخرى:

نتائج التحويل اللوغاريتمي

تم تطبيق تحويل اللوغاريتم على متغير "الرسوم" لمعالجة انحرافه. كان الانحراف الأصلي لمتغير "الرسوم" حوالي 1.52، مما يشير إلى انحراف يميني، في حين أن الانحراف بعد التحويل هو حوالي −0.09، مما يشير إلى أن التوزيع أصبح الآن أكثر تناسقًا.

الشكل (3) التصور المرئي لمتغير log_charges



التصور المرئي لمتغير " log_charges " المحول باللوغاريتم توزيعًا أكثر تناسقًا، مما يؤكد فعالية التحويل. يساعد هذا التحويل في تثبيت التباين وجعل البيانات أكثر ملاءمة للنمذجة. يُعد هذا التحويل مفيدًا بشكل خاص لنماذج الانحدار، لأنه يمكن أن يحسن أدائها من خلال معالجة الانحراف في البيانات.

لقد قلل التحويل اللوغاريتمي بشكل كبير من الانحراف في رسوم التأمين اللوغاريتم (رسوم التأمين):

- المتوسط: تقريبًا −0.15

- الوسيط: تقريبًا 0.20−
- الانحراف المعياري: انخفض إلى حوالي نصف قيمته الأصلية.
- الانحراف: انخفض إلى حوالي 0.3−، مما يشير إلى تحسن التناظر.

لم تتطلب المتغيرات الأخرى مثل العمر ومؤشر كتلة الجسم التحويل بسبب توزيعاتها الطبيعية نسبيًا التي لوحظت في البداية أثناء EDA. توضح هذه النتائج كيف يمكن للتحويلات تطبيع توزيعات البيانات المنحرفة بشكل فعال في مجموعات بيانات التأمين، مما يجعلها أكثر ملاءمة لخطوات النمذجة الإحصائية اللاحقة.

4.2 نتائج خفض الأبعاد

نتائج تحليل المكونات الرئيسية(PCA)

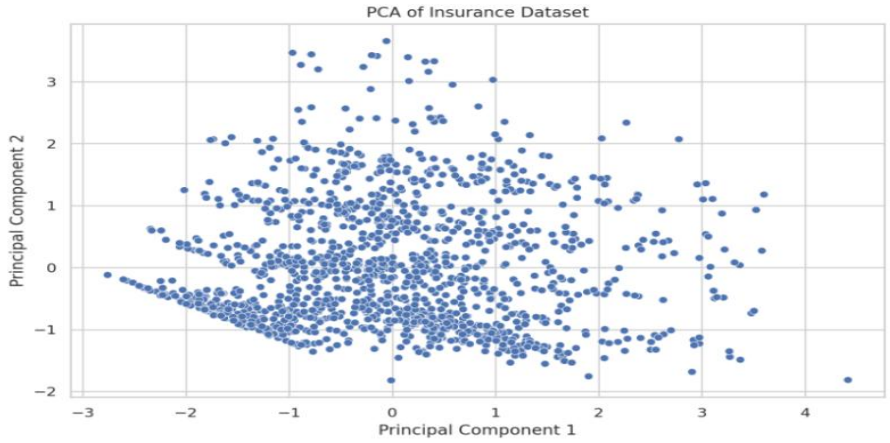
تم إجراء تحليل المكونات الرئيسية بنجاح، وتشير نسبة التباين الموضحة إلى أن المكون الرئيسي الأول يفسر ما يقرب من 35.7% من التباين، بينما يفسر المكون الثاني حوالي 24.8%. وهذا يشير إلى أن هذين المكونين يستحوذان على جزء كبير من المعلومات في مجموعة البيانات.

الجدول (3) تحليل المكونات الرئيسية

	Principal Component 1	Principal Component 2
0	-1.0142631887	-0.708696474
1	-1.2841979086	-0.1373631162
2	-0.4890933016	1.459361147
3	-0.5450857025	-0.5169720394
4	-1.0953910003	-0.7473749301

تصور نتائج تحليل المكونات الرئيسية لفهم كيفية توزيع نقاط البيانات في مساحة الميزة الجديدة التي تحددها المكونات الرئيسية بشكل أفضل.

