מיני פרוייקט בקורס מבוא לניתוח נתונים בפייתון

חלק 4 - Communication and Reflection

QUESTIONS:

השאלות הבאות עניינו אותנו כשבחרנו בdataset:

1. האם יש קשר בין כמות הקורבנות בתאונה לבין היום בו היא התרחשה?
2. האם תאונות המתרחשות בסופי שבוע קטלניות יותר מאלה המתרחשות באמצע שבוע?
3. האם לשעת הנסיעה ביום יש השפעה על הסיכוי להיקלע לתאונה?
4. מהו המספר השכיח ביותר לכמות כלי רכב המעורבים בתאונה אחת?
5. האם יש שכונות בברצלונה בהן הסיכוי להיקלע לתאונה הינו גדול יותר?
6. האם תאונות שמתרחשות ביום גובות יותר קורבנות מאשר תאונות המתרחשות בלילה?
7. האם באמצעות איסוף מידע על שעת התאונה, כמות ואופי הקורבנות ומספר כלי הרכב המעורבים בה, ניתן לחזות באיזה חלק בשבוע היא התרחשה?

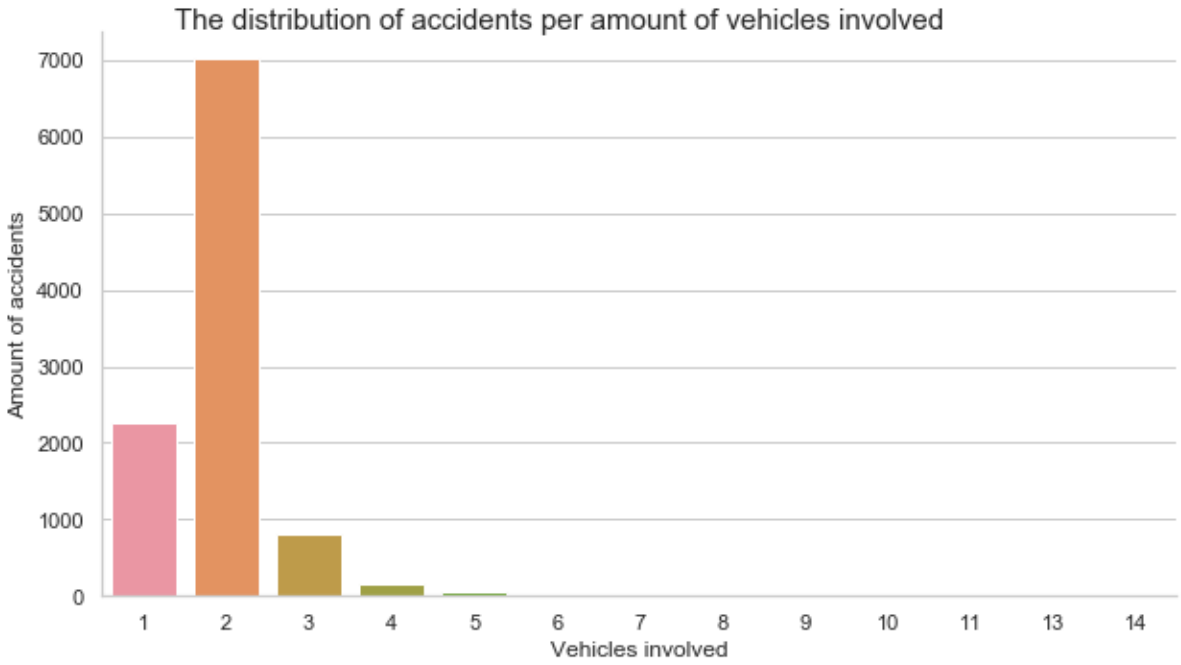
שאלות אלו עוררו אצלנו עניין כיוון שבמידה וניתן יהיה לענות עליהן ברמת דיוק מספקת, יש סיכוי טוב כי באמצעות שימוש נכון במידע הנאסף, יתאפשר להוריד את מספר הקורבנות של תאונות דרכים, ואף למנוע את התאונה הבאה. למשל, במידה ונסיק כי בשעות מסוימות מתרחשות תאונות רבות יותר , הרשויות יוכלו לפרסם הנחיות והמלצות לשעות בהן עדיף שלא לעלות על ההגה.

DATASET:

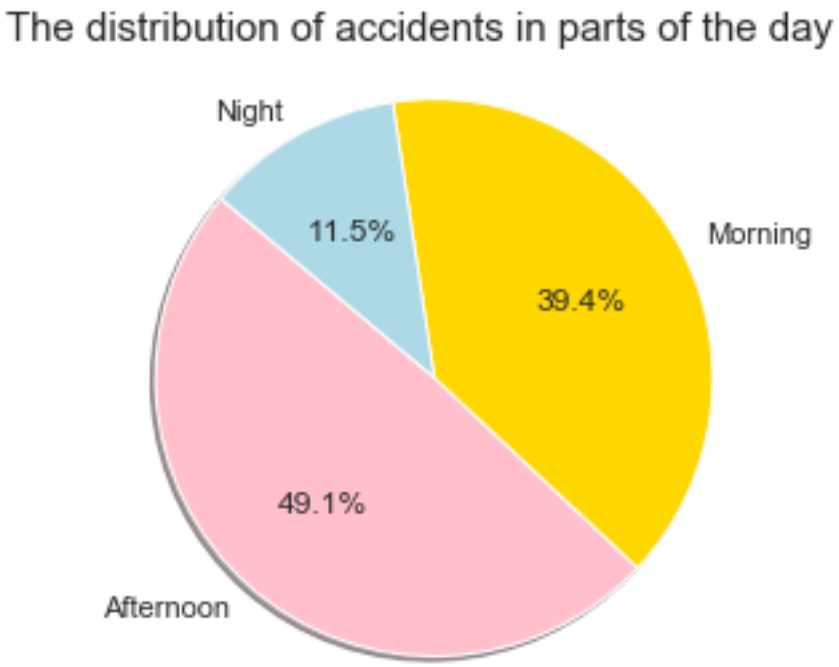
ה dataset שנבחר הוא על תאונות הדרכים שטופלו בידי המשטרה בברצלונה בשנת 2017.
ב dataset נמצא פירוט על מועד התרחשות התאונה (שעה, חלק ביום, יום בשבוע, יום בחודש וחודש), מספר הפצועים (על פי חומרת הפציעה: קלה, קשה), מספר הקורבנות, מספר כלי רכב מעורבים, מיקום התרחשות התאונה (רחוב, מחוז, שכונה, קווי אורך וקווי רוחב) ומספר הזהות של התאונה.

ANALYSIS & FINDINGS:

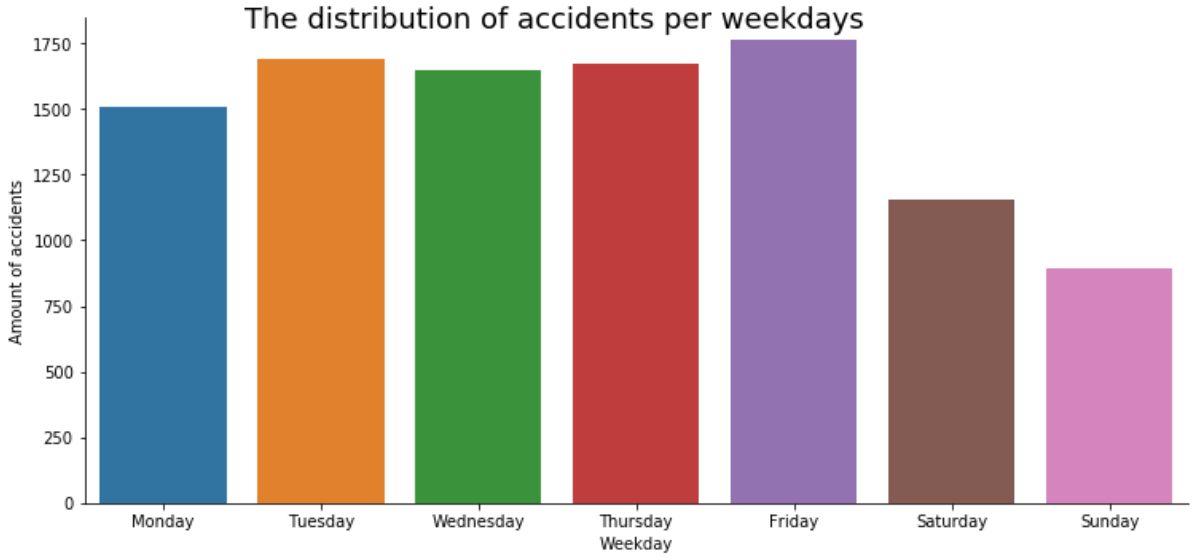
1. אספנו מידע על התאונות לפי כמות רכבים מעורבים. ניכר כי תאונות בהן מעורבים שני רכבים בלבד הן השכיחות ביותר.



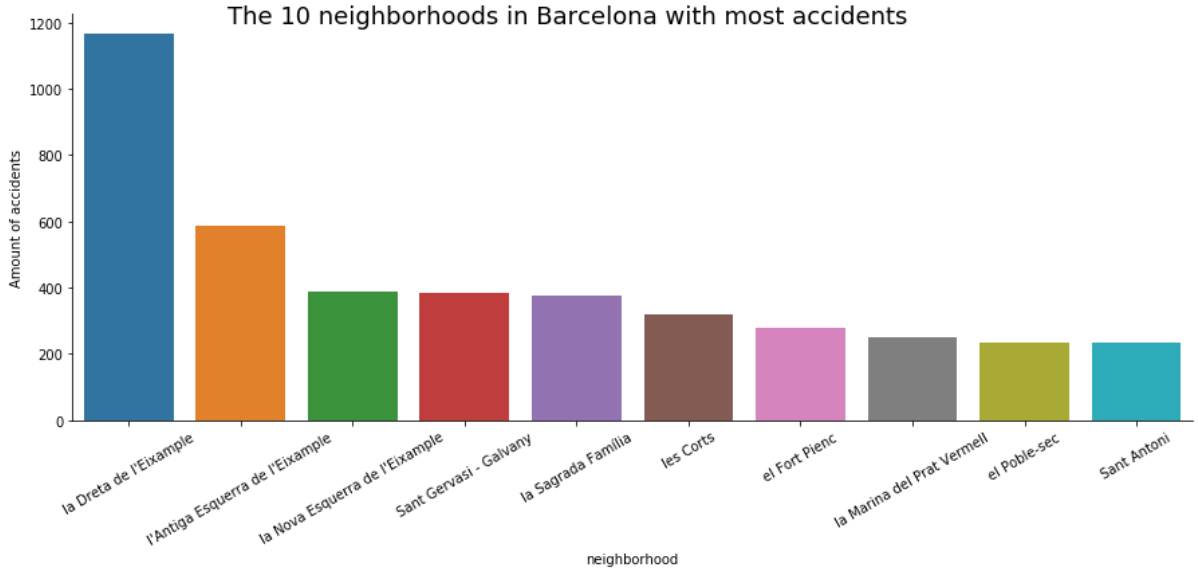
2. מגרף זה ניתן להסיק כי הסיכוי הכי גדול להיקלע לתאונה הוא בשעות אחה״צ.



