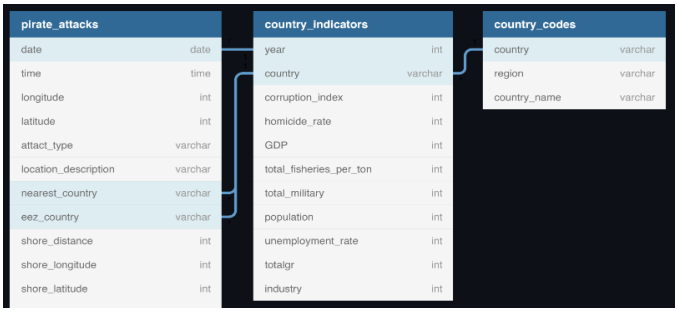
פרויקט סיום בקורס מבוא למדעי הנתונים 71253

מתן ספקטור וסער פרדו

מבוא

בחרנו לנתח בסיס נתונים שמקורו בgithub ומספק תיעוד גלובלי של תקיפות פיראטים בין השנים 1993-2020. בבסיס הנתונים יש 7511 "תקיפות" לפי תאריך, מיקום, מדינה קרובה ביותר ואזור כלכלי בלעדי קרוב. הקובץ מכיל 3 בסיסי נתונים לפי נתוני תקיפות, מאפייני מדינות ושיוך מדינה קרובה. בתמונה משמאל מוצגות רשימות המשתנים המקורית ותיאור השדות המקשרים בין הבסיסים. ביצענו איחוד של שלושת בסיסי הנתונים לפי מדינה ושנה, על מנת לקבל בסיס אחד המכיל את כלל המידע (בסה"כ 40 משתנים). כאמור, כל תצפית מספקת מידע על תקיפה ספציפית וכוללת את המאפיינים של המדינה הקרובה ביותר כמו גם את מאפייני המדינה השולטת באזור הכלכלי הימי (EEZ) שבו התבצעה התקיפה.



תחילה, זיהינו תצפיות עם ערכים חסרים בקוד מדינה או בשם מדינה בשני הקשרים: מדינה קרובה ביותר (Nearest Country) ומדינה השולטת באזור הכלכלי הימי (EEZ Country). עבור המדינות הקרובות ביותר, זיהינו את כל המדינות שיש להן ערכים חסרים והשלמנו את המידע על ידי חיפוש בשמות מדינות קיימים עבור אותו קוד מדינה, או על ידי שימוש במפה מתוקנת מראש של שמות מדינות עבור קודים שזוהו כחסרים. לדוגמה, הקצינו את השם " Cocos (Keeling) Islands" עבור הקוד CCK, ו-"Taiwan" עבור הקוד TWN, במקרים בהם היה קיים קוד מדינה ללא שם מדינה תואם. עבור מדינות השולטות באזור הכלכלי הימי, ביצענו תהליך דומה על מנת להבטיח שכל התצפיות יכילו שמות מדינות וקודים תקפים. ערכים שנותרו `Unknown` או `NA` לאחר שלבי ההשלמה נשארו כפי שהם, במקרים בהם לא ניתן היה למצוא מידע תואם. לאחר מכן בחנו את רמת ההתאמה בין מדינה קרובה למדינה השולטת באזור הכלכלי הימי.

Type	N	Percentage
Total Attacks	7375	100.00%
Exclusive Economic Zone == Nearest	6845	92.81%
Exclusive Economic Zone != Nearest	241	3.27%
No Exclusive Economic Zone	289	3.92%

שאלת מחקר

אחרי שהרכבנו את בסיס הנתונים מחדש ויצרנו פרמטר תוצאתי, אנחנו רוצים לראות איך אפשר להתמודד עם מתקפת פיראטים בצורה טובה ולא להיקלע לסיטואציה מזיקה. לשם כך נרצה לזהות קשרים בין הסבירות לניסיון תקיפה לא מוצלח ובין המשתנים השונים הנוגעים לזמן, מיקום ומאפיינים אחרים שכלי השיט.

משתנה תלוי - תוצאת התקיפה

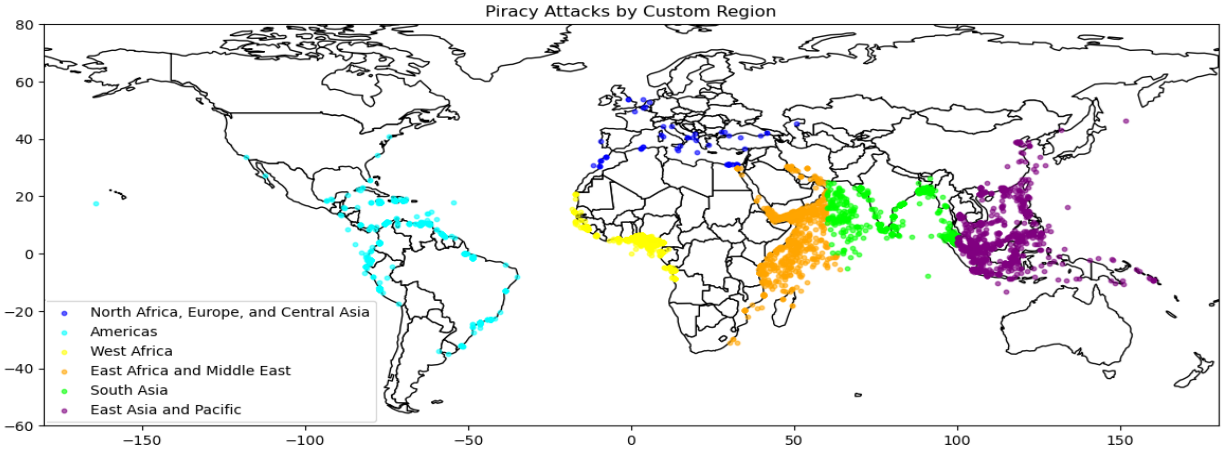
status	Count	Our status	Count
Boarded / Boarding	4788	Successful	5303
Hijacked	511		
Explosion	3		
Detained	1		
Attempted	1999	Unsuccessful	2072
Fired Upon	73		

בשלב זה, ביצענו איחוד של ערכים דומים כדי לשמור על עקביות ויצרנו משתנה קטגוריאלי חדש בשם (,attack outcome)אשר מסווג את התקיפות לשתי קטגוריות: "מוצלחת" (Successful) או "כושלות" (Unsuccessful). קטגוריה מוצלחת פירושה ברוב במקרים, עלייה של התוקפים לכלי השייט, לצד מעט מקרים של חטיפות ונזקי פיצוץ על הסיפון. קטגוריית התקיפות הלא מוצלחות בעיקרה מבוססת על ניסיונות לתקיפה ובמעט מקרים אירועים שהסתכמו בירי מרחוק על כלי השייט. ערכים חסרים בקטגוריית התקיפה המקורית סוננו החוצה. בכך הורדנו את מספר התצפיות הכולל מ-7511 ל-7375, והמשכנו רק עם תצפיות שבהן תוצאת התקיפה ידועה.