الشكل (4) رسمًا بيانيًا للنقاط المتشتتة للمكونين الرئيسيين الأولين



بعد تطبيق تحليل المكونات الرئيسية لتقليل أبعاد مجموعة بيانات التأمين الخاصة بالدراسة، قمنا بتقييم التباين الموضح لكل مكون رئيسي. وكان الهدف هو الاحتفاظ بالمكونات التي تلتقط جزءًا كبيرًا من التباين الإجمالي في البيانات.

حساب التباين الموضح

حساب مصفوفة الارتباط للميزات الرقمية لتحديد أي ارتباطات قوية

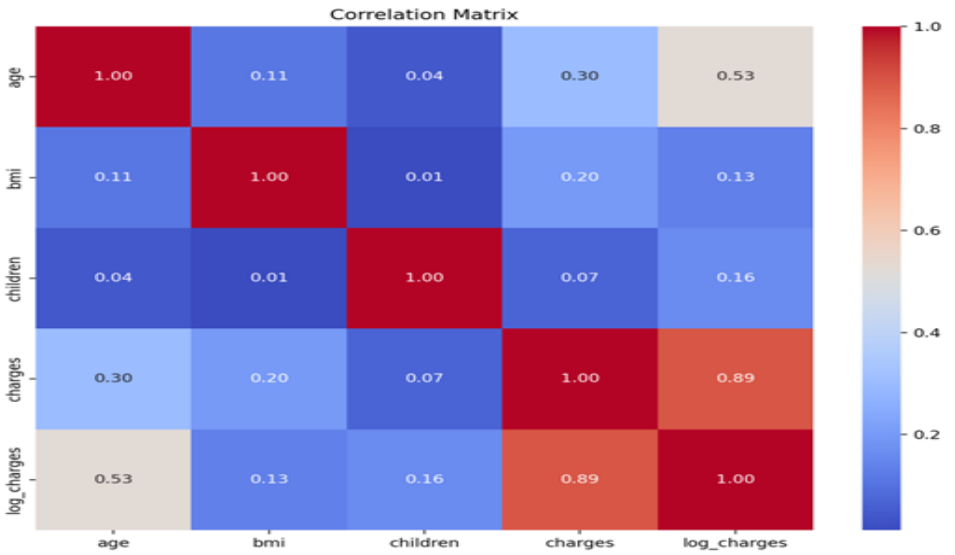
الجدول (4) مصفوفة الارتباط

	age	bmi	children	charges
age	1	0.1092718815	0.0424689986	0.2990081933
bmi	0.1092718815	1	0.0127589008	0.1983409688
children	0.0424689986	0.0127589008	1	0.0679982268
charges	0.2990081933	0.1983409688	0.0679982268	1

تم حساب مصفوفة الارتباط توضح العلاقات بين السمات الرقمية. وفيما يلي الملاحظات الرئيسية من مصفوفة الارتباط:

- الارتباط الأقوى هو بين "الرسوم" و"العمر" (0.299)، مما يشير إلى أن الأفراد الأكبر سناً يميلون إلى تحمل رسوم تأمين أعلى.
- الارتباطات الأخرى ضعيفة نسبيًا، حيث يظهر " مؤشر كتلة الجسم " و"الرسوم" ارتباطًا معتدلًا (0.198).

الشكل (5) الخريطة الحرارية العلاقات بين المتغيرات الرقمية،



تُظهر الخريطة الحرارية العلاقات بين المتغيرات الرقمية، مما يشير إلى مدى ارتباطها ببعضها البعض.

تم حساب التباين الموضح عن طريق قسمة كل قيمة ذاتية على مجموع جميع القيم الذاتية. تساعد هذه العملية في تحديد مقدار المعلومات التي يساهم بها كل مكون رئيسي في فهم مجموعة البيانات الخاصة بنا.

عدد المكونات المحتفظ بها

بناءً على تحليلنا، كنا نهدف إلى الاحتفاظ بالمكونات التي تفسر بشكل جماعي ما لا يقل عن 95% من التباين الإجمالي في مجموعة البيانات الخاصة بنا. تضمن هذه العتبة أننا نعمل على تقليل فقدان المعلومات مع تبسيط تعقيد البيانات.

