תיאור הפרויקט

ראשית, עברנו על המשימה (HU.BER) ועל מאפייני השדות השונים ושאבנו 5% מהמידע לקובץ נפרד.

משימה 1-

בתור התחלה לקחנו שרירותית כמה שדות שנראו לנו אינטואיטיבית הגיוניים ו"קלים" (לא דורשים שינוי והתאמות) בשביל ליצור אלגוריתם למידה ראשוני (**נספח ד'**).

החלטנו כי נמיר את כל המאפיינים למספרים בשביל שנוכל לנסות לעשות השוואות קורלציה לעמודה "passengers_up" בצורה הבאה:

- המרנו את שעת ההגעה לדקות ביום.
  במידה והשדה היה ריק הכנסנו את הזמן הממוצע בדקות.
- המרנו את שעת סגירת הדלתות לכמה זמן הדלת הייתה פתוחה (שעת סגירה – שעת הגעה).
  במידה והשדה היה ריק הכנסנו את הזמן הממוצע בדקות.
- המרנו את חלק הקו לקבוצה סגורה של {1,2,3} במקום {א,ב,ג} בשביל שנוכל להשוות ולאמן.
- הסקנו כי אין צורך בעמודות "trip_id_unique_station" ו-"trip_id_unique" כיוון שמובעות באמצעות עמודות אחרות בטבלה.
- הסקנו כי אין צורך בעמודות "station_id" ו-"station_name" כי אינן קשורות לקו ספציפי.
- הסקנו כי אין צורך בעמודת "alternative" עקב חוסר קורלציה (**נספח א'**).
- באמצעות השוואה של הנ"צ לכמות האנשים שעלו ראינו כי במרכז גוש דן קיים משמעותית יותר עומס (**נספח ב'**). ולכן החלטנו להשאיר את העמודה של "cluster" (אם כי לא נראה שהיווה שינוי משמעותי בחישוב ה-Loss- **נספח ג'**).

ולאחר מכן, באמצעות מודל של רגרסיה ליניארית וחישוב Loss בשיטת MSE. ביצענו זאת על 75%
מה-5% של המידע ללא עמודת "cluster" ואיתה (**נספח ג'**).
ניתן לראות כי היה שיפור בביצועי ה-Loss לאחר התאמת הפיצרים.

משימה 2-

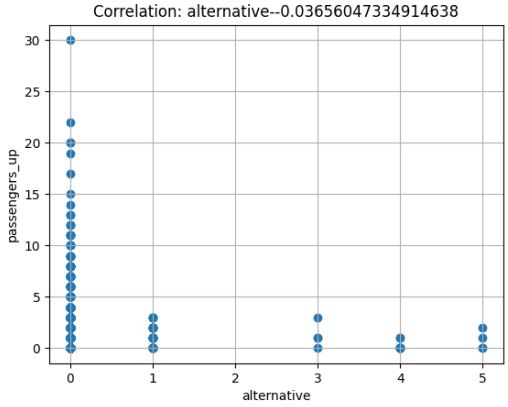
לאחר שכבר התנסינו והבנו טיפה את המידע מהמשימה הראשונה הגענו למשימה השנייה. הבנו כי בכדי לקבל את המידע שאנחנו רוצים על כל "trip_id_unique" עלינו לקבץ שורות בעלי ערכי זהים ולחשב לכל מקבץ שורות אלו מהו זמן הנסיעה הכולל. לאחר מכן, בדומה למשימה הקודמת, ננסה באמצעות כלים אנליטיים וחשיבה רציונלית להבין אילו פיצרים נוספים עשויים להועיל ללמידת האלגוריתם.
בכדי לגרום לדגימת המידע וכלל המידע לעבוד בשלב זה ביצענו שליפה של המידע בקבוצות לפי "trip_id" בגודל שנסכם לכ-5% מכלל המידע.

- ראשית קיבצנו את השורות לפי "trip_id_unique".
- השארנו את העמודות "trip_id_unique", "cluster", "line_id" כפי שהן ("cluster" לאחר העיבוד שלו בדומה למשימה הקודמת)
- חישבנו את העמודות הנוספות-
  - "departure_time": שעת יציאה מהתחנה המינימלית (למקרה שלא קיבלנו את תחנה מספר 1)
  - "num_of_staitions": מספר התחנות בנסיעה.
  - "total_passengers": כמות הנוסעים הכוללת שהייתה בנסיעה.
  - "total_distance": מרחק מצטבר בין כל התחנות
  - "total_time": הזמן שעבר (בדקות) מהיציאה מהתחנה המינימלית ועד הגעה לתחנה האחרונה.
- לצערנו, נראה כי הקורלציה בין העמודות המעובדות ל-"total_time" לא גבוה ולכן לא נראה כי הקשר הוא אכן ליניארי (**נספח ה'**).

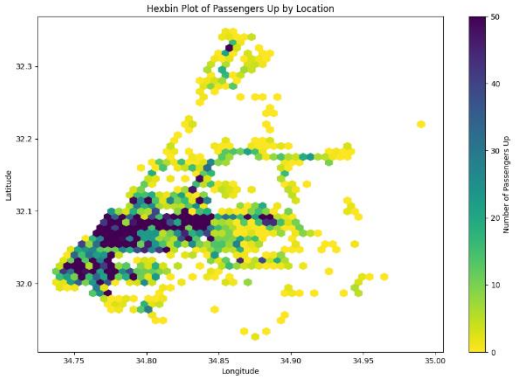
לאחר מכן, הפעלנו מודל של רגרסיה ליניארית וחישוב Loss בשיטת MSE. את מודל הפעלנו ראשית על
כ-5% מהמידע ולאחר מכן בשלב הסיום (עקב קשיים לשלוף את המידע לפי קבוצות של "trip_id") על כלל המידע (**נספח ו'**).

נספחים

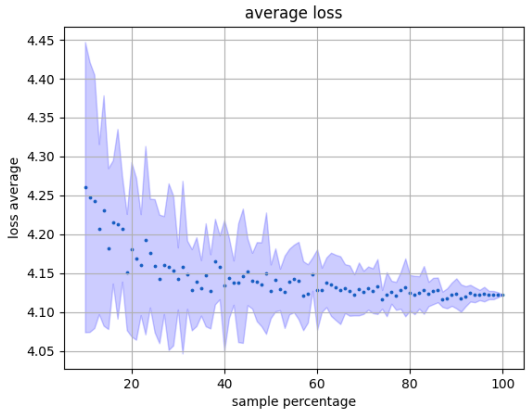
נספח א' – קורלציה של עמודת "alternative" ו-"passengers_up":

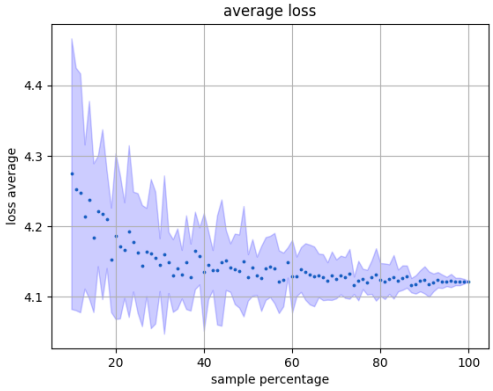


נספח ב' – גרף כמות האנשים שעלו לאוטובוס לפי נ"צ:

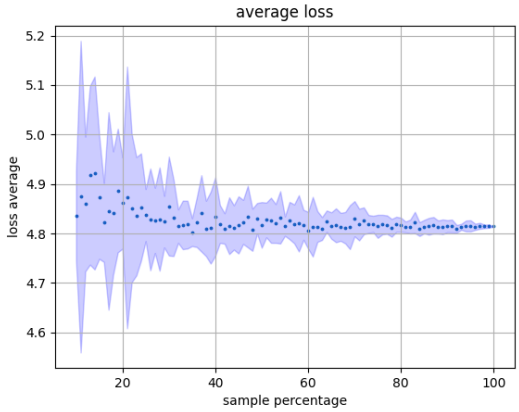


נספח ג' – גרף ה-Loss של המשימה הראשונה על 75% מ-5% מהמידע – ללא עמודת "cluster" משמאל ועם מימין:

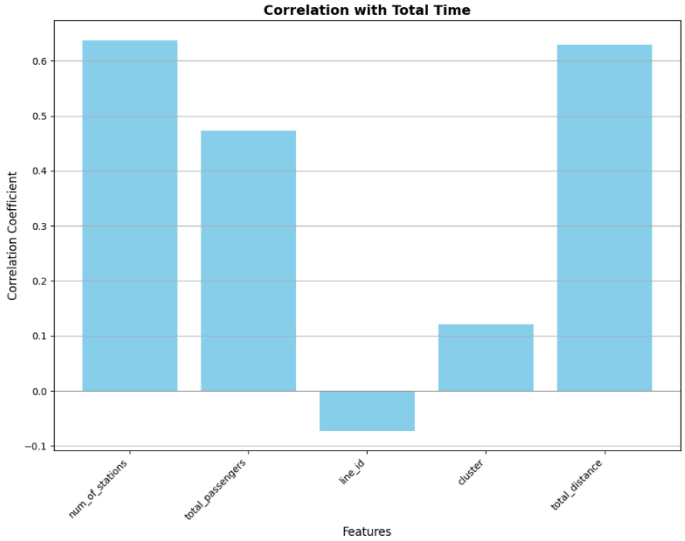




נספח ד' – גרף ה-Loss של האלגוריתם הבסיסי (עמודות הבאות ללא שינוי בלבד – "arrival_is_estimated", "passengers_up",
"passengers_continue","mekadem_nipuach_luz", "passengers_continue_menupach")
על 75% מ-5% מהמידע



נספח ה' – קורלציה בין העמודות המעבדות של משימה 2 ל-"total_time:"



נספח ו' – Loss משימה 2 (בדקות):