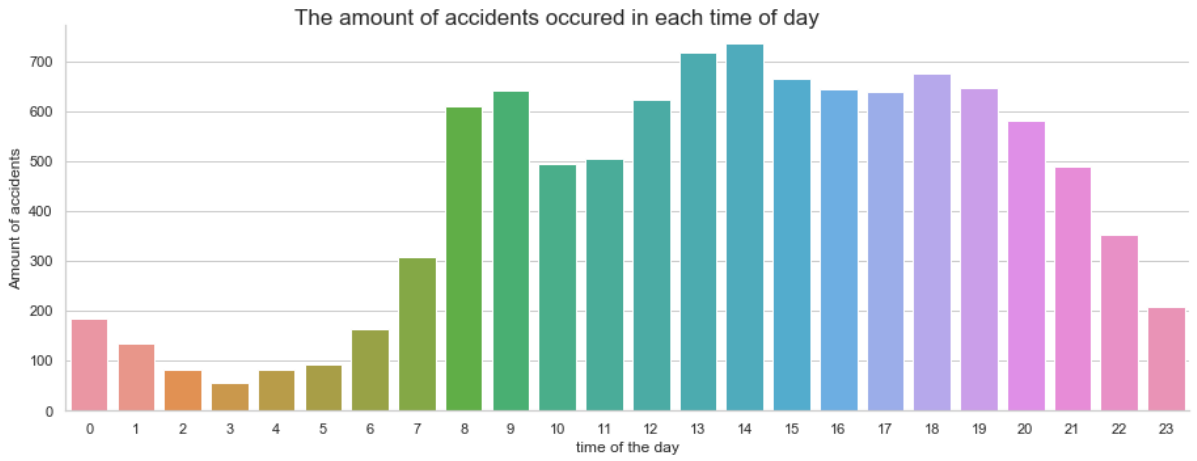
3. גרף זה מצביע על ירידה במספר התאונות המתרחשות בסופי שבוע לעומת אלה המתרחשות במהלכו.



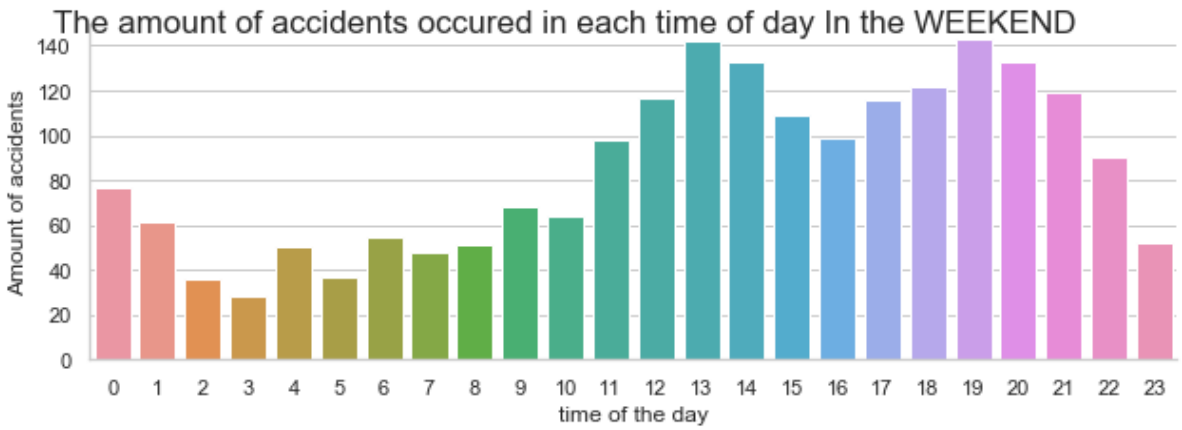
4. נראה כי יש שכונה אחת שבאופן קיצוני מתרחשות בה יותר תאונות מבשאר השכונות. הדבר נראה לנו מעורר ספק, כיוון שייתכן שזה כי זו שכונה גדולה במיוחד, או שנאספו עליה יותר דגימות. מאידך, ייתכן כי הנתון הזה משקף את המציאות וזו אכן שכונה עם תאונות רבות יותר עקב נהיגה לא זהירה או בעיית תשתיות.