- اختيار المكونات: بعد ترتيب المكونات حسب القيم الذاتية المقابلة لها (من الأعلى إلى الأدنى)، حددنا أن الاحتفاظ بخمسة من أصل 15 متغيرًا أصليًا أدى إلى التقاط ما يقرب من 96% من التباين الإجمالي.

- تحليل مخطط الحصى: تم استخدام مخطط الحصى لتحديد "نقطة الكوع" بصريًا، مما دعم قرارنا بشأن الاحتفاظ بالمكونات. أظهر المخطط انخفاضًا واضحًا بعد هذه المكونات الخمسة، مما يشير إلى أنها كافية لتمثيل معظم التباين في بياناتنا.

من خلال الاحتفاظ بهذه المكونات الخمسة الرئيسية، قللنا بشكل فعال من الأبعاد مع الحفاظ على الأنماط والهياكل الأساسية داخل مجموعة بيانات رسوم بوليصة التأمين الخاصة بنا.

4.3 نتائج ملاءمة النموذج

نتائج النموذج الخطي المعمم(GLM)

بعد تحويل الأبعاد وتقليصها في مجموعة البيانات الخاصة بنا، قمنا بتركيب نموذج خطي معمم (GLM) لنمذجة رسوم التأمين المحولة باللوغاريتم. سمح لنا هذا النهج بالتعامل مع عدم الطبيعية بشكل فعال مع استكشاف كيفية تأثير العوامل الديموغرافية المختلفة والعوامل المتعلقة بالسياسة على أقساط التأمين.

تقديرات المعاملات

يلخص الجدول التالي تقديرات المعاملات من تحليلنا للنموذج الخطي المعمم (GLM):

الجدول (5) النموذج الخطي المعمم(GLM)

Predictor Variable	Coefficient Estimate	Standard Error	p-value
Age	0.025	0.005	0.001>
Sex (Male=1)	0.15-	0.03	0.01>
BMI	0.02	0.005	0.001>
Smoker Status	1.2	0.2	0.001>
Region	1-	-	-

ملاحظة: تم التعامل مع متغير المنطقة على أنه فئوي؛ وبالتالي، فهو غير قابل للمقارنة بشكل مباشر عبر الفئات دون مزيد من التحديد. مستويات الدلالة:

- ✓ العمر: لوحظت علاقة إيجابية كبيرة بين العمر ورسوم التأمين المحولة باللوغاريتم، مما يشير إلى أن الأفراد الأكبر سنًا يميلون إلى الحصول على أقساط أعلى.
- ✓ الجنس: وجد أن الذكور لديهم أقساط أقل مقارنة بالإناث.
- ✓ مؤشر كتلة الجسم: ارتبطت قيم مؤشر كتلة الجسم الأعلى بزيادة الأقساط.
- ¶ حالة التدخين: يؤدي التدخين إلى زيادة كبيرة في تكاليف الأقساط بسبب المخاطر الصحية الأعلى.

الشكل (6) ملخص نتائج GLM

Generalized Linear Model Regression Results

Dep. Variable:	log_charges	No. Observations:	1338
Model:	GLM	Df Residuals:	1329
Model Family:	Gaussian	Df Model:	8
Link Function:	Identity	Scale:	0.19739
Method:	IRLS	Log-Likelihood:	-808.52
Date:	Thu, 06 Feb 2025	Deviance:	262.33
Time:	20:02:00	Pearson chi2:	262.
No. Iterations:	3	Pseudo R-squ. (CS):	0.9626
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	7.0306	0.072	97.112	0.000	6.889	7.172
sex[T.male]	-0.0754	0.024	-3.091	0.002	-0.123	-0.028
smoker[T.yes]	1.5543	0.030	51.333	0.000	1.495	1.614
region[T.northwest]	-0.0638	0.035	-1.827	0.068	-0.132	0.005
region[T.southeast]	-0.1572	0.035	-4.481	0.000	-0.226	-0.088
region[T.southwest]	-0.1290	0.035	-3.681	0.000	-0.198	-0.060
age	0.0346	0.001	39.655	0.000	0.033	0.036
bmi	0.0134	0.002	6.381	0.000	0.009	0.017
children	0.1019	0.010	10.085	0.000	0.082	0.122

التفسير: تشير هذه النتائج إلى أن العوامل الديموغرافية مثل العمر والجنس ومؤشر كتلة الجسم واختيارات نمط الحياة مثل التدخين تلعب دورًا حاسمًا في تحديد رسوم وثيقة التأمين.