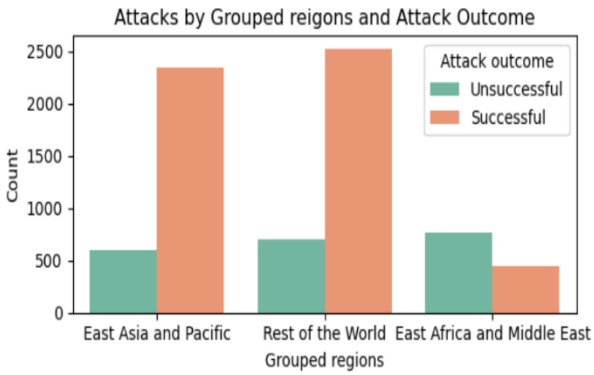
משתנים קטגוריים

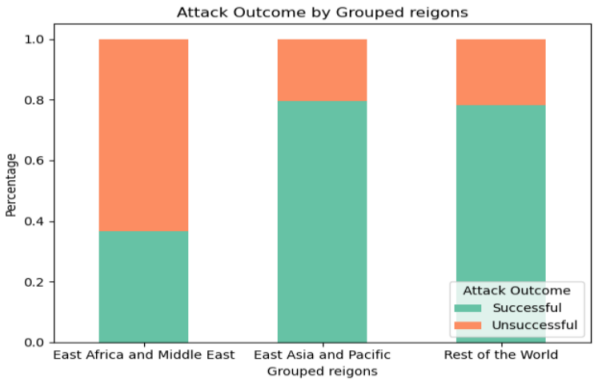
אזור גיאוגרפי

בסיס הנתונים מסווג לכל תקיפה את האזור הגאוגרפי לו שייכת המדינה בעלת השליטה על האזור הכלכלי הימי. ביצענו שינויים מעטים בהגדרת הקבוצות. שילבנו אזורים עם מעט תקיפות כדוגמת צפון אמריקה (5 תצפיות) אשר יחד עם אמריקה הלטינית והקריביים שולבו לקטגוריה "אמריקה". את אירופה (34 תצפיות) צירפנו לאזורי המזרח התיכון וצפון אפריקה לקטגוריה "המזרח התיכון ואירופה". משתנה זה כולל כ 6 קטגוריות: אמריקה, המזרח התיכון ואירופה, מערב אפריקה, מזרח אפריקה והמזרח התיכון, דרום אסיה ומזרח אסיה כפי שניתן לראות במפה. להמשך העבודה ראינו לנכון לצמצם את הקטגוריות לשלוש בלבד: "מזרח אפריקה והמזרח התיכון", "מזרח אסיה" ו"שאר העולם" (גרף 2).



גרף 1: מפת תקיפות הפיראטים בסיווג לפי אזור גיאוגרפי (custom region)



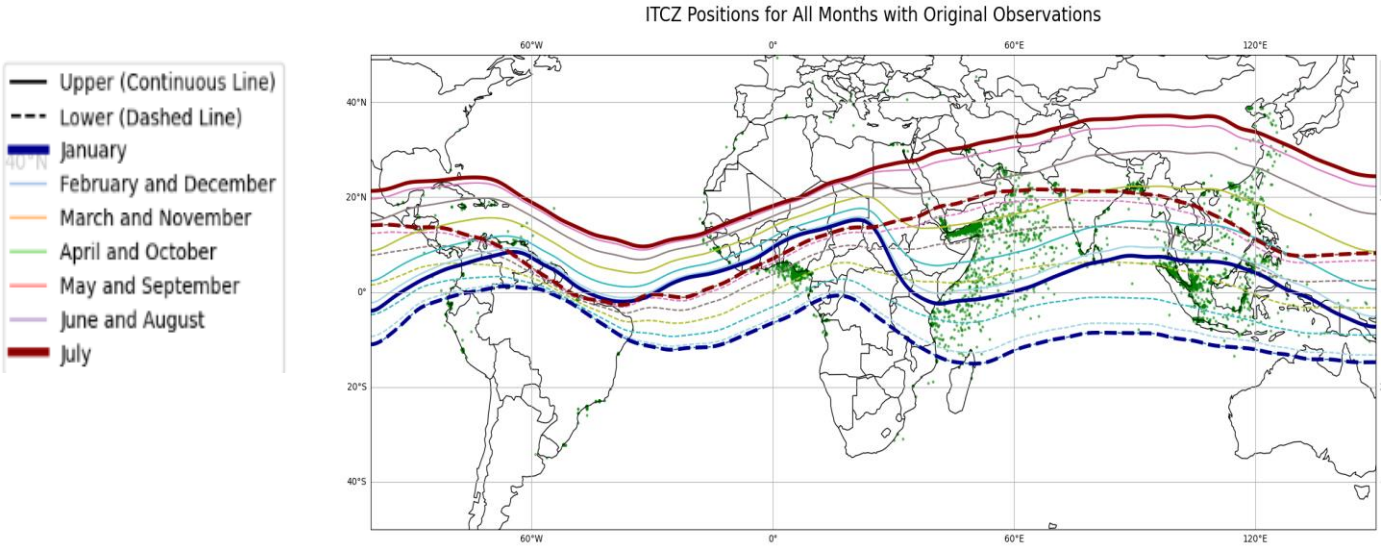


גרף 2: תקיפות פיראטים לפי תוצאת תקיפה ואזור גאוגרפי מקובץ (custom region2). בגרף הימני שיעור התקיפות המוצלחות והכושלות לכל אזור.

אזור אקלימי

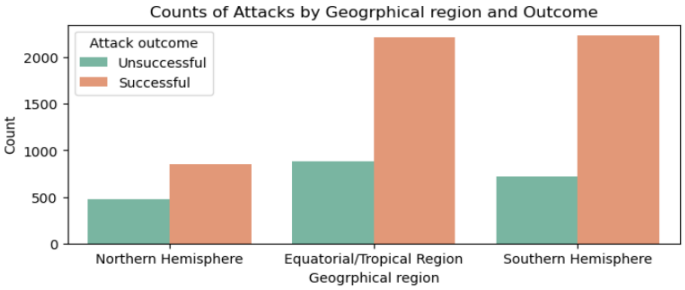
לטובת משתנה לייצוג הבדלי אקלים, יצרנו את המשתנים Geographical region המפריד לשלושה אזורי אקלים - אזור טרופי/קו משווה, חצי הכדור הצפוני וחצי הכדור הדרומי. החלטה זו נבעה מההכרה כי חלוקה לעונות לפי חודשים גם אם בשילוב עם המיקום ביחס לקו המשווה אינה משקפת נאמנה את המציאות האקלימית המורכבת.

השתמשנו ב"אזור ההתכנסות הבין-טרופי" - Inter Tropical Convergence Zone (ITCZ) - כבסיס לחלוקה גיאוגרפית עונתית. באזור זה, הידוע גם בשם "דולדרומים", הרוחות מחצי הכדור הצפוני והדרומי מתכנסות בקרבת קו המשווה נפגשות ו"מבטלות" זו את זו, וכך יוצרות אזור של רוחות חלשות מאוד או היעדר רוחות. אזור זה היווה מוקד חרדה עבור פיראטים וימאים סוחרים בתקופת המפרשיות בשל תנאי הרוח והאקלים הייחודיים שלו.



גרף 3: מיקום אזור ההתכנסות הבין-טרופי לפי חודשי השנה.

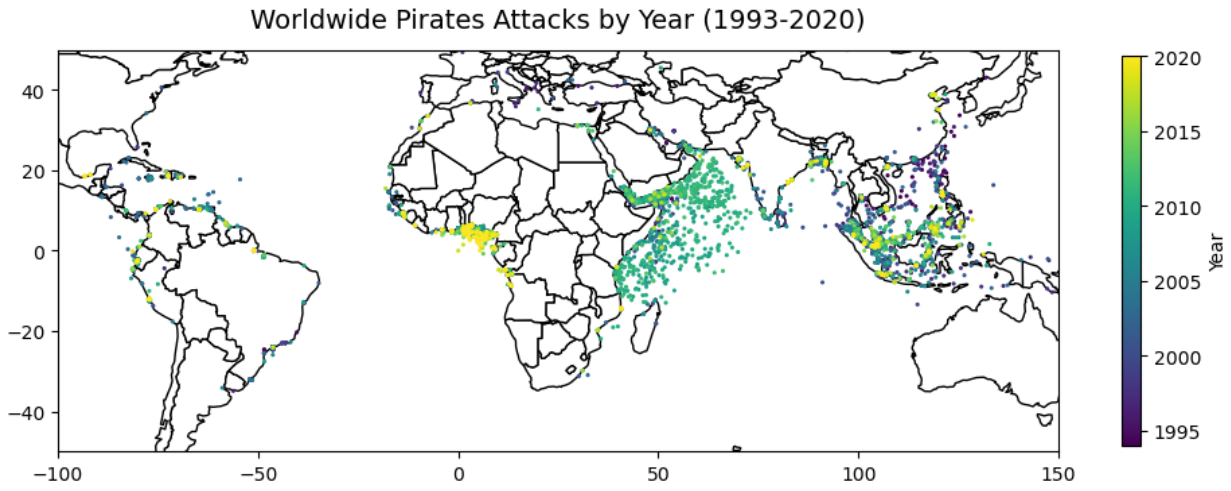
מיקומו של ה-ITCZ משתנה עונתית מכיוון שהוא עוקב אחר השמש, הוא נע צפונה מאזור יולי בקיץ של חצי הכדור הצפוני ודרומה מינואר כאשר בחצי הכדור הצפוני חורף (גרף 3). השמש חוצה את קו המשווה פעמיים בשנה, בחודשים מרץ וספטמבר, וה-ITCZ שעוקב אחריה גורם לשתי עונות רטובות בכל שנה. את נתוני ה-ITCZ הפקנו ממפות אקלימיות בעזרת חישוב הקואורדינטות לטווחי ה-ITCZ הממוצעים המוערכים לינואר וליולי. על מנת ליצור הערכה למיקום ה-ITCZ בחודשי הביניים עשינו שימוש באינטרפולציה, שיטה מתמטית להערכת ערכים בין נקודות מדידה ידועות. במקרה שלנו, אנו מנסים להעריך את מיקום ה-ITCZ בחודשים שבין ינואר ליולי. בעוד שאינטרפולציה לינארית מניחה שינוי קבוע בין שתי נקודות, פונקציה סינוסואידלית מייצרת עקומה חלקה יותר שמדמה את התנועה הטבעית וההדרגתית של ה-ITCZ. זה מאפשר לנו לקבל תמונה מדויקת יותר של מיקום ה-ITCZ הממוצע בכל חודש, תוך התחשבות בעובדה שצורתו וקצב התנועה שלו אינם אחידים לאורך השנה. מתוך כל אלה יצרנו טווחים המשמשים אותנו לסיווג האזורים הגיאוגרפיים בהתאם למיקומם הממוצע בכל חודש. בבחינת אזורים אלו על תוצאות התקיפה (גרף 4) ניתן לראות כי רוב התקיפות מתרחשות באזורים טרופים ובדרום שם גם שיעור התקיפות המוצלחות מעט גבוהה יותר בהשוואה לחצי הכדור הצפוני.



גרף 4: התפלגות התקיפות לפי אזור אקלימי.

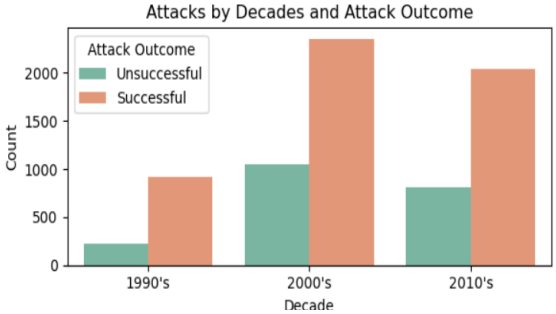
עשור (קטגוריאלי זמן)

יצרנו מיפוי של ציר הזמן לתקיפות (גרף 5). נראה כי במערב אפריקה (מפרץ גינאה) פשיטות פיראטים הינה תופעה יותר עכשווית, עיקר פעילות הפיראטים היא במזרח אפריקה (מפרץ עדן) תועדה לפני כעשור ובים הסיני רוב התקיפות התרחשו בשנות ה90. לעומת זאת, במזרח הרחוק, אינדונזיה היא אזור שורץ פיראטים לאורך כל התקופה שבה הנתונים זמינים לנו.



גרף 5: פריסה גלובלית של תקיפות פיראטים לפי השנה בה התרחשו.

עקב ממצאים אלו, יצרנו משתנה זמן "עשור" בעל 3 קטגוריות זמן: 1994-1999, 2000-2009, 2010-2020. ניתן לראות כי פחות בשנות ה90 (גרף 6).



גרף 6: התפלגות התקיפות על פני עשורים.

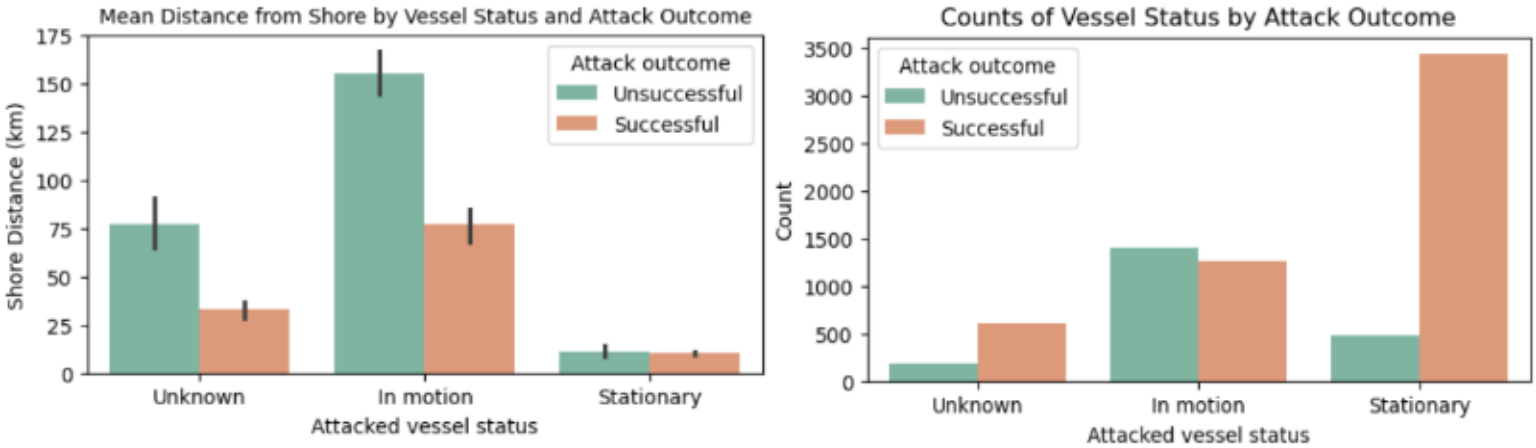
סטטוס כלי שיט

את המשתנה "סטטוס כלי שיט" המרנו לסיווג על בסיס מצבי תנועה אפשריים של כלי שיט המותקף, בין אם בתנועה כלשהי במים, מצב סטטי (עגינה) או שהסטטוס לא ידוע. המשתנה המקורי מכיל תיאור מצבים שונים המעידים על תנועה (הפלגה, דיג, גרירה, ועוד) לבין אלו המעידים על כלי שיט במצב סטטי (עגינה, חניה בחוף ועוד). ערכים שהיו חסרים או לא ידועים נשארו תחת קטגוריית " לא יודע"(Unknown).

Vessel status	Count	Our status	Count
Anchored	3250	In motion	3920
Berthed	653		
Stationary	14		
Moored	2		
Grounded	1		
Steaming	2582	Stationary	2662
Underway	56		
steaming	11		
Drifting	10		
Towed	1		
Fishing	1		
Bunkering operatior	1		
NaN	793	Unknown	793

בבחינת ההתפלגות לפי תוצאת תקיפה (גרף 7), ניתן להבחין כי ישנו הבדל משמעותי ביחס התקיפות ה"מוצלחות" לאלו הכושלות בין כלי שיט בתנועה לבין אלו במצב סטטי. בעוד אצל כלי השיט שבתנועה קיים הבדל קטן לטובת ניסיונות כושלים, אצל כלי השיט הסטטיים מעל 80% מהתצפיות במדגם הן תקיפות מוצלחות.

בעוד שאנו מצפים כי בממוצע כלי שיט בתנועה יהיה רחוק יותר מהחוף, אנו מעוניינים לבדוק האם זה שכלי שיט נמצא במצב סטטי זה בהכרח אומר שהוא עוגן בחוף. בהתבוננות על התפלגות סטטוס כלי שיט לפי מרחק מהחוף (גרף 7) נשים לב כי כלי שיט סטטי הוא במרחק ממוצע מהחוף של פחות מ 25 ק"מ, מה שמעיד כי קיימים מקרים רבים בהם כלי השיט אינו עוגן בצמוד לחוף. לעומת זאת, כלי שיט בתנועה נמצא בממוצע במרחק של 160 ק"מ. קו השגיאה בשחור מתאר את רווח הסמך, בחרנו להציג אותו כדי לנסות לזהות בצורה פשטנית הבדל בין תוצאות תקיפה בהתייחס למצב כל השיט.



גרף 7: בימין: התפלגות התקיפות לפי תוצאה ומצב כלי השיט הנתקף, בשמאל: שילבנו ממוצע ורווח סמך למרחק מהחוף לכל אחת מהקבוצות.

משתנים כמותיים

משתנים כמותיים נוגעים למרחק מהחוף בק"מ (נתון), מרחק מקו המשווה בק"מ (יצרנו מתוך קווי רוחב) או לפרמטרים של מדינה בשליטה על האזור הכלכלי הימי. "אזור הכלכלי הימי" ידוע גם בתור "מים כלכליים", הוא אזור ימי המשתרע עד 200 מיילים ימיים (370.4 ק"מ) מחוף מדינה, ובו למדינה זכויות מיוחדות להגנת הטבע כמו גם את הזכות הבלעדית לנצל או לשמור על כל המשאבים הנמצאים בתוך המים, על קרקעית הים או מתחת לקרקעית הים. משתנים השייכים למדינה כוללים מדדים כלכליים כמו תוצר מקומי גולמי (GDP) והכנסות ממשלתיות, מדדים דמוגרפיים כמו גודל אוכלוסייה, ומדדים חברתיים כמו מדד השחיתות ושיעור הרצח. בנוסף, ישנם משתנים הקשורים ישירות לתעשיית הדיג, לכוחות הצבא, ולתרומת התעשייה לתמ"ג. משתנים אלה נמצאים בסדרי גודל שונים מאוד - למשל, ה-GDP נמדד במיליארדי דולרים, בעוד שמדד השחיתות נע בין 0 ל-100, והתוצר שמקורו בתעשייה הוא שבר עשרוני (קטן מ1).

