פרויקט לימוד מכונה – חלק א'

תביעת ביטוח רכב

מרצה: פרופ' בועז לרנר

תאריך הגשה: 27/04/2022

מגישים:

אופיר דוד 315695643

משה כהן 316161694

תוכן עניינים

הגדרת הבעיה

מספר הנהגים הפרטיים גדלה משנה לשנה וכתוצאה מכך עולה הצורך לאתר את הנהגים שבסבירות גבוהה יגרם נזק לרכבם (תאונת דרכים, פריצה ועוד) ויתבעו את חברת הביטוח. נרצה לאתר את אותם המאפיינים שיעזרו לנו לחזות בצורה הטובה ביותר מיהם אותם הנהגים שעתידים לתבוע את חברת הביטוח הרכב בשנה הקרובה. בכך, חברות הביטוח יוכלו להתאים את מסלול הביטוח אל כל אחד מהנהגים.

במהלך השנים בוצעו מחקרים רבים המבוססים על אלגוריתמים שונים המנתחים את נתוני הנהגים. דוגמאות לאלגוריתמים אלו, המתבססים על כלים סטטיסטיים ולימוד מכונה, הם K-NN, naive Bayes, decision trees, random forest, logistic regression, XGboost, gradient lifting tree (GBDT), lifting machine algorithm (LightGBM) (1). המשותף לכלל האלגוריתמים הוא הניסיון לחזות או לסווג את הנהגים לקבוצות בעלות סיכון שונה לתביעת ביטוח הרכב עפ"י מאפייניהם, ובכך לענות על הצורך של חברת הביטוח.

נרצה ליישם את הכלים הסטטיסטים ואלגוריתמי לימוד מכונה שנחקרו בעבר על סט הנתונים שברשותנו, המאופיין במשתנים מגוונים המאפיינים את הנהג ומאפשרים לנו ללמוד על התנהגותו ועל הפוטנציאל שלו לתבוע את הביטוח על סמך נתוני עבר של נהגים עם מאפיינים דומים.

הבנת הנתונים

1. תיעוד מקורות הנתונים ומשמעותם

משתנה	סוג המשתנה - רציף/בדיד/קטגוריאלי	הסבר קצר על המשתנה	הדרך בה נאספו הנתונים
age	קטגוריאלי	חלוקה לארבעה טווחים המאפיינים נהגים עם מאפייני נהיגה ומאפיינים פיזיים שונים. קטגוריות: 16-25, 26-39, 40-64, 65+	משתנים המתייחסים למאפיינים אישיים בסיסיים על הנהג כאדם. נתונים אלו נאספים באופן ידני בזמן שיחה עם הנהג.
gender	קטגוריאלי	מין הנהג. 1 - גבר, 0 - אישה	
married	קטגוריאלי	משתנה המייצג האם הנהג נשוי. 1 - נשוי, 0 - אינו נשוי.	
children	קטגוריאלי	משתנה המייצג האם לנהג יש ילדים. 1 - יש, 0 - אין.	
education	קטגוריאלי	משתנה המייצג את רמת ההשכלה של הנהג. קטגוריות: תיכונית, אוניברסיטאית, ללא השכלה.	משתנים המתייחסים למצב הסוציואקונומי של הנהג. גם נתונים אלו נאספים באופן ידני בשיחה עם הנהג, אך יש צורך
income	קטגוריאלי	משתנה המייצג את רמת ההכנסה של הנהג. קטגוריות: עוני, מעמד פועלים, מעמד ביניים, מעמד גבוה.	
credit score	רציף	משתנה המייצג את דירוג האשראי של הנהג וטווח ערכיו הוא 0-1. ככל שערך הינו גבוה יותר כך קטן הסיכון לא לעמוד בהתחייבויות. כלומר, סיכוי גבוה יותר שנקבל הלוואה בתנאים טובים יותר.	

postal code	קטגוריאלי	משתנה המייצג את אזור המגורים של הנהג. בבסיס הנתונים שלנו קיימות ארבע קבוצות.	באימותם אל מול גורמים חיצוניים.
driving experience	קטגוריאלי	משתנה המייצג את מספר שנות הניסיון של לנהג בנהיגה. משתנה מחולק לארבעת הטווחים הבאים (בשנים): 0-9, 10-19, 20-29, 30+.	משתנים המתייחסים למאפייני הרכב וניסיון הנהג. נתונים אלו נאספים באופן ידני בשיחה עם הנהג, וחלקן מתעדכנים באופן שוטף במידה והנהג החליף את רכבו או ביצע תאונה.
vehicle ownership	קטגוריאלי	משתנה המייצג האם הרכב בבעלות הנהג. 1 - כן, 0 - לא.	
vehicle year	קטגוריאלי	משתנה המייצג את שנת ייצור הרכב. מחולק לשתי קטגוריות: לפני 2015, אחרי 2015.	
vehicle type	קטגוריאלי	משתנה המייצג את סוג הרכב. מחולק לשתי קטגוריות: רכב משפחתי ורכס ספורט.	
annual mileage	בדיד	משתנה המייצג את כמות המיילים השנתיים שהרכב נסע. בבסיס הנתונים מיוצג בקפיצות של 1000 מיילים (אינו רציף).	
past accidents	בדיד	משתנה המייצג את מספר התאונות שביצע הנהג.	
speeding violations	בדיד	משתנה המייצג את מספר עבירות המהירות שביצע הנהג	
outcome	משתנה מטרה - קטגוריאלי	משתנה המטרה מייצג האם הנהג יתבע את חברת הביטוח בשנה הקרובה (1- יתבע, 0- לא יתבע). בסט האימון משתנה זה מייצג האם הנהג תבע את הביטוח בשנה האחרונה.	

בנוסף למאפיינים אלו לכל נהג (תצפית), ישנו מזהה חד-חד ערכי בסט הנתונים (ID).

2. הסתברויות אפריוריות

26-39	0.34026
40-64	0.29091
16-25	0.21558
65+	0.15325

Age - ניתן לראות שיותר ממחצית הנדגמים הם בני 26-64 (כ-61%). ניתן לתמוך בממצא זה בעזרת שתי העובדות הבאות: טווח הגילאים השכיחים להנפקת הרישיון הוא בתחילת שנות העשרים ובגיל מבוגר אנשים מפסיקים לנהוג בעקבות מגבלות פיזיות.

male	0.55584
female	0.44416

Gender - המשמעות היא שעל פי בסיס הנתונים שלנו, מרבית הנהגים הינם גברים. זאת, בהתאמה למידע מחקרי שאספנו אודות אוכלוסיית הנהגים (2).