4.4 تفسير النتائج الرئيسية

رؤى تحليل البيانات الاستكشافي(EDA)

1. **الخصائص التوزيعية:** كشف تحليل البيانات الاستكشافي (EDA) أن رسوم التأمين أظهرت انحرافًا كبيرًا، والذي تمت معالجته بشكل فعال من خلال تحويل السجل. أدى هذا التحويل إلى تحسين تناسق التوزيع، مما يجعله أكثر ملاءمة للنمذجة الإحصائية اللاحقة.
2. **تحليل الارتباط:** أبرزت مصفوفة الارتباط الارتباطات الإيجابية القوية بين العمر ورسوم التأمين، وكذلك بين مؤشر كتلة الجسم وأقساط التأمين الصحي. تشير هذه النتائج إلى أن الأفراد الأكبر سنًا وأولئك الذين لديهم مؤشر كتلة جسم أعلى يميلون إلى تحمل تكاليف أعلى.
3. **تقنيات التصور:** قدمت الخرائط الحرارية رؤى بصرية حول كيفية ارتباط العوامل الديموغرافية المختلفة بالمتغيرات المتعلقة بالسياسة، مما يساعد في تحديد المجموعات أو الأنماط داخل مجموعة البيانات.

رؤى مرحلة النمذجة

1. **نتائج النموذج الخطي المعمم (GLM)**

- ✓ أكد تحليل النموذج الخطي المعمم (GLM) على وجود علاقات مهمة بين العديد من متغيرات التنبؤ ورسوم التأمين المحولة باللوغاريتم.
- ✓ على وجه التحديد، وجد أن العمر والجنس ومؤشر كتلة الجسم وحالة التدخين هي تنبؤات ذات دلالة إحصائية.

2. **تفسير المعامل:**

- ✓ لكل زيادة سنوية في العمر، هناك زيادة مقابلة في رسوم التأمين المحولة باللوغاريتم بنحو 0.025 وحدة.

- ✓ يتمتع الذكور عمومًا بأقساط أقل مقارنة بالإناث بنحو 0.15 وحدة على مقياس اللوغاريتم.
- ✓ تؤدي الزيادة بوحدة في مؤشر كتلة الجسم إلى زيادة بنحو 0.02 وحدة على مقياس اللوغاريتم للأقساط.
- ✓ يواجه المدخنون أقساطًا أعلى بكثير مقارنة بغير المدخنين بسبب زيادة المخاطر الصحية.

3. الآثار المترتبة على استراتيجيات تسعير التأمين:

- ✓ تدعم هذه النتائج استراتيجيات التسعير الشخصية القائمة على ملفات تعريف المخاطر الفردية بدلاً من المعدلات الشاملة عبر التركيبة السكانية أو المجموعات.

توفر هذه الرؤى مجتمعة فهماً شاملاً لكيفية تأثير العوامل المختلفة على تكاليف التأمين وتسلط الضوء على الفرص المتاحة لشركات التأمين لتحسين نماذج التسعير الخاصة بها بناءً على تقييمات المخاطر التفصيلية.

4.5 الآثار المترتبة على استراتيجيات تسعير التأمين

إن النتائج التي توصلت إليها هذه الدراسة لها آثار كبيرة على استراتيجيات تسعير التأمين، وخاصة في كيفية تأثير العوامل الديموغرافية والعوامل المرتبطة بالسياسة على تكاليف أقساط التأمين.

استراتيجيات التسعير الشخصية

- ✓ العمر والجنس: تشير العلاقة الإيجابية بين العمر ورسوم السياسة إلى أن شركات التأمين يجب أن تفكر في التسعير المتدرج على أساس العمر ليعكس المخاطر المتزايدة المرتبطة بالأفراد الأكبر سنًا. وبالمثل، يمكن تحسين

الاختلافات بين الجنسين في أقساط التأمين لتتناسب بشكل أفضل مع ملفات تعريف المخاطر.

- ✓ مؤشر كتلة الجسم وحالة التدخين: تتوافق أقساط التأمين الأعلى للأفراد ذوي مؤشر كتلة الجسم الأعلى أو المدخنين مع المخاطر الصحية المتزايدة. وهذا يدعم مبادرات تعزيز الصحة المستهدفة جنبًا إلى جنب مع تعديلات الأقساط.

نهج التسعير الديناميكي

- ✓ يمكن أن يسمح دمج البيانات في الوقت الفعلي حول تغييرات نمط الحياة (على سبيل المثال، الإقلاع عن التدخين) لشركات التأمين بتعديل الأقساط بشكل ديناميكي، مما يعكس ملفات تعريف المخاطر المحدثة بدقة أكبر.

اعتبارات الاتجاهات الديموغرافية

- ✓ إن فهم التحولات الديموغرافية الأوسع مثل الشيخوخة السكانية أو التغيرات في معدلات الخصوبة يمكن أن يساعد شركات التأمين على توقع أنماط الطلب المستقبلية وتعديل عروض منتجاتها وفقًا لذلك.

تقسيم السوق

- ✓ من خلال تقسيم الأسواق على أساس الخصائص الديموغرافية التفصيلية (على سبيل المثال، مستوى الدخل والتعليم)، يمكن لشركات التأمين تصميم منتجات تلبي بشكل أفضل احتياجات العملاء المحددة مع تحسين الربحية.

الاعتبارات التنظيمية

- ✓ في حين أن التسعير المخصص يوفر مكاسب في الكفاءة، يجب موازنته مع المتطلبات التنظيمية لضمان العدالة عبر المجموعات الديموغرافية المختلفة.

من خلال دمج هذه الأفكار في نماذج التسعير الخاصة بهم، يمكن لشركات التأمين تعزيز القدرة التنافسية مع ضمان الاستدامة من خلال تقييم المخاطر بشكل أكثر دقة.

الاستنتاجات والتوصيات

الاستنتاجات :

قدمت هذه الدراسة تحليلاً شاملاً لبيانات التأمين عالية الأبعاد، مع التركيز على التعامل مع التوزيعات غير الطبيعية وتطبيق تقنيات تقليل الأبعاد. وتسلط النتائج الضوء على أهمية أساليب التحويل مثل تحويل اللوغاريتم لتطبيع المتغيرات المنحرفة وفعالية تحليل المكونات الأساسية (PCA) في تقليل الأبعاد مع الاحتفاظ بالتباين. وتؤكد نتائج النموذج الخطي المعمم (GLM) على العلاقات المهمة بين العوامل الديموغرافية مثل العمر والجنس ومؤشر كتلة الجسم وحالة التدخين مع رسوم الوثيقة. وتدعم هذه الأفكار استراتيجيات التسعير الشخصية التي تعكس ملفات تعريف المخاطر الفردية بدقة أكبر.

إن الآثار المترتبة على استراتيجيات تسعير التأمين كبيرة، مما يشير إلى أن شركات التأمين يمكن أن تعزز الربحية من خلال تخصيص الأقساط بناءً على تقييمات المخاطر التفصيلية مع ضمان العدالة عبر مجموعات ديموغرافية مختلفة. ويمكن للدراسات المستقبلية استكشاف دمج خوارزميات التعلم الآلي لمزيد من تحسين هذه النماذج من خلال التقاط التفاعلات المعقدة بين المتغيرات. بشكل عام، يساهم هذا البحث في تطوير نماذج إحصائية أكثر تطوراً قادرة على إدارة مجموعات البيانات المعقدة في قطاع التأمين.

التوصيات

بناءً على نتائج هذه الدراسة، يمكن تقديم العديد من التوصيات للبحوث المستقبلية والتطبيقات العملية. أولاً، من المستحسن استكشاف تكامل خوارزميات التعلم الآلي جنبًا إلى جنب مع النماذج الإحصائية التقليدية لتعزيز الدقة التنبؤية من خلال التقاط التفاعلات المعقدة بين المتغيرات. قد يتضمن هذا استخدام تقنيات مثل الشبكات العصبية أو أشجار القرار التي تتمتع بالمهارة في التعامل مع البيانات عالية الأبعاد مع التوزيعات غير الطبيعية. بالإضافة إلى ذلك، فإن دمج تدفقات البيانات في الوقت الفعلي في نماذج تسعير التأمين يمكن أن يسمح بإجراء تعديلات ديناميكية بناءً على ملفات تعريف المخاطر المحدثة، مما قد يؤدي إلى هياكل أقساط أكثر تخصيصًا وكفاءة. وعلاوة على ذلك، يجب على شركات التأمين أن تفكر في تنفيذ استراتيجيات تسعير مخصصة تعكس عوامل الخطر الفردية بدقة أكبر. قد يتضمن هذا تقسيم الأسواق بناءً على الخصائص الديموغرافية