בטבלה מטה אנו מציגים סטטיסטיקה תיאורית של סט המשתנים הכמותיים. נראה כי קיימת שונות גדולה בתוך המשתנים עצמם, כפי שמשתקף בסטיות התקן הגבוהות. לדוגמה, במשתנה האוכלוסייה יש פער עצום בין המינימום למקסימום, מה שמעיד על הבדלים משמעותיים בגודל המדינות המעורבות. מעניין לציין כי במרבית המשתנים הממוצע גבוה מהחציון, דבר המצביע על התפלגות א-סימטרית עם זנב ארוך לכיוון הערכים הגבוהים, ומדגיש את הרגישות של הממוצע לערכי קיצון. זה נכון במיוחד למשתנים כמו GDP, דייג וגודל הכוחות הצבאיים.

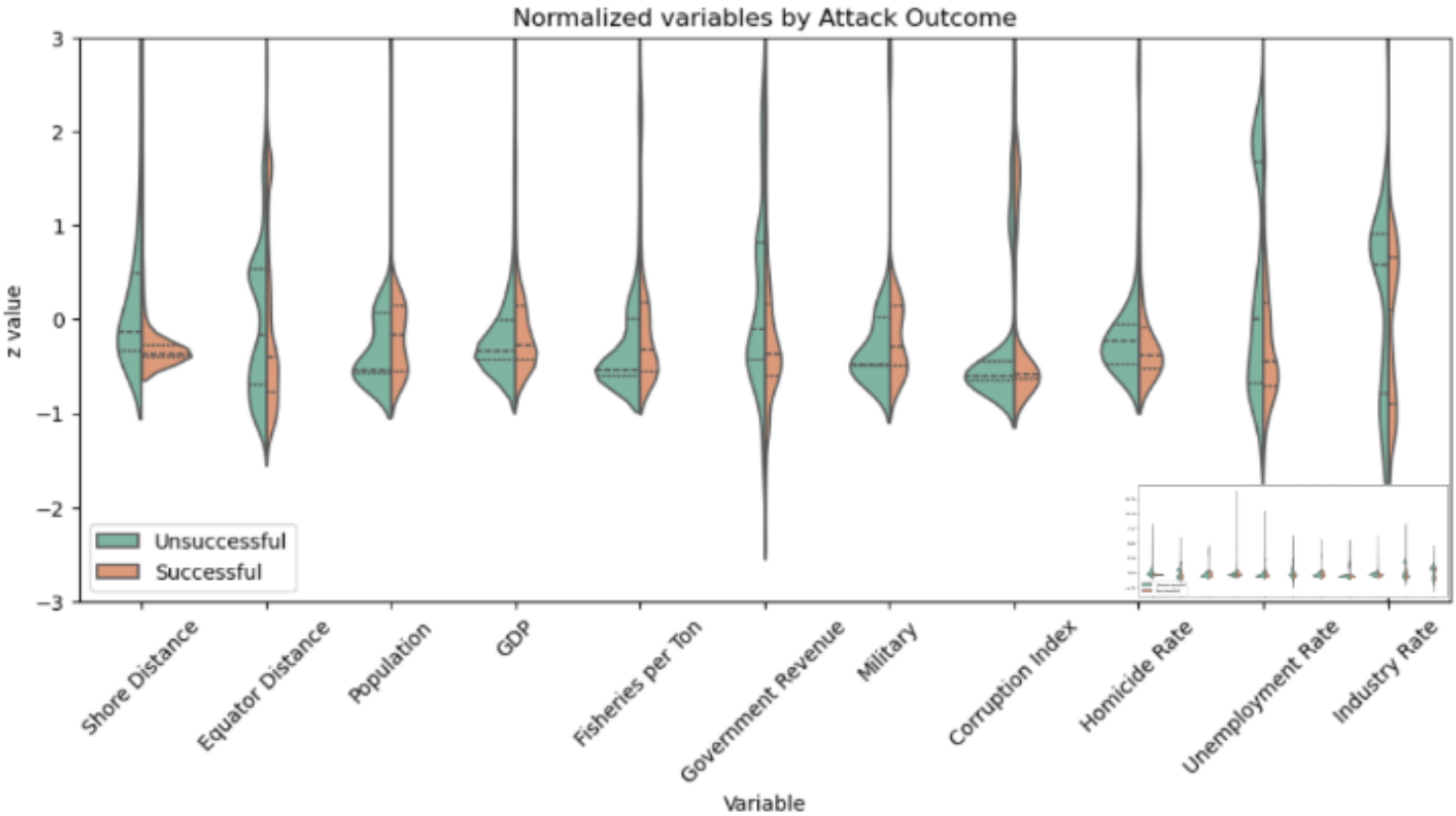
Successful Attacks	N	Mean	Median	std	Min	Max
corruption index	4449	12.52	3	15.34	0.4	87
homicide rate	3211	6.6	2.38	11.23	0.16	69.45
GDP	4945	2866.49	1573.88	4767.92	102.6	66188.78
total fisheries per ton	4645	5354969	2328545	8016169	320	81500000
total military	4657	516757.1	280000	719897.8	50	4135000
population	5062	197891400	136986400	289518300	76417	1.398E+09
unemployment rate	5046	5.62	4.51	3.06	0.58	31.84
government revenue	4143	0.18	0.16	0.08	0.01	0.67
industry of gdp	4927	0.35	0.37	0.11	0.04	0.84
shore distance	5303	28.65	7.06	80.17	0.04	1006.51
equator distance	5303	1028.33	677.79	923.12	0	5977.09

Unsuccessful Attacks	N	Mean	Median	std	Min	Max
corruption index	1583	8.77	2.65	12.33	0.4	85
homicide rate	1192	6.02	3.99	9.46	0.3	69.45
GDP	1731	2735.4	1229.25	4443.28	137.17	65233.28
total fisheries per ton	1702	3101570	636901	6190093	800	78800000
total military	1666	375534.8	149000	637541.3	200	4135000
population	1816	139024200	29866559	250833500	17955	1.393E+09
unemployment rate	1792	7.19	6.08	4	0.63	32.46
government revenue	1447	0.21	0.18	0.09	0.02	0.61
industry of gdp	1687	0.38	0.43	0.13	0.08	0.77
shore distance	2072	114.93	36.27	187.76	0.08	1024.03
equator distance	2072	1047.62	888.7	764.05	0	5975.27