0	0.52208
1	0.47792

Married - ניתן לראות שקיימת פרופורציה דומה המייצגת כל אחד מהמצבים.

1	0.64675
0	0.35325

Children - במדגם קיימים יותר נהגים עם ילדים מאלו ללא ילדים.

high school	0.43117
university	0.36883
none	0.20000

Education - החלוקה לרמות ההשכלה בבסיס הנתונים תואמת למידע מחקרי שאספנו על רמת ההשכלה באוכלוסייה (3).

upper class	0.38701
middle class	0.24416
poverty	0.20260
working class	0.16623

Income - כפי שניתן לראות ככלל, כאשר רמת ההכנסה עולה גם הפרופורציה מהמדגם עולה. עובדה שיכולה לתמוך בכך שלאנשי אוכלוסייה עשירה יותר יש רכבים.

10238	0.68052
32765	0.25974
92101	0.04675
21217	0.01299

Postal Code - רוב הלקוחות שבמדגם נדגמו מאזור מיקוד מסוים, לכן ייתכן וללקוחות אלו מאפיינים כלכליים-חברתיים דומים.

0-9y	0.41558
10-19y	0.30909
20-29y	0.18701
30y+	0.07273

Driving Experience - ניתן לראות שנדגמו פחות נהגים עם מספר רב של שנות ניסיון. נתון זה תואם למשתנה הגיל בו נדגמו יותר צעירים ממבוגרים.

1	0.63377
0	0.36623

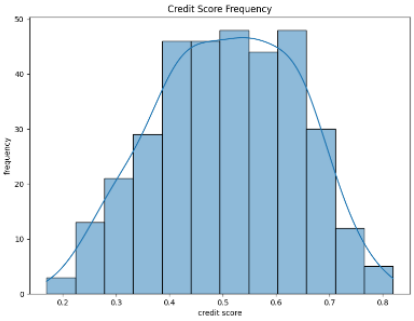
Vehicle Ownership - מרבית האנשים המעוניינים לבצע ביטוח לרכב מבצעים זאת על רכב שנמצא בבעלותם.

before 2015	0.74545
after 2015	0.25455

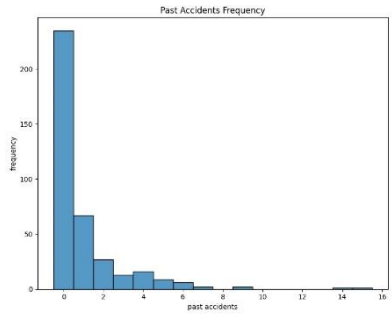
Vehicle Year - מרבית מהרכבים המבוטחים הם רכבים שיוצרו לפני שנת 2015.

sedan	0.93247
sports car	0.06753

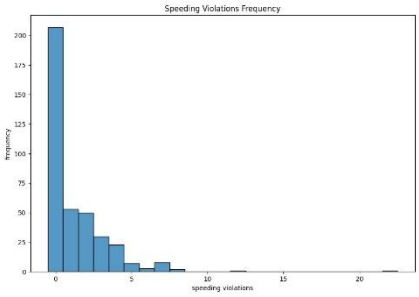
Vehicle Type - ניתן לראות כי הרוב המוחלט של הרכבים המבוטחים הם מדגם רכב משפחתי. זאת בהתאמה לכך שרכבים משפחתיים שכיחים יותר מרכבי ספורט שעלותם גבוהה יותר.



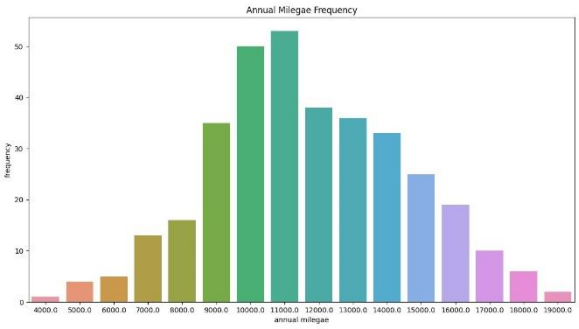
Credit Score - ניתן לראות שהתפלגות דומה להתפלגות נורמלית, כך שרוב האוכלוסייה ממוקמת קרוב לתוחלת (דירוג אשראי ממוצע -כ-0.5) וכלל שמתרחקים ממנה יש פחות תצפיות.



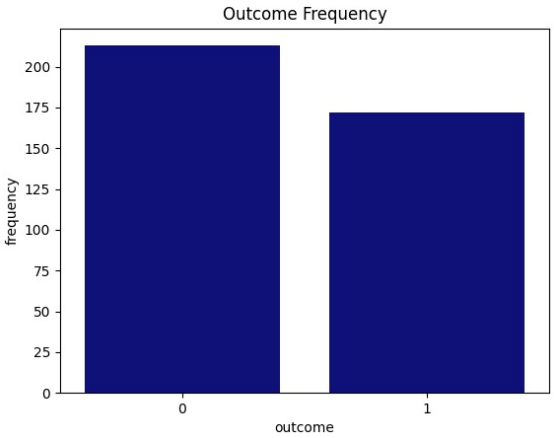
Past Accidents - רוב הנדגמים לא ביצעו תאונות כלל. ככל שמספר התאונות עולה כך השכיחות יורדת. במספר מצומצם של רשומות הוזן הערך -999, זהו ערך שלילי ואינו מייצג כמות אפשרית של תאונות ולכן אינו נכלל באפיון המשתנה.



Speeding Violations - רוב הנדגמים לא ביצעו עבירות מהירות כלל. ככל שמספר עבירות המהירות עולה כך השכיחות יורדת.



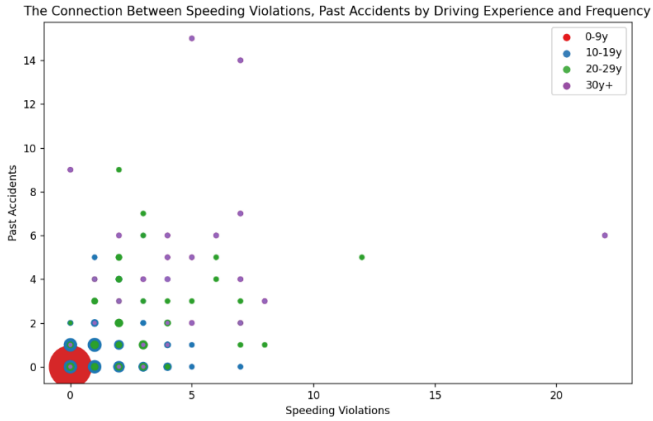
Annual Mileage - למרות שהמשתנה בדיד ניתן לראות שהתפלגות הנתונים מזכירה התפלגות נורמלית. זאת, משום שבוצעה דיסקרטיזציה לנתונים בעת איסופם. במצב זה רוב האוכלוסייה ממוקמת קרוב לתוחלת (כמות מיילים שנתית ממוצעת) וכלל שמתרחקים ממנה יש פחות תצפיות.



