# ויזואליזציה של מידע: נסועה בקווי אוטובוס שנת

# הקדמה:

בעבודה זאת נעסוק בתיאור גרפי של קבצי נסועה בקווי אוטובוס במדינת ישראל בין השנים 2022-2024. קובץ נתונים זה, מפורסם ע"י משרד התחבורה ומדעכן אחת לרבעון. הקובץ מתעד את קווי האוטובוס השונים ברמה שבועית במצב שגרה. מכיל תיאור גיאוגרפי של הקו, סכימה שבועית של נוסעים עלויות ורווחים ממוצעים.

בעבודה ננסה להשוות בין חברות האוטובוסים השונות, ננסה למפות התרחשויות בתחום האוטובוסים ברחבי המדינה.

# תיאור נתנונים

קבצי הנסועה בקווי אוטובוס זמינים להורדה באתר הממשלתי ומחולקים לשלושה קבצים לפי שנים. המידע הזמין בקבצים אלו הינו מהרבעון האחרון של שנת 2022, הרבעון הראשון והשני של שנת 2023 והרבעון הראשון של שנת 2024. הנתונים נאספים ע"י תיעוד התיקופים באוטובוסים ומכיל את המידע על כלל הקווים בארץ ישראל. מידע זה לא נמדד בזמני חירום ולכן שנת 2023 לא תועדה במלואה.

בשנת 2023 נוספו השדות "מהירות ממוצעת", "משך הנסיעה הממוצע" ו"עלות תפעול ממוצעת" שלא היו קיימים בוקבץ של שנת 2022.

קובץ תיאורי של כלל המשתנים מצורף לעבודה זו. הקובץ מכיל את תיאור המשתנים כפי שמופיע בתיאור המשתנים של משרד התחבורה. בנוסף מכיל את השדות: ממוצע, סטיית תקן, חציון, ערך מקסימלי ומינימלי.

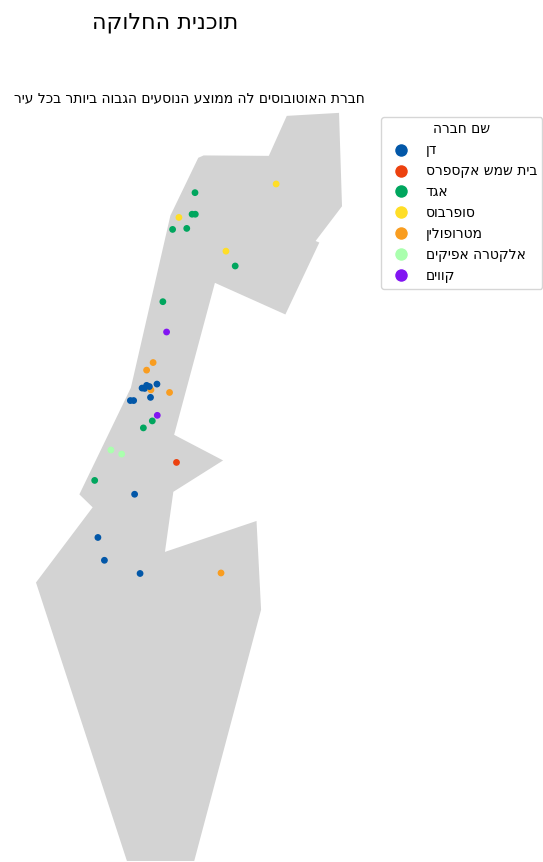
בנוסף לקובץ הנתונים המקורי. נעשה שימוש גם בקובץ נתונים גיאוגרפיים של הערים בשיראל.

# השיטה

בעבודה זו נעשה שימוש בשפת התכנות Python. נעשה שימוש בספריות pandas, numpy לעיבוד הנתונים.matplotlib plotly geopandas כדי לבצע את הויזואליזציות ו-bidi כדי להתמודד עם השפה העברית.

# חברות האוטובוסים בישראל ותרבות הנסיעה באוטובוס

המוטיבציה של עבודה זו הינה מצד אחד להשוות בין תפקוד ו נתח השוק של חברות האוטובוסים השונות במדינה. ומצד שני, להשוות ולהסתכל על ערים נבחרות בישראל ולנסות להסתכל על ההבדלים בניהן.

