**הסבר תהליך עבודה:**

**חלוקת הדאטה הגולמי:**

חילקנו את הדאטה בקובץ train\_bus\_schedule לשלושה קבצים csv חדשים לפי החלוקה הבאה: 75% מהדאטה ניתן לקובץ train, 5% ניתן לקובץ validation ו 20% לקובץ test.

**שלב Preprocess:**

בהתאם למידע הניתן בקובץ bus\_column\_description והסקת מסקנות משלב EDA קיבלנו את ההחלטות הבאות:

עיבוד משותף לשתי המשימות:

תחילה אנו מוודאים שכל שורה מכילה נתונים הגיוניים בעמודותיה לפי הגדרתם, אחרת מסירים אותה. לאחר מכן נבצע עדכון פיצ'רים (עמודות) כדלהלן:

עמודות שהוסרו:

* trip\_id, part – המידע של עמודות אלה מוכל בעמודה trip\_id\_unique ולכן אנו מאמינים כי הם תלויים בה לינארית ולכן אינם מוסיף מידע חדש למודל.
* Station\_name – קיים מזהה ייחודי מספרי לכל תחנה בעמודה station\_id ולכן מזהה מילולי לתחנה אינו מוסיף מידע חדש.
* Cluster – לפי ניתוח ערכי העמודה קיימת חלוקה של המדינה ל 11 אזורים שונים. לפיכך, אנו מאמינים כי כל cluster מכסה שטח רחב מידי ולכן המידע שהוא מסוגל לספק על התחנה וקווי האוטובוס המוכלים בו כללי מידי. על מנת לשפר את איכות המידע יצרנו עמודה חדשה zone שתפקידה זהה לשל cluster, אך מספקת מידע איכותי יותר מאופן יצירתה.

עמודות שנוספו:

* Stay\_time - משך שהיית האוטובוס בתחנה, חושב על ידי:  
   door\_closing\_time - arrival\_time.
* Zone – חילקנו את מדינת ישראל ל 2000 אזורים לפי longtitude ו altitude. כל דגימה מתויגת לאזור המתאים לפי ערכי ה longtitude ו altitude שלה.

עיבוד נתונים ייחודי למשימת passengers up:

עבור שאלה זו דגימה היא שורה בטבלת הנתונים המייצגת תחנת אוטובוס במסלול אוטובוס מסוים. על מנת למקסם את המידע הרלוונטי למודל הגענו להחלטות הבאות:

עמודות שהוסרו:

* Longtitude, altitude – כל אחד בפני עצמו אינו מספק מידע רב, אך שניהם יחד עוזרים לנו לשייך את התחנה לאזור בארץ לפי איך שהוגדרה בעזרתם עמודה zone.
* door\_closing\_time - כל אחד בפני עצמו אינו מספק מידע רב, אך שניהם יחד עוזרים לנו לחשב את זמן שהיית האוטובוס בתחנה, מידע שהכנסנו לעמודה החדשה stay\_time.

עיבוד נתונים ייחודי למשימת Trip duration:

עבור שאלה זו הדגימה אינה שורה בודדת בקובץ הדאטה המקורי, אלא איחוד כל השורות המתארות מסלול trip\_id\_unique. חשבנו כיצד ניתן לאחד את שורות המסלול יחד עם שמירה על מידע רלוונטי למודל והגענו להחלטות הבאות:

נצמצם את כל השורות המאפיינות את trip\_id\_unique לשורה יחידה אשר עמודותיה יהיו כדלהלן:

עמודות שהוסרו:

* Arrival\_time – השתמשנו בערכו לחישוב העמודות החדשות: trip\_duration\_in\_minutes, hold\_time, trip\_start\_time
* Longtitude, altitude – השתמשנו בהם לחישוב עמודה חדשה distance\_traveled.

עמודות שנוספו:

* trip\_duration\_in\_minutes - מכילה את ערך ה arrival\_time של השורה בעלת station\_id הגדול ביותר במסלול של trip\_id\_unique פחות arrival\_time של השורה בעלת station\_id השווה לאחד.
* Trip\_start\_time – שעת תחילת המסלול שערכה הוא arrival\_time של השורה בעלת station\_id השווה לאחד.
* Hold\_time – סכום ערכי stay\_time של כל התחנות במסלול.
* Distance\_traveled – חישוב סכום המרחקים מתחנה לתחנה במסלול בעזרת ערכי longtitude, altitude.

שאר העמודות:

* Zone - נקבע לפי ה zone של תחנה 1 במסלול.

**שלב EDA:**

**בחירת האלגוריתם:**

**i**

**שלב validation:**

**Test error expactions:**

**גרפים מצורפים:**

גרף 1:

גרף 2:

גרף 3:

גרף 4:

גרף 5:

גרף 6:

גרף 7:

גרף 8:

גרף 9:

גרף 5: