**מסמך הסבר:**

אני אסביר פה איך הקובץ XML\_DataPrep.ipynb עובד. הרעיון מרכזי הוא היה לכתוב אוטומציה שבעצם בסופו של כל יום תביא לנו דאטה מסדורת ויפה שאפשר לעבוד איתה בצורה נוחה, כי גם קושי המרכזי של הפרויקט היא לעסוף דאטה שהיא זמן אמת ואין מאגרי נתונים שאפשר למצוא דאטה מהסוג הזה. אז איך זה עובד?

תחילת כתבתי פונקציות עזר כדי שיעזרו בפירוק דאטה בצורה תקנית וצורה הקלה ביותר. פונקציות עזר:

def first(doc) – בחלק הזה לא היה שימוש, פיצלתי לחלב אחר של קוד, כי לא היה לי IP סטטי לזה

def secondFunc(doc) – מאותה סיבה כמן פונקציה first

def dateDiff(date1, date2) – פונקציה שעוזרת לחשב יחס בין זמנים שאנו מקבלים דרך XML

def coorDist(latO,lonO,latP,lonP) – חישוב מרחק בין שתי קוארדינטות

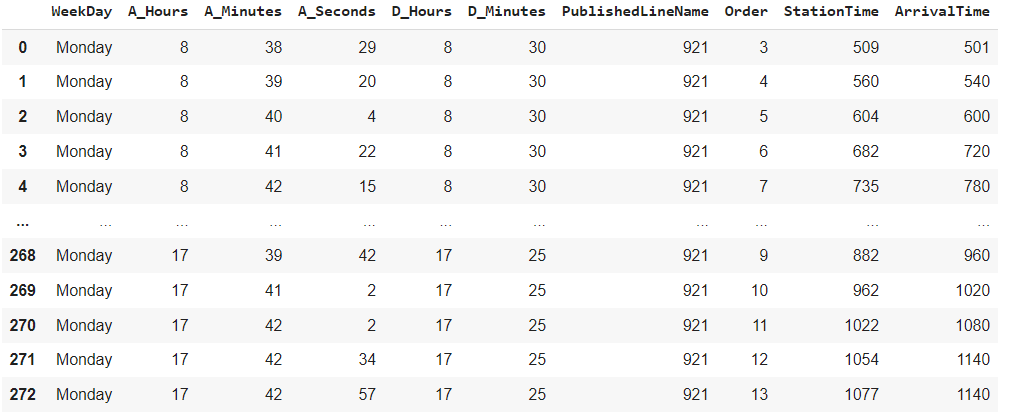
בתיקיה יש שתי קבצים אקסל שאיתם אני משתמש בקובץ פייטון, אחד מהם זה קובץ שלקתי אקראי (*mon 8-20 may 10.xls*) לצורך לדגום את התוצאה סופית מקובץ שהוא אחרי XML, וקובץ שני (*921\_13Station.xlsx*) הוא קובץ של קואדינטות לתחנות שאנחנו דוגמים, איתו אנחנו עושים תיקון קטן (מסביר בהמשך).

התאליך מהסוג הזה אנחנו עושים לכל קובץ יומי של דאטה שעספנו אותה מ XML יומי(של התאליך אוטומטי).

אנו לוקחים קובץ שקבלנו אותו בסוף יום, קובץ אקסל, ומריצים קוד הזה כדי לנקות אותו מדאטה שהיא לא רלוונטית לפרויקט, וגם להוריד פיצירים שאנחנו לא נעבוד איתם בסוף, כל התאליך לקוח לא יותר מדקה ובסופו של דבר אנחנו מקבלים קובץ קטן עבור דאטה שהיא רלוונטית אלינו. תוצאה סופית אפשר לראות בסוף.

דברים מרכזיים שנעשים פה זה פירוק זמן לשעות, דקות, שניות, גם מציאת יום בשבוע, שזה בעצם לקחת תאריך ובעזרת פונקציה להגיד מה הוא יום בשבוע. תאליך נוסף שנעשה פה הוא חישב זמני הגעה לתחנות אוטובוס בעזרת פונקציה dateDiff שמקבלת שתי תאריכים, שזה מתי אוטובוס יצאה מתחנה מרכזית ופניה אחרונה לשרת(אנו פונים כל 15 שניות לשרת) בעצם שינוי אחרון או רשומה אחרונה עבור תחנה ספציפית שאנו בודקים זמן,פלוס אנו עושים תקון קטן מוספים זמן שיש לו להגיע עד תכנה, כי במערכת של משרד התחבורה אין אינדקציה מתי הוא נכנס לתכנה וגם במערכת לפעמים רואים שאוטובוס בדרך תחנה 4, אבל עדיין לא הגיע לתחנה 3 או כמעט הגיע לתחנה 3, השינוי הקטן הזה נותן דיוק יותר טוב והוא פשוט חישוב של מרחק חלקי מהירות שתתן לנו זמן שאותו אנו מוספים, וכל זה אנו רושמים בעמודה של StationTime . עמודה נוספת שבדקנו אותה היא עמוד שבתוכה שמנו זמן של מודל או סוג של מודל שיש להם במערכת, בעצם לקחנו זמן יציאה מתחנה מרכזית וזמן הגעה לתחנה לפי מה שהם חושבים ובעזרת פונקציה dateDiff מוצאים זמן בשניות. זה לא נלקח בשימוש בהמשך של פרויקט כי היו טעיות בתוצאות סופיות כמו אותו זמן הגעה לשתי תחנות שונות, או באגים של מערכות שלהם.

דוגמה לתוצאה סופית של טבלה מקבלים אותה:



WeekDay – יום בשבוע

A\_Hours – שעת הגעה לתחנה

A\_Minutes – דקה הגעת לתחנה

A\_Seconds – שניה הגעה לתחנה

זמן הגעה לתחנה הוא בפורמט של:*A\_Hours:A\_Minutes:A\_Seconds*

D\_Hours - שעה יציאה מתחנה מרכזית

D\_Minutes – דקה יציאה מתחנה מרכזית

זמן יציאה מתחנה מרכזית הוא בפואמט של: D\_Hours:D\_Minutes

PublishedLineName – מספר קו(אפשר להוריד, יותר חשוב אם יש יותר מקו אחד)

Order – מספר תחנה בסדר עולה לפי מיקומו שהוא נכנס אליה מ 2 עד n , כך ש n היא תחנה אחרונה

StationTime – זמן שמצאנו בין פניה אחרונה לזמן מוצא מתחנה מרכזית

ArrivalTime – זמן שמצנו בין מודל של שרת לזמן מוצא לתחנה מרכזית.

בקובץ פייטון יש הסבר עבור כל קטע קוד, קוד נבדק ועובד תקין.