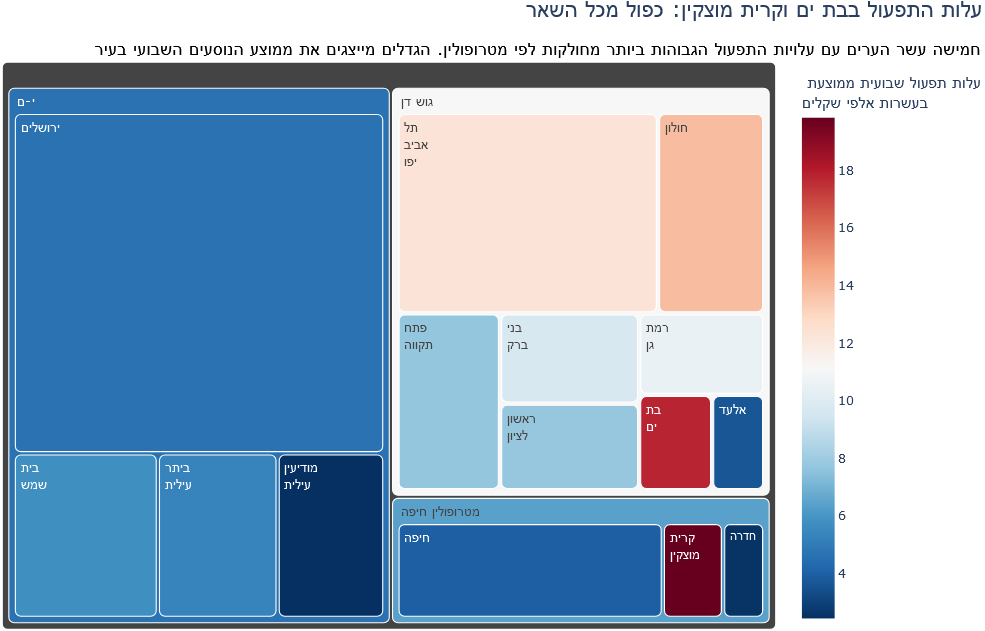
בגרף הראשון ניתן לראות את מפת ישראל ואת חברת האוטובוסים שלה הממוצע הגבוה ביותר של נוסעים באוטובוסים היוצאים בן עיר מסויימת. הגרך מראה כי ישנה חלוקה גאוגרפית של חברות האוטובוסים בארץ. ניתן לראות שחברת אגד היא המועדפת ביותר על הקהל בצפון, חברת דן פעילה במרכז ובדרום וחברת מטרופלין מועדפת בעיקר בשרון.

קידוד הצבעים נעשה בהתאם לצבעים המייצגים את החברות השונות. פרט לסופרבוס ואלקטרה אפיקים שצבע האוטובוסים והלוגו שלהן הוא כחול ומתנגש עם הצבע של דן.

הכנת הנתונים:

* סינון של תת חברות כמו אגד תעבורה ודן באר שבע והכללה שלהן בתוך חברות האב שלהן. נעשה כדי להקל על העומס בגרף.
* קיבוץ הנתונים לפי עיר וחברה. סכימה של מספר הנוסעים הממוצע.
* איחוד של נתוני הצבעים והקורדינטות על הגרף וזריקת הערים שלא נמצאו בקובץ הערים.
* סינון של 70 הערים להן הממוצע הגבוה ביותר. נעשה כדי להקל על העומס.

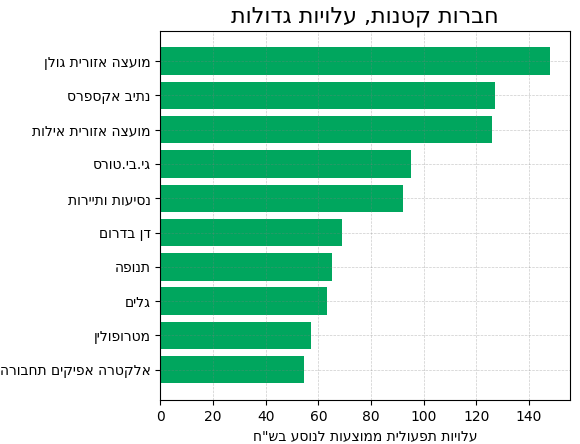
בגרף הנמצא בקובץ ההרצה ניתן להצביע עם הסמן על עיר ולראות את שם העיר ואת שם החברה. בנוסף, ניתן לגלול פנימה והחוצה (zoom) בעזרת הגלגלת. פיטצ'רים אלו הוספו כדי לקבל מידע על העיר אך מסר הגרף ברור גם בלעדיהם.