בתקיפות המוצלחות (5,303) לעומת הלא מוצלחות (2,072), קיימים הבדלים משמעותיים במספר משתנים מרכזיים. למשל, מדד השחיתות הממוצע גבוה יותר בתקיפות מוצלחות (12.52 לעומת 8.77), מה שעשוי לרמוז על קשר בין רמת השחיתות במדינה להצלחת התקיפות. המרחק הממוצע מהחוף בתקיפות מוצלחות קטן משמעותית (28.65 ק"מ לעומת 114.93 ק"מ), מדגיש את חשיבות הקרבה לחוף ל"הצלחת" התקיפה. גם משתנים כלכליים ודמוגרפיים מראים הבדלים מעניינים. שיעור האבטלה נמוך יותר בממוצע במדינות בהן התקיפות מוצלחות (5.62% לעומת 7.19%). בתחום הדיג, ממוצע התפוקה גבוה משמעותית בתקיפות מוצלחות, מה שעשוי להצביע על קשר בין פעילות דיג אינטנסיבית לסיכויי הצלחה של תקיפות פיראטים.

חשוב לציין כי בכל המשתנים, הפער בין הממוצע לחציון גדול יותר בתקיפות המוצלחות, מה שמצביע על פיזור רחב יותר של הנתונים ואולי על השפעה חזקה יותר של ערכי קיצון. למשל, בתקיפות מוצלחות, החציון של מדד השחיתות הוא 3, בעוד הממוצע הוא 12.52, פער המדגיש את הא-סימטריות בהתפלגות. השונות הגדולה וקני המידה השונים של המשתנים מחייבים התייחסות מיוחדת בוויזואליזציה ובניתוח. לדוגמה, התמ"ג נע בין מאות למיליארדי דולרים, בעוד ששיעור האבטלה נמדד באחוזים בודדים. כדי להתמודד עם אתגרים אלו, נשתמש בפרק זה בזי סקור של המשתנים ובפרק הניתוח נפרט על טרנספורמציית לוג. אנחנו מוצאים את שיטות נרמול האלו כחיוניות במיוחד עבור משתנים כמו GDP, אוכלוסייה ותפוקת הדיג, שמראים פערים גדולים בין המינימום למקסימום ובין הממוצע לחציון.

באשר לצורות ההתפלגות של המשתנים הכמותיים (גרף 8). הדבר הבולט ביותר הוא התארכות הפעמונים כלפי מעלה, מה שמעיד על ערכים מאוד רחוקים מהחציונים. אנחנו משערים שמדובר בתקיפות שאירעו בשטחן של סין והודו שמובילות את המדינות במדדים הדמוגרפיים. בשטחה של אינדונזיה תועדו הכי הרבה ניסיונות תקיפה, לכן ייתכן וריכוז התצפיות בכל המשתנים למעט מרחק מייצג את המיקום היחסי של אינדונזיה בהשוואה לשאר המדינות ה"מותקפות". עבור משתני שיעור תעשיה, הכנסות הממשלה ואבטלה ההתפלגות די מפוזרת מה שאומר שיש הבדלים גדולים בין המדינות השונות במדגם, אך אין הבדלים מהותיים בין תקיפות מוצלחות ללא מוצלחות. במשתני המרחק לעומת זאת כן אפשר לראות שוני בצורת הפיזור של מרחקי התקיפות כאשר תקיפות מוצלחות מרוכזות יותר (כבר ראינו לפני בפרק המשתנים המסבירים כי יש יותר תקיפות מוצלחות בקרבת החוף).



גרף 8: בימין: חתך הגרף של התפלגות משתנים כמותיים לאחר נרמול. בגרף "ויולין" הקווים מציינים חציון ורביעים שני שלישי כך שחצי מהמדגם נמצא בין הקו המקווקו עליון לתחתון. בפינה הימנית מוצג הגרף במלואו (ערכו המקסימלי של זי סקור הוא 12.5).

ניתוח

פרמוטציות וקורלציות

מבחן הפרמוטציה זו שיטה סטטיסטית לא פרמטרית לבחינת השערות המאפשרת לנו לבחון האם יש הבדל בין שתי קבוצות על ידי נטרול הקשר המקורי בין המשתנה הנבחן למשתנה התלוי. במקרה שלנו, אנו בוחנים את ההבדל בין התקפות פיראטים מוצלחות לניסיונות התקפה כושלים - משתנה תוצאות התקיפה שלנו. הניתוח מתרכז במשתנים מספריים (כמו מרחק מהחוף, תמ"ג) על בסיס הפרש בין ממוצעים. לכל משתנה כמותי אנו מחשבים את ההפרש הנצפה בין הממוצעים לכל קבוצה מתוצאות התקיפה. לאחר מכן, מערבבים את ערכי המשתנה תוצאות תקיפה (מוצלחות\כושלות) בצורה אקראית בעוד מחזיקים את ערכי המשתנים המספריים קבועים. עבור כל ערבוב אנחנו מחשבים מחדש את ההפרש בין הממוצעים וחוזרים על כך כ10,000 פעם. לבסוף, אנו בוחנים בעזרת ה p-value באיזו תדירות אותם הפרשים היו זהים או קיצוניים יותר (מבחן דו-צדדי) מההפרשים הנצפים המקוריים.

Variable	Successful Attacks	Unsuccessful Attacks	Successful Attacks Mean	Unsuccessful Attacks Mean	Means difference	Perm p.value
shore distance	5303	2072	28.65	114.93	-86.29	0.0000
equator distance	5303	2072	1028.33	1047.62	-19.29	0.3935
GDP	4945	1731	2866.49	2735.4	131.08	0.3065
population	5062	1816	197891400	139024200	58867159.45	0.0000
corruption index	4449	1583	12.52	8.77	3.74	0.0000
homicide rate	3211	1192	6.6	6.02	0.58	0.1070
total military	4657	1666	516757.1	375534.8	141222.3	0.0000
unemployment rate	5046	1792	5.62	7.19	-1.57	0.0000
total gov revenue	4143	1447	0.18	0.21	-0.03	0.0000
industry of gdp	4927	1687	0.35	0.38	-0.03	0.0000
total fisheries production	4645	1702	5354969	3101570	2253398.95	0.0000

ניתן לראות כי עבור המשתנים מרחק מקו המשווה, תמ"ג ושיעור הרצח ה p-value הגבוהה מרמז כי הקשר הינו חלש או לא קיים שכן ניתן לקבל הפרשים זהים או קיצוניים יותר במקרים רבים בצורה אקראית. לעומק זאת לשאר המשתנים: מרחק מהחוף, אוכלוסייה, מדד שחיתות, כוחות צבא", אבטלה, הכנסות הממשלה, "חס התעשייה בתמ"ג ודייג, ה p-value שווה אפס ומרמז על כך שהקשר אינו אקראי וכי משתנים אלו עשויים להיות מהותיים לתוצאות התקיפה.