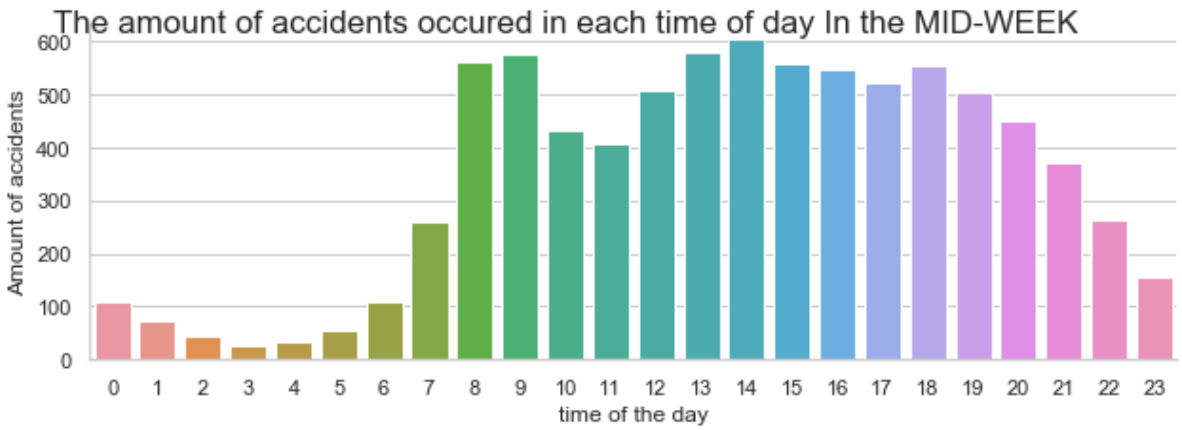


5. אנחנו רואים כי בשעות הבוקר - 7 עד 8 יש עלייה חדה במספר התאונות, ולאחר מכן ירידה, ואז שוב עלייה חדה החל מהשעה 12 בצהריים. הכמות יורדת משמעותית לאחר השעה 20 בערב, כאשר בלילה מתרחשות הכי מעט תאונות.
מעניין לראות האם יש הבדל בין שעות מרובות תאונות באמצע השבוע ובסוף השבוע.



החלטנו להפריד את הנתונים ולבדוק מה הן השעות מרובות התאונות בסוף השבוע לעומת באמצע השבוע:



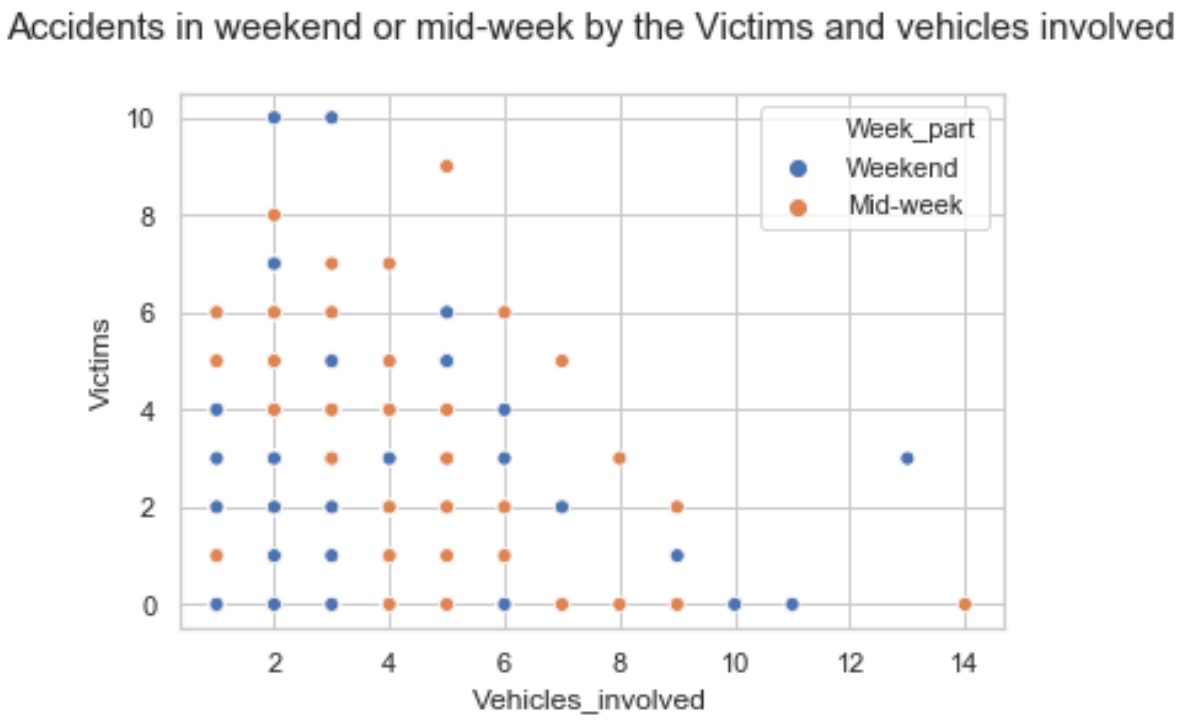


אנחנו רואים הבדל בשעות מסויימות - למשל השעות 7-8 בבוקר והשעה 16:00 אחה"צ שרוויות תאונות במהלך השבוע אך לא בסוף השבוע.