ניתוח העלויות בוצע גם בהסתכלות על המרחב הגאוגרפי. בולט לעין שבבת ים וקרית מוצקין העלויות מאוד גבוהות.

הכנת הנתונים:

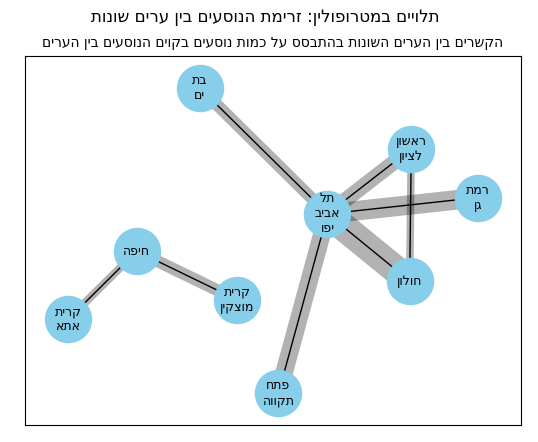
* סינון של שנת 2024 בלבד.
* חישוב עלות שבועית ממוצעת ע"י הכפלת הערכים ממוצע נוסעים שבועי ו-עלות התפעול הממוצעת פר נוסע (OperatingCostPerPassenger , WeeklyPassengers)
* איחוד המטרופולינים מזרח ירושלים וי-ם.
* סינון שאר המטרופולינים (שאינם גוש דן, חיפה וירושלים). שאר המטרופולינים מתנהגים פחות או יותר כמו ירושלים ולכן סינון שלהן לא גורע ממטרת הגרף.
* קיבוץ לפי השדות מטרופולין ושם העיר וחישוב הממוצע של עלות השבועית הממוצעת.
* התחשבות ב15 הערים בעלות העלויות הגבוהות ביותר כדי לא ליצר עומס על הגרף.
* חלוקה של הערך עלות שבועית ממוצעת ב10,000 כדי להקל על עומס המספרים במקרא

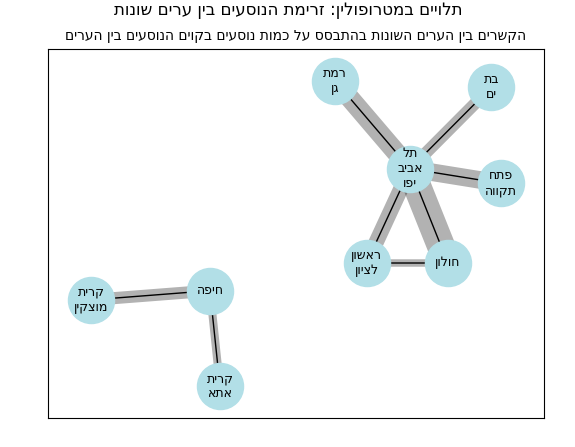


בגרף, "חברות קטנות, עלויות גבוהות" ניתן להבין כי בחברות ההיסעים הקטנות עלויות התפעול פר נוסע גדולות.