קורלציות בין משתנים

בשלב הבא אנו מציגים מפת חום של קורלציות ספירמן (גרף 9). בחרנו במבחן זה שכן אינו מניח התפלגות נורמלית, אינו ליניארי אלא מונוטוני ואינו רגיש לערכי קיצון , בניגוד לקורלצית פירסון. אנו בוחנים התאמה בין המשתנים השונים כדי לאתר קשרים שיסייעו לנו לצמצם את מספר המשתנים המסבירים במודל. ע"פ הטבלה ניתן לראות קורלציה חזקה בין אוכלוסייה, גודל צבא ותפוקה מדייג (90%-81%). לכן בהמשך נתייחס אל אוכלוסייה כמשתנה מייצג מאחר ורמת ההתאמה לאחרים היא החזקה ביותר (90% ,88%).

בסיכום פרק זה החלטנו להתמקד במשתנים המספריים הבאים לטובת ניבוי תוצאת תקיפה לא מוצלחת הרגרסיה הלוגיסטית: מרחק מהחוף, גודל האוכלוסייה, מדד שחיתות, שיעור אבטלה, הכנסות הממשלה ושיעור התוצר שמקורו בתעשייה. בנוסף, יצרנו משתני דמי עבור קבוצות קטגוריאליות שאנחנו רוצים לבדוק קשר בינם לבין המשתנה התלוי שלנו. המשתנים שעברו טרנספורמציה דיכוטומית

גרף 9: טבלת קורלציות לאינדיקטורים של המדינות במדגם תקיפות. מבוסס על קורלציות ספירמן לבחינת טיב קשר מונוטוני בין משתנים.

הם : עשור, עונה גיאוגרפית, אזור גיאוגרפי וסטטוס של כלי השיט המותקף. קטגוריות הייחוס שהגדרנו עבור כל אחד מהמשתנים הקטגוריאליים שציינו הם העשור הראשון "1990", עונת ה"אביב", אזור "מזרח אפריקה ומזרח תיכון" וסטטוס כלי שיט "בתנועה" בהתאמה.

מודל

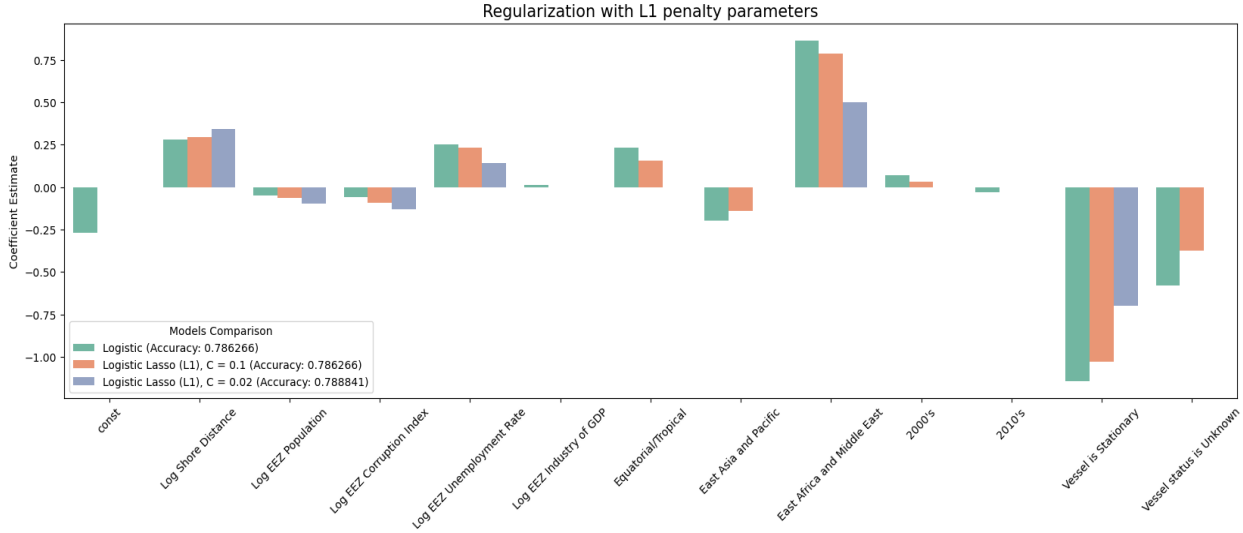
רגרסיה לוגיסטית לניבוי תקיפה כושלת

על מנת לחזות את הסיכוי לכישלון של תקיפת פיראטים ביצענו רגרסיה לוגיסטית בעזרתה נרצה להבין אילו גורמים קשורים לתוצאת התקיפה. המשתנה התלוי `outcome`, מקבל את הערך 1 עבור ניסיון תקיפה כושל ומקבל 0 עבור תקיפה "מוצלחת". על מנת להתמודד עם הטיות אפשריות ופערים הנובעים מקנה מידה בין המשתנים, השתמשנו בטרנספורמציה הלוגריתמית log)) למשתנים רציפים. אנחנו מאמינים שהטרנספורמציה הלוגריתמית דומה להמרה לערכי z בהיבט היכולת שלנו לבחון את המודל בהמשך תוך השוואת ערכי המשתנים השונים .

```
                           Results: Logit
==============================================================================
Model:                  Logit               Method:              MLE
Dependent Variable:     outcome             Pseudo R-squared:    0.180
Date:                   2024-08-15 14:03    AIC:                 5432.4598
No. Observations:       5821                BIC:                 5519.1598
Df Model:               12                  Log-Likelihood:      -2703.2
Df Residuals:           5808                LL-Null:             -3297.0
Converged:              1.0000              LLR p-value:         8.0884e-247
No. Iterations:         6.0000              Scale:               1.0000
------------------------------------------------------------------------------
                             Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
------------------------------------------------------------------------------
const                       -0.7674   0.4985  -1.5395 0.1237 -1.7444  0.2096
Log Shore Distance           0.2721   0.0278   9.8035 0.0000  0.2177  0.3266
Log EEZ Population          -0.0431   0.0254  -1.6950 0.0901 -0.0930  0.0067
Log EEZ Corruption Index    -0.0809   0.0441  -1.8372 0.0662 -0.1673  0.0054
Log EEZ Unemployment Rate    0.2456   0.0756   3.2462 0.0012  0.0973  0.3938
Log EEZ Industry of GDP     -0.0467   0.0921  -0.5072 0.6120 -0.2272  0.1338
Equatorial/Tropical          0.1793   0.0702   2.5555 0.0106  0.0418  0.3169
East Asia and Pacific       -0.1220   0.1015  -1.2016 0.2295 -0.3209  0.0770
East Africa and Middle East  0.9186   0.1235   7.4362 0.0000  0.6765  1.1607
2000's                       0.1575   0.1306   1.2059 0.2279 -0.0985  0.4136
2010's                       0.0918   0.1534   0.5987 0.5494 -0.2088  0.3925
Vessel is Stationary        -1.1529   0.0880 -13.1027 0.0000 -1.3253 -0.9804
Vessel status is Unknown    -0.5091   0.1310  -3.8871 0.0001 -0.7657 -0.2524
==============================================================================
```