ההבדל העיקרי בין הגרפים הוא כמות התאונות המתרחשות, זאת בהמשך למה שראינו בגרפים הקודמים - הרבה פחות תאונות מתרחשות בכל השעות בסוף השבוע.

* בחרנו להשאיר את הסקאלות של הגרפים באופן מותאם לכל גרף ולא לאזן אותם. הבחירה נובעת מהרצון לראות איזה שעות רוויות תאונות ביחס לשעות אחרות באותם ימים ולא להעמיס על העיניים בכמות התאונות היחסית לשאר הימים.

6. בגרף הזה אנחנו רואים תאונות לפי כמות הרכבים המעורבים בתאונה לעומת כמות הנפגעים בתאונת הדרכים (הנפגעים = פצועים קל + פצועים קשה), מסווגות לפי מתי מרבית התאונות עם מספרים כאלה התרחשו : אמצע השבוע או סוף השבוע. מעניין אותנו לבדוק האם אכן ניתן לסווג תאונות אם הן התרחשו באמצע/סוף השבוע לפי נתונים אלו ואחרים.



LIMITATIONS:

1. אחת המגבלות בניתוח שביצענו היא הימים באמצע שבוע שהם למעשה ימי חג ובהם אנשים, בדומה לסופי שבוע, אינם יוצאים לעבודה. מצב זה מגדיר ימים אלו כאמצ"ש ולכן עשוי להחליש את הסיווג לפי סוף השבוע או אמצע שבוע.
2. מגבלה נוספת הינה הסיווג לפי בוקר, אחה"צ ולילה. מה DATASET נקבע כי החלוקה בוצעה באופן הבא: בוקר: 06:00-13:00, אחה"צ: 14:00-21:00 ולילה: 05:00-21:00. היות וחלוקה זו אינה אחידה עבור כל בני האדם (אחד יכול להגדיר בוקר עד השעה 12:00 למשל) היא עלולה להטעות בני אדם בהקשר של החלק ביום בו הכי מסוכן לנהוג.

FUTURE DIRECTIONS:

מגרף 5 ניתן לראות כי הפער בין מספר התאונות מהשעה 07:00 עד 20:00 אינו גדול במיוחד משעה לשעה. עם זאת, לא נוכל להסיק על אחוז המכוניות שנקלעות לתאונה ביחס לכלל המכוניות שנמצאות באותה שעה על הכביש. כלומר, נצטרך DATASET נוסף שיצביע על כמות הרכבים שיש על הכביש בכל שעה ביום בכדי להבין האם שעה מסוימת ביום מסוכנת יותר לנהג משעה אחרת.

שאלה חדשה שעלתה לנו היא האם מעצם זה כי בשכונה מסוימת מתרחשות תאונות רבות יותר מאשר בשכונות אחרות, נוכל להבין, באמצעות הצלבה עם מידע על התוכנית ההנדסית של השכונה, האם תכנון הנדסי אחד בטיחותי יותר בהיבט התחבורתי, מאשר תכנון הנדסי אחר. לדוגמא, אם ניקח DATASET נוסף בו מידע על רוחב וסוג הכבישים בשכונה (חד סטרי, מהיר, ללא מוצא וכו'), מספר מעברי החצייה, צמתים מתומרים או מרומזרים וכדומה, נוכל למצוא קשר בין מידת ההתאמה של השכונה לנהיגה לבין מספר התאונות המתרחשות בה.

חלק בונוס

במהלך חקר הנתונים, ראינו כי יש שכונה אחת שבה כמות התאונות חריגה במיוחד. את הניתוחים בעבודה עשינו על כל ברצלונה, אבל כל הזמן סיקרן אותנו כיצד ייתכן שכ9% מהתאונות התרחשו בשכונה אחת בלבד.

לכן, רצינו להבין מה הם המאפיינים של השכונה la Dreta de l'Eixample ומה מייחד אותה, ולאחר מכן לתת המלצות לעירייה או למשטרה כיצד להפחית את כמות התאונות שמתרחשות בה.

שלב 1:

העלינו השערות לסיבות שבגללן כמות התאונות בשכונה גבוהה כל כך:

1. מקום מרכזי או עמוס בבני אדם - מקומות תעסוקה רבים, כמות גדולה של אוכלוסייה.
2. איכות ירודה של תשתיות.
3. טעות באיסוף הנתונים - חוסר בנתונים על תאונות בשכונות אחרות מה שגורם להטייה לגבי כמות התאונות בשכונה.