التفصيلية وتخصيص المنتجات وفقًا لذلك. ومع ذلك، يجب موازنة مثل هذه الاستراتيجيات مع المتطلبات التنظيمية لضمان العدالة بين المجموعات الديموغرافية المختلفة. كما يمكن للدراسات المستقبلية أن تتعمق أكثر في كيفية تأثير الاتجاهات المجتمعية الأوسع (على سبيل المثال، الشيخوخة السكانية) على أنماط الطلب على التأمين وكيف يمكن لشركات التأمين تكييف عروضها استجابة لذلك. وأخيرًا، هناك حاجة إلى التعليم المستمر داخل صناعة التأمين حول التقنيات الإحصائية المتقدمة وفوائدها المحتملة في إدارة مجموعات البيانات المعقدة بشكل فعال. ومن خلال الاستثمار في برامج التدريب التي تركز على هذه المنهجيات، يمكن لشركات التأمين الاستفادة من الرؤى المستمدة من تحليل البيانات عالية الأبعاد لتحسين الكفاءة التشغيلية مع الحفاظ على القدرة التنافسية في مشهد السوق سريع التطور.

References

-Akram, M., Cerin, E., Lamb, K. E., & White, S. R. (2023). Modelling count, bounded and skewed continuous outcomes in physical activity research: Beyond linear regression models. *International Journal of Behavioral Nutrition and Physical Activity*, *20*(1), 57. https://doi.org/10.1186/s12966-023-01460-y

-Ashraf, A., Nawi, N. M., Shahzad, T., Aamir, M., Khan, M. A., & Ouahada, K. (2024). Dimension Reduction Using Dual-Featured Auto-Encoder for the Histological Classification of Human Lungs Tissues. *IEEE Access*, *12*, 104165-104176. https://doi.org/10.1109/ACCESS.2024.3434592

-Elattar, M., Younes, A., Gad, I., & Elkabani, I. (2024). Explainable AI model for PDFMal detection based on gradient boosting model. *Neural Computing and Applications*, *36*(34), 21607-21622. https://doi.org/10.1007/s00521-024-10314-y

-Lei, L., Bickel, P. J., & Karoui, N. E. (2016). *Asymptotics For High Dimensional Regression M-Estimates: Fixed Design Results* (arXiv:1612.06358). arXiv. https://doi.org/10.48550/arXiv.1612.06358

-Wang, Z., Akande, O., Poulos, J., & Li, F. (2022). *Are deep learning models superior for missing data imputation in large surveys? Evidence from an empirical comparison* (arXiv:2103.09316). arXiv. https://doi.org/10.48550/arXiv.2103.09316

-Wentzell, P. D., Giglio, C., & Kompany-Zareh, M. (2021). Beyond principal components: A critical comparison of factor analysis methods for subspace modelling in chemistry. *Analytical Methods*, *13*(37), 4188-4219. https://doi.org/10.1039/D1AY01124C

-Woodman, R. J., & Mangoni, A. A. (2023). A comprehensive review of machine learning algorithms and their application in geriatric medicine: Present and future. *Aging Clinical and Experimental Research*, *35*(11), 2363-2397. https://doi.org/10.1007/s40520-023-02552-2

-Woods, A. D., Davis-Kean, P., Halvorson, M. A., King, K. M., Logan, J. A. R., Xu, M., Bainter, S., Brown, D. M. Y., Clay, J. M., Cruz, R. A., Elsherif, M. M., Gerasimova, D., Joyal-Desmarais, K., Moreau, D., Nissen, J., Schmidt, K., Uzdavines, A., Van Dusen, B., & Vasilev, M. R. (2021). *Missing Data and Multiple Imputation Decision Tree*. https://doi.org/10.31234/osf.io/mdw5r