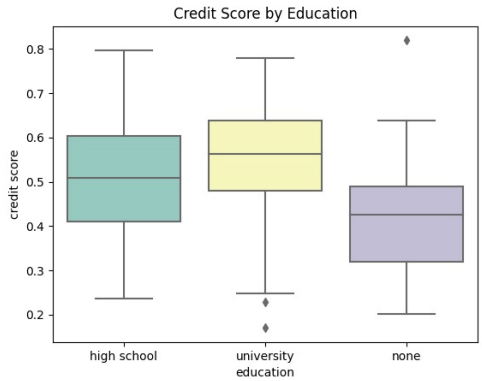
בהסתכלות על פרופורציית שתי המחלקות של משתנה המטרה ניתן לראות כי סט הנתונים מאוזן. 55.32% מהתצפיות שנדגמו הן של נהגים שלא תבעו את הביטוח בשנה האחרונה וכי 44.67% מהתצפיות הן של נהגים שתבעו את הביטוח. ממצא זה אינו מייצג את המציאות בה רק אחוז קטן ממבוטחי הרכב תובעים את הביטוח מידי שנה. ממחקר שביצענו מצאנו כי בשנת 2018 פחות מ-7% ממבוטחי הרכב תבעו את חברת הביטוח בארה"ב (4) ואנו מעריכים כי נתון זה משקף את האוכלוסייה כולה. למרות שכמות הנהגים שיתבעו את הביטוח מהווה אחוז קטן מהאוכלוסייה, נרצה לבחון מקרים רבים על מנת לאפשר למערכת לבצע למידה על בסיס מגוון מקרים, ולכן מדגם מאוזן ולא בהכרח מייצג כפי שקיבלנו עונה על הצורך.

3. קשרים בין מאפיינים (צפויים ולא צפויים)

<u>הקשר בין מספר תאונות ומספר עבירות המהירות בהתאם למספר שנות הניסיון של הנהג</u>

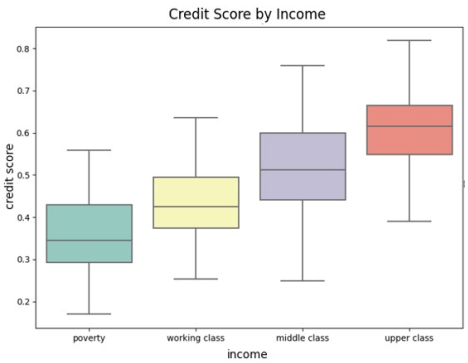


בתרשים זה כל נקודה מייצגת מספר תצפית או יותר של אותה קטגוריית שנות ניסיון. כך שככל והנקודה גדולה יותר הוא מאגדת מספר רב יותר של זהות עם אותן המאפיינים. ככל שמספר שנות ניסיון הנהיגה קטן יותר, כך הנהגים ביצעו פחות תאונות ועבירות מהירות.



<u>הקשר בין רמת ההשכלה לדירוג האשראי</u>

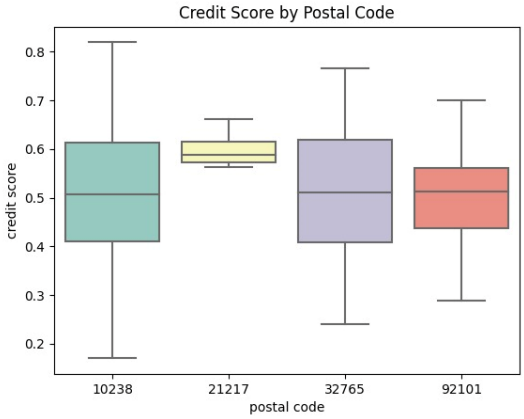
ככל שרמת ההשכלה של הנהג גבוהה יותר, כך בממוצע דירוג האשראי שלו גבוהה יותר. כמו כן, פיזור הנתונים של הנהגים ללא השכלה מצומצם ביחס לשאר רמות ההשכלה.



<u>הקשר בין רמת ההכנסה לדירוג האשראי</u>

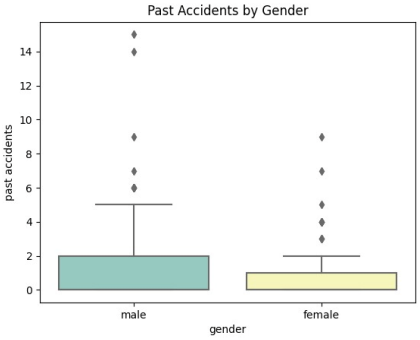
ניתן לזהות קשר בין רמת ההכנסה לגובה דירוג האשראי, כאשר ככל שרמת ההכנסה גבוהה כך דירוג האשראי דרג בהתאמה. כמו כן, הטווח הבין-רבעוני זהה בכל אחד מרמות השכלה ולכן ניתן ללמוד כי פיזור הנתונים דומה אך נע סביב טווח ערכים שונה בכל אחד מהרמות.

הקשר בין אזור המיקוד לדירוג האשראי



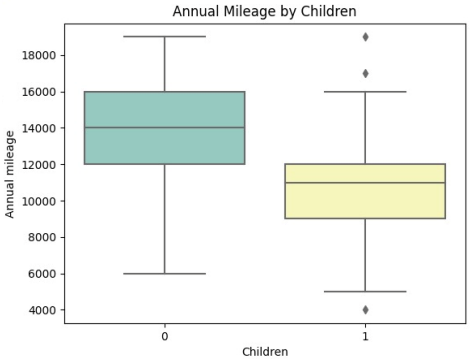
לא ניתן לזהות קשר בין אזורי המיקוד לגובה דירוג האשראי פרט לאזור המיקוד 21217 בו דירוג האשראי גבוה יותר. ייתכן כי עקב כמות מצומצמת של תצפיות עבור אזור מיקוד זה פיזור הנתונים מצומצם יותר ואינו משקף את הקשר בין המשתנים כראוי. כמו כן, נראה כי החציונים בכל אחד משאר אזורי המיקוד דומים להתפלגות של המדגם כולו (כ-0.5).

