בתרגיל זה תיחשפו למידע "מלוכלך" ומפוזר שתצטרך לנקות ולמזג לטבלה מרכזית אחת שמכילה מידע נקי ומקיף. באמצעות התרגיל תיחשפו לסוגי בעיות (אמיתיות) שיכולות להופיע במידע עליו תעבדו בעתיד, ותפתחו מודעות לחשיבות תהליך ניקוי המידע ולאיכות הנתונים עליהם תעבדו.

תרגיל זה מצריך גם יכולות אנליטיות של איסוף והצלבת נתונים ממקורות שונים שנחוצים גם לשלב האנליטיקה הבסיסית עליה למדתם בשלב הקודם.

לא בכדי אתם מתבקשים להשתמש ביכולות אלו בתרגיל זה: צעדים רבים של ניקוי מידע נוטים להתבצע תוך כדי האנליזה הבסיסית והראשונית של הנתונים. זאת מכיוון שרק בשלב זה אנו מתחילים לקבל הבנה ראשונית של הנתונים שבידינו שבאמצעותם ניתן להבין אם המידע הגיוני ודורש ניקוי.

רצוי כמובן לנסות ולצפות פגמים ובעיות אפשריות במידע בשלב כמה שיותר מוקדם, שכן גילוי מאוחר של בעיות יהיה יקר יותר לטיפול ויצריך הרצה חוזרת של כל התוצרים שפותחו עד לשלב הגילוי (אולם כאמור, אין זה הגיוני לצפות שנוכל לגלות ולטפל בכל הבעיות שיש לנו במידע לפני שנבצע אנליזה ראשונית שלו).

**חברת משאיות עזריאלי ובניו**

חברת משאיות עזריאלי ובניו משלמת לנהגיה בכל חודש על פי כמות הק"מ שהם נסעו באותו חודש, כאשר לכל ק"מ יש תעריף שונה בהתאם לחלק היום בו הוא התבצע (יום או לילה), זהות הספק שהזמין את הנסיעה ועוד שלל גורמים.

להלן פירוט המידע שקיבלתם מהחברה:

1. **Taarif.csv**: קובץ זה מכיל טבלת תעריפים לק"מ נסיעה, כאשר התעריף משתנה בהתאם לשלל פרמטרים הנוגעים לאופי הנסיעה.

להלן פירוט על שדות הטבלה:

* ***customer*** – שם החברה שהזמינה את הנסיעה. כל חברה משלמת תעריפים שונים בהתאם לחוזה שנחתם בינה ובין חברת עזריאלי ובניו.
* ***Basic\_taarif***– התשלום שמשלמת החברה המזמינה לכל ק"מ בתוך 200 הק"מ הראשונים בכל נסיעה.
* ***Extra\_milage*** – התשלום לכל ק"מ אחרי 200 הק"מ הראשונים בכל נסיעה.
* ***Night\_bonus*** – תוספת באחוזים לתעריף על כל ק"מ שמתבצע בלילה.
* ***Weekend\_bonus*** – תוספת באחוזים לתעריף על כל ק"מ שמתבצע בסופ"ש.

1. **Drivers\_with\_kviut.csv, new\_drivers.csv**: קבצים אלו מכילים מידע על הנהגים בחברה. להלן פירוט של השדות בטבלה:

* ***Id*** – ת.ז. של הנהג
* ***Birthdate*** – תאריך לידה של הנהג
* ***Gender*** – מגדר
* ***Vetek*** – ותק בחברה (זמן שהעובד עובד בחברה)

1. **Trips**: תיקייה זו מכילה קובץ לכל חודש ומשאית. שם הקובץ יהיה בפורמט <date>\_<truck\_id>. בכל קובץ שמורה טבלה שמתארת מידע על הנסיעות של המשאית באותו החודש. להלן פירוט השדות של הטבלה:

* ***Driver\_id*** – ת.ז. של הנהג בנסיעה
* ***Customer*** – שם הלקוח שהזמין את הנסיעה
* ***Start\_time*** – תאריך שמציין את זמן תחילת הנסיעה
* ***End\_time*** – תאריך שמציין את זמן סיום הנסיעה
* ***Km*** – מספר הקילומטרים שערכה הנסיעה

**המשימה**

עליכם לייצר מתוך הנתונים שקיבלתם טבלה בשם summary.csv . הטבלה תכיל עבור כל נהג וחודש מידע על הנהג ועל הביצועים של הנהג באותו החודש. השדות שצריכים להופיע בטבלה הם:

* ***Driver\_id*** – ת.ז. של הנהג
* ***Month*** –החודש, למשל 01/2015
* ***Total\_income*** – שכר כולל של הנהג באותו החודש
* ***Total\_km*** – מספר הקילומטרים הכולל שהנהג נסע באותו החודש
* ***Gender*** – מין הנהג
* ***Age*** – הגיל של הנהג בשנים (בדיוק של ספרה אחת אחרי הנקודה העשרונית.)
* ***Vetek*** – ותק של הנהג בחברה בשנים (בדיוק של ספרה אחת אחרי הנקודה.)

חישוב השכר הכולל של הנהג צריך להתבצע על פי טבלת תעריפים.

**דגשים והערות**

1. כמו בכל פרויקט, המידע הגולמי שמתקבל מהלקוח עשוי להיות מלוכלך ולהכיל הרבה בעיות. כיוון שמדובר בתרגיל בנושא של data cleaning הכנסנו הרבה מאוד בעיות בנתונים. לכן עליכם להיות חדים, לחוש את המידע ולחפש באופן שיטתי אחר בעיות אפשריות במידע.
2. תעד את כל הבעיות במידע בהן נתקלת במסמך וורד נפרד. הרבה פעמים, ריכוז הבעיות במידע עבור הלקוח הן תוצר מעניין בפני עצמו של הפרוייקט. כאשר אתה מתעד בעיות עשה זאת באופן כמה שיותר מסודר וברור, כך שהלקוח יוכל לצפות בהן בעצמו בקלות. עליך להגיש מסמך זה לחונך שלך בעת סיום התרגיל.
3. אל תהססו לבקש מהלקוח (קרי החונך) הסברים למידע שאינכם מבינים, או לבעיות במידע שמצריכות פירוש/הבהרה. עם זאת, נסו לגלות גם מידה של עצמאות, ואם יש דברים שאתם יכול לפתור ללא הלקוח עשו זאת. כמובן שכדאי לרכז צעדים אלו ולתעד אותם כך שהלקוח וכל מי שמתבונן בתוצרי הפרוייקט יהיה מודע אליהם.
4. בצעו בדיקה לפתרון שלכם.