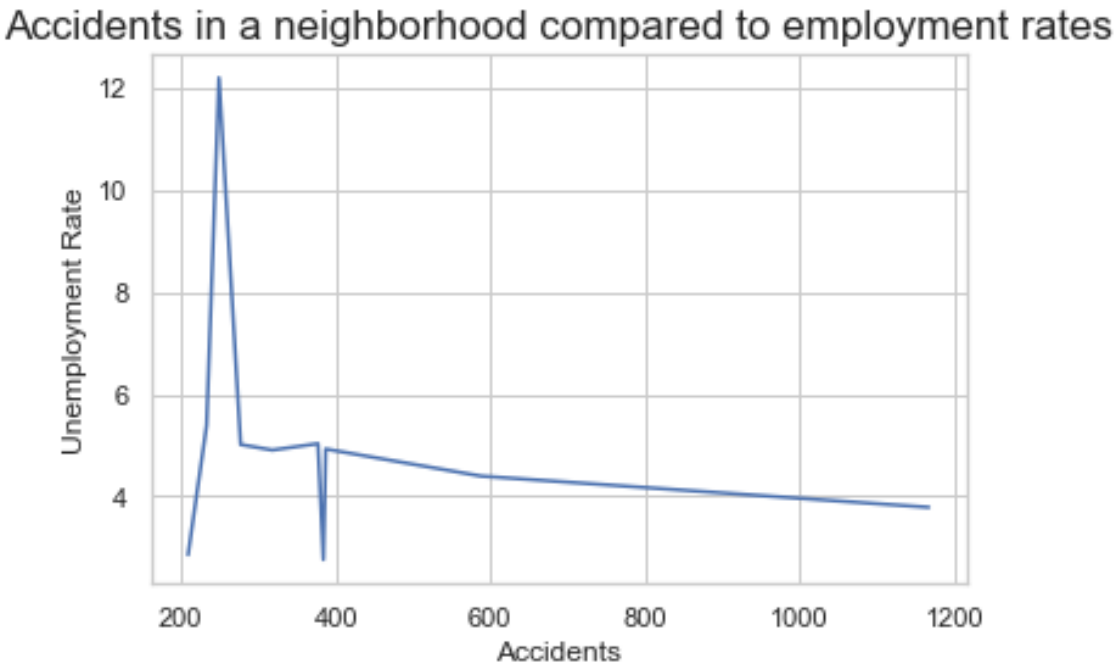
רצינו לבדוק את הסיבה הראשונה בגלל שהיו לנו נתונים על כמות אוכלוסיה בכל שכונה ונתוני אבטלה לאורך שנים בכל שכונה. את ניתוח הנתונים ביצענו ביחס לשכונות אחרות כדי שנוכל להסביר על נתון מסוים(אחוזי אבטלה) האם הוא נמוך או גבוה באופן יחסי.

השתמשנו בשני דטפריימים נוספים מאתר KAGGLE (מצ"ב בתיקייה) ואיחדנו אותם בעזרת עיבוד הנתונים לדטאפריים שכולל את הפיצ'רים הבאים:

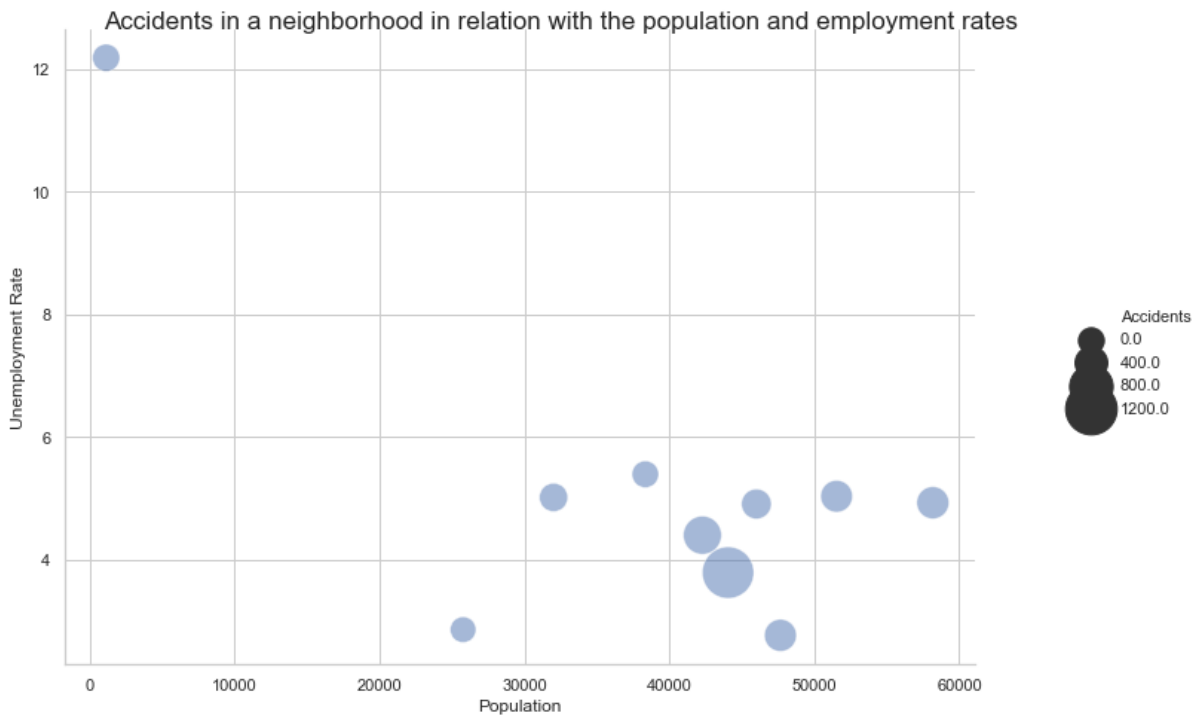
- Neighborhood_Name: משתנה מסוג STRING המתאר את שם השכונה.
- Population: משתנה מסוג INT המתאר את כמות אוכלוסייה בשכונה.
- Unemployment: משתנה מסוג INT המתאר את מספר המובטלים בשכונה.
- Accidents: משתנה מסוג DOUBLE המתאר את כמות התאונות שהתרחשו בשכונה.
- Unemployment rate: משתנה מסוג DOUBLE המתאר את אחוז האבטלה משכונה.