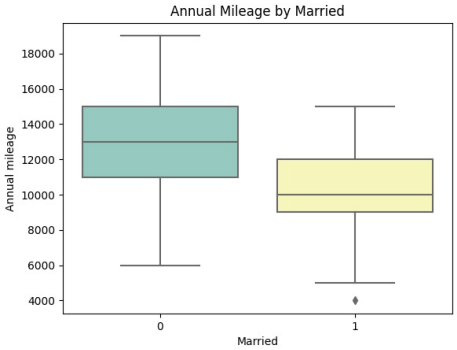
הקשר בין המגדר למספר התאונות ולמספר עבירות המהירות



בכל אחד מתרשימים אלו ניתן לזהות את הקשר בין המגדר של הנהג לבין מספר התאונות (תרשים שמאלי) או מספר דוחות המהירות (תרשים ימני). בתרשים השמאלי החציון עבור נשים וגברים זהה, לעומת תרשים הימני בו החציון עבור גברים גבוה יותר. בנוסף, בשני התרשימים פיזור הנתונים עבור נשים מצומצם יותר. כפי שניתן לראות כמות התאונות ודוחות המהירות של גברים גבוהה יותר משל נשים, ולכן ישנה השפעה של המגדר על מאפיינים אלו.

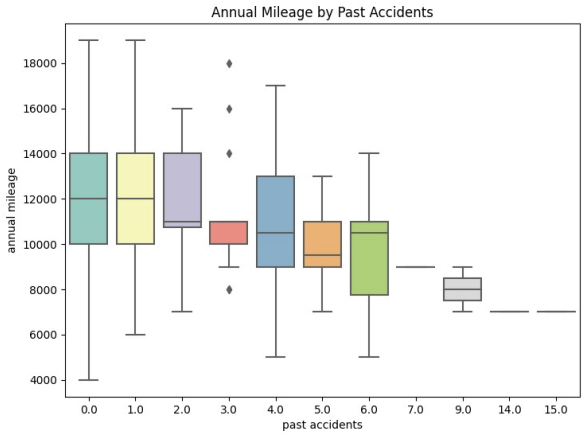
הקשר בין הסטטוס המשפחתי (נשוי, ילדים) של הנהג למספר המיילים השנתית של הרכב





ניתן לזהות קשר בין האם הנהג אב לילדים לכמות המיילים השנתיים שהרכב נוסע, כך שבמידה והנהג אב לילדים בממוצע כמות המיילים השנתיים קטנה יותר מאשר נהג ללא ילדים.

הקשר בין מספר התאונות למספר המיילים השנתית של הרכב



ניתן לזהות קשר בין כמות התאונות שביצע הנהג לכמות המיילים השנתיים שהרכב נסע, כך שרכבים שהיו מעורבים ביותר תאונות נסעו בממוצע פחות בשנה האחרונה.

4. קשרים אפשריים בין המאפיינים למשתנה המטרה

כאשר נסקור את כלל המאפיינים בבסיס הנתונים, נאתר את המאפיינים אשר ניתן לחשוד שיהיו בעלי השפעה על משתנה המטרה על פי הבנת עולם התוכן בו אנו נמצאים. הגורמים אשר יכולים להוביל נהג לתבוע את חברת הביטוח הם תאונת דרכים, פריצה לרכב, גניבת הרכב והשחתתו.

ותק נהיגה - ככל שלנהג יש יותר שנות ניסיון כך הוא יותר מיומן ולכן ישנה סבירות גדולה אשר הוא יצליח להימנע ממצבים מסוכנים אשר יכולים להוביל לתאונה.

מספר מיילים שנתיים - ככל שהרכב נוסע יותר כך גדל הסיכוי שיהיה מעורב בתאונת דרכים או שיהיה בלאי טבעי אשר עלול לגרור לתביעת הביטוח.

מיקוד - נצפה שיהיו הבדלים במספר תביעות הביטוח בין שכונות בטוחות יותר ממעמד סוציו-אקונומי גבוה לשכונות ממעמד סוציו-אקונומי נמוך, אשר ינבע מהבדלים בגניבות, פריצות והשחתות של הרכבים.

שנת הרכב - אנו מצפים שרכבים ישנים יותר יתבעו את הביטוח יותר מכיוון וקל יותר לפרוץ אליהם בעקבות מערכת אבטחה פחות טובה של הרכב או שיהיה בלאי טבעי אשר עלול לגרור לתביעת הביטוח.

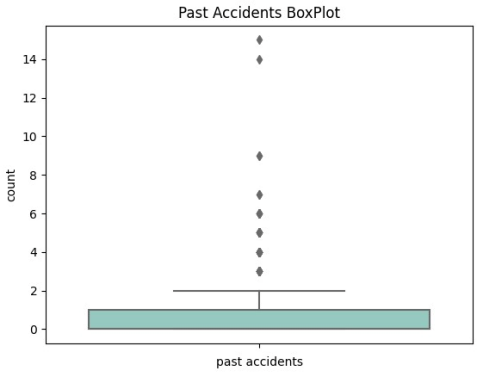
בנוסף, אנו מעריכים כי המאפיינים מספר עבירות המהירות ומספר תאונות הדרכים עשוי להשפיע על משתנה המטרה, אך יש צורך בשינוי אופן ההסתכלות עליהם. נציג את שינוי שלדעתנו יש לבצע בהמשך.

5. איכות הנתונים

בעת הסתכלות על טווח הערכים של כל אחד מהמאפיינים זהינו מספר מאפיינים בעלי נתונים חסרים, לא הגיוניים וחשודים כחריגים.

ראשית, עבור המאפיין speeding violations נמצא ערך חריג (22). מבדיקה שבצענו עולה כי רשומה זו מייצגת נהג מבוגר (+65), שביצע לא מעט תאונות בעבר (6), ולכן למרות שערך זה גדול ביחס לשאר התצפיות הוא אינו לא הגיוני.

גם במאפיין Past accident נמצאו ערכים חריגים (14,15). מבדיקה שביצענו כעולה כי רשומות אלט מייצגות נהגים מבוגרים (+65), שביצעו מספר רב של עבירות מהירות (7,5), ולכן למרות שערך זה גדול ביחס לשאר התצפיות הוא אינו לא הגיוני.



על כן, נבחר שלא לבצע שינוי ברשומות אלה.

עבור מאפיין past accident נמצאו שש רשומות עבורן הוזן הערך 999-. משום שזהו ערך שלילי ואינו יכול לייצג כמות אפשרית של תאונות, נסתכל על ערכים אלו כנתונים חסרים.

רשומות בעלות חוסר של שני מאפיינים או יותר מעידות על בעייתיות באיסוף הנתונים, ולכן בעקבות פער קריטי בנתונים וחשש שהנתונים הקיימים אינן משקפים כהלכה את הרשומה במלואה, ולכן נמליץ למחוק את רשומות אלו ממאגר הנתונים (שמונה רשומות סה"כ).

AGE	0
GENDER	0
DRIVING_EXPERIENCE	6
EDUCATION	0
INCOME	0
CREDIT_SCORE	40
VEHICLE_OWNERSHIP	0
VEHICLE_YEAR	0
MARRIED	0
CHILDREN	0
POSTAL_CODE	0
ANNUAL_MILEAGE	39
VEHICLE_TYPE	0
SPEEDING_VIOLATIONS	0
PAST_ACCIDENTS	6

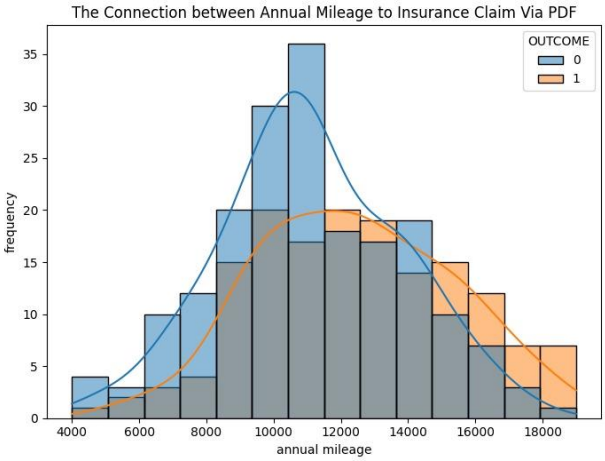
לבסוף, ישנם ארבעה מאפיינים עם שדות ריקים (שמות המאפיינים וכמות הרשומות החסרות מוצגים בטבלה) עבור חלק מהרשומות. בכדי לשפר את איכות הנתונים נרצה לשקול להשלים את הנתונים החסרים בעזרת מאפיינים אחרים אשר מצאנו קשר בינם לבין המשתנה עם הנתון החסר. זאת במטרה לאפשר שימוש ברשומות כחלק מתהליך הלמידה ובכך לא לצמצם את כמות הרשומות בבסיס הנתונים.

נרצה לאתר רשומות בעלות מאפיינים זהים למאפייני הרשומה עם הנתון החסר. בעזרת ערך החציון של רשומות אלו עבור ערך המשתנה החסר נשלים את הנתון ברשומות החסרות. עבור כל אחד מהמאפיינים החסרים נשתמש בסט מאפיינים שונה:

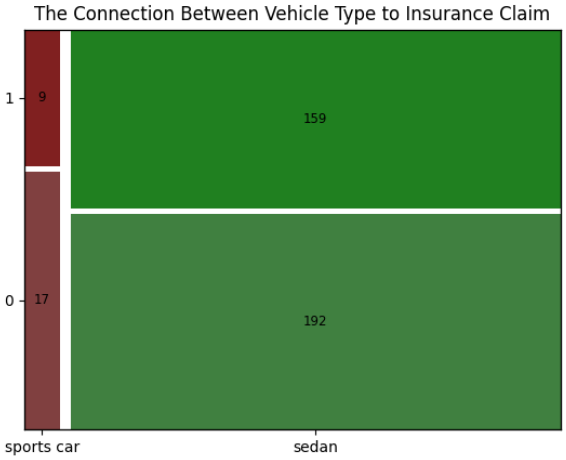
- Driving experience – גיל, מגדר, בעלות, עבירות מהירות.
- Credit score – הכנסה, מיקוד, סוג הרכב, השכלה.
- Annual mileage – ילדים, נשוי.
- Past accidents – מהירות, מגדר, מיילים, ניסיון נהיגה.

הכנת הנתונים

1. השמטת מאפיינים



כפי שניתן לראות בתרשים הצפיפות עבור מאפיין זה, לא קיים שוני משמעותי בהתפלגות הנתונים ביחס לכל אחד ממשתני המטרה (בתוחלת ובפיזור), ולכן קיים שטח רב החופף ביניהם. עקב כך, רוב התצפיות נופלות בשטח זה ולא ניתן לקבוע עבור איזה מחלקה היא עשויה להשתייך. על כן, בעזרת מאפיין זה לא ניתן להפריד בין המחלקות נבחר להסיר משתנה זה מבסיס הנתונים.

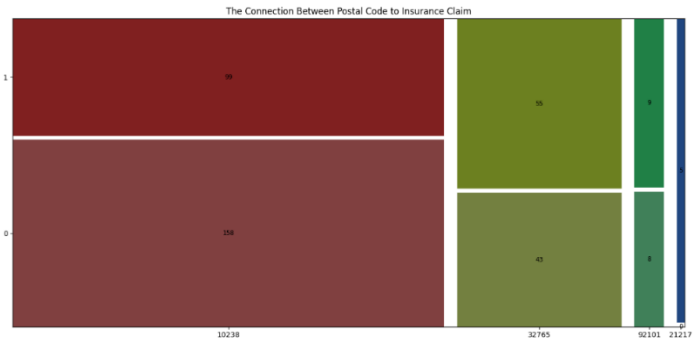


כפי שניתן לראות בתרשים היחס דומה בין קטגוריות ה-Vehicle type לבין המחלקות של משתנה המטרה דומה. בנוסף, היחס בין מספר התצפיות של כל אחת מהקטגוריות הוא גבוה (93.24% רכבים משפחתיים לעומת 6.75% רכבי ספורט). על כן, נחליט להסיר את משתנה זה במטרה לבנות מודל עם מאפיינים משמעותיים.

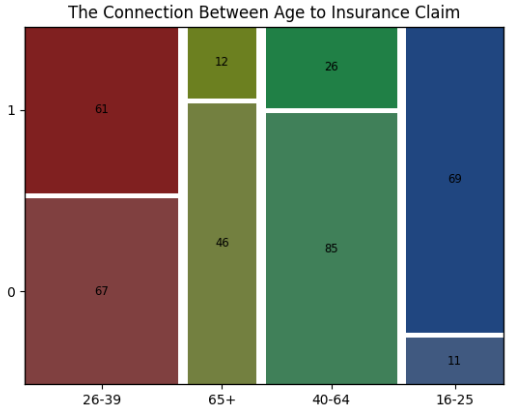
2. השמטת תצפיות

נבחר להסיר את הרשומות שבעלות שני מאפיינים חסרים כפי שהמלצנו קודם.
הרשומות שהוסרו מוצגות כנספח.

3. איחוד קטגוריות במאפיינים



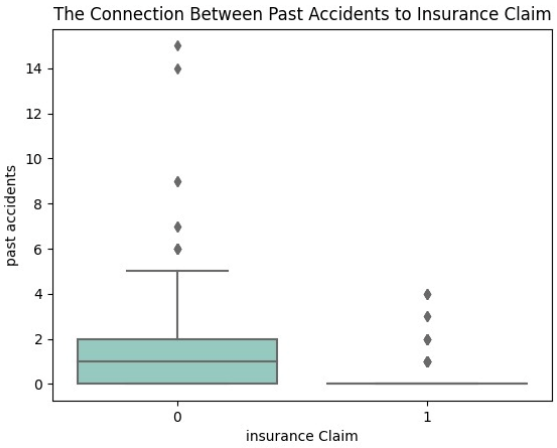
כפי שניתן לראות בתרשים היחס דומה בין המחלקות של משתנה המטרה עבור אזורי המיקוד 32765 ו- 92101 של מאפיין Postal Code, ולכן נחליט לאחד קטגוריות אלו לקטגוריה אחת בשם 32765+92101.



כפי שניתן לראות בתרשים היחס דומה בין המחלקות של משתנה המטרה עבור קטגוריות הגילאים 40-64 ו- 65+ של מאפיין Age, ולכן נחליט לאחד את קטגוריות אלו לקטגוריה אחת בשם 40+.

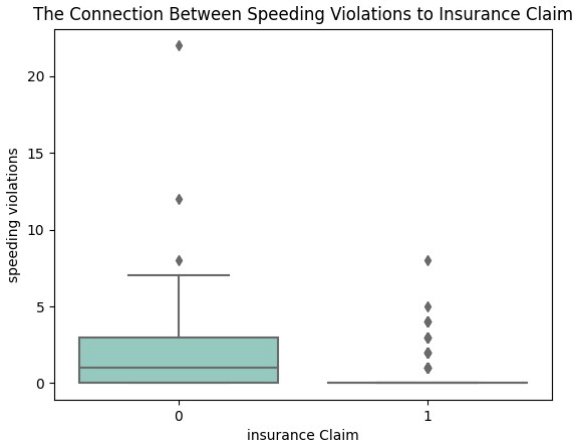
4. גזירת מאפיינים חדשים:

- <u>מספר תאונות ביחס לשנות ניסיון –ratio accident driving experience</u>



כפי שניתן לראות בתרשים, באופן יחסי נהגים עם מספר רב של תאונות לא תבעו את הביטוח בשנה האחרונה. זאת, בניגוד לצפייה שנהגים עם תאונות רבות שהיו מעורבים בתאונה יתבעו את הביטוח. מאפיין Past accident מייצג את מספר התאונות המצטבר שביצע הנהג. משום שלכל נהג מספר שנות ניסיון שונה המאפיין אינו מהימן ולא ניתן להשוות על בסיסו בין נהגים עם מספר שנות ניסיון שונה. לכן, נרצה לנרמל את עמודה זו כך שתייצג את היחס בין מספר התאונות למספר שנות הניסיון של הנהג.

- <u>מספר עבירות מהירות ביחס לשנות ניסיון – speeding violations driving ratio experience</u>



כפי שניתן לראות בתרשים, באופן יחסי נהגים עם מספר רב של עבירות מהירות לא תבעו את הביטוח בשנה האחרונה. זאת, בניגוד לצפייה שנהגים עם תאונות רבות שהיו מעורבים בתאונה יתבעו את הביטוח. מאפיין speeding violations מייצג את מספר עבירות המהירות המצטבר שביצע הנהג. משום שלכל נהג מספר שנות ניסיון שונה המאפיין אינו מהימן ולא ניתן להשוות על בסיסו בין נהגים עם מספר שנות ניסיון שונה. לכן, נרצה לנרמל את עמודה זו כך שתייצג את היחס בין מספר העבירות המהירות למספר שנות הניסיון של הנהג.

הנרמול יתבצע בצורה הבאה: מספר התאונות של הנהג חלקי ממוצע טווח שנות ניסיון הנהיגה שלו.

ביבליוגרפיה

1 - M.Hanafy , R.Ming (2). Risks. Machine Learning Approaches for Auto Insurance Big Data. Risks | Free Full-Text | Machine Learning Approaches for Auto Insurance Big Data (mdpi.com)

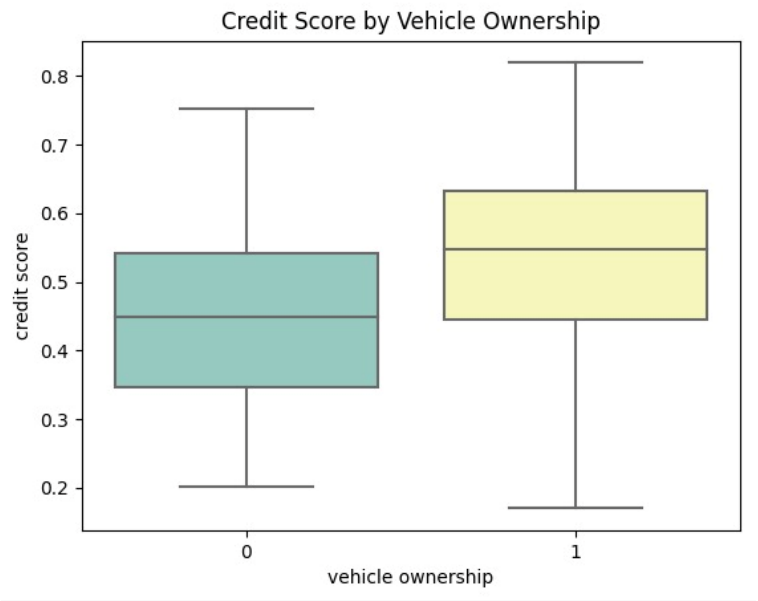
2 - הלשכה המרכזית לסטטיסטיקה (2020). מורשים לנהוג. עמ' 10. מורשים לנהוג, 2020 (cbs.gov.il).

3 - הלשכה המרכזית לסטטיסטיקה (2021). פני החברה בישראל, פערים לפי רמות השכלה. עמ' 127. פני החברה בישראל, דוח מס ,12 'שבט תשפ"א, ינואר 2021 – פערים לפי רמת השכלה(cbs.gov.il) .

