תיעוד הפרויקט

**חישובים והגדרות:**

1. קודם מחשבים את התשלום לקילומטר ועליו מוסיפים בונוסים.
2. נסיעה בלילה ובסופש – מחשבים את הבונוס כסכום האחוזים של הבונוסים בנפרד (לדוגמה אם נסענו בשבת בלילה והחברה מביאה לנו 30 אחוז בונוס לילה ו50 אחוז בונוס סופ"ש אז הבונוס לכל קילומטר בלילה בסופ"ש יהיה 80 אחוז.
3. שעות לילה – 22:00 – 6:00
4. שעות של סופ"ש – 16:00 ביום שישי עד 20:00 בשבת.

**חוסרים והשלמות:**

1. **חברת חשמל וחברת ביטוח לאומי** – אין מידע ולכן סימלצתי ממוצע של שאר החברות. (אצל חברת חשמל היה חלק מהמידע אבל הוא היה טעות)
2. **חברות שמשלמות בונוס של 0 בלילות\סופ"שים** – ברוריה אמרה להעלות ל20 אחוז ישנם מספר חברות שהופיעו בנסיעות אך לא בקובץ התעריפים והם בעצם היו שם נרדף לחברה שכן בקובץ. החברות הן עמינח, יס והוט.

Aminach -> mizranei\_kfar\_saba  
Yes -> dbs\_satelites  
Hot -> telecommunication\_ltd

1. הורדתי 0 מהסוף בתשלום של הבונוסים של חיל האוויר כי הוא שילם יותר מדי.



1. **חוסר במגדר אצל נהג** – משאיר כחסר מגדר
2. היה מגדר גבר ואישה כתוב בדרכים שונות ואיחדתי לפורמט אחד של m,f
3. **חוסר בתאריך לידה** – השארתי ללא גיל (בתמונה אלה ללא תאריך לידה)
4. **קבצים כפולים** בנסיעות - הסרתי
5. **כפילויות בתוך קבצי הנסיעות** של נסיעה כפולה - הסרתי

**תיקון שגיאות וחוסרים:**

1. יצרתי מודל שבהינתן כמות קילומטרים ומידע על היום ושעה שבו נסע נהג מוציא זמן הגיוני שלקחה הנסיעה (בעצם מוציא קמש ממוצע שזה בגדול אותו דבר)
2. **נסיעות ללא זמן התחלה וללא זמן סיום** - שמתי את הזמן התחלה שלהם יום ראשון בשעה 6 בבוקר (מינימום כסף שאפשר לשלם להם לפי החישובי בונוסים – כלומר יקבלו בונוס רק אם נסעו עד אחרי 22 בלילה)
3. **נסיעות ללא זמן התחלה או סוף** – שימוש במודל נפרד לכאלה ללא זמן התחלה ולכאלה ללא זמן סיום אשר אומן על הדאטה ללא outliers לפי מבחן z עם threshold מסוים שנבחר לפי הסתכלות על איך המידע מתפלג בהינתן הthresholds השונים. השימוש במודל נעשה על מנת לחזות כמה זמן נסעו בהינתן היום והשעה של הזמן התחלה\סיום.
4. **A graph of a graph showing a number of blue dots

   Description automatically generated with medium confidenceתיקון outliers עם יותר מדי זמן נסיעה** -  
   אפשר לראות בתמונה – הגושי נקודות הדלילים בפינות של צד ימין.  
   א) את הoutliers תיקנתי באמצעות כך שאלה שנסעו נסיעות מאוד ארוכות (4000-8000 קילומטר), הקטנתי את הנסיעה שלהם לממוצע הנסיעות שלהם.  
   ב) לאחר מכן השתמשתי במודל על מנת לחשב כמות זמן הגיונית לפי הקילומטרים והזמן התחלה שהם נסעו בו.  
   ג) לoutliers שנסעו כמות קילומטרים הגיונית (הoutliers בצד ימין למטה בתמונה) חוץ מהעניין של להקטין את הכמות קילומטרים שהם עשו מכיוון שכנראה הטעות שם זה ששכחו את המונה זמן דולק והפסיקו לנסוע, לכן הכמות קילומטרים כנראה תקינה.

**נוספים:**

1. קביעות – אין לזה משמעות מבחינת מר עזריאלי