שלב 2:

רצינו לבדוק את הקשר בין אחוזי האבטלה בכל שכונה לבין כמות התאונות המתרחשות בה.



מגרף זה ראינו שאומנם ב6 השכונות בעלות מספר התאונות הגבוה ביותר ככל שאחוז האבטלה נמוך יותר שיעור התאונות גבוה יותר, אבל אין משמעות רבה למסקנה זו כיוון שמדובר במספר קטן מדי של שכונות והשינויים באחוזי האבטלה משכונה לשכונה הינו שולי(ירידה מ5% ל3%).

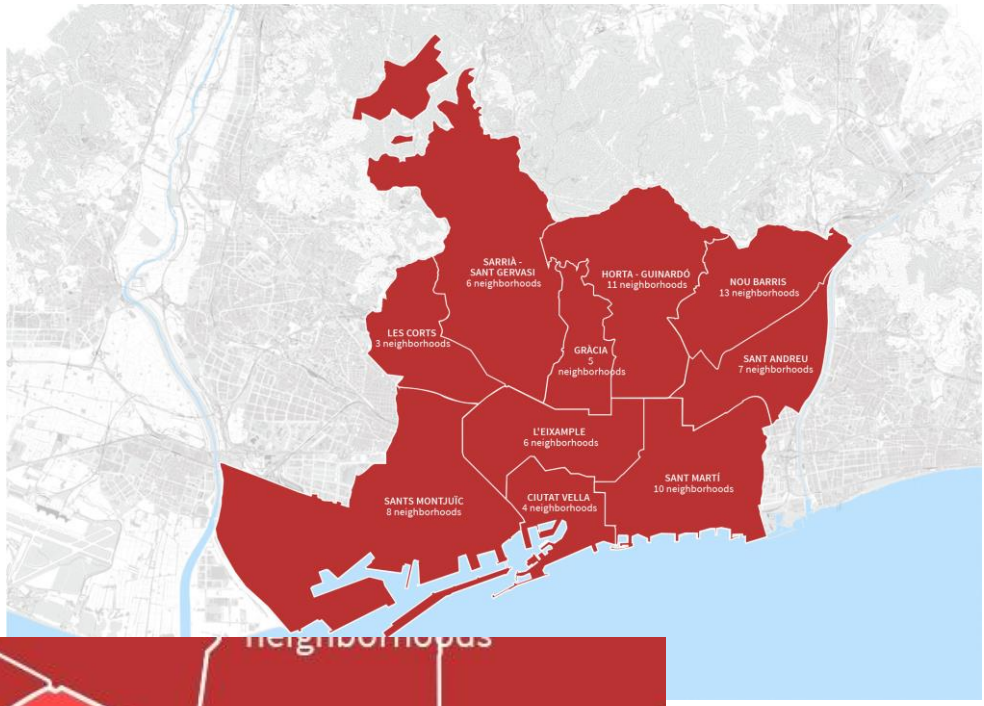


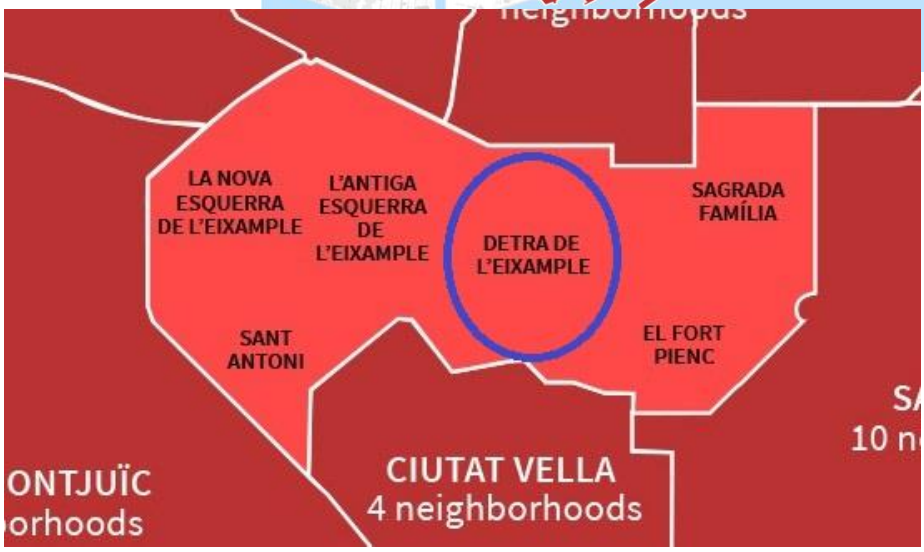
מהגרף ניתן לראות כי השכונה la Dreta de l'Eixample היא מהשכונות היותר גדולות בברצלונה ובעלת אחוזי האבטלה מהנמוכים ביותר.