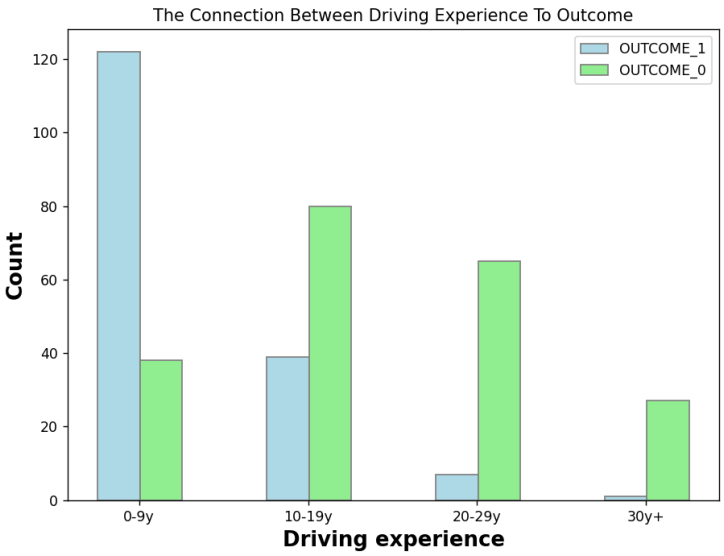
4 - PolicyAdvice (2022). Auto Insurance Statistics for 2022. Notable Auto Insurance Statistics for 2021 | Policy Advice.

נספחים

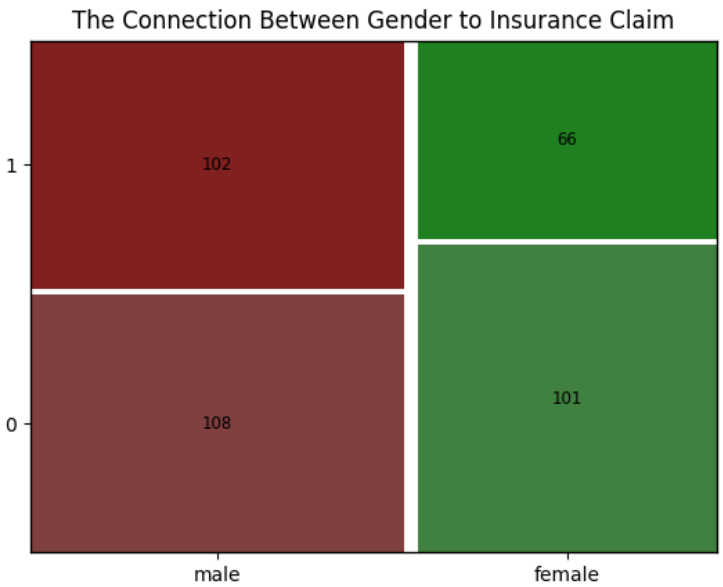
קשרים בין מאפיינים



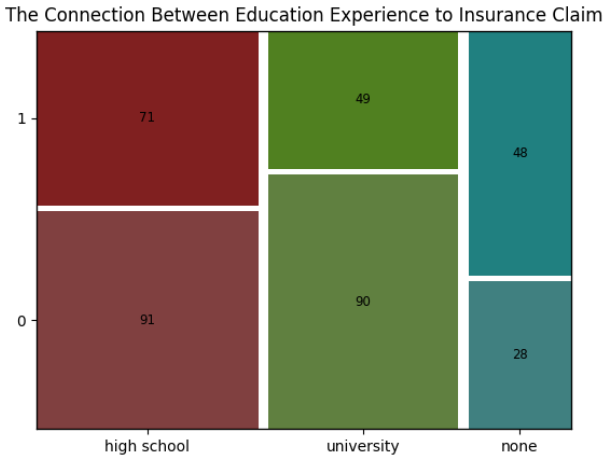
השמטת מאפיינים



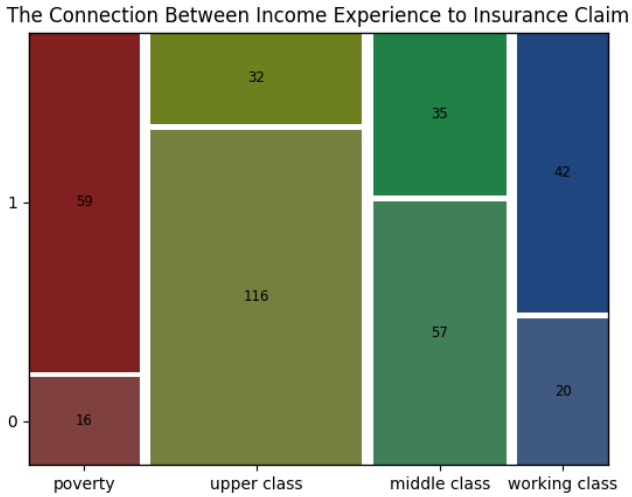
לא לאחד כי כבר איחדנו את הגיל . משלים את הקטגוריה המאוחדת שם.



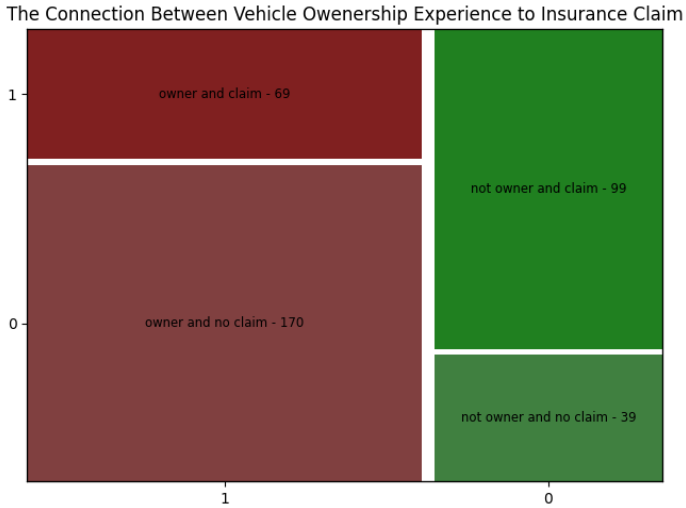
כפי שניתן לראות בתרשים היחס דומה בין קטגוריות ה-Gender לבין המחלקות של משתנה המטרה דומה. עם זאת, זהינו קודם קשרים בין מאפיין זה למאפיינים נוספים. לכן, נחליט לא לאחד את קטגוריות אלו מתוך הבנה שייתכן וקיימים תתי קבוצות יחד עם מאפיינים נוספים הנותנים אינדיקציה משמעותי על משתנה המטרה.