הכנת הנתונים:

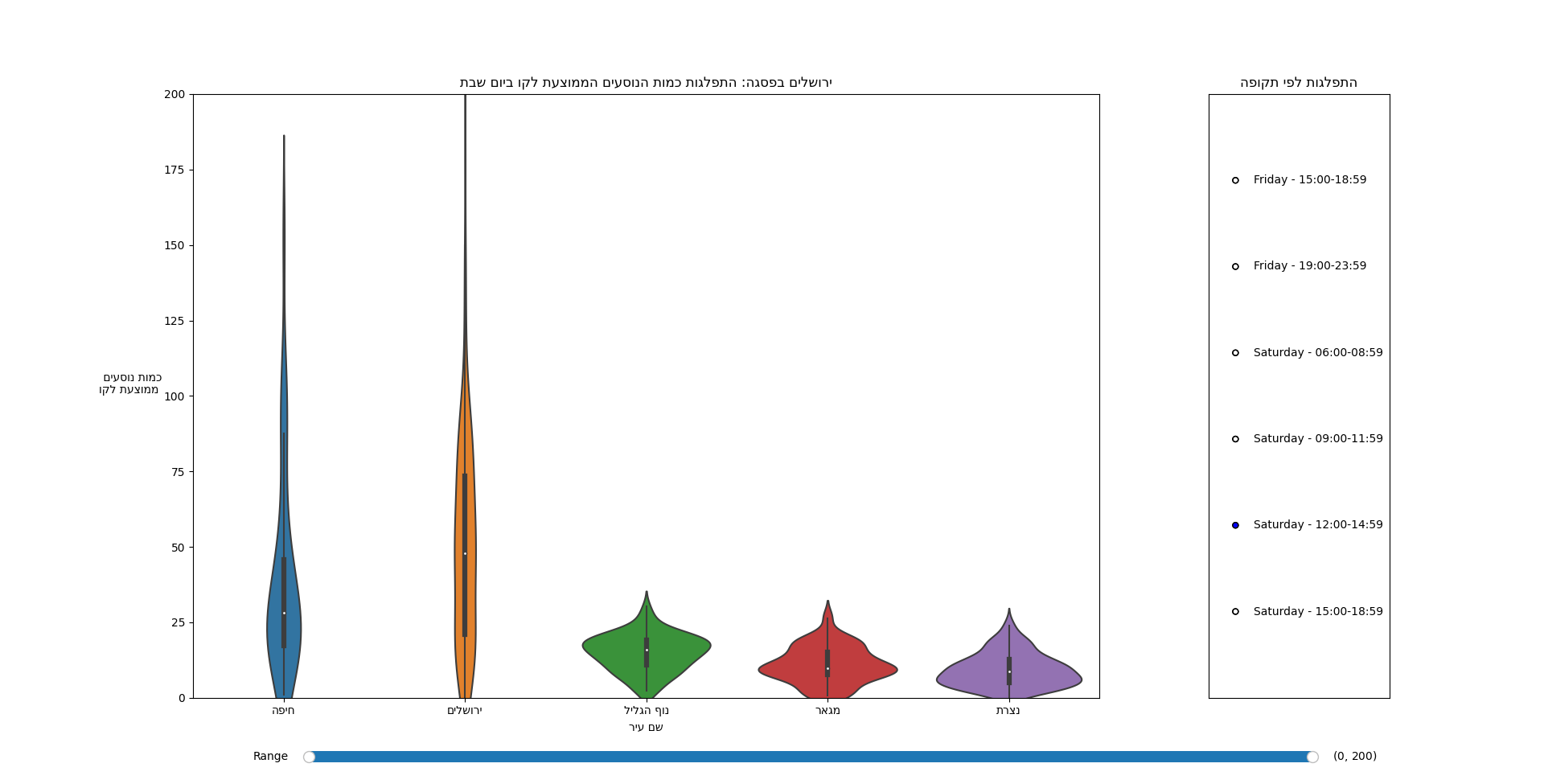
* קיבוץ הנתונים לפי וחברה. מיצוע של השדה OperatingCostPassenger.
* התחשבות ב10 החברות בעלות העלויות הגבוהות ביותר כדי לא ליצר עומס על הגרף.
* סיבוב הצירים בגרף נעשה כדי להקל על הקורא בקריאת שמות חברות ההסיעים.





לאחר התבוננות בגרף זה אפשר לראות שגם לקרית מוצקין וגם לבת ים יש קשרים חזקים עם המטרופולינים שלהם. אולי נותן הסבר לעלויות התפעול הגבוהות שלהן. כנראה לתחזק קוי אוטובוסים המעבירים את הנוסעים הלוך וחזור מערים גדולות, עולים יותר לתחזוק, ואחוז הנוסעים שיוצא מעירים אלו למטרופולינים גבוה.