שלב 3:

רצינו להצליב את המידע עם מידע נוסף מאתרי אינטרנט. חשבנו לראות ביקורות על השכונה ולנתח את הטקסטים. הבנו שהביקורות באנגלית (מאתרים כמו VISIT BARCELONA,TRIP ADVISOR) הן בהכרח חיוביות ומדברות על אטרקציות בשכונה ולא נוכל להפיק מהם מסקנות.

לכן חיפשנו מידע באתר האינטרנט של העירייה. אתר זה הינו בספרדית ולכן אם נסתמך על התרגום וניתוח הטקסט, כנראה שלא נוכל להגיע למסקנות חותכות על התשתיות העירוניות בשכונה. החלטנו לחקור את השכונה באמצעות מפות.





ראינו שזו שכונה באיזור המרכזי של ברצלונה והיא גם השכונה המרכזית בעיר.

מסקנתנו מהחקר הנ"ל היא שזו השכונה מרכזית וגדולה ועל כן ניתן להניח כי היא בעלת מרכזי תעסוקה רבים וצירי תנועה עמוסים.

שלב 4:

נרצה לסייע בפתרון הבעיה. החלטנו לנתח את מיקומי התאונות ולהביא מסקנות על צמתים מסוכנות בשכונה על מנת שהעירייה תטפל בהם וכך נצליח להוריד את כמות התאונות.

עבור כל תאונה ראינו את מידת החומרה (כמות רכבים מעורבים ומספר קורבנות). אנחנו רואים כי התאונות מתרחשות בכל השכונה ולא באיזור אחד.

החלטנו לבצע קיבוץ של הנקודות לצמתים בגודל של 10*10 מטר, ולספור את כמות התאונות בצומת.

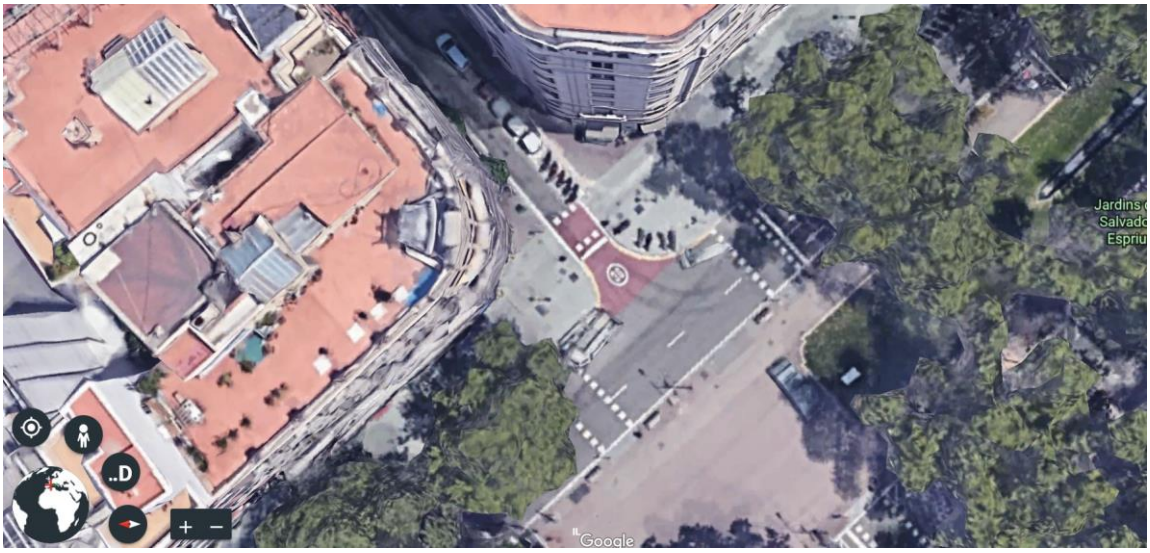
הצלחנו להביא למספר מצומצם של צמתים מסוכנות ביותר שבהן התרחשו מעל 15 תאונות השנה.

נרצה להביא את המידע הנ"ל לעירייה ולמחלקת התשתיות על מנת שיגשו לכל צומת ויבחנו את הסכנות בה.

	Longitude	Latitude	Accidents
0	2.1595	41.3965	35
1	2.1699	41.3991	35
2	2.1649	41.3922	25
3	2.1683	41.3894	20
4	2.1631	41.3975	18
5	2.1720	41.3975	18
6	2.1733	41.3932	18
7	2.1615	41.3970	17
8	2.1675	41.3941	17
9	2.1634	41.3911	16
10	2.1664	41.3933	15
11	2.1710	41.3915	15

למשל, השורה הראשונה בטבלה היא הצומת:

ובתקריב :



רפלקציות:

אנחנו שמחים שהצלחנו לחקור את הנתונים ולהביא לתוצאות שאם ייושמו בשכונה, יוכלו להוריד את כמות התאונות ובכך להציל חיי אדם.