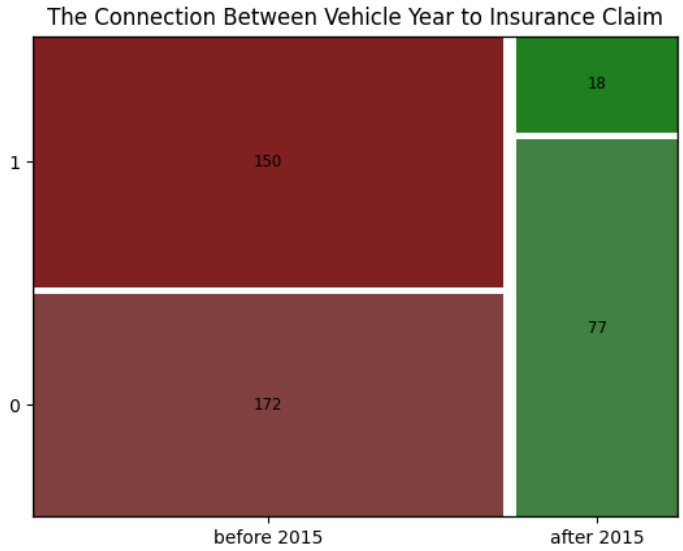
לא נאחד קטגוריות, משתנה רלוונטי



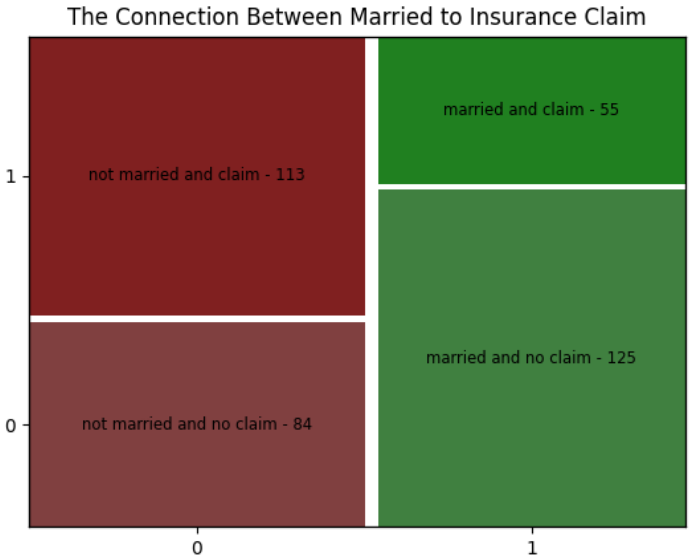
לא נאחד קטגוריות, משתנה רלוונטי



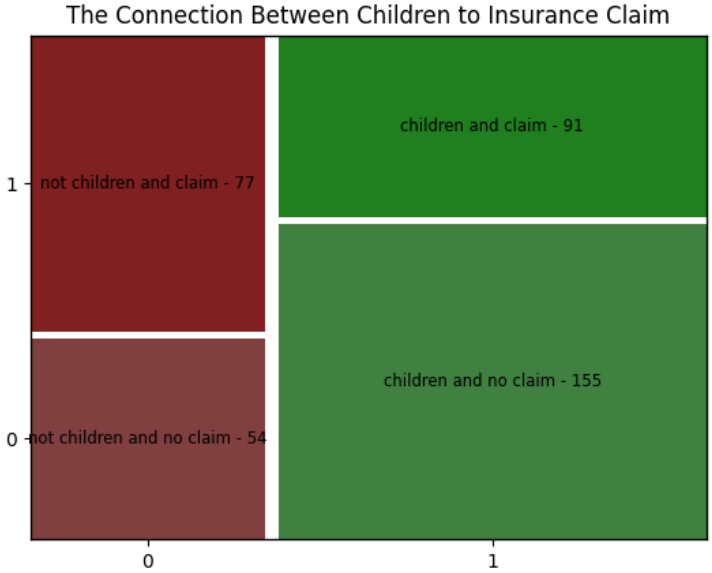
לא נאחד קטגוריות, משתנה רלוונטי



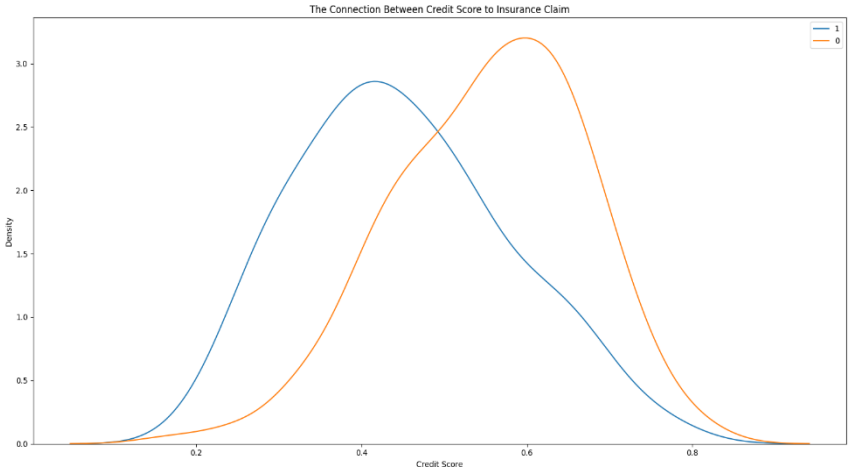
לא נאחד קטגוריות, משתנה רלוונטי



לא נאחד קטגוריות, משתנה רלוונטי



לא נאחד קטגוריות, משתנה רלוונטי



לא נאחד קטגוריות, משתנה רלוונטי

השמטת תצפיות

להלן התצפיות בעלות 2 מאפיינים חסרים אשר הושמטו מסט הנתונים

ID	AGE	GENDE	DRIVING_EXPERIENC	EDUCATIO	INCOME	CREDIT_SCOR	VEHICLE_OWNERSHI	VEHICLE_YEA	MARRIE	CHILDRE	POSTAL_CO	ANNUAL_MILEAG	VEHICLE_TYP	SPEEDING_VIOLATIONS	PAST_ACCIDENT	OUTCOM
14	16-25	male		high school	poverty		1	before 2015	0	0	10238	15000	sedan	0	0	1
115	40-64	female	20-29y	none	middle class		1	before 2015	1	1	10238		sedan	2	2	0
154	16-25	male	0-9y	university	poverty		1	after 2015	1	0	32765		sedan	0	0	1
189	26-39	male	10-19y	high school	working class		1	before 2015	1	1	32765		sedan	1	1	0
190	65+	female	30y+	university	upper class	0.626009704	1	after 2015	1	1	10238		sedan	0		0
205	16-25	female	0-9y	high school	poverty		0	before 2015	0	0	92101		sedan	0	0	1
243	26-39	female	10-19y	university	working class	0.517732144	0	before 2015	0	0	10238		sedan	1		0
384	26-39	male		high school	middle class		0	after 2015	0	0	10238	15000	sedan	0	0	1