הכנת הנתונים:

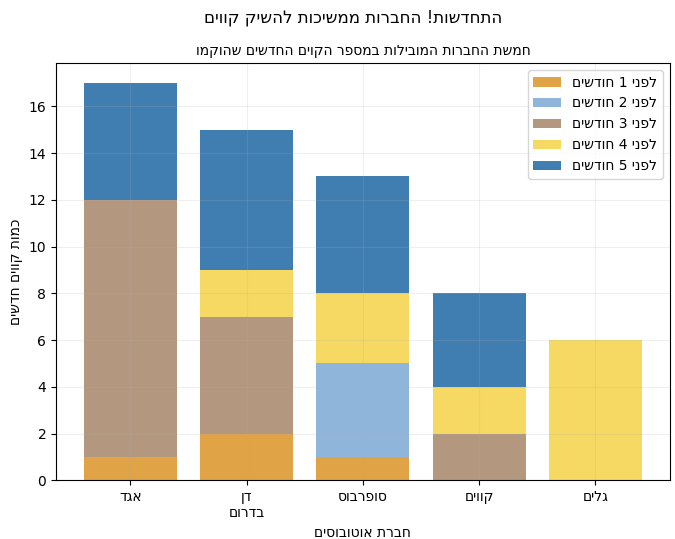
* קיבוץ של הנתונים לפי עיר מקור ועיר יעד. ביצוע של ממוצע הנוסעים השבועי.
* סינון המידע כך שלא תהיה עיר שמקושרת לעצמה. עניין לראות אילו ערים הכי מקושרות לערים אחרות ולא לעצמן.
* סינון 15 הקשרים החזקים ביותר. נעשה כדי להקל על העומס בגרף.
* ציור ה-nodes וציור ה-edges לפי אלגוריתם spring. כדי שהאלגוריתם לא ייתכנס יותר מידי הובל מקסימום סה"כ האיטרציות של האלגוריתם ל-10.

בגרף הכינור אפשר לראות את התפלגות של המשתנה Saturday - 12:00-14:59. המשתנה הוא ממוצע לתיקופים בנסיעה חלקי מספר הנסיעות בלו"ז.

בנוסף, הגרף מכיל גם בחירה של התקופה ביום (כדי לשנות את הנתונים שבטבלה) וגם שליטה על התצוגה של מקסימום כמות הנוסעים הממוצעת המוצגת.

הכנת הנתונים

* קיבוץ לפי ערים. סכימת כמות הנוסעים הממוצעת ומציאת 5 הערים המובילות
* סינון הנתונים לפי הערים הנבחרות
* זריקת שורות עם נתונים חסרים
* הגבלת ערכי כמות הנוסעים הממוצעת בין 0 ל-140. תחום תחתון הוגבל בשל חוסר עניין במשתנה אם הוא מתחת ל-0. והגבול העליון הוגדר בשביל ראות יותר טוב של החציון.
* הוספת מסנן radio button לפי קטגוריות הזמנים הנתונות בקובץ הנתונים. הוספה זו נתנה שליטה למשתמש. המסקנת הגרף לא משתנה כאשר משנים את הקטגוריה.
* הוספת סנן המאפשר משחק עם הגבולות של ציר ה-y. נעשה בשל הצגת התפלגויות שונות עם טווחים שונים.



גרף זה מתאר את מספר הקוים החדשים שנוספו עד כ-5 חודשים לפני הדוח, משמע אוקטובר - פברואר. תקופה זו עניינה אותי במיוחד. ניראה שלמראת הקשיים חברות האוטובוסים ממשיכות להשיק קוים חדשים.

הכנת הנתונים:

* סינון של שנת 2024 בלבד
* יצירת 5 עמודות חדשות המציינות האם הרשומה נמצאת באחת מהקטגוריות ע"פ המשתנה OperationSince.
* קיבוץ לפי שם חברת האוטובוסים וספירה של כמות הקווים החדשים.
* סינון 5 החברות הטובות ביותר. כדי להקל על עומס הצבעים.