רגולריזציה

על מנת להימנע מהתאמת יתר של המודל, עשינו שימוש במודל רגולריזציה של פונקציה לוגיסטית, הגדרנו את גודל פרמטר הענישה כדי לבחון את השינוי במקדמי הביטא של כל משתנה (גרף 10). מאחר רמת הדיוק של המודל כמעט ולא נפגעה (0.786 = Accuracy) בחרנו להשתמש בענישה חזקה יותר כדי להשאיר מחוץ למודל משתנים שהשפעתם חלשה כמו משתני הדמי לעשורים, חלק התעשייה בתוצר של מדינה שולטת, מצב לא ידוע של כלי השיט (בהשוואה לכלי שיט בתנועה) ומשתנים גיאוגרפים חלשים.



גרף 10: השוואת את ערכי המקדמים בכל מודל; בטורקיז המודל הרגיל ללא קנס. בכתום מודל l1 עם "קנס" נמוך

תוצאה סופית ומסקנות

המודל הסופי לניבוי ניסיון תקיפה כושל של פיראטים לוקח בחשבון פרמטרים של מדינה אשר בשטחה הכלכלי יתרחש ניסיון התקיפה. ע"פ תוצאות המודל בטבלה מטה, אנחנו מוצאים קשר שלילי בין גודל האוכלוסייה לסיכוי שניסיון התקיפה יכשל. מצאנו גם קשר שלילי בין מדד שחיתות במדינה לסיכוי לניסיון כושל לתקיפה. באופן מפתיע, ישנו קשר חיובי בין שיעור האבטלה במדינה לבין אותו סיכוי לתקיפה כושלת, זאת על אף ששיעור האבטלה נמוך יותר בממוצע ובחציון במדינות בהן התקיפות מוצלחות . באשר לפרמטרים הקשורים במיקום ומצב כל השיט, יש קשר חיובי בין מרחק מהחוף לסיכוי לכישלון התקיפה. להמצאות כלי השיט שלכם במצב סטטי בים או על החוף יש קשר חיובי לתקיפה מוצלחת. מפתיע לגלות שהאזור הכי סימפטי לשוט בו הוא מפרץ עדן ומזרח אסיה שעל אף ריבוי מקרי התקיפה מצא עם קשר חיובי לתקיפה כושלת.

Results: Logit

Model:	Logit	Method:	MLE
Dependent Variable:	outcome	Pseudo R-squared:	0.180
Date:	2024-08-15 13:44	AIC:	5432.4598
No. Observations:	5821	BIC:	5519.1598
Df Model:	12	Log-Likelihood:	-2703.2
Df Residuals:	5808	LL-Null:	-3297.0
Converged:	1.0000	LLR p-value:	8.0884e-247
No. Iterations:	6.0000	Scale:	1.0000

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	-0.4870	0.3585	-1.3584	0.1743	-1.1896	0.2157
Log Shore Distance	0.3034	0.0210	14.4669	0.0000	0.2623	0.3445
Log EEZ Population	-0.0503	0.0190	-2.6511	0.0080	-0.0875	-0.0131
Log EEZ Corruption Index	-0.1445	0.0257	-5.6232	0.0000	-0.1949	-0.0942
Log EEZ Unemployment Rate	0.1276	0.0615	2.0754	0.0379	0.0071	0.2481
East Africa and Middle East	0.8875	0.0896	9.9019	0.0000	0.7118	1.0631
Vessel is Stationary	-0.9812	0.0726	-13.5154	0.0000	-1.1235	-0.8389

מסקנותינו מהניתוח הם שלמכלול הפרמטרים הנוגעים למיקום יש קשר ברור לתוצאת התקיפה בין אם למיקום הפיזי של כלי השיט ביחס ליבשה ובין אם ביחס למדינה ויכולתה להתמודד עם התופעה בשטחה. רצוי שלא לשוט בקרבת מדינות הנתפסות כמושחתות ואם אין ברירה אז כדאי לשמור מרחק ביטחון מהחוף ולהמשיך בתנועה מתמדת. גילינו גם שמפרץ עדן בחצי האי ערב הוא מקום יחסית לא מאיים לשוט בו, אנחנו יכולים לשער שהסיבה לכך טמונה בהיות המפרץ נתיב סחר עולמי חשוב ומאובטח על ידי גורמים שלא נוגעים למדינות האזור בהכרח. התגלית המפתיעה מזכירה לנו שיש המון מידע שאין לנו ותקיפת פיראטים נובעת מצורך מסוים. לכן סטטיסטית נכון יותר לחקור את הסיכוי למתקפה על פני דאטה הכולל תיעוד של ספינות שהותקפו לצד ספינות שלא הותקפו. היבט אחר ממנו למדנו המון במהלך העבודה נוגע לפן הגיאוגרפי כאשר חקרנו את אזור הICTZ אשר תנאי השיט בו קשים ויוצרים חשש של ממש של ספינות רבות להיתקע באזור ללא רוח.

ובנימה קלילה יותר, אנו חבים תודה גדולה לקפטן פנחס על מקור נוסף למוטיבציה למחקר

https://www.youtube.com/watch?v=SPfwZ9-tBxE